Schriften des Forschungszentrums Jülich
Reihe Materie und Material/Matter and Materials Band/Volume 19

Forschungszentrum Jülich GmbH
Institut für Festkörperforschung

Lecture notes of the
35th Spring School of the Institut für Festkörperforschung

Gerhard Gompper, U. Benjamin Kaupp, Jan K.G. Dhont,
Dieter Richter, Roland G. Winkler (Eds.)

Physics meets Biology

From Soft Matter to Cell Biology

This Spring School was organized by the
Institut für Festkörperforschung
in the Forschungszentrum Jülich
on March 22 – April 2, 2004.

In collaboration with
universities, research institutes and the industry.

Schriften des Forschungszentrums Jülich
Reihe Materie und Material / Matter and Materials Band/Volume 19

ISSN 1433-5506 ISBN 3-89336-348-3

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Publisher and Distributor:	Forschungszentrum Jülich GmbH Zentralbibliothek 52425 Jülich Phone +49 (0) 24 61 61 53 68 . Fax +49 (0) 24 61 61 61 03 e-mail: zb-publikation@fz-juelich.de Internet: http://www.fz-juelich.de/zb
Cover Design:	Grafische Betriebe, Forschungszentrum Jülich GmbH
Printer:	Druckerei Gehler, Düren-Birkesdorf
Copyright:	Forschungszentrum Jülich 2004

The editors cannot accept any responsibility or liability for the accuracy of any statements or information given in the papers.

Schriften des Forschungszentrums Jülich
Reihe Materie und Material / Matter and Materials Band/Volume 19

ISSN 1433-5506
ISBN 3-89336-348-3

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Contents

Preface

I Introduction

I1 Introduction: Physics meets Biology
G. Gompper and U. B. Kaupp

A General Concepts

A1 Statistical Mechanics
G. Schütz

A2 Translational Brownian Motion
J. K. G. Dhont

A3 Self-Assembly
W. Sager and E. M. Blokhuis

A4 Molecular Components of the Cell
A. Baumann

B Techniques and Methods

B1 Scattering Experiments
W. Schweika

B2 Protein Crystallography: X-Rays
J. Labahn

B3 NMR Analysis of Biopolymer Structures
D. Willbold and B. W. Koenig

B4 Atomic Force Microscopy & Laser Tweezers
A. Engel and M. Hegner

B5 Introduction to Computer Simulations
G. Vliegenthart

B6 Single Molecule Fluorescence
J. Enderlein

B7 Modern Applications of Fluorescence Methods in Biology
T. Gensch

B8 Exploring the Inner Space of Cells by Cryoelectron-Tomography
W. Baumeister et al.

C Polymers, Biopolymers, and Proteins

C1 Flexible Polymers
H. Frielinghaus

C2 Polymer Dynamics
D. Richter

C3 Polyelectrolytes and Semiflexible Polymers
R. G. Winkler

C4 Protein-Polymer Mixtures
R. Tunier

C5 The Protein Glass Transition
V. Kurkal and J. C. Smith

C6 Protein Folding: Theory
P. Grassberger

C7 Protein Folding: Experiment
T. Kiefhaber

C8 Motor Proteins
J. Howard

C9 Physics of the Cytoskeleton
A. Bausch

C10 Retinal Proteins
G. Büldt, J. Heberle, and V. Gordeliy

C11 Viruses as Colloidal Model Systems
P. Lettinga

D Membranes

D1 Biomembranes
B. Hoffmann

D2 Mechanics and Statistical Mechanics of Membranes
G. Gompper

D3 Fluid Vesicles
U. Seifert

D4 Physics of Multicomponent Membranes
R. Zorn

D5 Lipid Rafts in Multicomponent Membranes
K. Simons

D6 The Mechanical Strength of Bioadhesion
R. Merkel

D7 Ion Channels and Pumps
R. Seifert

D8 Membrane Receptors
M. Heck and K. P. Hofmann

D9 Exo- and Endocytosis: An Ubiquitous Mechanism for Secretory Processes and Membrane Trafficking
E. Neher

E Index

Preface

In the last two decades, we have witnessed breathtaking advances in molecular biology. In particular, the elucidation of several genomes — including humans, mice, worms, flies and plants — have provided the sequence of hundreds of thousands of different genes and proteins. In the years to come, it will be a formidable task to elucidate the location, structure, function, and interactions of these cellular components. Key requirements for a successful investigation and a detailed understanding of these enormously complex systems are:

- The development and application of an array of modern physical techniques to study the structure and dynamics of macromolecules and their assemblies over a wide range of length and time scales. Such techniques include scattering methods, nuclear magnetic resonance (NMR), atomic force techniques, single-molecule fluorescence and electron tomography.

- A detailed understanding of macromolecules and their assemblies. Soft Matter physics investigates the behavior of complex mixtures consisting of several different components of polymeric, colloidal or amphiphilic character, with the aim to understand the cooperative behavior of systems with a large number of interacting degrees of freedom. A profound understanding of the fundamental mechanisms underlying biological function and self-organization may be fostered by investigations of simpler systems which are composed of synthetic soft materials.

- Analytical and synthetic organic chemistry. The principles of biological structure formation and function may be used to design and build similar synthetic molecules with tunable properties. Markers and labels have to be attached to biological macromolecules to visualize them and trace their motion. Finally, synthetic chemistry is required to provide model systems for macromolecular assemblies.

In order to cope with the enormous challenges posed by understanding the structure and function of the cellular machinery, biologists and physicists have to come together and to share their expertise. Biologists have to master the state-of-the-art technologies required to study and understand biological processes. In a similar vein, physicists and chemists have to learn about and to be inspired by the large variety and diversity of exciting biological problems, and the immense research opportunities in the Life Sciences in general, and Cell Biology in particular.

With the IFF spring school 2004, we want to contribute to the integration process of soft matter physics, biophysics and biology of the cell.

As in previous years, the IFF spring school has been organized in collaboration with the RWTH Aachen and the Universities of Cologne, Düsseldorf and Münster, and is integrated in

the curricula of these universities.
We are very grateful to our colleagues

Dr. A. Baumann	Institut für Biologische Informationsverarbeitung (IBI-1)
Prof. G. Büldt	Institut für Biologische Informationsverarbeitung (IBI-2)
Dr. J. Enderlein	Institut für Biologische Informationsverarbeitung (IBI-1)
Dr. T. Gensch	Institut für Biologische Informationsverarbeitung (IBI-1)
Prof. P. Grassberger	Neumann Institute for Computing (NIC)
Dr. B. Hoffmann	Institut für Schichten und Grenzflächen (ISG-4)
Dr. J. Labahn	Institut für Biologische Informationsverarbeitung (IBI-2)
Prof. R. Merkel	Institut für Schichten und Grenzflächen (ISG-4)
Dr. R. Seifert	Institut für Biologische Informationsverarbeitung (IBI-1)
Prof. D. Willbold	Institut für Biologische Informationsverarbeitung (IBI-2)

from other institutes of the Forschungszentrum Jülich, which conduct research in the field of Soft Matter, Biological Physics, Biophysics, and Biology for their willingness to participate in this spring school.
We are very glad that several colleagues from universities and research institutes have agreed to contribute to the program of the school:

Prof. W. Baumeister	Max-Planck-Institut für Biochemie, Martinsried
Prof. A. Bausch	Technische Universität München
Dr. E. M. Blokhuis	Universiteit Leiden, The Netherlands
Prof. A. Engel	Biozentrum Basel, Schweiz
Prof. K. P. Hofmann	Humboldt Universität Berlin
Prof. J. Howard	Max-Planck-Institut für Molekulare Zellbiologie und Genetik, Dresden
Prof. T. Kiefhaber	Biozentrum Basel, Schweiz
Prof. E. Neher	Max-Planck-Institut für Biophysikalische Chemie, Göttingen
Prof. U. Seifert	Universität Stuttgart
Prof. K. Simons	Max-Planck-Institut für Molekulare Zellbiologie und Genetik, Dresden
Prof. J. C. Smith	Universität Heidelberg

Without the participation of all these colleagues, the program would not be as interesting, versatile and attractive. We would like to express our thanks to all of them for the effort and enthusiasm which they have put into the preparation and presentation of their lectures and manuscripts. We are very grateful to the board of directors of the Forschungszentrum Jülich

for the organizational and financial support, which we have received — as in previous years — for the realization of the IFF spring school and the production of this book of manuscripts. We also thank all the many people in the IFF who have contributed to this school. Our special thanks go to Dipl.-Ing. R. Hölzle for the general management, to Prof. U. Buchenau for his valuable comments during the preparation of the talks, to M.-L. Göcking for her determined help with the formating of the manuscripts, and to H. Paffen for handling most of the correspondence.

Gerhard Gompper, Benjamin Kaupp, Jan Dhont, Dieter Richter and Roland Winkler

February 2004

I 1 Introduction: Physics meets Biology

G. Gompper[1] and U. B. Kaupp[2]

[1] Institut für Festkörperforschung

[2] Institut für Biologische Informationsverarbeitung

Forschungszentrum Jülich GmbH

Contents

1 Can Physics meet Biology? 2

2 The Basic Laws of Physics 2

3 Complex Behavior of "Simple" Systems 4
- 3.1 Phase Transitions . . . 4
- 3.2 Neural Networks . . . 6

4 A little History 7

5 Bridging the Culture Gap 8
- 5.1 What Physics can do for Biology . . . 8
- 5.2 Bringing physics, not just physicists, to biology and medicine . . . 12
- 5.3 Biological Relevance — Biologically Inspired Physics . . . 13

6 From Soft Matter to Cell Biology 14
- 6.1 Protein Crystallization . . . 15
- 6.2 Packing DNA into Viruses . . . 17
- 6.3 Micromechanics of the Cell . . . 19

7 Literature 21

1 Can Physics meet Biology?

What is at the heart of physics; and what characterizes biology? Does biology study the make-up of the living world, and physics is concerned with the inanimate and inorganic? Is physics theory-driven and biology merely a descriptive and empirical science? Are biologists dealing with exceedingly complex systems and physicist with less complex systems? Does physics provide the foundation for the development of other disciplines, notably chemistry and biology? What sets the two disciplines apart from each other? In most cases, the answers are not simply "no" or "yes"; rather there is a whole spectrum of more-or-less answers, sometimes more, sometimes less.

2 The Basic Laws of Physics

From the results of many careful and quantitative experiments, physicists have been able to derive fundamental laws of nature, which seem to be universally valid. Physics uses a mathematical language to represent these laws. In fact, the number of basic physical laws is quite small.

Classical Mechanics – is described by Newton's law. It relates the acceleration of a particle of mass m at position $\mathbf{r}$ as a function of time t to an external force $\mathbf{f}$ by

$$m\frac{d^2}{dt^2}\mathbf{r} = \mathbf{f} \tag{1}$$

This law has been generalized later by Einstein to describe motion for speeds close to the speed of light.

Electrodynamics – is described by Maxwell's equations. The electric field $\mathbf{E}(\mathbf{r},t)$ and magnetic field $\mathbf{B}(\mathbf{r},t)$ in a system of electric charges q and currents are determined by the electric charge density $\rho(\mathbf{r},t)$ and current density $\mathbf{j}(\mathbf{r},t)$ via the partial differential equations

$$\nabla\cdot\mathbf{E}(\mathbf{r},t) = 4\pi\rho(\mathbf{r},t) \tag{2}$$

$$\nabla\times\mathbf{B}(\mathbf{r},t) - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{E}(\mathbf{r},t) = \frac{4\pi}{c}\mathbf{j}(\mathbf{r},t) \tag{3}$$

$$\nabla\cdot\mathbf{B}(\mathbf{r},t) = 0 \tag{4}$$

$$\nabla\times\mathbf{E}(\mathbf{r},t) + \frac{1}{c}\frac{\partial}{\partial t}\mathbf{B}(\mathbf{r},t) = 0 \tag{5}$$

In turn, the Lorentz force

$$\mathbf{f} = q\left(\mathbf{E} + \frac{1}{c}\frac{d}{dt}\mathbf{r}\times\mathbf{B}\right) \tag{6}$$

describes how the electric and magnetic fields act on charged particles. The only parameter in these equations is c, the speed of light.

Quantum Mechanics – is described by the Schrödinger equation. The main difference to classical physics is that the location of a particle is described by a wave function $\Psi(\mathbf{r},t)$, which is a complex number. The probability to find a quantum particle at the point $\mathbf{r}$ at time t is given by $|\Psi(\mathbf{r},t)|^2$. The Schrödinger equation determines the temporal evolution of the wave function as

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t) = H\Psi(\mathbf{r},t) \tag{7}$$

where H is the Hamilton operator. For a particle of mass m in an external potential $V(\mathbf{r})$, the Hamilton operator is

$$H = \frac{\mathbf{p}^2}{2m} + V(\mathbf{r}) \tag{8}$$

This looks like the classical Hamilton function; however, in quantum mechanics, the momentum $\mathbf{p}$ becomes an operator, $\mathbf{p} \rightarrow -i\hbar\nabla$. Therefore, the variables $\mathbf{p}$ and $\mathbf{r}$ no longer commute,

$$[\mathbf{p}, \mathbf{r}] \equiv \mathbf{pr} - \mathbf{rp} = -i\hbar \tag{9}$$

which is the mathematical form of Heisenberg's uncertainty principle that position and momentum cannot be measured simultaneously with arbitrary precision.

Thermodynamics – is described by the first and second law. The first law requires energy conservation; different forms of energy like mechanical, electrical and magnetic energy and heat can be transformed into each other, but the total energy U can only change by a flux across the boundaries of a system under consideration. The second law asserts the existence of an extensive variable, the entropy S, which for any closed system always increases and reaches its maximum value in thermal equilibrium. For a system at temperature T and of volume V, changes in U, S and V for a quasi-static process are related by

$$dU = TdS - pdV \tag{10}$$

A system at constant temperature and volume is described by the free energy $F(T, V) = U - TS$, with

$$dF = -SdT - pdV \tag{11}$$

Statistical Physics – provides the microscopic foundation of thermodynamics, and allows more detailed insights into the collective behavior of multi-particle systems. At constant temperature, the probability to find a system in a microstate $\{s\}$ with energy $H(\{s\})$ is given by the Boltzmann weight

$$P(\{s\}) = \frac{1}{Z} \exp(-H(\{s\})/k_B T) \tag{12}$$

where k_B is the Boltzmann constant, and the normalization factor

$$Z = \sum_{\{s\}} \exp(-H(\{s\})/k_B T)\,, \tag{13}$$

which is called the partition function. The connection to thermodynamics is provided by

$$F = -k_B T \ln(Z) \tag{14}$$

While the description of the basic laws of physics given above is certainly somewhat incomplete, in particular on the sub-atomic level, it covers all the essential aspects that are necessary to understand the behavior of biological (macro-)molecules and their interactions. These laws have been tested in considerable detail, so that their validity is without doubt.

The basic laws of physics are surprisingly simple – in fact, deceivingly simple. Thus, complex behavior arises not from complicated laws of nature, but from a large number of interacting degrees of freedom.

3 Complex Behavior of "Simple" Systems

The essence of physics is to simplify. The principal idea is that a model may be simplified considerable but still retain the essential properties of the original complex system.

3.1 Phase Transitions

We want to illustrate this important point by a paradigmatic example. Let us consider a two-component liquid mixture, like methanol (CH_3OH), which as water has distinctly polar molecules, and hexane (C_6H_{14}), a non-polar substance [1]. The molecular interactions between the two molecules are very complex. However, an important feature is that the attractive interaction E_{MM} between two methanol molecules, as well as E_{HH} between two hexane molecules, is stronger than the interaction E_{HM} between a hexane and a methanol molecule. Thus, a very simplified model can be constructed for this mixture, in which the molecules occupy the sites of a hyper-cubic *lattice*, and have a nearest-neighbor interaction

$$E = -\sum_{\alpha=H,M}\sum_{\beta=H,M} E_{\alpha,\beta}\sum_{<ij>} n_i^{(\alpha)} n_j^{(\beta)} \tag{15}$$

where $\langle ij\rangle$ indicates a sum over pairs of nearest-neighbor sites, and $n_i^{(\alpha)}$ is an occupation variable, which is unity if lattice site i is occupied by component α, and zero otherwise. Since $n_i^{(M)} + n_i^{(H)} = 1$, the energy can be expressed in terms of a "spin variable" $s_i = \pm 1$, where $s_i = 1$ if the site is occupied by a methanol molecule, $s_i = -1$ if occupied by a hexane molecule. Thus, $n_i^{(M)} = (1 + s_i)/2$ and $n_i^{(H)} = (1 - s_i)/2$. The energy (15) then becomes

$$E = -J\sum_{\langle ij\rangle} s_i s_j \tag{16}$$

where

$$J = \frac{1}{4}[E_{MM} + E_{HH} - 2E_{HM}] \tag{17}$$

The model described here is the famous Ising model, which has first been proposed in 1905, almost 100 years ago [1].

The *qualitative behavior* of this model as a function of temperature is derived easily. For high temperatures, the energy plays no role, thermal fluctuations are so large that the interactions plays no role, and the system is completely disordered. At low temperature, neighboring spins prefer to adopt the same value in order to minimize the energy (16). This leads to a phase separation into a phase in which all spins are positive, and another phase in which all the spins are negative. These two regimes are separated by a critical point. The critical temperature T_c can be calculated within the model; however, it cannot be easily related to the experimental values in the real system, because of the simplifications which lead to the energy (16).

A *more detailed analysis*, which took more than 80 years to be completed, and which involved three Nobel prize winners (Onsager 1968, Wilson 1982, de Gennes 1991), reveals a much richer and more interesting behavior. The four main points are:

- The behavior in the vicinity of the critical point is characterized by power laws. For example, the concentration difference between the 'plus' phase and the 'minus' phase disappears as $|T - T_c|^\beta$. The value of the exponent is calculated to be $\beta = 1/8$ in two

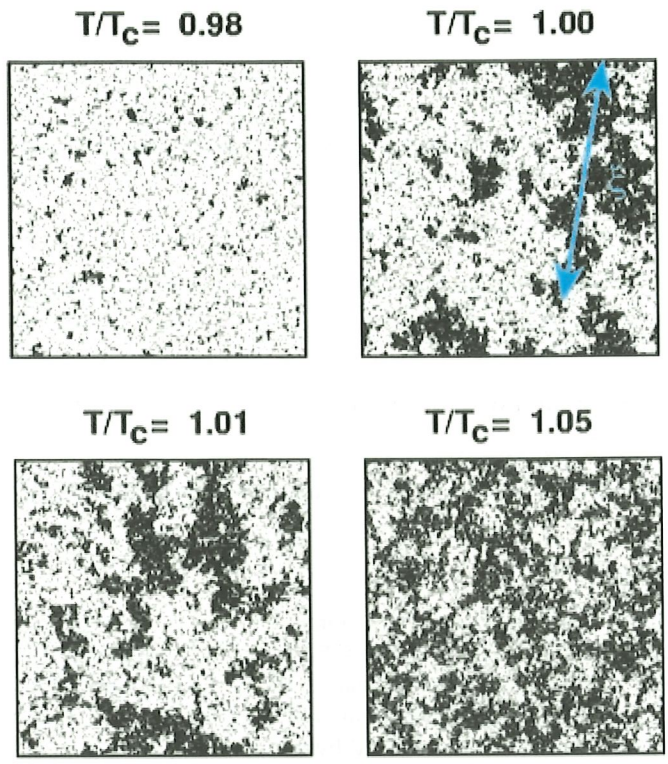


Fig. 1: *Typical configurations of the Ising model near the critical temperature T_c. Plus spins are shown by white, negative spins by black squares. The correlation length ξ, the range over which spins are correlated, diverges at T_c in an infinite system. From Ref. [2].*

dimensions, and $\beta = 0.33$ in three dimensions. Furthermore, these values are predicted to be *universal*, i.e. they should not change if an attractive next-nearest-neighbor interaction is taken into account in addition to the nearest-neighbor interaction.

- The values of the critical exponents have been measured extremely carefully, and perfect agreement with the theoretical predictions has been found. Furthermore, the same values are found for completely different systems, like various binary liquid mixtures, different liquids near the liquid-vapor critical point, or uniaxial magnets. Thus, the idea of *universality* is confirmed beautifully.

- At the critical temperature T_c, snapshots of the typical configurations show an intricate, *self-similar* structure, as shown in Fig. 1. Here self-similar means that by zooming into the picture with some magnification factor, the pictures on the larger and smaller scale look exactly the same. Only when the magnification becomes so large that the lattice constant becomes visible, scaling breaks down.

- Self-similarity of spatial configurations is equivalent to a power-law decay of correlations,

i.e. the correlation function at $T = T_c$ behaves as

$$\langle s(\mathbf{r})s(\mathbf{r}')\rangle - \langle s(\mathbf{r})\rangle\langle s(\mathbf{r}')\rangle \sim |\mathbf{r} - \mathbf{r}'|^{-(d-2+\eta)} \tag{18}$$

where η is another universal critical exponent, which takes the value $\eta = 1/4$ in two dimensions and $\eta = 0.03$ in three dimensions.

We can draw three very important conclusions from this example:

- Despite of dramatic simplifications, the analysis of model systems can provide exact results for complex systems. Thus, *simplifications are possible* while retaining the essential degrees of freedom.
- *Simplifications are necessary* in complex systems in order to extract universal laws on mesoscopic scales.
- Universal behavior *emerges* due to a large number of interacting degrees of freedom.

3.2 Neural Networks

We want to illustrate this point by a second example. In order to understand how information is stored and retrieved in the brain, Hopfield in 1982 suggested a simple model of neural activity [3]. The neurons are placed on a hypercubic lattice, and have two states of activity. Each neuron can interact with all other neurons. The strength of the interaction between any two neurons, which is typically asymmetric, is determined by the number of synapses and their strengths. The stable states of the network are then obtained by minimizing the function

$$E = -\sum_{i,j} J_{ij} s_i s_j \tag{19}$$

where $s_i = \pm 1$ describes the activity level of the neuron at site i, and J_{ij} the effect of the neuron at site i on the neuron on site j. Note that this model bears a strong similarity with the Ising model (16), except that now *all* spins can interact, and the interactions can be all different. For coupling constants J_{ij} selected randomly from a bimodal or from a Gaussian distribution, such a model is called an Ising spinglass.
The important point of the Hopfield model is that the network can learn a number of P patterns $\{s_1^{(1)}, ..., s_N^{(1)}\}$ to $\{s_1^{(P)}, ..., s_N^{(P)}\}$. This learning process is done by adjusting the interaction strengths according to the rule

$$J_{ij} \rightarrow J_{ij} + \frac{1}{N} s_i^{(I)} s_J^{(I)} \tag{20}$$

for $I = 1, ..., P$. After going through all learning patterns, the network has memorized them. When a new pattern is now presented to this neural network, which vaguely resembles a pattern I_0 of the learning set, a minimization of the energy recovers this pattern (almost) exactly.
We can draw the following conclusions from this example:

- The information is not stored in a single neuron, but is stored in the whole set of interaction strengths. The recollection of any pattern is a collective property of the whole network. This is sometimes called the absence of a "grandma" neuron that extracts a particular feature from a picture (e.g. a grandma).

- Due to this collective behavior, the information retrieval is *robust*, i.e. a failure of some part of the network does not immediately lead to a complete loss of all affected information.

- Simplification is again the key to understand the collective properties. The details of structure and chemistry of the real synapses are not important to understand the network behavior.

Today, physicists are exploring applications for network theory in molecular biology. Network of gene regulation may have parallels in neural networks [4]. Just as neurons activate and repress other neurons, gene products can activate and repress other genes, directly or indirectly. The parallels may even extend to similarities between learning, which alters the firing pattern of individual neurons and hence the network's overall behavior, as explained above, and the way in which evolutionary pressure alters the patterns of gene expression. Hopfield has proposed that discrete biological functions rarely lie with individual genes or proteins but instead with "modules" comprising many interacting molecules [5].

4 A little History

Physics and its techniques have been playing an important role in biology for several decades. For example, X-ray crystallography and nuclear magnetic resonance are essential tools for structural biologists. Electron microscopy makes protein assemblies and cell organelles visible, and has become an invaluable technique for cell biologists.
Therefore, the role of physicists in biology and medicine has occasionally been portrayed as merely developers of tools of measurements that allow biomedical scientists to do the really important work. This is expressed, for example, in a 1967 commentary of Sergej Feitelberg, a physicist from Mount Sinai Hospital in New York, who noted that while such "spectacular developments created a clear and unequivocal need for physicists and their help, the role of the physicist was that of a glorified technician engaged in methodology and instrumentation, dignified only by the strangeness of his doings and the mysteriousness of his tools" [6].
Fortunately, this view is not shared by all biologists. Harold Varmus, director of the National Institute of Health (NIH) of the United States, notes for example in 1999 that there is "a deeper set of contributions that physics makes to biology — through the efforts of physicists who themselves seek to understand the rules of living systems" [6].
In 1949, in a speech entitled "A physicist looks at biology", **Max Delbrück**, a student of Niels Bohr and a leading physicist who made a conversion to biology a few years earlier, attempted to describe the transition: "A mature physicist, acquainting himself for the first time with the problems of biology, is puzzled by the circumstance that there are no 'absolute phenomena' ... The animal or plant or micro-organism he is working with is but a link in an evolutionary chain of changing forms, none of which has any permanent validity. Even the molecular species and the chemical reactions which he encounters are the fashions of today to be replaced by others as evolution goes on. The organism he is working with is not a particular expression of an ideal organism, but one thread in the infinite web of all living forms, all interrelated and all interdependent. The physicist has been reared in a different atmosphere. The materials and phenomena he works with are the same here and now as they were at all times and as they are on the most distant star" [6].

Delbrück and his friends were fascinated by some fundamental questions: What is the physical form in which hereditary information is stored? How is it reproduced when a cell divides, or when a single virus particle invades a cell and makes hundreds of thousands of copies of itself? How does the information change when mutations occur?
Answers to many of these questions came from the "phage school" that Delbrück founded. The phage school was a group of former physicists and some biologists who shared his passion for reducing the problem of heredity to simple rules, physical entities and conserved energy by studying the replication and genetic behavior of bacterial viruses — also called bacteriophage or "phage" — in their bacterial hosts. The studies culminated in findings that form the pillars of modern molecular biology: The identification of deoxyribonucleic acid (DNA) as genetic material, a description of the physical organization of DNA through X-ray crystallography, the deduction of the principles of base pairing and the strategy of replication from the organization of the double helix, and the deciphering of the genetic code as triplets chosen from a set of four nucleotides [6].
Delbück and his phage school were important, but there were, in fact, multiple intellectual lineages connected with physics that helped to create the modern world of molecular biology. For instance, Warren Weaver was a mathematical physicist turned science administrator, who in 1932 first used the term "molecular biology". He chose that phrase because he foresaw "that the moment would arrive when the distinction between chemistry and physics and even mathematics on the one hand and biology on the other would be so illusory and in fact so unfortunate" that he did not want to use the word "biology" to describe the programs he was supporting at Rockefeller foundation [6].
British scientists with a strong physical bent, such as Astbury, Bragg and others, used X-ray diffraction to study the organization of fibers of many kinds, mainly proteins found in textiles. This lead to the crystallography of DNA by **Maurice Wilkins** and **Rosaline Franklin**, and ultimately to the first model of the double helix by the physicist **Francis Crick** and the biologist **Thomas Watson** (Nobel prize in Medicine 1962).
In the last quarter of the 20th century, the physicist **Erwin Neher** in collaboration with **Bert Sakmann**, a medical doctor, revolutionized our understanding of the electrical excitability of cells by introducing the patch-clamp technique that allows the recording of minute ion currents flowing through single ion channels in the cell membrane of virtually every cell. In 1991, Neher and Sackmann received the Nobel prize in Physiology for their ground-breaking achievements.

5 Bridging the Culture Gap

5.1 What Physics can do for Biology

Although physics has been important for biology, the notion that physics directed the development of biology by providing the basic principles does not withstand critical examination. Certainly, physical laws also apply to the molecules of life and physicists — in particular their principles, methods, and instrumentation — have been illuminating our views of virtually every living thing. However, not every discipline that uses a laser diode is necessarily driven by physics. In fact, the breathtaking advance of molecular biology in the past 50 years was a genuinely biological accomplishment. The tools as well as the emerging concepts and models were ingenious but simple. Although physicists like Max Delbrück contributed their analytical skills to the birth of modern molecular biology, they were not really doing physics. Many of them

did not use computers, imaging tools or other physical gadgets. Delbrück and his colleague Salvador Luria were counting virus infections by hand and eye, just like any other biologist.

Biological Function and Stamp Collection — The essence of physics is to extract the laws of nature from bewildering riches of observations and to use mathematics to represent it. For the soft-matter physicist, the chemical details of a polymer are less important; what matters is that all polymers are long flexible chains that share common features. In contrast, biologists are interested in the structural and functional details of a molecule, e.g. a strand of DNA or a protein. They strive to elucidate the smallest detail that sets apart one protein from another related protein. The passion to turn up every stone on the ground is fueled by the gratifying assurance that something interesting is discovered below.

The most important concept in biology is that of *function* of a molecule inside a cell. Proteins serve as enzymes that catalyze thousands of different chemical reactions, proteins serve a plethora of different structural functions, proteins serve as adaptor molecules that organize macromolecular complexes etc. etc. The list of distinct functions is almost as long as the list of molecules which take on these cellular responsibilities. To some spectators, the hunting for new molecules serving new functions has become an exercise in molecular stamp-collecting.

The most fundamental insights into the function of a molecule often did not require a whole lot of systematic quantification and mathematical equations. The essence of a molecule's function could be stated in simple words. Thus, biologists seemingly do not rely so heavily on mathematics and quantification. Instead they construct diagrams and pathways to describe the function of living things. The word *function* does not exist in the vocabulary of physicists. However, it seems important for us in order to bridge the cultural gap that physicists learn about function; otherwise, they may be playing in a sandbox all by themselves.

What is a Model in Biology? — However, the advent of molecular biology and the identification of hundreds of thousands of different genes and proteins, has transformed biology; now, different issues are on the agenda of the biological sciences. We will develop this idea further along the following example of a typical "biological" model.

The egg and sperm meet through a process called chemotaxis. The egg releases a chemical substance that forms a chemical gradient surrounding the egg. The sperm utilizes cues of the chemo-attractant gradient to reorient its swimming behavior and thereby locate the egg. During the past 30 years or so, a wealth of experimental findings has resulted in a detailed model of the cellular and physiological reactions underlying chemotaxis, see Fig. 2. The model is by no means unique and cohorts of similar models can be found in the biological literature. For the sake of the argument it is not necessary to understand the model and its constituents in each detail. The important point is that a chemo-attractant binds to a receptor protein on the surface of the sperm. Binding initiates a series of cellular events that eventually cause the influx of Ca^{2+} from the outside. The increase of the Ca^{2+} concentration inside the cell somehow changes the beating pattern of the sperm tail and thereby reorients the swimming trajectory. At birth, the model was simple; the ingredients were a receptor and its ligand. But over the years, new components and new interactions were slotted into ever more complex biochemical pathways. Each amendment resulted from a unique set of experiments that seemed to suggest that a particular molecule or biochemical reaction was involved in chemotaxis. At first, the model was assembled from a series of consecutive reactions. Later, feedback loops and the "crosstalk" between different cellular signaling pathways (cyclic AMP versus cyclic GMP) were incorporated. However, recent evidence suggests that the model, except for the first step, may be not correct [9, 10]. What went wrong? The analysis of the failure reveals

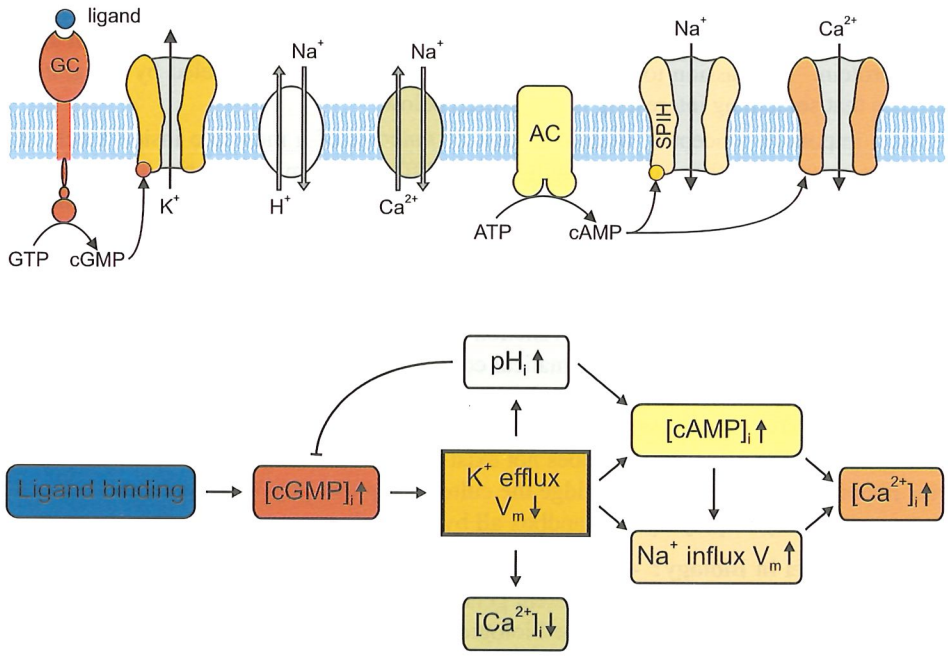


Fig. 2: *Model of the cellular pathway involved in chemotactic signaling of sperm (from Ref. [7]).* **Top:** *molecular components;* **Bottom:** *sequence of cellular events. Binding of peptides to a receptor guanylyl cyclase (GC) increases the intracellular cGMP concentration. The sperm hyperpolarizes by opening cGMP-regulated K^+ channels. Two transporters — a Na^+/H^+ and a Na^+/Ca^{2+} exchanger — export protons and Ca^{2+}, respectively, in response to the hyperpolarization; accordingly pH_i rises and $[Ca^{2+}]_i$ falls. The rise in pH_i activates an adenylyl cyclase. The increase in cAMP concentration causes an influx of Ca^{2+}. In principle, cAMP can stimulate the Ca^{2+} entry by two different mechanisms. First, the hyperpolarization and cAMP activate HCN channels (SPIH; see Ref. [8]); the ensuing depolarization of the cell activates voltage-dependent Ca^{2+} channels. Second, cAMP could activate a Ca^{2+} channel directly. From Ref. [9].*

specific weaknesses in biological thinking and reasoning and highlights where physicists, and also chemists for that matter, can help.

It was known from the work of Howard Berg, a well-known physicist, that bacteria can sense chemical gradients in less than a hundred milliseconds and can respond accordingly [11]. It is plausible that sensing of chemical gradients by sperm also proceeds on a millisecond time scale. Despite this insight, experiments in the past 30 years were lacking the time resolution required to delineate the sequence of cellular events inside sperm. Thus, the experimental support for the arrows that signify a causal relationship in the model was at best weak. Moreover, stimulation of cells by light, chemicals, or mechanical forces often proceeds in two stages: The initial excitation of the cell is followed by an array of cellular reactions that adjust the cell's sensitivity to the stimulus, collectively called adaptation. Because the rapid primary reaction of sperm was missed in those early experiments, relatively high concentrations of the chemo-attractant were required to observe a cellular reaction. In fact, many studies used concentrations that were several orders of magnitude larger than the concentration needed to saturate the primary reaction, i.e. the chemo-attractant concentrations were unphysiological. When the kinetics of several cellular reactions had been determined by time-resolving techniques, it was discovered that sperm can respond to binding of a single molecule [9].

The *first résumé* is that models often serve a different purpose in biology than in physics. For most biologists, models provide frameworks to incorporate the ever increasing stream of new observations rather than hypotheses that should be subjected to a critical test. Often, these models do not allow deriving a set of quantitative predictions that can be tested experimentally. This practice may survive as long as experimental results are of qualitative rather than of quantitative nature. The remedy is obvious: we need to establish a culture of quantitative thinking. How much, how fast, where and when?

The Art of Experimentation — This brings us to a related problem — the art of experimentation. Physicists are trained in devising new methods and building new instruments. They are familiar with the physical principles underlying a measurement; they know the strengths and limits of their physical devices; and they are aware of potential experimental artifacts. It is part of the physics culture to recognize failing components of an instrument, and — where possible — to take apart the instrument and to repair it. This culture, although existent in some areas, is less developed in the biological sciences. Biologists are less acquainted with physical methods and are more hesitant to open up an instrument, leaving alone to repair it. In recent years, biological research has been rearmed with new physical methods. Fancy electro-physiological recordings combined with a multitude of different spectroscopic and microscopic techniques, and a perplexing armamentarium of chemical tools to label and modify bio-molecules has become a biologist's dream. Yet, are biologists sufficiently well prepared to utilize these novel techniques? Do they know about the opportunities and limitations of these techniques? Are biologists in a position to critically evaluate and analyze the results?

Another weakness in the model of Fig.2 may illustrate the point. In the model, a change in membrane potential represents a key event. This change is brought about by the opening of ion channels in the membrane envelope; ions can traverse membranes by flowing through microscopic pores in the center of ion channels. The ion flow redistributes the charge on both sides of the membrane and thereby changes the membrane potential. The change in membrane potential has been monitored with fluorescent dyes that partition into the membrane phase and distribute between inside and outside. The distribution equilibrium and the fluorescence quantum yield of the dye change with the membrane potential and give rise to small fluorescence signals. However, the redistribution of dye molecules across membranes is inherently slow,

taking many seconds, whereas the gating of the channel and the redistribution of ions is likely to be completed within a few milliseconds. Moreover the signals have been recorded in the presence of unphysiological ion gradients across the cell envelope. Therefore, the origin of the fluorescence signals is obscure. Notwithstanding these flaws, the notion that the membrane potential of sperm becomes first more negative and than more positive has been the quintessential constituent of all models of sperm chemotaxis. The question might be appropriate: In a physics culture, would have researchers used an inherently slow detection principle to record an inherently rapid process?

The *second résumé* is that we must become concerned about how students should be trained in universities and in graduate studies to pursue biological problems over the next 30 years. Preparing students for their future work using exceedingly complex and sensitive physical methods will be an essential part of any new curriculum in biology. We recognize that this plea for the art of experimentation could give rise to misinterpretations. We do not want to portray biologists as simple-minded "beetle counters" and physicists as merely the developers of tools that enable biologists to resolve the really important scientific riddles.

Towards a "Chemistry and Physics" of the Cell — In the past 30 years, mainly through efforts to identify the genes and proteins that control the cell response to external stimuli, in particular hormones, researchers have constructed many hundreds so-called signaling pathways similar to that shown in Fig. 2. While most investigators acknowledge that the often linear pathways are oversimplified, it is not clear how to depart from here. For example, the pathways have unidentified components. Is it, therefore, just a matter of stamp-collecting more molecules? The pathways branch and information is certainly flowing between, not just along, the several pathways. Finally, the pathways are regulated in complicated ways through feedback and feedforeward mechanisms on time scales ranging from milliseconds to many hours or even days and on length scales of nanometers to many hundreds of micrometers. Biologists are faced with a flood of facts that need an explanation. Turning up still more details and facts will worsen rather than resolve the problem. What has been all too rare in biology is the symbiosis between theory and experiment, whereas most physicists bring a strong background in theory and the modeling of complexity.

Thus, the *third résumé* is that biology has become a science that requires more mathematical and physical analysis than biologists have been accustomed to. It is this crossroad between analytical thinking, computational models, and theory where physicists can help. What needs to be established is the "chemistry and physics of cellular systems".

5.2 Bringing physics, not just physicists, to biology and medicine

This change was clearly foreshadowed in Max Delbrück's lecture in 1949 in Connecticut. He first describes his awe at the complexity of biology: "The closer one looks at [the] performances of matter in living organisms the more impressive the show becomes. The meanest living cell becomes a magic puzzle box full of elaborate and changing molecules, and far outstrips all chemical laboratories of man in the skill of organic synthesis...".

But Delbrück also sounds a warning: "Biology is a very interesting field ... [because of] the vastness of its structure and the extraordinary variety of strange facts ... but to the physicist it is also a depressing subject, because ... the analysis seems to have stalled around in a semi-descriptive manner without noticeably progressing towards a radical physical explanation ... we are not yet at the point where we are presented with clear paradoxes and this will not happen until the analysis of the behavior of living cells has been carried out in far greater detail." [6]

In the past 50 years, and especially in the past 20 years, molecular and cell biologists have moved much closer to the "radical physical explanation" of cell behavior that Delbrück was looking for. Certainly, the chemical elements — especially the genes, the ribonucleic acid (RNA), and the proteins — and some of the basic functions come into view. What is lacking is a sense of how these functions are integrated to allow cells to manifest their physiological traits.

In his Physics World article [6], Harold Varmus sees three arenas of biology, where he believes the skills of physicists can be most productively used:

- Methods are now available for examining the physical and chemical properties of **single macromolecules** and single complexes of large molecules. These advances are important because they avoid the need to synchronize a population of molecules to measure function. Several of these methods have been developed [12–16]. A more detailed explanation of the measurement of the behavior of single molecules or small ensembles of molecules, so called signaling complexes, can be found in Chaps. B.7 and D.8 of this book.

- The **computational experience** of physical scientists is needed to help interpret complex data sets and the process of "gene expression". One of the consequences of the project to sequence the genomes of humans and many other species is the opportunity to understand the process by which the genes of an organism are expressed. New methods, called "expression arrays", build on the availability of a piece of DNA from each gene, and allow the measurement of the extent to which genes are read to form RNA and subsequently proteins in different tissue and under different environmental conditions. Some progress has been made through a computer-based "cluster analysis" [17] to begin to interpret the voluminous data that such experiments generate, but biologists are generally unused to such complex data sets.

- Through efforts to identify the genes and proteins that control cell growth and responses to hormones, many so-called **signaling pathways** have been constructed that link molecular interactions at the cell surface to changes inside the cell [18, 19]. It will be a formidable task to describe every one of these signaling pathways and their interactions on both a molecular but also cellular, i.e. functional level in precise chemical and physical terms.

5.3 Biological Relevance — Biologically Inspired Physics

Biologists and physicists tend to ask different questions when presented with the same biological system. For example, phosphate groups are often added or removed from proteins to modulate their activity. Biologists will typically be interested which proteins are switched on or off in this way. But physicists might ask the more general question: why phosphate, as opposed to some other chemical group? [4] Another example is the "purple membrane", a two-dimensional crystalline phase of the membrane protein bacteriorhodopsin (compare Sec. 6.1 below). For a biologist, the protein crystal is the basis for doing X-ray crystallography to obtain detailed information about the protein structure and function [20]. For a physicist, the purple membrane is a very nice model system to study fluctuations and melting of membranes with shear stiffness [21].

By asking different questions, physicists can bring fresh and useful insights to biology. However, there is always a risk of wasting time on issues that have no biological meaning. Not every feature of a biological system has a functional importance, nor is every biological system likely to operate with optimal efficiency. Evolution does not chose the best option, just the best

available option. Physicists stumbling into this minefield risk wasting their time on apparently interesting but ultimately biologically unimportant problems [4].
Hans Frauenfelder, the well-known nuclear physicist who began studying proteins three decades ago, is unfazed by such concerns. He is interested in the oxygen-transport protein myoglobin as a system in which to study the behavior of complex matter. Frauenfelder developed the concept of the "energy landscape" , in which valleys represent the stable forms of the protein, and hills are barriers to changing shapes. He likes to quote Stanislaw Ulam, a mathematician who worked together with Enrico Fermi and helped to develop the hydrogen bomb, who once said to him: "Ask not what physics can do for biology, ask what biology can do for physics" [4].
Indeed, in recent years a new branch of physics and biophysics has developed, which is called *Biological Physics* or *Biologically Inspired Physics*. This field is distinguished from classical Biophysics in the sense that the physical aspects of biological systems are emphasized more strongly. This implies that simple model systems are investigated experimentally and theoretically, and detailed quantitative studies of biological systems are performed. The systems are studied to an essential part for their interesting physical properties. This must not lead immediately to new biological insight. However, there is of course the hope that this curiosity-driven research will ultimately benefit the understanding of biological processes.

6 From Soft Matter to Cell Biology

No area of physics seems to be more relevant to molecular and cell biology than Soft Condensed Matter. After all, living things are made of complex fluids — liquids that contain mesoscopic structures with length scales of several nanometers to a few micrometers — and for decades, physicists have been studying complex fluids such as colloids, polymers, liquid crystals and solutions containing amphiphilic molecules [22,23]. *Colloids* are small solid particles, typically with a diameter in the micrometer range. Most classical colloidal particles are hard spheres, such as silica or latex spheres, but colloidal hard rods have also been investigated. *Polymers* are long chain molecules, in which the same molecular building block, the monomer, is repeated several thousand times. *Amphiphiles* are molecules with a two-sided architecture, where one side, the polar head, is hydrophilic and prefers a polar solvent like water, while the other side, a hydrocarbon tail, is hydrophobic and wants to avoid water contact. Therefore, amphiphilic molecules self-assemble at water-oil interfaces to form monolayers, and in pure water to form bilayers, where the contact of the tails with water is shielded by the head-group surface.
In recent years, an increasing number of molecules have been investigated, for which the classification into colloids, polymers and amphiphiles fails. This is shown schematically in Fig. 3. Consider, for example, the left-hand side of the triangle. As colloidal rods get longer and longer, their flexibility becomes more and more important. The fd-virus, for example, with a length of 880 nm and a diameter of 6.6 nm displays pronounced thermal fluctuations of its shape. Its persistence length — the length over which the molecule points roughly in the same direction — is 2200 nm, much larger than its length. Even more flexible is DNA, which is a semiflexible polymer with a persistence length of about 50 nm. Finally, we reach the triangle apex of the classical synthetic polymers like polyethylene, which have a very short persistence length on the order of the monomer size, and are therefore very flexible. The preferred state of a flexible polymer is a random coil, because this is the state of maximum entropy.
In fact, soft condensed matter physics has been fueled, in part, by advances in synthetic chemistry. These developments have led to the kind of well-characterized materials which physicists

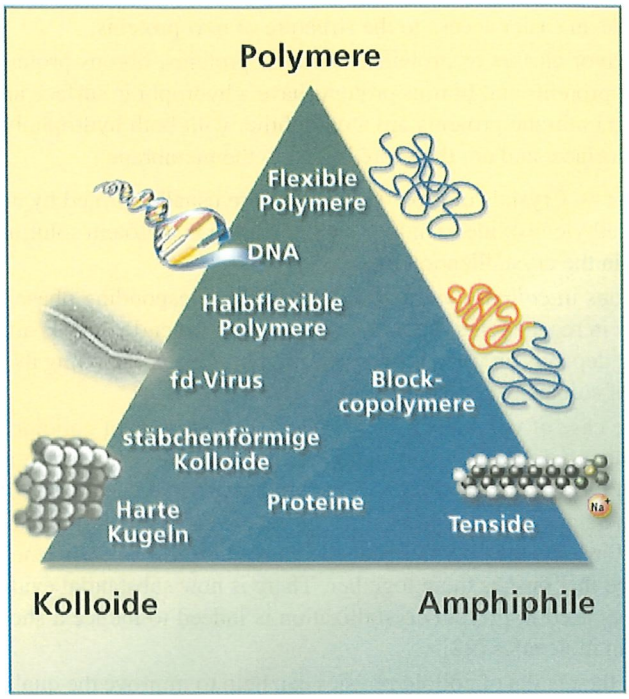


Fig. 3: *Soft matter systems with a single macromolecular component have different polymeric, colloidal and amphiphilic character. Biological macromolecules such as DNA and proteins, as well as viruses, are indicated.*

love to work with. Another important boost has come from the realization that many of the mathematical tools developed for other areas of theoretical physics can be used to describe soft matter. A third essential factor is the array of new experimental techniques that are available for the study of mesoscopic samples.
We want to illustrate the close connection of soft matter physics and cell biology be the following examples.

6.1 Protein Crystallization

The structure of a protein is determined by its amino acid sequence, which is encoded in the genes. This *primary structure* determines in a highly non-trivial way the *secondary structure* of α-helices and β-sheets, which are finally packed together to form the *tertiary structure* of the functional molecule. The protein molecules can perform their biological function only when they fold up in a particular three-dimensional structure. Knowledge of this structure therefore provides the starting point for understanding how proteins work.
A standard technique to determine the structure of proteins is X-ray crystallography. It requires large single crystals of a protein. The larger the crystal, the higher the resolution with which the structure can be determined. The crystallization process is currently an art that is based almost entirely on practical know-how. A better understanding of crystallization is therefore highly

desirable to provide an easier access to the structure of new proteins.

There are three mayor classes of proteins: Globular proteins, fibrous proteins, and membrane proteins. Globular proteins and fibrous proteins have a hydrophilic surface and are found in the cytoplasm, while membrane proteins are amphiphilic, with both hydrophobic and hydrophilic domains on their surface, and are therefore bound to the membrane.

Globular Proteins — Crystals of globular proteins are usually formed by adding salt or polymers, such as polyethylene oxide to induce crystallization in a protein solution. What is the role of these additives in the crystallization process?

Effective interactions in *colloidal suspensions* and the corresponding phase behavior has been studied intensively in recent years. In such a system, the suspended solid particles interact via a potential $U(r)$ that depends on various properties of the solvent, including its dielectric constant and the presence of solutes like ions and polymers.

Let us consider the case of polymers in more detail. In a mixture of colloidal spheres of radius R and inert polymer chains with a coil radius R_g, the polymer can lead to an effective attractive force between the colloidal particles. This attractive force is called *depletion interaction*, because the polymers are excluded from the region between the colloids when their separation is less than $2R_g$. However, the polymers outside this region still bombard the colloidal particles to create a net force that pushes them together. There is now substantial evidence that the main role of the additives used in protein crystallization is indeed to induce a short-range attraction between the protein molecules [22].

More importantly, the results of colloid physics can help to improve the quality of protein crystals. Obviously, a sufficiently strong attraction is necessary in order to induce crystallization. On the other hand, if the attraction is too strong, crystallization starts at several different nucleation centers and proceeds very rapidly. Both effects lead to crystals with a large number of defects, which are ill-suited for diffraction analysis. Colloidal physics suggests that there is an optimal choice for the interaction [24]. A study of the crystallization kinetics by molecular dynamics simulations showed that good crystals only form when the interaction is sufficiently weak that no there is no coexistence between a very dilute and a dense (fluid) protein solution, but sufficiently strong that such a coexistence would appear with a slight increase of the interaction strength [24]. In this case, the crystal nucleus is expected to be surrounded by a layer of the fluid-like protein phase, which lowers the energy barrier for nucleation.

Membrane Proteins — The crystallization of membrane proteins is even more difficult than of globular proteins. The reason is that because of their amphiphilic architecture, a local membrane environment has to be provided. Therefore, the first structure of a trans-membrane protein, the photosynthetic reaction center of the bacterium *Rhodopseudomonas viridis*, a light-driven photon pump, could only be determined in 1985 by J. Deisenhofer, R. Huber and H. Michel (Nobel prize in Chemistry 1988).

Membrane proteins have to be crystallized in the lamellar phase. Amphiphilic bilayers, which contain the proteins, form a regular stack of planar membranes, which are all oriented in the same direction. The proteins crystallize within each membrane, but also interact sufficiently strongly between the layers to from a three-dimensional crystal structure.

However, hydrophobic surfaces and anisotropic orientation of membrane proteins are serious obstacles in producing well-ordered crystals. A possibility of improving this situation has been suggested recently on the basis of the phase behavior of amphiphilic systems [25]. Phase diagrams of lipid-water mixtures have been studied intensively in recent years [26,27]. In particular, it was found that in addition to the lamellar phase, other phases are stable, such as the cubic

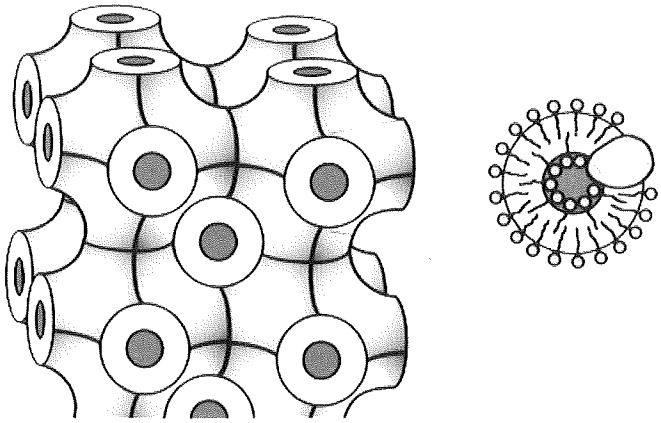


Fig. 4: *Schematic model of a bicontinuous cubic phase composed of monoolein, water, and a membrane protein. The matrix consists of two compartments, a membrane system with an infinite three-dimensional periodic minimal surface (Left), interpenetrated by a system of continuous aqueous channels (shown in black). The enlarged section (Right) shows the curved lipid bilayer (with an inserted membrane protein molecule) enveloping a water conduit. In a cubic phase consisting of 60-70% (wt/wt) monoolein or monopalmitolein and water, hydrophobic proteins diffuse laterally in the bilayer, while water-soluble components diffuse freely through the intercommunicating aqueous channel system. From Ref. [25].*

bicontinuous phase shown in Fig. 4. The advantage of a bicontinuous phase is that all parts of the membrane are connected, so that it provides a very flexible matrix for facilitating crystal nucleation and growth. When the cubic and the lamellar phase are nearly in coexistence, the cubic phase acts as a reservoir for both the protein and the lipid, which can diffuse laterally in the bilayer [25].
This method has been used, for example, to obtain large crystals of bacteriorhodopsin, a light-driven proton pump of the *Halobacterium salinarum*, both in the ground state [25] and in an activated intermediate state [20]. Within the membrane, bacteriorhodopsin forms trimers, which are packed into a hexagonal array, see Fig. 5.

6.2 Packing DNA into Viruses

DNA is a negatively charged molecule with a persistence length — the length over which the molecule behaves effectively as a weakly bendable rod — of about 50 nm. Therefore, squeezing a 6.6 μm long double-stranded DNA chain, which is the genetic material of a virus like the bacteriophage ϕ29, into the viral capsid with diameters of only 42 and 54 nm, obviously is a difficult task [29]. DNA does not spontaneously pack into such a small volume; instead, it is compressed into the viral shell by a molecular motor, see Fig. 6.
Biomolecular motors are proteins that undergo shape changes to generate force or torque. Acting like tiny engines, biomolecular motors come in a wide assortment and perform a broad range of tasks, many involving movement and transportation. One such task is the packing of the coiled lengths of DNA into the protective external shell or "capsid" of a number of viruses

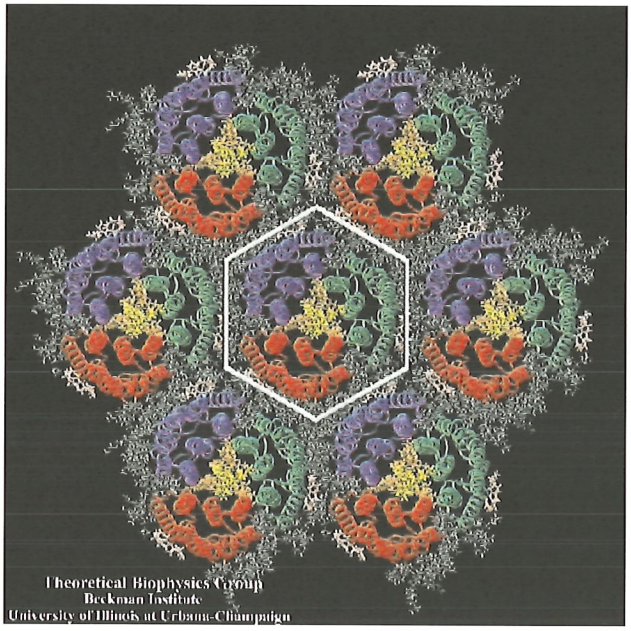


Fig. 5: *Two-dimensional crystal of bacteriorhodopsin. The protein molecules consist of seven transmembrane α-helices and are organized into 44 Å diameter trimers which form a regular hexagonal lattice in the membrane plane. From Ref. [28].*

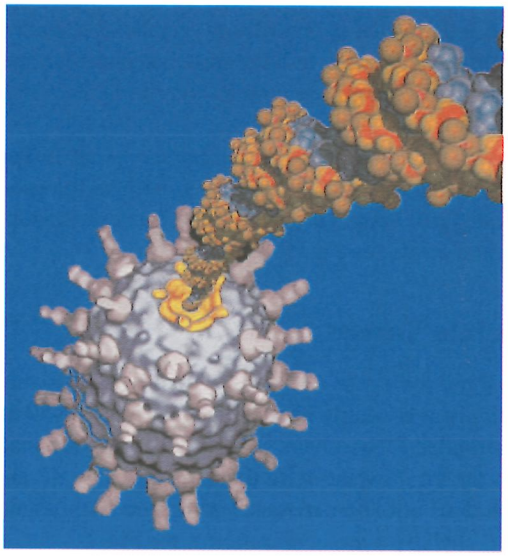


Fig. 6: *The biomolecular portal motor of the bacteriophage $\phi 29$ (yellow) compresses DNA into the viral capsid. From Ref. [30].*

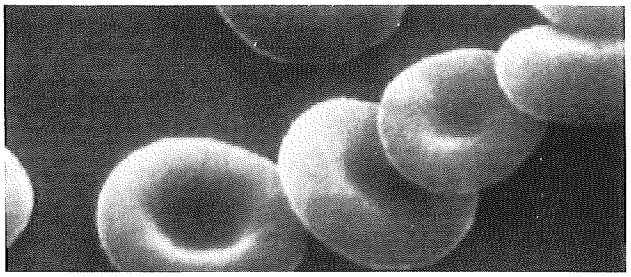


Fig. 7: *Red blood cells (erythrocytes) have a characteristic, discocyte shape.*

including those that cause herpes, chicken pox and shingles. The force of a portal motor was measured for the first time in 2001 for the bacteriophage $\phi 29$, a virus that infects and destroys soil bacteria and is considered an excellent model system for studying viral assembly.
"The portal motor of bacteriophage $\phi 29$ compresses the DNA to 6000 times its normal volume" says Carlos Bustamante, head of the lab in which the motor force was measured. "This generates an internal pressure of about 60 atmospheres, which is about ten times that in a champagne bottle". This internal pressure inside the bacteriophage's capsid forcibly injects the viral DNA into the infected cell, when the virus attaches itself to the soil bacterium [31].
The strength of the bacteriophage $\phi 29$ portal motor was measured with the help of an "optical tweezer". The unpacked end of DNA of a partially packed capsid as well as the capsid itself were attached to a pair of micron-sized polystyrene beads [32]. While the capsid-attached bead was held in place by a micro-pipette, the DNA-attached bead was captured by the optical tweezer — a laser beam that can be used to grasp and move the beads.
In the presence of adenosine triphosphate (ATP), the fuel that powers many biomolecular motors, the packaging of the viral DNA can be observed in this way in real time, and the force being applied by the motor can be measured. The packaging rate has also been determined. With about 100 base pairs per second, the packaging rate is maximal when the capsid is nearly empty, and continuously decreases until it reaches zero when 105% of the genome is packaged [32].

6.3 Micromechanics of the Cell

In order to understand the properties of the cell, different approaches have to be used. One approach, which is pursued by biologists, is to modify or knock-out single proteins, and to study the change of the cell behavior as as result of this protein modification. Another, more physical approach is to investigate the material properties of various cellular components, either in a simplified model system or in the living cell [33].
The cell consists of three subsystems with very different mechanical properties: the shell, the cytoplasm, and the nucleus. The shell is made from a lipid bilayer containing proteins — the plasma membrane —, together with an associated macromolecular network — the actin cortex —, which is part of the cytoskeleton. The cytoplasm is a densely packed colloidal suspension, which is interlaced with a macromolecular network of rod-like microtubules and actin filaments.
The mechanical properties of *lipid-bilayer membranes* have been studied intensively in recent years. It has been shown that the elastic and thermal properties can be understood very well by a

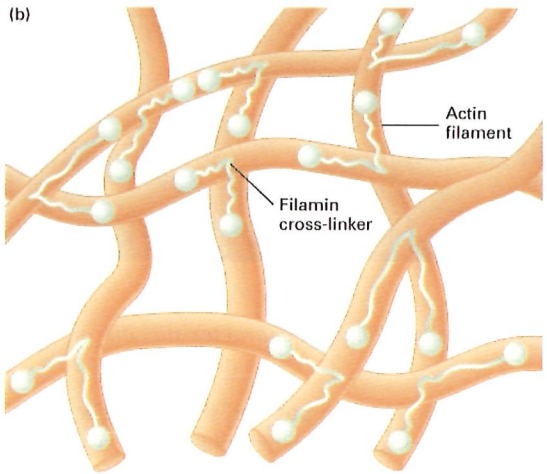


Fig. 8: *In the actin network of the cytoskeleton, the persistence length is typically larger than the mesh size. The actin filaments are semiflexible polymers, which can be crosslinked by motor proteins. From Ref. [34].*

theory based on the curvature elasticity of thin shells. For example, the characteristic, discocyte shape of a red blood cell, see Fig. 7, can be obtained by minimizing the curvature energy with the constraints of fixed membrane area and a particular fixed enclosed volume.

The actin cortex of the cytoskeleton is a network of semiflexible polymers, see Fig. 8. The general viscoelastic properties of this network are not unlike the properties of melts of synthetic polymers (compare Chap. C.2). The time dependence of the shear modulus shows three regimes [35]: (i) a short-time regime, in which the modulus is determined by the entropic tension of the macromolecules, (ii) the plateau regime, where the modulus is independent of time (rubber plateau), and (iii) the terminal regime, in which the network behaves like a fluid. The physical origin of the rubber plateau is that the relaxation of the polymer chains perpendicular to its contour is strongly hindered by the entanglements with other polymer chains. Therefore, only a quasi one-dimensional motion is only possible within a tube around the contour, see Fig. 9, which is called "reptation" [35].

However, if the dynamical properties of a network are studied quantitatively, it becomes evident that the shear modulus of a network of semiflexibile polymers like actin within each regime is governed by different power laws — as a function of time, temperature, density, and persistence length — than networks of synthetic, flexible polymers [33, 37].

Another interesting property of actin networks is that they are crosslinked by motor protein like myosin II, see Fig. 8. If the motors are passive, they behave as any other chemical crosslink. However, when the motors are active, they drive each of the actin filaments in a certain direction [33]. This implies that the relaxation dynamics changes drastically, because the diffusion out of the tube of Fig. 9 is now replaced by a driven motion, which proceeds much faster.

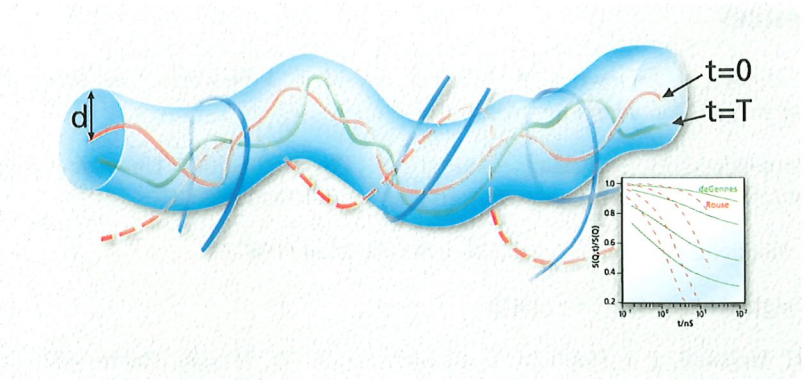


Fig. 9: *Polymer chains in a melt can move rather easily in a tube along their contour. In contrast, the walls of the tube, which are formed by the neighboring polymers, are nearly impossible to be penetrated. From Ref. [36].*

7 Literature

A number of books have appeared in recent years in the fields of cell biology [34, 38, 39], biochemistry [40, 41], biological physics [42–45], and soft matter physics [46–49], which we strongly recommend for further reading.

References

[1] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman, *The Theory of Critical Phenomena* (Clarendon Press, Oxford, 1992).

[2] Superconductivity and Magnetism group, Physik Institut, Universität Zürich. See http://www.physik.unizh.ch/groups/groupkeller/fluctuations/.

[3] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[4] J. Knight, Nature **419**, 244 (2002).

[5] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, Nature **402**, Suppl. C47 (1999).

[6] H. Varmus, Physics World **September**, 27 (1999).

[7] A. Darszon *et al.*, Dev. Biol. **240**, 1 (2001).

[8] R. G. R. Seifert and U. B. Kaupp, Nature **393**, 583 (1998).

[9] U. B. Kaupp *et al.*, Nature Cell Biology **5**, 109 (2003).

[10] J. C. Kirkman-Brown, K. A. Sutton, and H. M. Florman, Nature Cell Biology **5**, 93 (2003).

[11] H. C. Berg and E. M. Purcell, Biophys. J. **20**, 193 (1977).

[12] A. D. Mehta *et al.*, Science **283**, 1689 (1999).

[13] W. E. Moerner and M. Orrit, Science **283**, 1670 (1999).

[14] S. Weiss, Science **283**, 1676 (1999).

[15] C. Bustamante, S. B. Smith, J. Liphardt, and D. Smith, Curr. Opin. Struct. Biol. **10**, 279 (2000).

[16] C. Bustamante, Z. Bryant, and S. B. Smith, Nature **421**, 423 (2003).

[17] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. USA **95**, 14863 (1998).

[18] U. S. Bhalla and R. Iyengar, Science **283**, 381 (1999).

[19] G. Weng, U. S. Bhalla, and R. Iyengar, Science **284**, 92 (1999).

[20] H. J. Sass *et al.*, Nature **406**, 649 (2000).

[21] I. Koltover *et al.*, Phys. Rev. Lett. **82**, 3184 (1999).

[22] W. Poon, T. McLeish, and A. Donald, Physics World **May**, 33 (2001).

[23] G. Gompper, J. K. G. Dhont, and D. Richter, Physik in unserer Zeit **Januar**, 12 (2003).

[24] P. R. ten Wolde and D. Frenkel, Science **277**, 1975 (1997).

[25] E. M. Landau and J. P. Rosenbusch, Proc. Nat. Acad. Sci. USA **93**, 14532 (1996).

[26] J. M. Seddon, Biochim. Biophys. Acta **1031**, 1 (1990).

[27] U. S. Schwarz and G. Gompper, in *Spatial Statistics and Statistical Physics*, Vol. 600 of *Lecture Notes in Physics*, edited by K. R. Mecke and D. Stoyan (Springer, Berlin, 2002), pp. 107–151.

[28] Theoretical Biophysics Group, Beckman Institute, University of Illinois at Urbana-Champain. See http://www.ks.uiuc.edu/Resarch/newbr/.

[29] Y. Tao *et al.*, Cell **95**, 431 (1998).

[30] Bustamante Lab, Department of Molecular and Cell Biology and Department of Physics, University of California, Berkeley. See http://alice.berkeley.edu/.

[31] "ScienceBeat", Lawrence Berkeley National Laboratory. See http://www.lbl.gov/Science-Articles/Archive/infectCells.html.

[32] D. E. Smith *et al.*, Nature **413**, 748 (2001).

[33] E. Sackmann, Physik Journal **Februar**, 35 (2004).

[34] H. Lodish *et al.*, *Molecular Cell Biology*, 4th ed. (Freeman, New York, 2000).

[35] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, Oxford, 1986).

[36] G. Gompper, J. K. G. Dhont, and D. Richter, Physik in unserer Zeit **Januar**, 19 (2003).

[37] E. Frey, Phys. Blätter **Februar**, 63 (2001).

[38] B. Hille, *Ion Channels of Excitable Membranes* (Sinauer, Sunderland, MA, 2001).

[39] B. Alberts *et al.*, *Molecular Biology of the Cell*, 4th ed. (Garland, New York, 2002).

[40] D. Voet and J. G. Voet, *Biochemistry*, 2nd ed. (Wiley, New York, 1995).

[41] K. E. van Holde, W. C. Johnson, and P. S. Ho, *Principles of physical biochemistry* (Prentice Hall, Upper Saddle River, NJ, 1998).

[42] *Structure and dynamics of membranes - from cells to vesicles*, Vol. 1 of *Handbook of biological physics*, edited by R. Lipowsky and E. Sackmann (Elsevier, Amsterdam, 1995).

[43] J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland, MA, 2001).

[44] D. H. Boal, *Mechanics of the Cell* (Cambridge University Press, Cambridge, 2002).

[45] P. Nelson, *Biological Physics* (Freeman, New York, 2004).

[46] S. A. Safran, *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* (Addison-Wesley, Reading, MA, 1994).

[47] *Micelles, Membranes, Microemulsions, and Monolayers*, edited by W. M. Gelbart, A. Ben-Shaul, and D. Roux (Springer-Verlag, Berlin, 1995).

[48] *Soft Matter — Complex Materials on Mesoscopic Scales*, Vol. 10 of *Matter and Materials*, edited by J. K. G. Dhont, G. Gompper, and D. Richter (Forschungszentrum Jülich, 2002).

[49] *Statistical Mechanics of Membranes and Surfaces*, 2nd ed., edited by D. Nelson, T. Piran, and S. Weinberg (World Scientific, Singapore, 2004).

GENERAL CONCEPTS



A 1 Statistical Mechanics

G. M. Schütz
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Statistical Mechanics **2**
1.1 What is Statistical Mechanics? 2
1.2 Equilibrium statistical mechanics 4
1.3 Fluctuations in equilibrium 17
1.4 Nonequilibrium systems 23
1.5 Conclusions 32

A Some combinatorics: The degeneracy of microstates **33**

B Derivation of the fluctuation-dissipation theorem **33**

C Solution of the lattice random walk equation **35**

1 Statistical Mechanics

1.1 What is Statistical Mechanics?

It is known for more than three centuries that physical properties of macroscopic objects can be described in terms of quantities such as pressure or temperature which can be linked quantitatively to basic properties like volume and different kinds of energy, in particular heat and mechanical energy. As technology and science dealing with these properties developed, their treatment grew into an autonomous branch of physics, named thermodynamics. It was realized that a unifying framework for all known laws of thermodynamics could be set up on the basis of the three fundamental laws. These encompass

(1) energy conservation
(which appears in an non-obvious manner, e.g., when the heat and mechanical energy added to a system equals its increase in internal energy)

(2) entropy increase
(which essentially means that in an isolated macroscopic system only such processes are observed which correspond to a state of highest probability)

(3) positivity of temperature

(which asserts that the lowest possible temperature $T = 0$ cannot be reached).

These are empirical laws which have proved to be of very general validity – independent of any specific properties of the material they describe – as long as the underlying assumptions, in particular the absence of a flow of energy through the system apply. A system which has reached its state of highest entropy, i.e., its most likely state, is said to be a system in thermal equilibrium.

While the success of classical thermodynamics in the 19th and early 20th century was tremendous in the field of applied science, the emergence and establishment of the atomistic view of matter left the foundations of thermodynamics in a rather unsatisfactory state. If ultimately any macroscopic object consists of small particles, moving according to the laws of quantum mechanics or newtonian mechanics, and hence determine in which state a system is found, then the basic laws of thermodynamics cannot be fundamental laws of nature. There must be some way of *deriving* these laws which govern the physical properties of macroscopic objects from the basic laws describing microscopic particles. This problem gave rise to the development of statistical mechanics. It employs a statistical description of many-body systems on a microscopic basis in order to elucidate the emergence of the macroscopic laws describing a thermodynamic system as a whole.

Clearly, the same philosophy applies to other areas of macroscopic physics. E.g. one may ask which microscopic interactions between the particles of a fluid give rise to its viscosity, or in the case of solids, to its elasticity. In soft matter systems and in biological systems materials composed of polymers display viscoelasticity, i.e., they behave like solids on short time scale and like fluids on long time scales. Again one may ask which microscopic properties give rise to such behavior. Indeed, the same line of thinking can be extended to dynamical properties. How do e.g. the Navier-Stokes equation for the motion of a fluid emerge from intermolecular interactions? What is the microscopic origin of Brownian motion? How can randomness as observed in Brownian motion arise from deterministic fundamental laws of motion? Why can ensembles of interacting Brownian particles display pattern formation on a macroscopic scale?

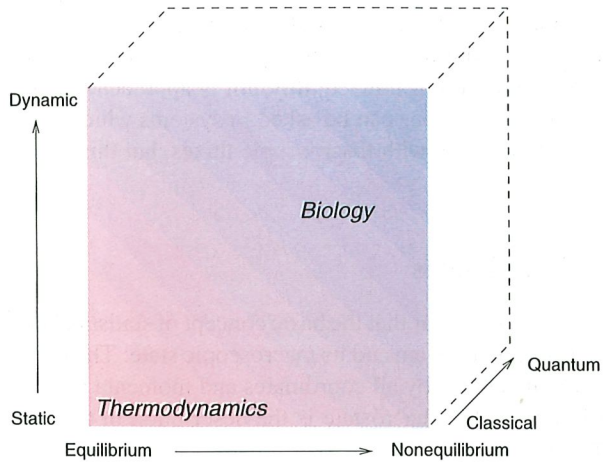


Fig. 1: *Methodically distinct aspects of statistical mechanics. Biological problems are often nonequilibrium phenomena.*

The number of questions of this type that one may ask appears to be unlimited. All fall at least to some extent into the realm of statistical mechanics.

This explains why we need statistical mechanics when biology meets physics. Soft matter – which appears in biological systems as membranes, polymers or complex fluids – is condensed matter consisting of a large number of microscopic particles. Thus statistical mechanics is one of the foundations of soft matter physics and hence enters biology. Moreover, typical thermodynamic phenomena such as phase transitions occur in biological systems. A well-known example is the denaturation of DNA. Dynamical statistical phenomena, in particular diffusion, and pattern formation play a crucial role in various biological systems. This makes the application of ideas from statistical mechanics inevitable.

At this point it is imperative to introduce some general structure into this convoluted set of problems, to give at least some rough sense of orientation. It is useful to make three distinctions (Fig. 1). They are given by the opposites (i) classical vs quantum, (ii) equilibrium vs nonequilibrium, (iii) static vs dynamical. This distinction requires some comments. (i) In this lecture we focus exclusively on classical particles as they are the most relevant for the statistical mechanics of biological systems. Here the basic entities are often long polymers rather than single atoms and their interactions are well approximated by classical laws rather than quantum mechanics which applies to much smaller particles. (ii) From what we discussed above it is clear that not all systems encountered in nature are equilibrium systems, as we explicitly required the absence of any macroscopic current of energy, particles, etc. passing through the system. In fact, much of what see in biological systems is not in equilibrium. Nevertheless we shall focus on equilibrium systems as they serve as a basis from which to start discussing simple nonequilibrium phenomena. Moreover, there is no general framework for nonequilibrium physics comparable to the three laws of thermodynamics and therefore nonequilibrium systems must be discussed on a case-to-case basis. Hence equilibrium systems also serve as a good and well-defined reference point to build theories for systems out of, but not so far from equilibrium. (iii) The third distinction arises from the desire to study also time-dependent phenomena which brings the parameter time into consideration in an explicit fashion. Dynamical properties are almost

completely ignored in thermodynamics or in the study of stationary states of nonequilibrium systems. However, following the time evolution of a system provides important information e.g. on how fluctuations decay in equilibrium or how equilibrium is approached from a state far from equilibrium. Similar dynamical questions can be asked in systems which do not relax to an equilibrium state, but to a stationary state with macroscopic fluxes, but this is not a topic discussed in this lecture.

1.2 Equilibrium statistical mechanics

From the introductory remarks it has become clear that the basic concept of statistical mechanics is the distinction between the microstate of a system and its macroscopic state. The microstate is the description on an atomistic level, it is given by all coordinates and momenta of all particles in the system at a given instant in time.[1] The macrostate is the description in terms of a few macroscopic observables as outlined above.
In principle one could derive all macroscopic equilibrium properties of a system by studying the solutions of Newton's (microscopic) equation of motion. E.g. one would then just add all energies of the individual particles to obtain the macroscopic total internal energy. However, a macroscopic system typically comprises 10^{23} particles (maybe somewhat less in a biological setting). This obviously leads to an unsurmountable problem: There is no way how one could solve 10^{23} coupled differential equations, let alone determine experimentally the initial values which enter the solution. But this is not all. Even if one could, one would not have gained anything in terms of understanding. The solution would just give what one observes anyway, but would not explain why the macroscopic behavior of the system can be understood in terms of just a few parameters like temperature or pressure and how these parameters are linked with each other e.g. through the equation of state which makes no reference to microscopic properties. Hence this approach is not only practically impossible for large systems, it is not even desirable from a theoretical point of view. It is conceptually useful only as a substitute for expensive and time-consuming experiments.[2]
Instead, since the number of particles involved is large, one approaches the problem in an entirely different way by investigating the statistical properties of ensembles of particles. Given the values of some external parameters and given some basic and very general assumptions on the dynamics one asks how relations between thermodynamic quantities arise.

The microcanonical ensemble

The conceptually simplest statistical ensemble is the so-called microcanonical ensemble. This describes an isolated system with some volume V which has no exchange of energy U or mass with its environment. Hence it may be specified by the conserved quantities U, V and particle number N. These quantities are called extensive because they are the sum of microscopic quantities and hence by extending a given system their values grow proportionally. Of course, these quantities do not uniquely define the microstate of a given system. There may be many microstates (configurations of particles) with the same values of U, V, N. To illustrate this let us consider a very simple toy system of cubic bricks (Fig. 2).

[1] In a quantum mechanical treatment the analog of a microstate is the full many-body wave function of the system.

[2] Numerically solving Newtons equation of motion with molecular dynamics simulations indeed provide an invaluable tool for some problems involving many, but not too many, particles. One obtains data for which real

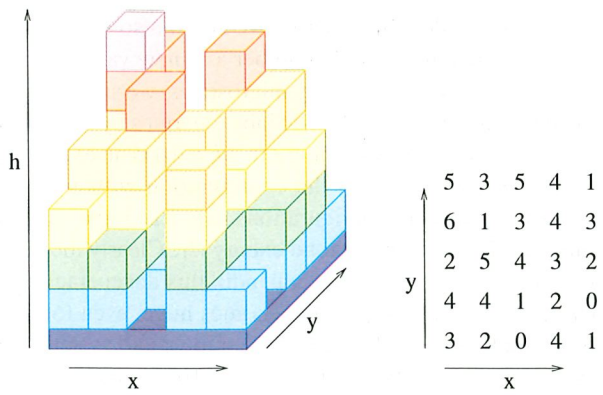


Fig. 2: *Random configuration in the toy model of coins on towers. Each height level is displayed in the same color. The array of numbers shows the height of each tower. The coins are not shown.*

To describe the system, we imagine that we divide an area into a square lattice with N sites. Each square may be covered by a tower of n cubes. Their length is a, their weight we assume to be negligible. (One can imagine these cubes as made from styrofoam.) There is a bottom layer of half cubes, on which these towers rest. The number $n \geq 0$ of cubes is arbitrary. On each tower we put a coin. Due to the gravitational potential a coin on a tower of n bricks is located at a height $h_n = (n + 1/2)a$ and therefore has a potential energy $\epsilon_n = mgh_n = (n + 1/2)\epsilon_0$. The quantity $\epsilon_0 = mga$ is a basic energy scale of our system, depending on the weight mg of the coin and the length a of the cubes. The length a is a basic length scale.

To make contact with statistical mechanics we first remark that the quantities ϵ_0 and a are the analog of the microscopic parameters of a real system. They can be used as natural scales for the quantities length and energy.

Since we use the styrofoam cubes only as a device to give potential energy to the coins, they do not play a further role in our considerations. The system can therefore be viewed as an effectively two-dimensional array of coins with different potential energies. The volume of the system is $V = Na^2$, the area which the system of coins cover. This may be visualized by looking at the system from above.

The coins play the role of particles, they are characterized by their individual energy ϵ_n. A given distribution of cubes corresponds to a microstate of the system. The set of all microstates is the statistical ensemble, called the microcanonical ensemble.

The total internal energy U of the system for a fixed microstate is the sum of all individual energies

$$U = \epsilon_0 \sum_{i=1}^{N} (n_i + 1/2) \equiv \epsilon_0 (M + N/2) \tag{1}$$

where n_i is the number of cubes on lattice site i. The energy is a multiple of the basic energy scale, and is given by the sum M of local cube numbers n_i, which is the total number of cubes, plus a contribution proportional to the volume, coming from the bottom layer. For simplicity we have assumed that each square is covered by a coin. Therefore the total particle number is

experiments are not feasible. This is the subject of the lecture by Vliegenthart, see Sec. B.4.

N and not an independent variable of our model. In principle we could also study the system with a different number of coins and thus treat the particle number as a new variable.
To see that for a given energy U and volume V there are many microstates, we simply enumerate the number of configurations with all possible values of U. We label energies by the number M. We see the lowest possible energy $U_0 = N\epsilon_0/2$ (corresponding to $M = 0$) is an exception from the rule. There is only one configuration with $M = 0$, the configuration without any cubes. For $M = 1$ there are evidently exactly N different configurations, since a single cube may be placed at any of the N different sites, leading to the same total energy U_1. Clearly the number of microstates with given M grows with M. As shown in the appendix this degeneracy $\Gamma(M, N)$ is given by the binomial coefficient C_M^{N+M-1} which quickly becomes huge, even for moderate values of N and M. It grows exponentially in the volume. For the quite small system shown in Fig. 2 one has $M = 70$ and $N = 25$. There are $\Gamma(70, 25) = C_{70}^{94} \approx 10^{22}$ microstates with the same energy!
Notice that in a closed physical system internal energy, volume and particle number are conserved quantities. Indeed, the first law of thermodynamics rests on energy conservation. The basic assumption of statistical physics as applied to the microcanonical ensemble is that *in an equilibrium system each microstate with fixed U, V, N can be found with equal probability.* This probability is the inverse $1/\Gamma$ of the degeneracy.[3] As a measure of this degeneracy one introduces the *entropy*

$$S = k_B \ln \Gamma = S(U, V, N). \tag{2}$$

The Boltzmann factor k_B is a scale factor chosen by convention. The logarithm is taken to make S an extensive quantity as usually (but not by any basic law of nature) Γ grows exponentially in the volume of a system. The assumption that all microstates are equally likely essentially means that a system tries to maximize its entropy, as asserted in the second law of thermodynamics. From an information theoretical point of view it implies maximal "ignorance" concerning which microstate a given system is found in.[4]
The entropy contains all statistical knowledge we have about the system and all thermodynamic properties can be calculated from it. The entropy is (usually) a monotic increasing and convex function of its arguments U, V, N.[5] By construction S is additive for two statistically independent systems,

$$S_{12} = S_1 + S_2 \tag{3}$$

as the number of microstates of two independent systems is the product of the number of states of each individual system. Extensitivity (which is an empirical fact for most systems) implies

$$S(bU, bV, bN) = bS(U, V, N) \tag{4}$$

for a scale factor b. Therefore one may write $S = Vs(u, \rho)$ where

$$u = U/V, \quad \rho = N/V \tag{5}$$

[3]This version of the basic assumption applies to systems with discrete energy levels. For a continuous energy spectrum it has to modified: Each microstate on the infinitesimal energy shell with constant U has equal statistical weight.

[4]There is no reference to any dynamics in the statistical description of the ensemble. In a dynamical picture one assumes that when the system goes to one microstate to another (as a result of the fundamental microscopic laws of motion) it does so in such a manner that after a sufficiently long time each microstate is visited equally often. This assumption is usually referred to as ergodicity hypothesis.

[5]We shall not discuss systems where this does not hold and drop in all future discussion the word "usual".

are the energy density and particle density respectively. A function with the property (4) is said to be a homogeneous function in all its arguments. Since S is convex it may be inverted with respect to its arguments. Specifically one may fully describe the microcanonical ensemble by its energy

$$U = U(S, V, N) \tag{6}$$

as a function of the extensive parameters S, V, N. It is also homogeneous and has the form $U = Vu(s, \rho)$.

We stress that within our description of the microcanonical ensemble there is no notion of temperature yet. In fact, temperature is not an external parameter in this ensemble, but is given by how the system responds to a change of energy. This corresponds to our intuition of temperature; the precise relation is

$$T = \frac{\partial U}{\partial S} = \left[\frac{\partial S}{\partial U}\right]^{-1} \tag{7}$$

where the first equality expresses T as a function of S, V, N while the second equality gives T as a function of U, V, N. Because of the convexity of S the temperature is strictly positive, in agreement with the third law of thermodynamics.

Similarly the pressure is given by

$$p = -\frac{\partial U}{\partial V} = T\frac{\partial S}{\partial V}. \tag{8}$$

The analogous quantity

$$\mu = \frac{\partial U}{\partial N} = -T\frac{\partial S}{\partial N}. \tag{9}$$

describes the response of the system to a change in the number of particles. It is called the chemical potential. In (7) - (9) the first equality yields the functional dependence on S, V, N, while the second yields the same quantities as function of U, V, N. When taking a derivative it is always assumed that the other arguments are kept constant. These definitions may be summarized in the total differential

$$dU = TdS - pdV + \mu dN \tag{10}$$

From their definition as derivatives of an extensive quantity w.r.t. another extensive quantity is follows that T, p, μ are intensive, i.e., they do not change under a rescaling of the extensive quantities. This is well-known: Putting together two identical systems with identical temperature doubles the volume, total internal energy and particle number, but certainly not the temperature! The quantity U is called a thermodynamic potential. As we shall see below, there are other, equivalent, thermodynamic potentials $\Phi(x_1, x_2, \ldots, x_n)$ for systems characterized by n macroscopic observables. A derivative $y_i = \partial\Phi/\partial x_i$ of a thermodynamic potential and the quantity x_i are called conjugate to each other. In our case (T, S), (p, V) and (μ, N) resp. are pairs of conjugate variables.

The formalism developed so far is very abstract and appears to be rather remote from anything having to do with either physics or biology. It is what the mathematician Pablo Ferrari, would perhaps call (with a friendly smile in his face) "abstract nonsense". However, not only he himself, but many scientists have occasionally solved difficult problems by applying abstract nonsense to specific problems. As an example consider first the ideal gas without internal degrees of freedom. From microscopic considerations it can be derived [5] that

$$U(S, V, N) = \frac{3}{2}aV\left(\frac{N}{aV}\right)^{5/3}\exp\left(\frac{2S}{3k_B N}\right) \tag{11}$$

The derivation requires some approximations, in particular that molecules in the gas do not interact for any positive distance and hence have only kinetic energy. We cannot review all these assumptions and their significance here. We just point out that calculating T from (7) and p from (8) immediately yields for the ideal gas with entropy (11) the well-known equation of state

$$pV = Nk_BT. \tag{12}$$

One may also express the internal energy as a function of temperature and finds the equally well-known expression $U = 3/2Nk_BT$. The interpretation is that each translational degree of freedom with microscopic mean kinetic energy $\bar{p}^2/2m$ contributes a thermal energy $k_BT/2$ to the macroscopic internal energy. The factor of three results from the three space dimensions. From the internal energy given as a function of temperature one obtains the specific heat (at constant volume) by taking the derivative

$$C_V = \frac{\partial U}{\partial T} = \frac{3}{2}Nk_B. \tag{13}$$

For the ideal gas (which becomes an unphysical approximation at low temperatures) and in some other simple model systems C_V is independent of T, but generally it is a function of T. It is a measure for the response of the internal energy to a small change in temperature.

In order to calculate of the entropy from the microscopic properties of a given system and thus to really illustrate the workings of the basic concepts of statistical mechanics it is better to return to the toy model of coins distributed on cubes. The reason for defining this model is of course purely pedagogical. By *not* considering a real system on a microscopic scale there is no danger that physically motivated approximations (which are not always immediately plausible) obscure the conceptual simplicity of the approach. Moreover, it becomes clear that statistical mechanics may be applied not only to physical systems, but to any system that may be characterized by a suitably defined ensemble. This includes biological systems, but also stylized facts of economical and sociological systems become accessible. Even traffic of automobiles on roads has become a focus of statistical mechanics in the last decade. A third reason will become apparent below.

For notational simplicity we introduce the dimensionless internal energy per site

$$q = \frac{U}{N\epsilon_0} = \frac{M}{N} + \frac{1}{2} \tag{14}$$

which is not to be confused with the energy density u which has a dimension of energy per unit area and is related to q by $u = q\epsilon_0/a^2$. Moreover we follow standard practice and introduce the quantity

$$\beta = 1/(k_BT). \tag{15}$$

as a measure for the inverse temperature. $1/\beta = k_BT$ has the dimension of an energy. It appears often in thermodynamic expressions as a natural scale of the thermal energy of a system. We shall consider the so called thermodynamic limit, i.e., the limit where N and M become large numbers. This allows us to use Stirlings formula

$$k! \sim (k/\mathrm{e})^k \tag{16}$$

to approximate the factorials that appear in Γ (106). In terms of of q this yields the entropy

$$S(U,V) = k_BV/a^2[(q+1/2)\ln(q+1/2) - (q-1/2)\ln q - 1/2]. \tag{17}$$

Since we fix the number of coins to equal to N there is no separate N-dependence of the entropy. From the expression (17) one verifies that S is extensive. One also finds from a simple calculation

$$\beta = \frac{1}{\epsilon_0} \ln \frac{q+1/2}{q-1/2} \tag{18}$$

$$p = \frac{\epsilon_0}{2a^2} \frac{\ln\left[(q+1/2)(q-1/2)\right]}{\ln\left[(q+1/2)/(q-1/2)\right]} \tag{19}$$

$$= k_B T/a^2 \ln\left(1 - \mathrm{e}^{-\beta\epsilon_0}\right)^{-1} - \epsilon_0/(2a^2) \tag{20}$$

The internal energy as a function of T therefore becomes

$$U = \frac{V\epsilon_0}{2a^2} \coth\left(\frac{\beta\epsilon_0}{2}\right). \tag{21}$$

Notice that at low temperatures S vanishes and U approaches its minimal value $V\epsilon_0/(2a^2)$. This is in agreement with the notion that low temperatures correspond to configurations with low internal energy and (relatively) small degeneracy. The specific heat is strictly positive for $T > 0$ and vanishes at $T = 0$. This is in agreement with the third law of thermodynamics asserting the inaccessibility of $T = 0$: In order to cool down to $T = 0$ some object by extracting its internal energy another object with non-zero specific heat at $T = 0$ would be required. Hence according to the third law no object with non-zero specific at $T = 0$ should exist.
For large T one finds $U \sim Nk_BT$ and therefore $C_V \sim Nk_B$. At high temperatures (corresponding to a large mean height of towers in our toy model) each coin with microscopic mean energy $mg\bar{h}$ contributes a unit k_BT of thermal energy to the internal energy.
From (21) one also obtains the average number of cubes $\bar{m} = M/N$ as a function of temperature. Since $\bar{m} = q - 1/2$ we find

$$\bar{m} = \frac{1}{\mathrm{e}^{\beta\epsilon_0} - 1} \tag{22}$$

Perhaps somewhat surprisingly this is the well-known Bose distribution for quantum mechanical harmonic oscillators. It gives the mean occupation number of an ensemble of harmonic oscillators consisting of elementary excitations (e.g. photons) with energy ϵ_0. Indeed, the energy levels of our classical toy model may be identified with the energy levels of a harmonic oscillator with frequency ν_0 by setting $\epsilon_0 = \hbar\nu_0$. A site with n cubes then corresponds to the n'th excited energy level of the oscillator. In electromagnetic language this would corresponds to an electromagnetic wave consisting of n photons of frequency ν_0. In an ensemble of such waves the mean occupation number of photons with frequency ν_j follows the Bose distribution $\bar{m}_j = 1/(\mathrm{e}^{\beta\epsilon_j} - 1)$. In our model only "photons" with frequency ν_0 occur. This gives rise to (22).
Here we find a third reason for studying our toy model. It tells us that entirely different systems (in our case real photons as compared to coins located on macroscopic classical cubes) may have identical statistical properties. This has two consequences of conceptual importance: (1) Given some thermodynamic function or relation one does not know yet what the underlying physical system is. Hence producing some result from a given model does not necessarily mean that this model captures all the main physical features of the real system that the model is supposed to describe. (2) This remark applies in particular to dynamical properties as in the framework of statistical mechanics we assume only ergodicity, nothing else. We can certainly not make any inference from our toy model on the propagation of electromagnetic waves.

This ambiguity, however, also has a positive side to it: Suppose an analytical calculation of the statistical properties of a given model is too difficult. (It often is in soft matter and biological systems). Then one has to resort to a numerical simulation of such system, e.g. one has to randomly generate samples with their proper statistical weight, i.e. for a microcanonical simulation one has to make sure that each state will be generated with equal probability. Such a simulation defines an artificial time evolution of the system. Our observation illustrates that the precise nature of this procedure is completely arbitrary, there is no need to devise a complicated time evolution that somehow mimics what one believes to happen in the real system. This is of great importance for numerical simulations. The strategy to develop a simulation procedure for thermal equilibrium needs to be based (besides ergodicity) only on numerical efficiency.

Canonical ensembles

The scenario described above – an isolated system with fixed energy – is not what one usually encounters, in particular not in a biological context. More often, a system under consideration can exchange energy with its environment so that in equilibrium the system has the same temperature as its environment. However, in equilibrium such a system will not have a strictly constant energy, as is assumed in the microcanonical ensemble. Fluctuations will persist and the statistical approach described above fails. Therefore we should attempt to describe a statistical ensemble with fixed temperature rather than fixed energy.
In order to do so we consider a small subsystem of a big system with fixed energy. The small subsystem is still assumed to be fairly large to justify a statistical description. We shall refer to the big environment as heat bath. We remember that dynamical processes occur which change the microstate of the whole system and hence of the subsystem. In the framework of our toy model such processes would amount to a reshuffling of cubes from one place to another by some mechanism. A subsystem is a subset of sites. Evidently in any (non-isolated) subsystem the energy will therefore not be constant as some reshufflings will change the number of cubes in the subsystem and due to ergodicity all conceivable reshufflings may occur.
In the microcanonical ensemble each microstate has the same statistical weight $1/\Gamma$. What is the statistical weight of a subsystem with energy E in a system of total energy Y?
Clearly, if the subsystem has energy E, the heat bath has energy $Y - E$. Hence the statistical weight $w(E)$ of a subsystem with energy E may be expressed as the ratio of the number of states of energy $Y - E$ with the number of states of total energy Y. Hence

$$w(E) = \frac{\Gamma(Y-E)}{\Gamma(Y)} = \mathrm{e}^{(S(Y-E)-S(Y))/k_B} \tag{23}$$

Since we assume $E \ll Y$ we may replace the difference in the exponential by the (negative) derivative w.r.t. E, which is the inverse temperature. We arrive at the conclusion

$$w(E) = \mathrm{e}^{-\beta E} \tag{24}$$

This is the Boltzmann distribution for a system of energy E.[6]
Strictly speaking the temperature that occurs in (24) is the temperature T_1 of the heat bath. For this formalism to make physical sense we have to show that the temperature T_2 in the subsystem

[6]The Boltzmann distribution can be obtained differently following Boltzmann and Planck. For statistically independent subsystems with energies $E_{1,2}$ one has for the joint system $w(E_1 + E_2) \propto w(E_1)w(E_2)$. This factorization implies $w(E) \propto \mathrm{e}^{-bE}$ with a constant b. One then shows that $b = \beta$.

is the same (as it should be for an physical equilibrium system). To prove this we make the following observations. (i) According to the assumption of the microcanonical ensemble the number of states in the full system (heat bath plus subsystem) may be written

$$\Gamma(Y) = \sum_E \Gamma_1(Y - E)\Gamma_2(E) \tag{25}$$

where Γ_1 refers to the heat bath and Γ_2 to the degeneracy in the subsystem. (ii) The degeneracy is a sharply increasing function of the energy. So Γ_2 increases sharply with E while Γ_2 decreases sharply with E. As a result the product of the two degeneracies in the sum has an extremely sharp peak at some specific value E^* of E. This peak value is exponentially large in the volume the system. Therefore the sum is dominated by this term which satisfies (since it is an extremum as a function of E)

$$\frac{d}{dE}\Gamma_1(Y - E)\Gamma_2(E) = 0 \qquad \text{for } E = E^*. \tag{26}$$

Using the definition (7) of the temperature as applied to the two subsystems, this is equivalent to $T_1 = T_2 \equiv T$.[7]

The statistical weight (24) is not a normalized probability. To complete the statistical description of the subsystem, i.e., to obtain a probability, we have to normalize by the sum

$$Z = \sum_{config} \mathrm{e}^{-\beta E} \tag{27}$$

over all configurations (microstates) of the subsystem. This sum is called the canonical partition sum or canonical partition function. Since for any E there are Γ microstates in the subsystem (we now drop the index two for the subsystem) we may write

$$Z = \sum_E \Gamma(E, V, N)\mathrm{e}^{-\beta E} = \sum_E \mathrm{e}^{-\beta(E-TS)} \tag{28}$$

where S is the entropy of the subsystem. Also the product $\Gamma(E, V, N)\mathrm{e}^{-\beta E}$ has a sharp maximum at an energy where the argument of the expontial is maximal, e.g. which satisfies $TdS/dE - 1 = 0$. Since the temperature has already been obtained above we realize that this energy is the same E^* that has been obtained above. It is the energy that maximizes the degeneracy $\Gamma_1(Y - E)\Gamma_2(E)$ (hence the entropy) of the joint system, subsystem plus heat bath. The energy E^* is the mean energy of the subsystem which is in exchange with the heat bath. Since the fluctuations around this mean are extremely small compared to E^* itself (see next section) it can be effectively identified with the internal energy $U(T, V, N)$ that is used in the microcanonical description of an isolated system. Thus one may write

$$Z(T, V, N) = \mathrm{e}^{-\beta(U-TS)} \tag{29}$$

We stress, however, that unlike in the microcanonical ensemble U is *not* an independent variable characterizing the state of the system. It is understood as a function of the temperature T and V, N of the subsystem. Correspondingly the entropy S is understood to be given as a function

[7]Moreover, the fact that we have a maximum (not just an unspecified extremum), implies negative second derivative of the summand which in turn implies that the energy is a monotically increasing function of the temperature.

of the triple (T, V, N) since according to (29) it is the entropy evaluated at the internal energy $U(T, V, N)$. From the partition function Z one defines an new energy by taking the logarithm

$$F(T, V, N) = -k_B T \ln Z = U - TS \tag{30}$$

This is the Helmholtz free energy, which fully characterizes the thermodynamic properties of a system at fixed (T, V, N). The total differential is

$$dF = -SdT - pdV + \mu dN. \tag{31}$$

Together with the second equation in (30) this follows also directly from (10). Given F, this total differential yields the entropy S, the pressure p and the chemical potential μ as functions of (T, V, N) by taking appropriate partial derivatives, keeping the other parameters fixed.
It is crucial to realize that $F(T, V, N)$ and $U(S, V, N)$ are equivalent, i.e., they contain the same amount of information about the system in question. We do not actually need a heat bath to derive F, we could have obtained this quantity directly by the so-called Legrende transformation $F = U - TS$ together with the definition of T (7). On the other hand, $U(T, V, N)$ does *not* provide an equivalent full description of a system. In order to calculate $U(S, V, N)$ from it one would have to integrate w.r.t. T, but this gives an undetermined integration constant which is an arbitrary function of V, N. In other words, very different systems with different microcanonical $\Gamma(U, V, N)$ may lead to the same $U(T, V, N)$, while two systems with the same $F(T, V, N)$ are statistically identical, i.e., have the same $\Gamma(U, V, N)$.
Most importantly we note this construction implies an extremal principle for the free energy which comes from the maximization of entropy. Given a choice between states with different free energies, any equilibrium system will always select the state with the lowest free energy. This is equivalent to stating that it will choose the state with the highest entropy.
Let illustrate these notions by applying them to the toy model. Since we know $\Gamma(U, V)$ explicitly we can calculate the partition function from (28). We obtain

$$Z = \sum_{M=0}^{\infty} \frac{(N+M-1)!}{M!(N-1)!} \mathrm{e}^{-\beta\epsilon_0(M+N/2)} = \left[\frac{\mathrm{e}^{\beta\epsilon_0/2}}{1-\mathrm{e}^{-\beta\epsilon_0}}\right]^N \tag{32}$$

Here we used $\sum_M C_N^{M+N-1} z^M = 1/(1-z)^N$. Hence

$$F = k_B T V/a^2 \ln\left(1 - \mathrm{e}^{-\beta\epsilon_0}\right) + V\epsilon_0/(2a^2) \tag{33}$$

which is manifestly extensive, as expected from the relation $F = U - TS$ which reproduces (33).
In the treatment above we assumed that the subsystem has fixed volume. However in a real system the volume may change and – in equilibrium – adjust itself to its environment. This is the case if the system is confined to an extensible membrane or to a container with a piston as in the classical steam engine. We know from experience that in such a situation the pressure in the subsystem and its environment is the same. This follows from the above considerations in the same way as the equality of temperatures popped out of the calculation (26). This is because Γ as a function of the volume is also sharply increasing. Repeating the same considerations as above, but with both E and V fluctuating, the analog of (23) becomes for a system with volume Z and subsystem with volume V'

$$w(E, V') = \frac{\Gamma(Y-E, Z-V')}{\Gamma(Y)} = \mathrm{e}^{(S(Y-E,Z-V')-S(Y))/k_B} \tag{34}$$

This leads to $w = \exp(-\beta E - \beta p V')$. One then proceeds to show that not only $T_1 = T_2 \equiv T$, but also $p_1 = p_2 \equiv p$ in the subsystem and heat bath respectively.
As a next step one normalizes w by calculating the partition function

$$Z(T,p,N) = \sum_{config} \mathrm{e}^{-\beta(E+pV')} \tag{35}$$

$$= \sum_{V',E} \Gamma(E,V',N)\mathrm{e}^{-\beta(E+pV')} \tag{36}$$

$$= \sum_{V'} Z(T,V',N)\mathrm{e}^{-\beta pV'} \tag{37}$$

The most significant contribution to the sum comes from that volume V which maximizes the entropy of the joint system (at which $p_1 = p_2 = p$) and one may write

$$Z(T,p,N) = Z(T,V,N)\mathrm{e}^{\beta pV} = \mathrm{e}^{-\beta(F+pV)} \tag{38}$$

where V is understood as a function of the triple (T,p,N). From this one defines a new thermodynamic potential, the Gibbs free energy

$$G(T,p,N) = F + pV = U - TS + pV \tag{39}$$

The corresponding total differential

$$dG = -SdT + Vdp + \mu dN \tag{40}$$

yields the quantities S, V, μ as function of the conjugate variables (T,p,N). Since the transformation from F to G is again a Legrende transformation the Gibbs free energy $G(T,p,N)$ is equivalent to both $F(T,V,N)$ and $U(S,V,N)$.
Notice that extensivity of G implies $G = Ng(T,p)$ with some function g that depends on the system under consideration. On the other hand, $\mu = \partial G/\partial N = g$. We arrive at the conclusion that

$$G(T,p,N) = N\mu(T,p). \tag{41}$$

This scaling argument can be adapted to our toy model. Since particle number is not a variable in our toy model one expects $F(T,V) = Vf(T)$. But according to (31) $\partial F/(\partial V) = -p$ and we conclude $f = -p(T)$. This gives $F(T,V) = -Vp(T)$, in agreement with the results (19), (33) found from direct computation.
Clearly the contents of this section makes physical sense only in the presence of dynamics. Without dynamical processes the energy of a subsystem could not fluctuate. This is obvious in the toy model: Without a dynamical mechanism for reshuffling cubes from one site to another any randomly chosen subsystem would just have the energy it happens to have. What general assumption do we have to make on dynamical processes describing an ensemble at constant temperature? As in the framework of the microcanonical ensemble we answer this question in the context of numerical simulation. Since energy is not conserved at constant temperature, a numerical calculation of the statistical properties of some model cannot be performed in the same way as in the microcanonical ensemble (where each microstate has to be generated with equal probability). Instead, one has to make sure that the Boltzmann distribution (24) is generated. This can be achieved by a random process on which one imposes the principle of *detailed balance*. This is a property of dynamical random steps which links the Boltzmann weights

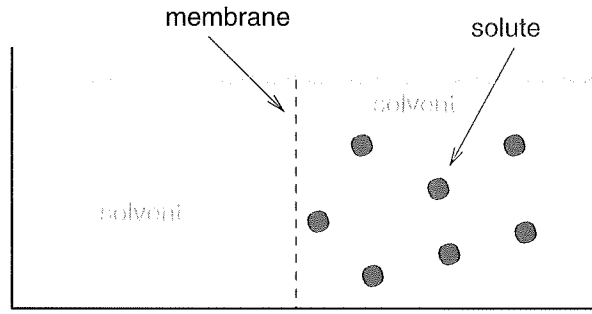


Fig. 3: *Container with a solvent and a semipermeable membrane (brown). Solute particles (red) in the right-hand part of the container cannot penetrate the membrane, but solvent particles (not shown) can.*

(24) with the probability $p_{\eta'\to\eta}$ of generating a microstate η out of a microstate η' and with its reversed probability $p_{\eta\to\eta'}$. The detailed balance condition reads

$$\frac{p_{\eta'\to\eta}}{p_{\eta\to\eta'}} = \frac{\mathrm{e}^{-\beta E}}{\mathrm{e}^{-\beta E'}} \tag{42}$$

where $E(\eta)$ is the energy of the microstate η. The theory of Markov processes guarantees that any computer algorithm for randomly generating subsequent microstates which satisfies detailed balance will generate samples of the system with their proper statistical weight (24). A simulation algorithm which randomly generates a new sample from a give sample is a Monte-Carlo algorithm, as opposed to e.g. molecular dynamics which numerically solves microscopic laws of motion. Which type of algorithm is preferentiable depends very strongly on the given problem.

Fluctuating particle number: The osmotic pressure

There is a well-known observation. Imagine a container with a solvent and a semipermeable membrane inside (Fig. 3). In one part of the solvent there is a solute with small concentration c. The other part, separated by the membrane from the first, contains only the solvent. The membrane has pores that allows solvent particles to penetrate, but not solute molecules. One finds that there is a pressure difference in the two regions called *osmotic pressure* p_{osm}. It is given by the law of van't Hoff

$$p_{osm} = c\rho k_B T \tag{43}$$

for a solute concentration c and particle density ρ of the solvent. Where does this peculiar phenomenon come from?

Let us first suppress our curiosity and return to abstract nonsense. It is clear that the game of letting conserved quantities fluctuate can be played with other variables, too. Rather than letting the volume fluctuate, one may consider a system with particle exchange with its environment, i.e., consider the so-called grandcanonical ensemble. This is biologically highly relevant, as mass transport through a membrane occurs not only in the experimental setting introduced above and in its technical applications, but also in biological cells.

One proceeds in a manner entirely analogous to the previous transformations from one ensemble to another. The intensive quantity conjugate to the particle number is the chemical potential.

One finds that the chemical potentials in the two systems coincide in equilibrium. Different chemical potentials across a membrane would lead to a particle flux through the membrane that tries to make the potentials equal until equilibrium is reached. This is entirely analogous to a pressure gradient that changes volume, or to a temperature gradient that leads to an energy flux. Before applying this insight to osmosis we write down for completeness the basic properties of the grandcanonical ensemble. The partition function for fluctuating particle number

$$Z(T,V,\mu) = \sum_{N'} Z(T,V,N')\mathrm{e}^{\beta\mu N'} = \sum_{N'} \mathrm{e}^{-\beta(F-\mu N')} \tag{44}$$

describing the equilibrium situation is called the grandcanonical partition function. The summand takes it maximum at that N that maximizes the entropy of the joint system. The corresponding thermodynamic potential is given by

$$\Omega(T,V,\mu) = -k_B T \ln Z(T,V,\mu) = F - \mu N \tag{45}$$

The corresponding total differential is

$$d\Omega = -SdT - pdV - Nd\mu. \tag{46}$$

From extensitivity we conclude that $\Omega = Vg(T,\mu)$. Since $p = -\partial\Omega/\partial V = g$ we deduce

$$\Omega = -pV. \tag{47}$$

In our toy model there is no particle number as independent variable and hence $F = \Omega$. However, in more complex systems one may have several particle numbers N_i which are independently conserved quantities. E.g. in a mixture of gases such as ordinary air one has as many conserved N_i as there are components of the gas. The same is true in a multicomponent solution which is described by a Helmholtz free energy $F(T,V,N_1,\ldots,N_k)$. In such a case there are as many conjugate chemical potentials μ_i and various grandcanonical ensembles can be defined. In order to discuss osmosis let us now first return to systems with fixed particle number. This is necessary firstly because the number of solute particles is fixed in regions 1 and 2 respectively. Secondly, one needs to understand the thermodynamic properties of a solution without semipermeable membrane for the derivation of (43). Generalizing (41) one finds that the Gibbs free energy for fixed $T, p, N_1, \ldots, N_k$ has the form

$$G = \sum_{i=1}^{k} \mu_i N_i \tag{48}$$

where k is the number of independent components. Consider now a dilute solution. Let $G_0(T,p,N) = N\mu_0(T,p)$ be the Gibbs free energy and chemical potential respectively of the pure solvent. We define by $\alpha(T,p,N)$ the small change of the Gibbs free energy experienced by adding a single solute molecule to the solvent. What is then the change of G when adding n molecules, with $1 \ll n \ll N$? Since $c = n/N$ is small the mean distance between solute particles is large and we can neglect interactions between them. Hence a first guess for G might be $G = G_0 + n\alpha$. However, this is wrong since all solvent particles are identical. The contribution $n\alpha$ would arise if the number of states $\Gamma(n)$ that the n particles can assume in the solution would be $\Gamma(1)^n$. This would be the case for distinct particles. For identical particles this number is reduced by all possible permutations that one can be perform with n particles, because

a permutation of identical particles represents the same state. Therefore $\Gamma(n) = \Gamma(1)^n/n!$ (see appendix). This effect, which is of very general nature, leads to an extra entropic change of $k_BT \ln n!$ in G and one finds

$$G = N\mu_0(T,p) + n\alpha(T,p,N) + k_BT \ln n! \tag{49}$$

We stress that the n-dependence of this expression is purely entropic, just originating in the counting of configurations. All the energy-dependence coming from the solute-solvent interaction is in the single-particle term α. This is because we assumed to have a dilute solvent where solute-solute interactions can be neglected.
Since n is large we may use Stirling's formula to obtain $\ln n! \approx n \ln(n/\mathrm{e})$. On the other hand G must be a homogeneous function in both N and n, i.e., $G(T,p,\lambda N,\lambda n) = \lambda G(T,p,N,n)$. Hence α must have the form $\alpha(T,p,N) = \psi(T,p) - k_BT \ln N$ and we arrive at

$$G(T,p,N,n) = N\mu_0(T,p) + nk_BT \ln\left(\frac{n}{\mathrm{e}N}\right) + n\psi(T,p) \tag{50}$$

with an unknown and system-dependent function ψ. Thus for small concentration $c = n/N$ we find the chemical potentials of the solvent

$$\mu = \frac{\partial \Phi}{\partial N} = \mu_0 - k_BTc \tag{51}$$

and solute

$$\mu' = \frac{\partial \Phi}{\partial n} = k_BT \ln c + \psi \tag{52}$$

respectively. For k distinct solutes the expression (50) straightforwardly generalizes to

$$G = N\mu_0(T,p) + \sum_{i=1}^{k} \left[n_i k_BT \ln\left(\frac{n_i}{\mathrm{e}N}\right) + n_i\psi_i(T,p) \right]. \tag{53}$$

We return now to the situation of a container with a dilute solution with two distinct regions of different concentrations $c_{1,2}$ as described in the experimental situation at the beginning to this subsection (Fig. 3). Since the membrane allows for both energy exchange and solvent exchange, the temperatures and chemical potentials of the *solvent* in both regions must be equal in equilibrium:

$$\mu_0(p_1,T) - c_1k_BT = \mu_0(p_2,T) - c_2k_BT \tag{54}$$

We therefore see that in equilibrium the pressure must be different in the two regions. This difference $p_{osm} = p_2 - p_1$ the *osmotic pressure*.
For a small difference of concentrations p_{osm} is small and we may write $\mu_0(p_2,T) \approx \mu_0(p_1,T) + p_{osm}\partial\mu_0/(\partial p)$. On the other hand

$$\frac{\partial \mu_0}{\partial p} = \frac{1}{N}\frac{\partial G_0}{\partial p} = \frac{V}{N} \tag{55}$$

since V is the variable conjugate to p. With $N/V = \rho$ we arrive at

$$p_{osm} = (c_2 - c_1)\rho k_BT. \tag{56}$$

For the special case $c_1 = 0$ this is the law (43) of van't Hoff. In essence it is an entropy effect. Only the purely entropic change (51) in the chemical potential of the solvent enters its derivation.

Summary

The upshot of all this is that there are several equivalent ways of describing a statistical ensemble. They correspond to various physical situations in which one or the other extensive quantity (that would be conserved in an entirely isolated system) is in exchange with its environment. As a result, this quantity has small fluctuations in the system and its conjugate intensive quantity is equal in the system and its environment. Therefore such a system is not any longer characterized by only conserved extensive quantities, but by the remaining conserved quantities and the intensive quantities conjugate to the fluctuating extensive quantities. This generalizes to systems which have other extensive conserved quantities. In magnetic systems, e.g., this could be the magnetization. The conjugate quantity would then be the magnetic field. In a biological setting one may have several distinct particle species with distinct conjugate chemical potentials. All thermodynamic potentials are governed by an extremal principle which tries to minimize the respective potential.

On a mathematical level one may move from one description to the other by the Legrende transformation, i.e., shifting a given thermodynamic potential by the product of two conjugate quantities. In order to retain the full information, the given quantity must be expressed in terms of its conjugate quantity. Statistical mechanics provides a microscopic basis for the derivation and relationship between these equivalent ensembles. The equivalence is ensured by vanishing relative fluctuations of macroscopic extensive variables.

From a modelling point of view there is no one-to-one correspondence between a given system and its thermodynamic properties. This leaves a huge amount of freedom for developing numerical procedures for generating samples with their proper statistical weight. Moreover, one can derive with relative ease laws of very general validity (e.g. the equation of state for the ideal gas or the even more general law of van't Hoff). Few (ideal gas) or even no microscopic data at all (law of van't Hoff) enter these laws. However, one pays a price for this generality. A model which reproduces real data well, may fail completely to give real insight into the basic microscopic origin of the system. So some care has to be taken when drawing conclusions from a model for a real system.

1.3 Fluctuations in equilibrium

Linear response theory

So far we made reference to fluctuations only in a very qualitative way. Indeed, it may have appeared as if they were important only from a technical viewpoint. It is not so because what we used is not that fluctuations as such are negligible, but only that the magnitude of fluctuations of a variable compared to its mean value can be ignored. Fluctuations are in fact rather important. This is easily illustrated by studying energy fluctuations in the canonical ensemble. From the definition (27) of the partition function we find immediately the mean energy $\langle E \rangle$ as a logarithmic derivative

$$\langle E \rangle \equiv \frac{\sum_{config} E \mathrm{e}^{-\beta E}}{\sum_{config} \mathrm{e}^{-\beta E}} = -\frac{\partial}{\partial \beta} \ln Z = U \tag{57}$$

The average defined here is usually referred to as thermal average. Analogously the fluctuations

are given by the second derivative

$$-\frac{\partial \langle E \rangle}{\partial \beta} = \frac{1}{Z} \sum_{config} E^2 \mathrm{e}^{-\beta E} - \frac{1}{Z^2} \left(\sum_{config} E \mathrm{e}^{-\beta E} \right)^2 = \langle E^2 \rangle - \langle E \rangle^2 \tag{58}$$

For a nonfluctuating quantity the variance on the r.h.s. of (58) would simply vanish. The square root of the variance is a measure for the width of a distribution of a random quantity. Since the variance is the mean of a square of a number,

$$\langle E^2 \rangle - \langle E \rangle^2 = \langle (E - \langle E \rangle)^2 \rangle, \tag{59}$$

it takes positive values only, except for a state without fluctuations, where it vanishes.
On the other hand, the general expression (13) for the specific heat as a response to a change of internal energy in temperature yields

$$C_V = \frac{\partial U}{\partial T} = -\frac{1}{k_B T^2} \frac{\partial \langle E \rangle}{\partial \beta} \tag{60}$$

$$= \frac{1}{k_B T^2} \left(\langle E^2 \rangle - \langle E \rangle^2 \right) \tag{61}$$

Since C_V is extensive (derivative of an extensive quantity w.r.t. an intensive one), the variance of the distribution of energies must be of the order of the volume V. Hence the width Δ of the distribution is of the order $\sqrt{V}$. We remark that from a dynamical perspective an energy change is accompanied by a dissipation phenomenon, viz. equal distribution of the energy change over the whole system.
We learn from this statistical mechanics result several very profound properties of thermodynamic systems.

(i) The response of a thermodynamic quantity (here the energy dissipation) to a change of its conjugate variable (here the temperature) is related to the fluctuations of this quantity. This is an example for a fluctuation-dissipation relation.

(ii) The specific heat is given by a variance which is a positive quantity for a fluctuating state and which vanishes only in the absence of fluctuations, i.e., at $T = 0$. Therefore it is generally not possible to reach $T = 0$. This is the microscopic origin of the third law of thermodynamics.

(iii) Since the variance is extensive, the relative fluctuations $\Delta / \langle E \rangle \propto 1/\sqrt{V} \propto 1/\sqrt{N}$ are extremely small for macroscopic particle numbers of the order 10^{23}. This ensures the equivalence of ensembles in the thermodynamic limit.

Notice that the relationship (61) between response and fluctuations is easily generalized. Suppose a system is characterized by an extensive quantity A. Let h be the corresponding conjugate quantity. According to the considerations of the previous section the partition function for the ensemble with fluctuating A can be written

$$Z = \sum_{config} \mathrm{e}^{\cdots + \beta h A} \tag{62}$$

where the dots in the exponential depend on which ensemble precisely one wishes to study. In any case, the equilibrium state described by the ensemble (62) is characterized by the temperature and the intensive quantity h. Imagine now that the control parameter h is changed slightly, $h \to h' = h + \delta h$, and one waits until the new equilibrium state w.r.t. h' is reached. How does the mean of A responds to this change?
One has to linear approximation

$$\delta\langle A\rangle \equiv \langle A\rangle_{h'} - \langle A\rangle_h \approx \delta h \frac{\partial\langle A\rangle_h}{\partial h} \tag{63}$$

where $\langle A\rangle_h$ denotes thermal average w.r.t. the distribution that leads to the partition function (62). Because of the occurrence of $\beta h A$ in the exponential one has (to linear order)

$$\frac{\delta A}{\delta h} = -\beta\left(\langle A^2\rangle - \langle A\rangle^2\right). \tag{64}$$

The response of a quantity B which is *not* conjugate to h can be calculated in the same manner. It is given by the covariance

$$\frac{\delta B}{\delta h} = -\beta\left(\langle AB\rangle - \langle A\rangle\langle B\rangle\right). \tag{65}$$

If this covariance vanishes one says that the quantities A and B are *uncorrelated*. It means that they are statistically independent. Relations of the form (64),(65) (which are usually written in differential form) are called linear response relations.
Item (iii) furnishes us with a proof that our assumption of the previous section that led to the equivalence of ensembles is justified for macroscopic systems. Notice, however, that for small systems with just a few thousand particles (or even less) the error that one makes when looking at different ensembles may be significant and is not unlikely to be relevant in measurements. This can very well impact biological measurements where ensembles frequently do not contain huge numbers of objects. The fact that even for small N the exact treatment of our toy model led to the same expression for the Helmholtz free energy as the Legrende transformation did (which is assumed to be valid only asymptotically for large N) is a coincidence with no general significance. It just tells us that errors that may occur by changing ensembles of small systems are not by necessity large.

Fluctuation dynamics

So far we have treated dynamics only on a rudimentary and largely qualitative level. This is because the physical origin of the static fluctuations discussed above are microscopic dynamical processes which vary from system to system. Even though on a truly microscopic level these processes are deterministic (at least if a classical description with Newtonian mechanics is adequate) these processes are so complicated that they appear as random events. They contain immensely more information about a given system than only the static properties. We have shown that two absolutely unrelated systems may have the same static statistical properties. However, if one can handle statistical dynamical properties of a system, the danger of being mistaken with regard to which microscopic system one actually describes with a given model is greatly diminished. Of course, one pays a price: Results concerning dynamics are not as generally valid as static results. Moreover, simulations of dynamical events must be defined much

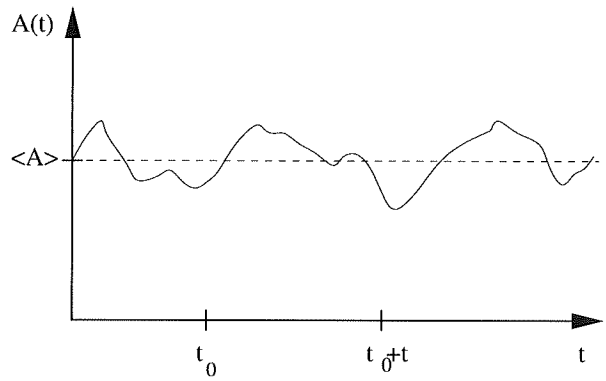


Fig. 4: *A stylized time series for some observable A fluctuating around its mean* $\langle A \rangle$. *Single measurements at time* t_0 *and* $t_0 + t$ *resp. produce unpredictable random numbers.*

closer to reality than simulations for static quantities. However, before entering into these intricacies, which is done in the last part of this lecture, we discuss some results about equilibrium dynamics which are still generally valid.

Suppose one measures some thermodynamic quantity A as a function of time. A very precise measurement device which can follow the usually extremely fast and tiny random fluctuations of $A(t)$ would then record a time series somewhat similar to what is shown in Fig. 4. In equilibrium one would find that A fluctuates around a mean value $\langle A \rangle$ which is just the thermal average of that quantity. Experimentally this can verified in two different ways: Either one makes many consecutive measurement $A(t_i)$ at times t_i (keeping external parameters constant) and divides the sum of the measured values by the number of measurements. Alternatively one could make single measurements A_k at many physically identical, but distinct samples of the same substance (subject to the same external conditions).[8]

Ergodicity together with the law of large numbers assert that a time average equals an ensemble average:

$$\langle A \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} A(t_i) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} A_k. \tag{66}$$

It is clear that this equilibrium quantity does not depend on time.

Imagine now that you make two measurements at times t_0 and $t_0 + t$. The product of the two measured quantities is again a random number, the average of which we write as

$$G_{AA} = \langle A(t_0 + t)A(t) \rangle \tag{67}$$

Since we assume $\langle A \rangle$ to be known by the first set of measurements the only quantity of interest for further study is the difference $A(t) - \langle A \rangle$. So we define the *correlation function*

$$C_{AA} = \langle (A(t_0 + t) - \langle A \rangle)(A(t) - \langle A \rangle) \rangle \tag{68}$$

[8] This discussion seems to contradict the fact that in order to get a precise measurement of e.g. the temperature of macroscopic object one needs neither many measurements nor many samples. Under well-controlled experimental conditions one always gets the same value as in a single measurement. This is because the fluctuations in a macroscopic object are so tiny that they cannot be recorded with a simple device like a mercury thermometer. Our discussion concerns more fancy measurements such as scattering by electromagnetic waves or neutrons or NMR measurements.

which is actually a covariance of A for different times.
It is of interest to study also cross correlations

$$C_{AB} = \langle\, (A(t_0+t) - \langle\, A \,\rangle)(B(t) - \langle\, B \,\rangle)\,\rangle. \tag{69}$$

More refined are local quantities which describe the state of the system in a very small neighborhood of the space point $\mathbf{x}$. Their covariance has the form

$$C_{AB}(\mathbf{x},\mathbf{y}) = \langle\, (A(\mathbf{x},t_0+t) - \langle\, A \,\rangle)(B(\mathbf{y},t) - \langle\, B \,\rangle)\,\rangle. \tag{70}$$

Here we assume spatial translational invariance of the thermal average $\langle\, Z(\mathbf{x},t)\,\rangle = \langle\, Z \,\rangle$ for any observable Z. Otherwise one would substract $\langle\, Z(\mathbf{x})\,\rangle$ in the definition of the correlation function. We also remark that in scattering experiments one measures the Fourier transform of local observables rather than the space dependence directly. If $\langle\, A \,\rangle$ gives the local density of some microscopic object then the Fourier transform of the correlation function C is called the dynamical structure function.
For the study of correlation functions space-time symmetries of the underlying laws of motion are of utmost importance. Because of translational invariance in space, a space dependent correlation function depends on the difference $\mathbf{r} = \mathbf{y} - \mathbf{x}$ rather than both coordinates separately. If in addition the system has rotational invariance, C_{AB} depends on the distance $r = |\mathbf{r}|$ only.
We are in equilibrium and so it does not matter at which time t_0 we start our measurement. An equilibrium system is (on average) time-translation invariant, $C_{AB}(t_0,t) = C_{AB}(t)$. In fact, we may set $t_0 = 0$ or alternatively $t_0 = -t$ and write

$$C_{AB}(t) = \langle\, A(t)B(0)\,\rangle - \langle\, A \,\rangle\langle\, B \,\rangle = \langle\, A(0)B(-t)\,\rangle - \langle\, A \,\rangle\langle\, B \,\rangle. \tag{71}$$

Equally important is time-reversal symmetry of the microscopic laws of interaction. This is reflected in the dynamics and allows us to replace t by $-t$. Thus

$$C_{AB}(t) = \langle\, A(0)B(t)\,\rangle - \langle\, A \,\rangle\langle\, B \,\rangle = C_{BA}(t) \tag{72}$$

for any observable, including local observables. This dynamical property is referred to as *reversibility*. It asserts that is does not matter which of the two quantities A, B is measured first. From reversibility and time-translation invariance it follows for any equilibrium correlation function

$$C_{AB}(t) - C_{AB}(0) = -\frac{1}{2}\langle\, (A(t) - A(0))(B(t) - B(0))\,\rangle. \tag{73}$$

There are some further general theoretical predictions about the generic behavior of correlation functions. At least for small times the correlation function does not vanish as the value of A does not change to a new value which is entirely independent of A at a small time interval earlier. Such a behavior would require unphysical infinitely fast changes of the microstate of the system. On the other hand, after a sufficiently long time one expects correlations to decay to zero since measurements taken after a long time become statistically independent (at least approximately). Phenomenologically it is known that in many systems the decay of the correlation follows an exponential law

$$C_{AB} \sim \mathrm{e}^{-t/\tau}. \tag{74}$$

The time scale τ – which determines how rapid the decay is – is called the relaxation time. One objective of a statistical mechanics treatment of such a system would be the derivation of this

quantity from the microscopic parameters of the model. Another frequently observed decay, particularly in glassy systems, has the form of a stretched exponential

$$C_{AB} \sim \mathrm{e}^{-(t/\tau)^{\beta}}. \tag{75}$$

The Kohlrausch exponent β is another parameter to be derived from a microscopic model. Particularly exciting is an algebraic decay

$$C_{AB} \sim t^{-\alpha} \tag{76}$$

with a so-called critical exponent α. In such a relaxation law there is no natural time scale for the decay of correlations. The system is scale invariant in time: $C(\lambda t) = \lambda^{-\alpha} C(t)$. This is often accompanied by a more general scaling behavior

$$C(\lambda r, \lambda^z t) = \lambda^{\nu} C(r,t) \tag{77}$$

referred to as dynamical scaling. Since in such systems natural length scales do not enter correlation functions, the critical exponents α, ν, z do not depend on microscopic details of the model which normally induce natural length scales. Thus there is a good chance for finding models that capture important physical features of the real system, but are simple enough to be amenable to numerical or analytical treatment. The exponents are often universal in the sense that the same exponents can be observed in very different classes of systems. Still they contain important information about hidden symmetries and elementary physical mechanisms that determine the macroscopic behavior of the system. Dynamical scaling is often associated with occurrence of phase transitions.

Ideally one would like to calculate the full correlation function for a given model and compare it with experimental data. However, even in the very simplest toy models this is an extremely difficult task. Fortunately, time-dependent linear response theory may be used to derive a result of wide importance. To state the result we first have to pose the problem. The general setting is the following: Consider a change in an external control parameter $h(t)$ (e.g. a chemical potential, electric field, ...) which is very slow compared to the typical time scale of microscopic fluctuations (so that after a change the system has time to relax to the new equilibrium). Otherwise we make no assumption on $h(t)$. The question is: How does the thermal average $\langle A \rangle$ follow the perturbation $h(t)$?

Let B be the (static) quantity conjugate to h. We assume that the perturbation has existed for a very long time and is switched off at time t. Generally the response at time t of the average of an observable to a small perturbation may be written

$$\langle A(t) \rangle = \langle A \rangle_0 + \int_{-\infty}^{t} dt' \, \mu(t-t') h(t') \tag{78}$$

This is the lowest order approximation in h to the real response which is of course non-linear. The expression *defines* the response function μ which – intuitively speaking – contains the information about the accumulated first-order effect of the perturbation. In appendix B we drive the fluctuation-dissipation theorem (FDT) which provides for an expression for the unknown response function in terms of a correlation function for the system. The theorem asserts

$$\mu(t) = -\frac{1}{k_B T} \frac{d}{dt} C_{AB}(t). \tag{79}$$

We use the FDT to obtain an important result originally due to Albert Einstein. To set the stage we remind ourselves that an ion that drifts in a fluid under the influence of a constant electric field travels with a constant velocity

$$v = F/\zeta \tag{80}$$

where F is the force acting on the ion and ζ is the friction coefficient. In the absence of a field the ion performs Brownian motion with a diffusion coefficient D which is a measure for the mean-square displacement of a Brownian particle

$$\langle\, (x(t) - x(0))^2 \,\rangle = 2dDt \tag{81}$$

Here d is the space dimension in which the Brownian particle moves.
The same formulae apply e.g. to a charged polymer drifting through a gel matrix in electrophoresis and to many other contexts. To keep notation simple we investigate one-dimensional motion. The generalization to three dimension is straightforward.
To make contact with linear response theory as described above we consider as observable A the position $x(t)$ of the ion. The force F due to the electric field plays the role of the conjugate variable $h(t)$ which we switch on at $t = 0$ and then keep constant. Alternatively one may think of the particle as a colloidal particle sinking down in a fluid under the influence of gravitation. In any case, the product Fx which enters the equilibrium distribution has indeed the dimension of an energy, as it should have.
Applied to the ion we have $h(t) = 0$ for $t < 0$ and $h = F$ for $t \geq 0$, moreover $A = B = x$ Hence

$$\langle\, x(t) \,\rangle_F = \langle\, x(t) \,\rangle_0 - \frac{F}{k_B T} \int_0^t dt' \, \frac{d}{dt'} C_{xx}(t') \tag{82}$$

$$= \langle\, x(t) \,\rangle_0 - \frac{F}{k_B T} (C_{xx}(t) - C_{xx}(0)). \tag{83}$$

Applying (73) to (83) gives

$$\langle\, x(t) \,\rangle_F - \langle\, x(t) \,\rangle_0 = \frac{F}{2k_B T} \langle\, (x(t) - x(0))^2 \,\rangle \tag{84}$$

On the other hand we know $\langle\, x(t) \,\rangle_F = x(0) + vt$ where v is given by (80). With (81) we find $Ft/\zeta = FDt/(k_B T)$ and hence the celebrated Einstein relation

$$\frac{1}{\zeta} = \frac{D}{k_B T} \tag{85}$$

which links two seemingly unrelated quantities, here the diffusion coefficient D (coming from fluctuations) and the friction coefficient (originating in the response to an external force). From the construction it is clear that similar results can be obtained in many other physical contexts.

1.4 Nonequilibrium systems

If the environment of a system is changed it will find itself in a nonequilibrium state and start to adapt to the new situation as discussed in the previous section. A different situation arises if the system is kept away from equilibrium permanently by maintaining external conditions which drive a current through the system, as indicated earlier. Such a system is necessarily open with

respect to its environment for in a closed system one cannot sustain a current forever. In this situation there is no basic principle analogous to ergodicity. The only statement one expects to remain true is that entropy will be maximal, in the sense that the system will be found in the most likely set of configurations, given the external conditions. There is, however, no general prescription how to calculate such an entropy.

Onsager relations

If one stays close to equilibrium in a sense similar to what was assumed in linear response theory, one can still argue (and verify experimentally) that some general relations for transport quantities should be valid. This is what Lars Onsager did during the middle of the 20th century and which earned him a Nobel prize.
Suppose one connects two systems at different temperatures $T_{1,2}$ and with different chemical potentials $\mu_{1,2}$ in such a way that a heat and particle transfer between the systems is possible. We assume the process to be so slow that both systems can be regarded as reservoirs where temperature and chemical potential stay essentially constant over the time of observation. One expects a heat flux J^Q and particle flux J^N resp. that maximizes the entropy of the whole system, i.e. the reservoirs and the connection. The reservoirs do not change their state, thus the entropy production occurs in the connection. Since the entropy production is directly linked with the current we express the current not in terms of temperature difference and difference in the chemical potential respectively, but rather as function of the difference in the entropy change w.r.t. internal energy

$$\delta\beta = \frac{\partial S_2}{\partial U_2} - \frac{\partial S_1}{\partial U_1} \tag{86}$$

and particle number

$$\delta(\beta\mu) = \frac{\partial S_2}{\partial N_2} - \frac{\partial S_1}{\partial N_1} \tag{87}$$

in the two reservoirs.
For small difference in T and N resp. also $\delta\beta$ and $\delta(\beta\mu)$ are small and we write in the spirit of linear response theory

$$J^Q = L_{QQ}\delta\beta + L_{QN}\delta(\beta\mu) \tag{88}$$
$$J^N = L_{NQ}\delta\beta + L_{NN}\delta(\beta\mu) \tag{89}$$

with the Onsager coefficients L. These coefficients appear also in a time-dependent setting. They can be expressed in terms of correlation functions in equilibrium (at $T_1 = T_2 = T$ and $\mu_1 = \mu_2 = \mu$) and satisfy the remarkable symmetry property

$$L_{QN} = L_{NQ}. \tag{90}$$

It has its origin in reversibility (72) of the equilibrium process which allows for interchanging the quantities that represent Q and N respectively in the correlation function.
These relations imply that even in the absence of a temperature gradient between the two systems ($\delta\beta = 0$) there will not only be a particle current, but also a heat current since $L_{QN} = L_{NQ} > 0$. (Of course, to sustain such a current for long times while keeping T fixed in both reservoirs, one reservoir has to be supplied with heat while the other has to be cooled.) Analogously there will be a particle current for different temperatures even if the two chemical potentials are equal. The Onsager symmetry (90) generalizes to situations with more than two

types of current. Generally one has $L_{XY} = L_{YX}$. Unless some of these coefficients vanish, a gradient in any single control parameters generates a current of *all* conjugate quantities.

Systems far from equilibrium

There are no known general principles that determine the properties systems far from equilibrium. The sad truth is that different non-equilibrium systems require different treatments, there is no unifying framework. Here we discuss only the master equation approach as it is not only enormously successful for theoretical analysis, but also the physical basis for one of the most widespread techniques of numerical simulation in statistical physics, viz. the Monte Carlo simulation.

The master equation governs the time evolution of the probability distribution, see (109). The master equation is a first order differential equation in time and defines a Markov process. The physical picture behind this approach is semi-microscopic. It is the notion that by observing a system not at each space-time point, but rather average over certain intervals in space and time the subsequent microstates of the system follow an effective random process. Furthermore one assumes that this process has no explicit memory of its own past, i.e., the microstate of a system at time $t + dt$ depends only on the microstate at the previous time instant t. This approach has been applied most successfully to many equilibrium problems, the simplest example being Brownian motion (see below). However, the underlying assumption of effective randomness without time-memory does not rely on any notion specific for equilibrium. So this approach may be applied to non-equilibrium phenomena as well. It often works but also often fails and one then has to resort to other ideas such as molecular dynamics simulations or adding noise to microscopic deterministic equations of motion as in the Langevin approach.

We start our discussion by reviewing the physical origin of Brownian motion and how it can be described in terms of a master equation. In 1827 Brown observed that small particles such as pollen particles immersed in a fluid perform an irregular type of motion, appearing to be essentially random in nature. In the absence of space-dependent external forces the probability density to find a Brownian particle at time t in some infinitesimal test volume is empirically found to be a Gaussian distribution. For notational simplicity we write this distribution for one-dimensional motion as in the previous section. Suppose that at the initial time $t = t_0$ the particle is located at position x_0. Then at time t the probability density P to find the particle in the infinitesimal interval $x, x + dx$ is given by

$$P(x, t|x_0, t_0) = \frac{1}{\sqrt{2\pi D(t - t_0)}} \exp\left(\frac{x - x_0 - v(t - t_0)}{2D(t - t_0)}\right). \tag{91}$$

This distribution has two parameters, v and D. The mean position

$$\langle\, x(t)\,\rangle = \int_{-\infty}^{\infty} dx\; xP(x, t|x_0, t_0) = x_0 + v(t - t_0) \tag{92}$$

illustrates the significance of the quantity v. It is the (mean) drift velocity of the Brownian particle. According to (80) $v = 0$ if there is no external force. But of course the particle still moves. Because of the randomness the actual position is not predictable, but the fluctuations around the mean

$$\langle\, x^2(t)\,\rangle - \langle\, x(t)\,\rangle^2 = 2D(t - t_0) \tag{93}$$

provides us with information about the width of the distribution. Eq. (93) shows that the parameter D is the diffusion coefficient. We stress that v, D completely specify the distribution

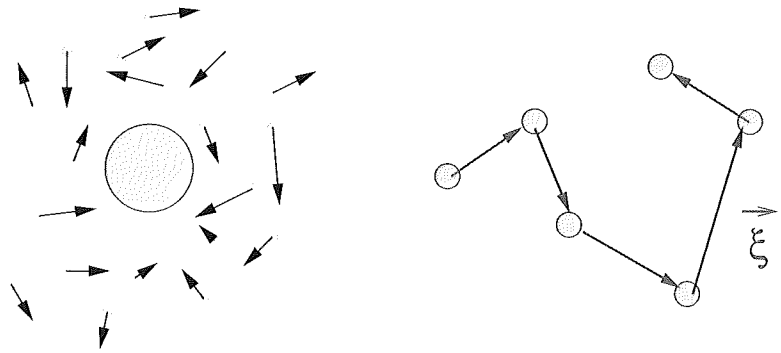


Fig. 5: *Brownian particle immersed in a fluid (left). The thermal motion of the fluid particles (small) causes collisions with the Brownian particle (large). Snapshots of the position of the Brownian particle taken after regular time intervals τ (right).*

of the Brownian particle. All averages that one might conceivably wish to measure depend only on v and D. We also stress that averages are the only predictable quantities and hence the only quantities that can be used as an experimental test for any theory of Brownian motion. Because of translation invariance in space and time we set $x_0 = 0$ and $t_0 = 0$ in all subsequent considerations.

The randomness is nowadays understood as originating in the incessant thermal collisions of the water molecules with the much larger pollen particles. The same mechanism drives the stochastic motion of colloidal particles in a fluid. In the simplest case of just a single particle one may therefore entirely disregard the properties of the fluid and just view it as the source of a random force. If one now takes snapshots of the particle position at regular time intervals τ one sees a picture similar to Fig. 5. After each time step the particle has moved by a random displacement $\vec{\xi}$.

Since the detailed properties of the random forces are in general not easy to determine we skip this problem here and adopt a phenomenological point of view as done by Einstein in his famous work on diffusion in 1905 (for original references see [2]). The essential idea for describing the random motion is what in modern physics is referred to as *coarse graining*: Rather than studying microscopic details of the dynamics with a very fast camera and a microscope with an ultra-high resolution (neither of which were available when Einstein wrote his paper) we consider time intervals θ that are large compared to all microscopic times such as collision times between fluid molecules, relaxation times for atomic or molecular excitations, correlation times between subsequent collisions of fluid molecules with the Brownian particle etc. The important insight by Einstein and independently also by Smoluchowski behind this coarse graining of time is that what is observed in experiments is not the perpetual change in the velocity of the Brownian particles due to collisions. Between two successive observations of the position of the Brownian particle the velocity has independently grown and decayed many times. Hence what is observed is the net displacement resulting after many variations of the velocity.

Since in our coarse graining the correlations between collision have long decayed, these displacements are independent random variables and one may write

$$x(t+\theta) = x(t) + \xi(t) \tag{94}$$

Here we have returned to one-dimensional motion. In three dimension this equation holds for

each space coordinate separately. The time coordinate $t = n\theta$ is an integer multiple of the basic observation time step θ. Eq. (94) has the form of a deterministic evolution equation (here for ballistic motion) with a random term. This is a discrete-time Langevin equation which is easy to solve by iteration. At time t the position of the particle is given by

$$x(t) = \sum_{k=0}^{n-1} \xi(k\theta) \tag{95}$$

which is again a random variable.

Due to the very simple nature of this random process we could conclude our discussion of Brownian motion already here by invoking a famous mathematical theorem, the central limit theorem. It states that a sum of independent random variables (here $x(t)$) where the random variables have finite mean v and finite variance $\sigma^2 = 2D$ converges (in distribution) after many time steps to the Gaussian distribution (91) with mean (92) and variance (93). This explains the origin of the Gaussian as resulting from the statistical independence of many collisions). It also relates the macroscopic parameters v and D to microscopic parameters, viz. the mean and variance resp. of the microscopic displacements.

The time evolution of a random process involving interacting particles is far more complex. To introduce the master equation which can formally handle also interacting systems we rederive the Gaussian as follows. We also coarse grain space by restricting our observation of the system to length scales that are of the order of the diameter r_0 of the Brownian particle.[9] This means that we introduce a discrete space coordinate $x = nr_0$ and say that the particle is located at site n if its center of mass is in the interval between $(n - 1/2)r_0$ and $(n + 1/2)r_0$. Thus Brownian motion becomes a random walk on a lattice defined by all integers. At each time step the random walker hops with some probability p_m from site n to $n + m$. The hopping probabilities p_m incorporate the microscopic details of the random forces which drive the Brownian motion. In our setting the particle is found with probability 1 at site 0 at the initial time $t = 0$. The question is how the probability distribution evolves in time.

We now invent two students who have been given the assignment to develop a microscopic theory for the random walk, i.e., they have to come up with a proposal what p_m should be. Student A figures that the simplest approach would be to just propose that $p_m = 0$ for $|m| > 1$ since anyway hopping by more than lattice unit is not so likely for sufficiently small θ. So the idea is to set

$$p_1 = p, \quad p_{-1} = q, \quad p_0 = 1 - p - q \tag{96}$$

with sojourn probability p_0. The difference $p - q$ of hopping rates should be proportional to the box size r_0, so A sets

$$p - q = \delta r_0 \tag{97}$$

and just for notational convenience also $\gamma = p + q$. The parameters δ, γ are to be fitted from experimental data since A could not handle the microscopic details of Brownian motion and therefore did not find any theoretical prediction for these quantities. It looks as if Student A is a rather laid back sort of person.

[9] One may further coarse grain to scales that are large not only compared to the size of the fluid particles but also large compared to the typical diameter of the Brownian particles. This voluntary imprecision (which would be easy to overcome by choosing a microscope with sufficiently large magnification) implies that one may replace the individual particle positions $\vec{x}_i(t)$ by a concentration field $c(\vec{x}, t)$: In the observed volume one always assumes to be so many particles that an average concentration is seen rather than the individual particles. This is the basis of the Smoluchowski equation discussed in the lecture by Jan Dhont.

Student B is far more sophisticated. He (or she) thinks that the approach of A is too simple to have anything to do with reality. Instead he (let us stick with the male form) suggests to keep at least 42 non-vanishing hopping rates and set only the remaining p_m for $|m| > 21$ to zero, just as the simple-minded A did for $|m| > 1$. He sets

$$p_m - p_{-m} = \delta_m r_0, \quad p_m + p_{-m} = \gamma_m \tag{98}$$

He thinks of some nice hypothesis concerning the value of the 42 parameters δ_m, γ_m which he is keen to have verified through an experiment. B feels very proud and superior to A.
Anything wrong with B? In order to pass judgment we have to analyze the predictions of the two models. Let $P_k(n)$ be the probability density to the find the particle at time step n at position k. According to the definition of the hopping process the time evolution of $P_i(n)$ is given by the *master equation*

$$P_k(n+1) = \sum_{m=-\infty}^{\infty} p_m P_{k-m}(n) \tag{99}$$

Notice that the Langevin equation (94) is stochastic evolution equation which yields the actual position of the particle, whereas the master equation is a deterministic evolution equation for the distribution of the random position.
This equation can be solved by a Fourier transformation (see appendix C). The result is

$$P_k(n) = \frac{1}{2\pi r_0} \int_{-\pi}^{\pi} d\alpha \, \mathrm{e}^{-i\alpha k} \Lambda^n(\alpha) \tag{100}$$

with the Fourier transformed hopping rates $\Lambda(\alpha)$ (118). For the interpretation of this solution we have to remind ourselves that our coarse graining procedure is arbitrary and therefore cannot carry much physical information for small lattice distances k and few time steps n where the precise nature of the discreteness plays a role. Only after many steps (in both space and time) the details of our coarse graining become immaterial because the size r_0, θ_0 of the discrete elementary space and time intervals becomes negligible compared to the large travel distance and large elapsed time. Indeed, only for macroscopic distances and time (meaning very large k and n respectively) do we expect to have an accurate description of space- and time-continuous Brownian motion. To recover continuous space and time we reintroduce x, t and take the limit $r_0 \to 0$ and $\theta \to 0$ such that the macroscopic values $x = kr_0, t = n\theta$ are fixed. Moreover, from the form of the Gaussian distribution we anticipate that we should consider space-time scales such that x^2/t is a constant, since the Gaussian depends on x only in this particular functional combination with time. Thus we take the limit $r_0 \to 0$ such that $r_0^2/\theta = 1$. The microscopic model is thus completely determined by the parameters r_0, δ_m and γ_m respectively.
With first and second moments of the hopping probabilities

$$u = \sum_{m=1}^{\infty} m\delta_m = 1/r_0 \sum_{k=-\infty}^{\infty} m p_m \tag{101}$$

$$\Delta = \sum_{m=1}^{\infty} m^2 \gamma_m = \sum_{m=-\infty}^{\infty} m^2 p_m \tag{102}$$

it is shown in appendix C that the expression (100) for the probability turns into the Gaussian distribution (91) with $u = v$ and $\Delta = D$. We conclude that all model parameters enter into

physically observable quantities only in terms of the first and second moment resp. of the distribution of hopping probabilities.
Therefore the approach of student B has absolutely no extra merit compared with the simpler model of student A. The model of B has not more predictive power than the model of A since all the 42 parameters enter the theoretical prediction only in the two combinations (101), (102) and hence cannot be measured individually. In fact, if one could not solve the model exactly, but would have to rely on numerical simulation, the more complicated model of B would be very stupid. Even though this model looks much closer to reality, there is no gain in information (his hypothesis for the 42 values cannot be verified experimentally), but considerable loss in efficiency.
This is an important conclusion of general practical importance when one tries to construct a model for a real system. Trying to be very precise and introducing many parameters is just a waste of time (and computer power) as long as these parameters are not relevant for what one can observe. Of course, it is not always clear beforehand which parameters are relevant and which are not. But it usually pays off to spend some thought on this issue rather than blindly taking into account any microscopic detail that indeed exists on a microscopic scale.

Universality

The observation of the last subsection leads to the theoretical concept of *universality*. It takes us to the heart of statistical mechanics and to a major reason for its success. This notion was originally developed for characterizing critical phenomena where it is intimately related with scale invariance as indicated above. Indeed the Gaussian (91) satisfies the scaling form (77) for dynamical scaling with dynamical exponent $z = 2$ and $\nu = 1/2$. The probability distribution of a Brownian particle looks identical, no matter whether it observed on a millimeter-scale or a kilometer-scale and no matter whether we observe a pollen particle, a colloidal particle or a biological particle in a cell.[10] Only when we zoom in to strongly, i.e., when the microscopic details of its motion become visible, universality breaks down and we have develop complicated models with many parameters to describe what we see.
Universality in different, more qualitative sense, applies to physical mechanisms such as hysteresis (which is associated with broken ergodicity), or nucleation which describes the initial stages of the growth of a crystal. These and many other mechanisms appear in a different guise in a large variety of systems, leading to similar physical laws for analogous, but different physical quantities. Even though caution is required with the interpretation of similar relations for different systems (as demonstrated in Sec. 2) one may learn what the dominant physical mechanisms are that determine macroscopic behavior, if a proper correspondence is established. The upshot is that universality allows us to construct very simple models for complex systems, as long as we are cautious with the interpretation and focus only on those properties of the model that are indeed universal. Simple models allow us to identify what is essential on a microscopic level for universal behavior.
In the spirit of universality we conclude this review by presenting a toy model that has been introduced in the late 60ies for the kinetics of protein synthesis. In a very (!) simplified manner this process may be described as follows (Fig. 6). A long biopolymer, the messenger-RNA (m-RNA), contains the information which sequence of amino acids the corresponding protein consists of. It consists of an ordered sequence of codons $1, 2, \ldots, N$ (each of which is itself

[10] Whether a particle that performs a random motion really performs Brownian motion at least in an approximate sense is, of course, an entirely different issue.

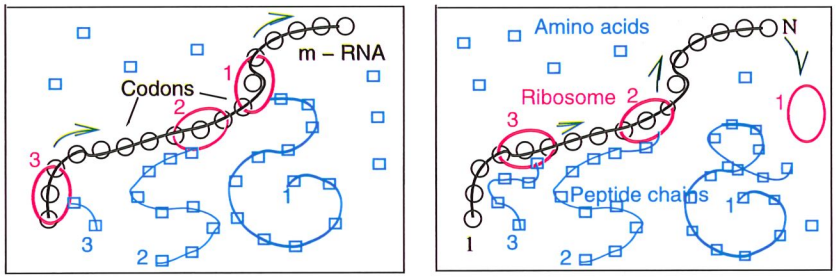


Fig. 6: *Cartoon of protein synthesis from a m-RNA template. The ribosomes (magenta) move forward stochastically along the m-RNA and synthesize a protein (cyan). Left: The initiation of the translation process takes place at the first codon (site 1). Right: At the terminal site N the ribosome is released after synthesis of a complete protein.*

a triplet) and each codon represents an amino acid. In the cell there are also ribosomes which attach at initial site 1 of the m-RNA. They "read" the genetic information the codon represents and "searches" for the corresponding amino acid. When a molecule of that particular amino acid is found it attaches to the ribosome. The mechanism for finding is very complicated and one expects the time required to be a random number. Having accomplished this task proceeds to the next codon, i.e., its motion is directed along the chain. The process of finding the corresponding amino acid is repeated and this molecule is appended to the first. This process of translation of the genetic information into a growing chain of amino acids goes on until the terminal site N is reached. The last amino acid is appended to the chain and the protein is completed. Subsequently the ribosome is released from the m-RNA. While the process of translation is still going on new ribosomes may attach to the initial site so that several ribosome may move along the same m-RNA. Of course they cannot be located at the same codon simultaneously. In physics terms this corresponds to a repulsive interaction, at least due to excluded volume interaction, but likely to be far more complicated.

In the early 60'ies it had been observed that ribosomes move faster at the beginning of the m-RNA than at the end. Indeed there seemed to a specific point or region on the m-RNA from which on they appeared to slow down. However, data points were few and noisy and no clear interpretation was found. One candidate for explanation was the control point theory which suggested that ribosomes would slow down when passing a specific codon, but no reason was suggested as to why that should happen.

This unsatisfactory state of affairs led a group of biophysicists to suggest a simple model for the process described above which later became known as the asymmetric simple exclusion process (ASEP). This is a random process which incorporates the above-mentioned ideas of

- random motion
- driving force
- repulsive interaction

in the simplest conceivable fashion. The m-RNA is represented by a lattice with N sites. The random motion is modelled by the simple lattice random walk with hopping only to the neighboring site, similar to the process described above, but in continuous time. The driving force is

represented by asymmetric hopping rates $p \neq q$ which are hopping probabilities per infinitesimal time unit. In fact, since ribosomes are totally directed one chooses $q = 0$. The repulsive interaction is modelled by site exclusion, i.e., each lattice site may be occupied by at most one particle.
In addition to these random processes for the bulk motion one has to specify a process for initiation and release. This corresponds to an open system which has particle exchange with its environment, but only at its boundaries. At site 1 particles are postulated to be created with rate αp, at site N particles are annihilated with rate βp.
This open system can never relax to equilibrium as the open boundaries allow the driving force to sustain a macroscopic steady state current. Hence none of the principles of usual thermodynamics apply. Also linear response theory must fail completely. What one needs to know is the local density ρ_n of particles (ribosomes) along the chain as well as the stationary particle current j. Given this quantities one can determine the mean velocity

$$v_n = j/\rho_n \tag{103}$$

of the ribosomes as a function of their position, but there seems to be no hope to derive this quantity from basic principles.
Fortunately an exact solution for the steady state of this particular model is possible and also a more general theory of boundary-induced phase transitions which allows us to apply principles valid for driven diffusive systems to the question at hand [4]. The particle flux for the ASEP

$$j = \rho(1 - \rho) \tag{104}$$

is nonlinear in the density ρ. In particular, it has been shown that for $\alpha = \beta < p/2$ (equal initiation and release rate) there is a discontinuous phase transition from phase with low bulk density to a phase with high density. Along the phase transition line one has phase coexistence with left domain (starting at site 1) of low density, separated by a shock – a sharp increase in the average density – from a domain of high density, extending up to site N. The position of the shock, which is a stable collective many-particle mode, performs an *undirected* random walk along the chain.
Before coming to an interpretation of this result we note that such a shock is a phenomenon that probably every reader is familiar with. Essentially the same occurs in traffic flow on a road. The end of a traffic jam (congested region of high density) is marked by a very sudden drop of the car density to a flow region of lower density. Most traffic jams are non-stationary and the end of the jam, i.e., the shock moves backwards. However, for low in-flow, just balancing the low flow inside the congested region, this shock is also expected to perform an unbiased random motion. The stability and occurrence of shocks is a universal phenomenon in the qualitative sense of universality. It is a characteristic feature of all system with a flux that has a nonlinear dependence on the density.
As applied to protein synthesis this result suggests that there is no static control point, but rather a dynamical breakdown of the speed of ribosomes, caused by a "traffic jam". This picture is consistent with experimental data available in the 60'ies and has laid the control point theory to rest. Also more recent microscopic images of ribosomes attached to an m-RNA give qualitative support to this picture of a traffic jam. Of course the ASEP does not answer any quantitative question concerning protein synthesis. As a quantitative model it is much too crude to allow for detailed comparison with experimental data. However, its very simplicity has helped to exhibit the relevant causes for the traffic jam effect: random motion, exclusion, and drift.

1.5 Conclusions

It would be dangerous to present a one-page summary of a 37-page summary of a subject that has occupied for about two centuries tens of thousands of physicists and chemists, plus an uncountable number of engineers, and which explains phenomena ranging from the functioning of a refrigerator to some simple properties of protein synthesis. Perhaps the most important conclusion is that statistical mechanics in general and nonequilibrium statistical mechanics in particular are relevant for a vast and still largely unexplored section of physical reality. Biology is one field where some notions of statistical mechanics as well as other branches of physics are likely to continue to turn out to be fruitful.

Appendices

A Some combinatorics: The degeneracy of microstates

For $M = 2$ we have to count the number of *distinct* possibilities of placing 2 identical cubes onto N sites which is a combinatorial problem. To work this out we imagine a linear array of squares. (Clearly the shape of our basic area does not matter for the number of possibilities of placing the cubes.) To count the number of configurations we initially represent each site by a 0, i.e., the whole array is initially represented a string of N zero's. For each distinct configuration we now look at the array. If there is a cube at site i we squeeze a 1 into the string to the left of i'th zero. For two cubes on the same site we add two consecutive one's. When all cubes have been accounted for we delete the last zero (which served only to define where to put 1's if the last site in the array was occupied.) We have thus generated a new array of length $N + 2 - 1 = N + 1$, consisting of zero's and one's.
This is a one-to-one mapping, i.e., to each such string corresponds a unique configuration of cubes and vice versa. Hence, in order to count the number of distinct configurations of one needs to find the number of possibilities of distributing $M = 2$ objects (the one's) in $L = N+1$ boxes. Counting all configurations for $M = 2$ yields the binomial coefficient

$$C_2^L = \frac{L!}{(L-2)!2!} = N(N+1)/2. \tag{105}$$

Generally, for M cubes one generates a string of length $L = N + M - 1$ with M one's. This is a well-known combinatorial problem (occurring e.g. in a lottery where one places 6 objects (crosses) into 49 boxes (number fields)). The number of states $\Gamma(M, N)$ with energy $E = \epsilon_0(M + N/2)$ is the binomial coefficient C_M^{N+M-1}, a quite large number even for only $M = 6$ cubes and a very small system of $N = 44$ squares.
In order derive the binomial coefficient we first assume that the M objects '1' to be arranged in a string of length L are not identical, but represented by the numbers $1, 2, \ldots, M$. To count the number of these configurations we simply observe the following: There are L possibilities to place the number 1. This leaves $L-1$ empty spaces for putting the number 2. Having placed the number 2 somewhere, there are $L - 2$ empty spaces where number 3 can be placed, and so on. Hence there are $L(L-1)\cdots(L-M+1)$ configurations of the distinct numbers $1, 2, \ldots, M$. Now we have to remember that we want to count the number of distinct configurations of *identical* objects. This can be achieved by renaming all numbers k to number 1. Thus each permutation of the M distinct numbers gives rise to the same arrangement of 1's. The number of permutations of M objects is $M!$ and the number we wish to find is therefore

$$\Gamma(M, N) = \frac{L(L-1)\cdots(L-M+1)}{M!} = \frac{L!}{M!(L-M)!} = C_M^L \tag{106}$$

with $L = N + M - 1$.
Notice that for $M \ll L$ one may use the approximation $C_M^L \approx L^M/M!$. For $M \approx L \gg 1$ one obtains (17)

B Derivation of the fluctuation-dissipation theorem

The fluctuation-dissipation theorem can be obtained in very general form for arbitrary time-dependent perturbations $h(t)$. Therefore, in general it is a theorem relating a time-dependent

response with a time-dependent correlation function of which (79) is only a special case. Here remain with this simplified scenario, corresponding to a perturbation with constant h up to time $t = 0$. We study how the system evolves after switching off the perturbation.
Within this framework we may write (78) for $t > 0$

$$\langle A(t) \rangle_h - \langle A(t) \rangle_0 = h \int_t^\infty dt' \, \mu(t') \tag{107}$$

By definition

$$\langle A(t) \rangle_h = \sum_\eta A(\eta) \Psi(\eta, t) \tag{108}$$

where η denotes all microstates of the system and $\Psi(\eta, t)$ is the probability distribution (normalized statistical weight) at time $t > 0$ for the microstates. We cannot say what $\Psi(\eta, t)$ looks like, but from general considerations which we do not discuss here we can write this quantity in terms of its initial distribution at time $t = 0$

$$\Psi(\eta, t) = \sum_{\eta'} F(\eta, t|\eta', 0) \Psi(\eta', 0) \tag{109}$$

where the *propagator* $F(\eta, t|\eta', 0)$ contains all the information about the dynamics of the system. It is the probability that the system moves from microstate η' at time $t = 0$ to microstate η at time t. Notice that since $t > 0$ there is no external perturbation any more. Hence this propagator describes the time evolution in the absence of the perturbation.
On the other hand, the constant perturbation has acted from the infinite past up to $t = 0$. Hence the distribution of the system $\Psi(\eta', 0)$ is the equilibrium distribution of the system with constant field h, conjugate to the quantity B. This yields

$$\Psi(\eta', 0) = \mathrm{e}^{-\beta E - \beta h B} / Z(h) \tag{110}$$

where E is the energy of the microstate η' in the absence of the field and $B = B(\eta')$ is the corresponding value of B. As usual, $Z_h = \sum_{\eta'} w(\eta')$ with the Boltzmann factor $w(\eta') = \exp(-\beta E - \beta h B)$.
Expanding to first order in h yields

$$\Psi(\eta', 0) = \Psi_0(\eta', 0) \left[1 + \frac{h}{k_B T} (B(\eta') - \langle B \rangle_0 \right] \tag{111}$$

with the zero-field distribution $\Psi_0(\eta', 0)$ and the zero-field thermal average $\langle B \rangle_0$. Plugging this into (109) yields

$$\Psi(\eta, t) = \sum_{\eta'} F(\eta, t|\eta', 0) \Psi_0(\eta', 0) + \beta h \sum_{\eta'} F(\eta, t|\eta', 0) \Psi_0(\eta', 0) (B(\eta') - \langle B \rangle_0) \tag{112}$$

$$= \Psi_0(\eta, t) + \beta h \sum_{\eta'} F(\eta, t|\eta', 0) \Psi_0(\eta', 0) (B(\eta') - \langle B \rangle_0). \tag{113}$$

To proceed further we have to remind ourselves of the definition of a correlation function $C_{AB}(t) = G_{AB}(t) - \langle A \rangle \langle B \rangle$. The quantity $G_{AB}(t)$ is the following: One measures B at $t = 0$ when it was in microstate η'. The system evolves in time to some microstate η at time t. Then A is measured. The expectation value $G_{AB}(t)$ is the average over many such measurements where in each case the microstates $\eta'(0)$, $\eta(t)$ were different. However, since we

are in equilibrium we know that $\eta'(0)$ occurs with equilibrium probability $\Psi(\eta', 0) = w(\eta')/Z$. We also know that given η' the microstate η is found at time t with probability $F(\eta, t|\eta', 0)$. Averaging means averaging over all η, η' and we find

$$G_{AB}(t) = \sum_{\eta',\eta} A(\eta)B(\eta')F(\eta, t|\eta', 0)\Psi(\eta', 0). \tag{114}$$

Normalization of probability means $\sum_\eta \Psi(\eta) = 1$. Thus putting everything turns (108) into

$$\langle A(t) \rangle_h = \langle A \rangle_0 + \beta h C_{AB}(t). \tag{115}$$

With (107) we then have the fluctuation-dissipation theorem (79) by taking the derivative w.r.t. t.

C Solution of the lattice random walk equation

The master equation (99) can be solved by a Fourier transformation

$$\hat{P}(\alpha, n) = \sum_{k=-\infty}^{\infty} \mathrm{e}^{i\alpha k} P_k(n). \tag{116}$$

Using (99) one immediately obtains by shifting the index of summation $k \to k + j$ the simple recursion for the Fourier modes

$$\hat{P}(\alpha, n + 1) = \Lambda(\alpha)\hat{P}(\alpha, n) \tag{117}$$

with the Fourier transformed hopping probabilities

$$\Lambda(\alpha) = \sum_{m=-\infty}^{\infty} p_m \mathrm{e}^{i\alpha m} \tag{118}$$

The solution of (117) is given by

$$\hat{P}(\alpha, n) = \Lambda^n \hat{P}(\alpha, 0) \tag{119}$$

where the initial value for $n = 0$ is determined by the initial distribution $P_k(0) = \delta_{k,0}/r_0$. Hence $\hat{P}(\alpha, 0) = 1/r_0$. With the back transformation

$$P_k(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\alpha \, \mathrm{e}^{-i\alpha k} \hat{P}(\alpha, n) \tag{120}$$

one then gets (100). Expressing this in terms of the space-time coordinates we replace $k = x/r_0$, $n = t/\theta$ and write $\Lambda^n = \exp(n \ln \Lambda)$. Substituting also the integration variable $\alpha \to \alpha/r_0$ turns (100) into

$$P_x(t) = \frac{1}{2\pi} \int_{-\pi/r_0}^{\pi/r_0} d\alpha \, \mathrm{e}^{-i\alpha x + (t \ln \Lambda(r_0\alpha))/\theta}. \tag{121}$$

Now we take the limit $r_0 \to 0$ such that $r_0^2/\theta = 1$. Taking the limit $r_0 \to 0$ extends the integration domain from $-\infty$ to ∞. To handle the expression in the exponent we first note that the sojourn probability is given by $p_0 = 1 - \sum_{m=1}^{\infty} \gamma_m$ and therefore

$$\Lambda(r_0\alpha) = 1 + ir_0 \sum_{m=1}^{\infty} \delta_m \sin(mr_0\alpha) - \sum_{m=1}^{\infty} \gamma_m (1 - \cos(mr_0\alpha)) \tag{122}$$

$$\approx 1 - r_0^2 \left(-iu\alpha + \frac{1}{2}\Delta\alpha^2 \right) + O(r_0^4) \tag{123}$$

The higher order terms in r_0 (which result from the Taylor expansion of the sine and cosine respectively) contain higher moments of the distribution of the hopping probabilities. They do not contribute to the integral (121) in the limit $r_0 \to 0$. Secondly we use

$$\lim_{x\to 0}(1 - cx)^{\frac{t}{x}} = \lim_{x\to 0} \mathrm{e}^{-t/x \ln(1-cx)} = \mathrm{e}^{-ct} \tag{124}$$

The second equality in (124) comes from the Taylor expansion of the logarithm.
Putting (122) and (124) together yields

$$\lim_{r_0\to 0} [\Lambda(r_0\alpha)]^{t/r_0^2} = \mathrm{e}^{iu\alpha t - \frac{1}{2}\Delta\alpha^2 t} \tag{125}$$

Inserting this in (121) leads to

$$P(x,t) = \lim_{r_0\to 0} P_k(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\alpha \, \mathrm{e}^{-i\alpha(x-ut) - \frac{1}{2}\Delta\alpha^2 t} \tag{126}$$

This is the Fourier representation of the Gaussian (91) with $x_0 = 0$, $t_0 = 0$, $v = u$, $D = \Delta$.
Notice that (91) is the propagator (in the sense of the previous appendix) for the probability distribution of a Brownian particle in a constant field. It gives the probability to move from microstate $x_0 = 0$ at time $t_0 = 0$ to microstate x at time t. Its time evolution is governed by the master equation which for continuous space and time turns into the diffusion equation

$$\frac{\partial}{\partial t} P(x,t) = -v \frac{\partial}{\partial x} P(x,t) + D \frac{\partial^2}{\partial x^2} P(x,t) \tag{127}$$

The same equation is also valid for heat transport, with a different interpretation for v, D.
From (126) we read off a continuous spectrum of relaxation times $\tau = 2/(D\alpha^2)$ for the Fourier modes of the propagator of Brownian motion. For small α (long wave lengths) the relaxation time diverges. This is a characteristic feature of dynamical critical phenomena which involve power laws for the decay of correlations. A simple exponential decay corresponds to a finite relaxation time even for large wave lengths. Stretched exponentials may arise as a superposition of infinitely many, but disordered finite relaxation times.

References

[1] M. Toda, R, Kubo and N. Saito, *Statistical Physics I* (Springer, Berlin, 1991)

[2] M. Doi and S.F. Edwards, *The Theory of Polymer Dynamics* (Oxford University press, Oxford, 1986)

[3] N.G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1992)

[4] G.M. Schütz, *Exactly solvable models for many-body systems far from equilibrium*, in: Phase Transitions and Critical Phenomena Vol. **19**, eds. C. Domb and J.L. Lebowitz (Academic Press, London, 2000)

[5] L.D. Landau und E.M. Lifschitz, Lehrbuch der theoretischen Physik, Band V (Statistische Physik I) (Akademie Verlag, Berlin, 1984)

[6] A.L. Lehninger, D.L. Nelson and M.M. Cox, Prinzipien der Biochemie (Spektrum Akademischer Verlag, Heidelberg, 1994)

A 2 Translational Brownian Motion

J. K. G. Dhont
Institut für Festkörperforschung (IFF)
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Free Brownian Motion **3**
2.1 The Langevin Equation 3
2.2 Time Scales 6
2.3 The Probability Density Function (pdf) for the Position Coordinate 8

3 Brownian Motion of Interacting Particles **10**
3.1 Self Diffusion 11
3.2 Collective Diffusion 13
3.3 Diffusion Towards Receptors 15
3.4 Fick's Diffusion Coefficient (without Hydrodynamic Interactions) 20
3.5 Diffusion in a Inhomogeneous Solvent : Brownian Motion of a Cell Towards Food-Rich Regions 24

4 The Generalized Diffusion Equation **27**
4.1 A Primer on Hydrodynamic Interactions (HI) 27
4.2 The Leading Order Expansion of Hydrodynamic Interaction Functions : The Oseen Matrix 28
4.3 The Smoluchowski Equation for Spheres 34
4.4 Fick's Law Revisited 35

5 Brownian Motion in External Fields : Sedimentation **38**

A Appendix : Evaluation of integrals in the calculation of Fick's diffusion coefficient. **40**

1 Introduction

Molecules and atoms always exhibit thermal motion, where the corresponding average kinetic energy for translational motion per molecule/atom is equal to $\frac{3}{2}k_BT$ (with k_B Boltzmann's constant and T the absolute temperature), independent of whether the molecule/atom is in the gas, liquid or solid state. This average kinetic energy is independent of the mass and size of the molecule. Thermal motion is still important for the collective behaviour of relatively large particles, in the size range of a few nanometers to several tens microns (typical sizes of simple molecules, like a water molecule, are about $0.1\,nm$). Examples of such large particles are inorganic particles, like SiO_2-spheres or $AlOOH$-rods, proteins, (bio-)polymers like $Polyethyleneoxide\,(PEO)$, $fd-virus$ and DNA, and entire cells. These large agglomerates are usually dissolved in water (in model studies, organic solvents are sometimes used). Since thermal motion is important for solutions of such "macromolecules", they obey the same laws of thermodynamics and statistical physics for simple molecular systems (except that quantum effects are not important for the collective behaviour of macromolecules). In particular, they exhibit a whole range of phase transitions and non-equilibrium phenomena that are also found for simple molecular systems. Contrary to simple molecular systems, where the electronic structure of molecules sets the *inter-molecular potential*, there is a large variety of *inter-macromolecular potentials*. For example, the surface of the macromolecules can be charged, where the screening of these charges can be changed by varying salt concentration in the solvent, the surface can be coated by either short or long polymers in either good or bad solvents and van der Waals interactions can be important depending on the differences of the refractive index of the solvent and the macromolecule. The interaction potential can thus vary from short-ranged attractive to long-ranged repulsive, depending on the system under consideration. This variation of interaction potentials gives rise to many phenomena and phase states that are not encountered in simple molecular systems.

Thermal motion of these large molecules is commonly referred to as *Brownian motion*, after the english botanist Robert Brown, who was the first to report on irregular motion of pollen grains in water as observed through a microscope, about 150 years ago. This irregular motion can be thought of as being due to collisions of water molecules with the surface of the macromolecule, and thus visualizes indirectly thermal motion of water molecules. At the time Brown reported his observations, the existence of real molecules was not yet generally accepted. This took another 70 years, where observations by Perrin on macromolecules and theory as developed by Einstein played an essential role.

In colloid science, where usually compact and relatively stiff macromolecules are studied, macromolecules are sometimes referred to as *Brownian particles*. For historical reasons these macromolecules are also referred to as *colloidal particles*, and their solutions are commonly referred to as *dispersions* or *suspensions*. It should be kept in mind, however, that polymers and bio-macromolecules exhibit Brownian motion as well.

In bio-systems there are essentially two ways to transport macromolecules. A macromolecule can be bound to a carrier that moves under the influence of an underlying "backbone" (like molecular motors), or the macromolecule can move due to Brownian motion. Brownian motion can have a preferred direction due to the existence of an external field or due to concentration gradients, giving rise to what is referred to as "*diffusion*". This lecture is a (certainly biased) treatment of Brownian motion. The lecture is restricted to translational Brownian motion of spherical macromolecules, that is, rotational Brownian motion is not considered.

More (partly introductory) material on macromolecular systems can be found in refs.[1]-[4].

These references cover part of the material treated in the present lecture.
Brownian motion of a single sphere in an otherwise unbounded solvent is considered in section 2 on the basis of Newton's equation of motion, where the random collisions of solvent molecules are accounted for by a random force. Einstein's diffusion coefficient is defined, and the Brownian (or diffusive) time scale is introduced. On this time scale inertial forces are shown to be very small in comparison to the remaining forces, which is important for further statistical mechanics developments later on in this lecture. Section 3 deals with Brownian motion of interacting particles, where a distinction between self and collective diffusion must be made, which are discussed on an intuitive level. Fick's law is derived, and applied to diffusion of a molecular compound towards receptors on a cell membrane. Fick's diffusion coefficient is then expressed in terms of the interaction potential between the macromolecules, on the basis of thermodynamic arguments, neglecting hydrodynamic interactions, which will be considered later on. The thermodynamic approach allows to calculate directed transport of a macromolecule in a inhomogeneous solvent (this describes the directed motion of a cell towards regions of high concentrations of food). A generalized diffusion equation is derived in section 4, where it turns out to be essential that inertial forces are negligible, as is discussed in section 2. This microscopic approach allows to include hydrodynamic interactions, which are shown to have a pronounced influence on the numerical value of Fick's diffusion coefficient. Finally, in section 5, transport under the influence of an external field is discussed, which finds its application in separation of compounds by means of, for example, sedimentation and electrophoresis.
For systems of interacting Brownian particles this is an important fact to device a statistical mechanics approach.

2 Free Brownian Motion

In this section, the fundamental physics of Brownian motion of a single spherical particle in an otherwise unbounded fluid is described. This is the type of Brownian motion of particles in a very dilute dispersion, where the average distance between Brownian particles is so large that they do not influence each other in their motion. Brownian motion is described on the basis of Newton's equation of motion where a random force accounts for collisions of solvent molecules with the macromolecule, which equation of motion is commonly referred to as the Langevin equation. This approach allows to calculate the mean squared displacement, introduce the various time scales of interest and to calculate the probability density function for the position coordinate of a Brownian sphere.

2.1 The Langevin Equation

Structural relaxation times for molecular fluids are known experimentally to be of the order $10^{-14}\ s$. As will be established shortly, relevant time scales for Brownian particles are at least of the order $10^{-9}\ s$. This separation in time scales is the consequence of the very large mass of the Brownian particle relative to that of a solvent molecule, and is essential for the validity of the Langevin description.
The interaction of the spherical Brownian particle with the solvent molecules is separated into two parts. First of all, there is a rapidly varying force $\mathbf{f}(t)$ with time t as the result of *random collisions* of solvent molecules with the surface of the Brownian particle. This force fluctuates on the fore mentioned solvent time scale of $10^{-14}\ s$. Secondly, as the Brownian particle attains

a velocity $\mathbf{v} = \mathbf{p}/M$ ($\mathbf{p}$ is the momentum coordinate of the Brownian particle and M its mass), there is a friction force due to *systematic collisions* with solvent molecules. When the volume of the Brownian particle is much larger than that of the solvent molecules, this systematic force equals the hydrodynamic friction force of a macroscopically large sphere. For not too large velocities, that friction force is directly proportional to the velocity of the Brownian particle, and the proportionality constant γ is the friction constant: *friction force* $= -\gamma\,\mathbf{p}/M$. The friction coefficient of such a macroscopically large sphere is equal to,

$$\gamma = 6\pi\eta_0 a, \tag{1}$$

with η_0 the shear viscosity of the solvent and a the radius of the Brownian spherical particle. The friction coefficient in eq.(1) is commonly referred to as *Stokes's friction coefficient*. Newton's equation of motion for a spherical Brownian particle is thus written as,

$$d\mathbf{p}/dt = -\gamma\,\mathbf{p}/M + \mathbf{f}(t). \tag{2}$$

The position coordinate $\mathbf{r}$ of the Brownian particle is, by definition, related to the momentum coordinate as,

$$d\mathbf{r}/dt = \mathbf{p}/M. \tag{3}$$

Since the systematic interaction with the solvent molecules is made explicit (the first term on the right-hand side of eq.(2), the ensemble average of the fluctuating force $\mathbf{f}$ is equal to zero,

$$< \mathbf{f}(t) > = \mathbf{0}. \tag{4}$$

Due to the fore mentioned large separation in time scales, it is sufficient for the calculation of the thermal movement of the Brownian particle to use a delta correlated random force in time (see Fig.1), that is,

$$< \mathbf{f}(t)\,\mathbf{f}(t') > = \mathbf{G}\,\delta(t - t'), \tag{5}$$

where δ is the delta distribution and $\mathbf{G}$ is a constant, 3×3-dimensional matrix, which may be regarded as a measure for the strength of the fluctuating force, and is referred to as the *fluctuation strength*. Such a delta correlated random force limits the description to a time resolution which is large with respect to the solvent time scale of $10^{-14}\,s$.

It should be mentioned that the delta-correlation in eq.(5) is not entirely correct due to hydrodynamic flow in the solvent around a moving sphere that impinges on the same sphere after a finite time (of the order of a ns). A discussion of the "memory effects" that are due to such a flow are outside the scope of this lecture.

Equation (2) is Newton's equation of motion for a macroscopic particle with a fluctuating random force added in order to account for the thermal collisions of the solvent molecules with the Brownian particle. Such an equation is called a *Langevin equation* [5],[2],[3]. It is a stochastic equation of motion in the sense that the momentum coordinate of the Brownian particle, as well as its position coordinate, are now stochastic variables. It makes no sense to ask for a deterministic solution of eqs.(2,3), since only ensemble averaged properties of the random force $\mathbf{f}$ are specified.

The simplest ensemble averaged quantity that characterizes the dynamics of a spherical Brownian particle is the so-called "mean squared displacement" $W(t)$, which is defined as,

$$W(t) = <|\mathbf{r}(t) - \mathbf{r}(t=0)|^2>\,, \tag{6}$$

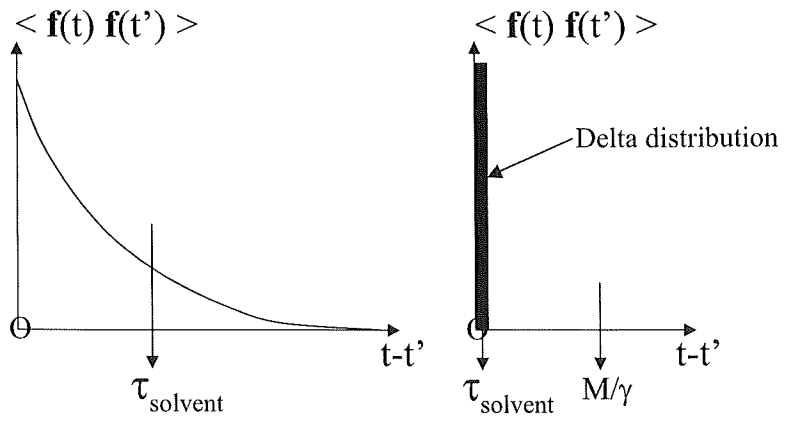


Fig. 1: *The correlation function of the random force in eq.(5) decays on the time scale* $\tau_{solvent} \approx 10^{-14}\ s$ *(left figure). On the time scale* $M/\gamma \approx 1\,ns \gg \tau_{solvent}$, *the correlation function can be regarded as being delta-correlated (right figure). The relevance of the time scale* M/γ *will be discussed in section 2.2.*

where the brackets $< \cdots >$ denote ensemble averaging with respect to the many possible realizations of the fluctuating force $\mathbf{f}$. The probability density function (pdf) for the position coordinate is also a quantity of interest. Once this pdf is know, it is possible to calculate additional ensemble averages besides the mean squared displacement.
Integration of eq.(2) yields,

$$\mathbf{p}(t) \;=\; \mathbf{p}(0)\,\exp\{-\frac{\gamma}{M}t\} \;+\; \int_0^t dt'\,\mathbf{f}(t')\,\exp\{-\frac{\gamma}{M}(t-t')\}. \tag{7}$$

This expressions allows us to calculate the fluctuation strength $\mathbf{G}$ in eq.(5) as follows. Using eqs.(144,145) it id found from eq.(7) that,

$$< \mathbf{p}(t)\,\mathbf{p}(t) > = \mathbf{G}\,\frac{2\gamma}{\beta}\left[1-\exp\{-\frac{2\gamma}{M}t\}\right] \;+\; \mathbf{p}(0)\,\mathbf{p}(0)\,\exp\{-\frac{2\gamma}{M}t\}. \tag{8}$$

On the other hand, from the equi-partition theorem we have (with $\hat{\mathbf{I}}$ the identity matrix),

$$\lim_{t\to\infty} < \mathbf{p}(t)\,\mathbf{p}(t) > = \hat{\mathbf{I}}\,\frac{M}{\beta}\,. \tag{9}$$

This is equivalent with the statement that the average kinetic energy of the Brownian particle in equilibrium with the solvent is equal to $\frac{3}{2}k_BT$. Comparing eqs.(8) and (9) it is immediately found that,

$$\mathbf{G} \;=\; \hat{\mathbf{I}}\,\frac{M}{2\gamma}\,. \tag{10}$$

Hence, eq.(7) can be written more explicitly as,

$$< \mathbf{p}(t)\,\mathbf{p}(t) > = \hat{\mathbf{I}}\frac{M}{\beta}\left[1-\exp\{-\frac{2\gamma}{M}t\}\right] \;+\; \mathbf{p}(0)\mathbf{p}(0)\,\exp\{-\frac{2\gamma}{M}t\}. \tag{11}$$

Integration of eq.(11), using eq.(3), yields, after an interchange of the order of two time-integrations,

$$\begin{aligned} \mathbf{r}(t) \;=\; \mathbf{r}(0) \;+\; & \frac{\mathbf{p}(0)}{\gamma}\left[1-\exp\{-\frac{\gamma}{M}t\}\right] \\ +\; & \frac{1}{\gamma}\int_0^t dt'\,\mathbf{f}(t')\left[1-\exp\{-\frac{\gamma}{M}(t-t')\}\right]. \end{aligned} \tag{12}$$

The mean squared displacement (in dyadic form) is thus found to be given by,

$$\begin{aligned} <(\mathbf{r}(t)-\mathbf{r}(0))\,(\mathbf{r}(t)-\mathbf{r}(0))> = & \frac{\mathbf{p}(0)\mathbf{p}(0)}{\gamma^2}\left[\exp\{-\frac{\gamma}{M}t\}-1\right]^2 \\ +\hat{\mathbf{I}}\,\frac{2M}{\beta\gamma^2}\,(\frac{\gamma}{M}t-\frac{1}{2}\left[\exp\{-\frac{2\gamma}{M}t\}-1\right]-2\left[1-\exp\{-\frac{\gamma}{M}t\}\right]). \end{aligned} \tag{13}$$

Note that for large times $t \gg M/\gamma$, this reduces to,

$$<(\mathbf{r}(t)-\mathbf{r}(0))\,(\mathbf{r}(t)-\mathbf{r}(0))> = \hat{\mathbf{I}}\,2\,\frac{k_BT}{\gamma}\,t\,, \tag{14}$$

and hence, by contraction,

$$W(t) \;=\; 6\,D_0\,t\,, \tag{15}$$

where "the diffusion coefficient" D_0 is defined as,

$$D_0 \;=\; \frac{k_BT}{\gamma}\,. \tag{16}$$

The index 0 on D is used to indicate that this diffusion coefficient relates to non-interacting particles, that is, Brownian particles which do not interact with each other. In an experiment, D_0 can be measured using very dilute suspensions, where such interactions can be neglected. We shall analyze time scales and discuss the result (15) in more detail in the following section.

2.2 Time Scales

In an experiment, the time scale is set by the time interval over which observables are averaged during a measurement. For example, taking photographs of a Brownian particle is an experiment on a time scale which is set by the shutter time of the camera. Subsequent photographs reveal the motion of the Brownian particle averaged over a time interval equal to the shutter time. Any theory considering the motion of the Brownian particle obtained in such a way should of course be aimed at the calculation of observables, averaged over that time interval. A time scale is thus the minimum time resolution of an experiment or theory, and observables are averaged over the time interval that sets the time scale.
Structural relaxation times of the solvent are of the order of $10^{-14}\,s$. As we shall see later, that the dynamics of the Brownian particles is much more slower. Since the solvent dynamics is of no interest here, an experimental time scale larger on which averaged solvent properties are probed suffices. This time scale is larger than $10^{-14}\,s$. The smallest time scale on which the specification (5) makes sense, is larger than the solvent time scale. This time scale is usually

referred to, for historical reasons, as the *Fokker-Planck time scale*, which we shall denote as τ_{FP}.
It is clear from the expressions (11,13) that an important characteristic time is M/γ. It follows from eqs.(11,13) that for times $t \ll M/\gamma$,

$$\begin{aligned} < \mathbf{p}(t)\,\mathbf{p}(t) > &= \mathbf{p}(0)\,\mathbf{p}(0)\ , \\ < (\mathbf{r}(t)-\mathbf{r}(0))\,(\mathbf{r}(t)-\mathbf{r}(0)) > &= \frac{\mathbf{p}(0)\,\mathbf{p}(0)}{M^2}\,t^2 = \mathbf{v}(0)\,\mathbf{v}(0)\,t^2\ . \end{aligned} \tag{17}$$

For such small times, the velocity of the Brownian particle is thus essentially equal to its initial velocity $\mathbf{v}(0)$. That is, friction with the solvent has not been effective in diminishing the velocity for such short times.
For times $t \gg M/\gamma$, on the other hand, the mean squared displacement varies linearly in time (see eqs.(14,15)), where the initial velocity is completely irrelevant. This is due to the many independent collisions of the Brownian particles with solvent molecules that occurred during this large time interval. Such motion is referred to as "*diffusive motion*".
Averaging of eq.(7), using eq.(4) immediately leads to,

$$< \mathbf{p}(t) > = \mathbf{p}(0)\,\exp\left\{-\frac{\gamma}{M}\,t\right\}\ . \tag{18}$$

The velocity thus relaxes for times $t \gg M/\gamma$, due to friction with the solvent. For time scales much larger than M/γ, the momentum coordinate is thus in equilibrium with the solvent. In many cases the velocity is not an interesting quantity. For example, when one is interested in the evolution of microstructural order, the time evolution of position coordinates is of interest, not the way in which velocities relax to equilibrium with the heat bad of solvent molecules. On time scales $\tau_D \gg M/\gamma$ velocities are irrelevant. This time scale is usually referred to as the *Brownian*, *Diffusive*, or *Smoluchowski time scale*. On that time scale a statistical description for the motion of the Brownian particle is feasible, without involving the momentum coordinate. We thus come to the following ordering of time scales,

$$10^{-14}s = \tau_{solvent} \ll \tau_{FP} \ll M/\gamma \ll \tau_D. \tag{19}$$

Using typical values for the mass and friction coefficient of a Brownian particle, one finds that $M/\gamma \approx 10^{-9}s$. A statistical description on the solvent time scale involves the position and momentum coordinates of both the solvent molecules and the Brownian particles. On the Fokker-Planck time scale, the solvent coordinates are long relaxed to thermal equilibrium (within the force field imposed by the Brownian particles), and only the momentum and position coordinate of the Brownian particle need to be considered. Finally, on the Brownian or diffusive time scale, in addition, the momentum coordinate of the Brownian particles relaxed to equilibrium with the heat bath of solvent molecules, and a statistical description involving just the position coordinate of the Brownian particle is feasible.
A coarsening of the time scale implies a coarsening of the length scale. On the diffusive time scale the spatial resolution is not better than the distance over which the Brownian particle moves during a time interval equal to the diffusive time scale. The ensemble average of that distance, the *diffusive length scale* l_D, is easily obtained from eq.(18) as,

$$l_D = \int_0^{\tau_D} dt\,\frac{|< \mathbf{p}(t) >|}{M} \approx |\ \mathbf{p}(0)\ |\ /\gamma\ , \tag{20}$$

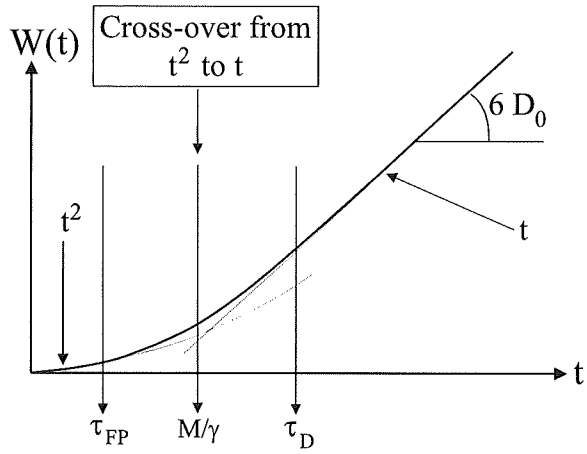


Fig. 2: *The time dependence of the mean squared displacement* $W(t) \equiv <|\mathbf{r}(t) - \mathbf{r}(0)|^2>$.

where it is used that $\tau_D \gg M/\gamma$. A typical value for $|\ \mathbf{p}(0)\ |$ is obtained from the equipartition theorem,

$$|\ \mathbf{p}(0)\ | \approx \sqrt{<|\ \mathbf{p}\ |^2>} = \sqrt{3Mk_BT}\ . \tag{21}$$

The diffusive length scale is thus estimated as,

$$l_D \approx \sqrt{3Mk_BT}/\gamma = \sqrt{3Mk_BT}/6\pi\eta_0 a\ . \tag{22}$$

Substitution of typical values for the various quantities yields,

$$\frac{l_D}{a} \approx 10^{-4} - 10^{-3}\ , \tag{23}$$

where a is the radius of the Brownian particle. The important conclusion is, that on the diffusive time scale the coarsening of the spatial resolution is only a tiny fraction of the size of the Brownian particle. *For the study of processes where a significant displacement of the Brownian particle is essential, a statistical description on the diffusive time scale is therefore sufficient.*
The results of the present section are summarized in Fig.2, where the mean squared displacement is plotted as a function of time. For small times, that is, on the Fokker-Planck time scale, the mean squared displacement is proportional to t^2, whereas for large times, on the diffusive time scale, the mean squared displacement is linear in t. It should be mentioned that the $\sim t^2$-behaviour of the mean squared displacement changes due to the memory effects mentioned in the previous section. A treatment of these memory effects is beyond the scope of this lecture.

2.3 The Probability Density Function (pdf) for the Position Coordinate

Consider an assembly of many non-interacting Brownian particles, for which a concentration profile is prescribed at time $t = 0$, say. How the concentration profile evolves can be predicted once an equation of motion for the density is known. Since the probability to find a particle at a certain position at some time is proportional to the density, the equation of motion for the

density is the same as for the probability density function $P(\mathbf{r},t)$ for the position coordinate of a single Brownian particle. The initial condition for the equation of motion for this pdf, where the position of the Brownian particle at $t = 0$ is prescribed, corresponds to a concentration profile where all particles are concentrated at that point. Hence, the pdf $P(\mathbf{r},t)$ can be regarded as a concentration profile, where at $t = 0$ all particles are concentrated within a small region. Furthermore, once this pdf is known, ensemble averages of functions f of $\mathbf{r}$ can be calculated as,

$$< f > (t) \;=\; \int d\mathbf{r}\, f(\mathbf{r})\, P(\mathbf{r},t)\;. \tag{24}$$

In case $f = r^2$, and $\mathbf{r}(t=0)$ is prescribed, the mean squared displacement can be calculated in this way, which should reproduce our earlier result 15 on the Brownian time scale.
The Langevin approach as described in the previous subsections allows to calculate the probability density function (pdf) of the position coordinate of a spherical Brownian particle, without having to solve an equation of motion. Consider the expression (12) on the Brownian time scale, where the diffusion length $|\,\mathbf{p}(0)\,|\,/\gamma$ can be neglected,

$$\mathbf{r}(t) \;=\; \mathbf{r}(0) \;+\; \frac{1}{\gamma}\int_0^t dt'\,\mathbf{f}(t')\left[1-\exp\{-\frac{\gamma}{M}(t-t')\}\right].$$

Now let τ be a time interval much larger than the solvent time scale of $10^{-14}s$. The random force evolves through many independent realizations during that time interval. On the other hand, let τ be so small that $\exp\{-\gamma t/M\}$ is almost constant over times of the order of τ, that is, we take $\tau \ll M/\gamma$. With this choice of τ, the above equation can be rewritten as,

$$\mathbf{r}(t) \;=\; \mathbf{r}(0) \;+\; \frac{1}{\gamma}\sum_{j=0}^{N-1}\left[1-\exp\{-\frac{\gamma}{M}(t-j\tau)\}\right]\int_{j\tau}^{(j+1)\tau} dt'\,\mathbf{f}(t')\;, \tag{25}$$

where $N = t/\tau$. Since the random force evolves through many independent realizations during the time interval τ, each integral in eq.(25) is a Gaussian variable with a mean equal to zero. This is a consequence of the central limit theorem, as each integral may be regarded as a sum of many statistically equivalent and independent terms (the central limit theorem states that a sum of many independent variables with mean zero is a Gaussian variable). Furthermore, since a sum of independent Gaussian variables is also a Gaussian variable, it follows from eq.(25) that $\mathbf{r}(t) - \mathbf{r}(0)$ is a Gaussian variable. The pdf is thus given by,

$$P(\mathbf{r},t) \;=\; \frac{1}{(2\pi)^{3/2}\,\sqrt{det\mathbf{D}(t)}}\;\exp\left\{-\tfrac{1}{2}\,(\mathbf{r}-\mathbf{r}(0))\cdot\mathbf{D}^{-1}(t)\cdot(\mathbf{r}-\mathbf{r}(0))\right\}\;, \tag{26}$$

where the so-called covariance matrix $\mathbf{D}(t)$ is equal to,

$$\mathbf{D}(t) \;=< (\mathbf{r}(t)-\mathbf{r}(0))\,(\mathbf{r}(t)-\mathbf{r}(0)) >\;, \tag{27}$$

and $det\mathbf{D}(t)$ is the determinant of $\mathbf{D}(t)$. We already calculated the dyadic mean squared displacement in eq.(14) on the Brownian time scale. It immediately follows that eq.(26) thus reads,

$$P(\mathbf{r},t) \;=\; \frac{1}{(4\pi\, D_0 t)^{3/2}}\;\exp\left\{-\frac{|\,\mathbf{r}-\mathbf{r}(0)\,|^2}{4D_0 t}\right\}\;. \tag{28}$$

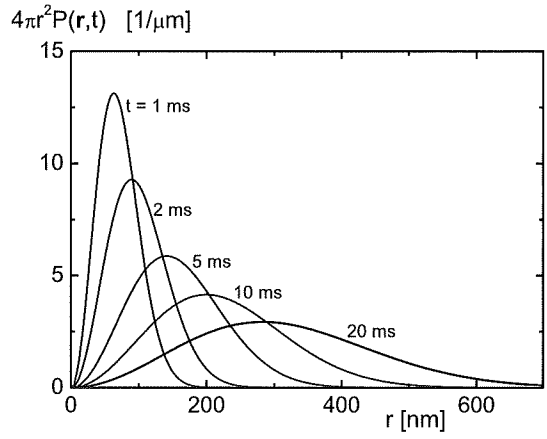


Fig. 3: *The function* $4\pi\, r^2\, P(\mathbf{r},t)$ *according to eq.(28) in units* $1/\mu m$ *for various times, as indicated in the figure. At time* $t = 0$ *all Brownian particles are concentrated at the origin, that is,* $\mathbf{r}(0) = \mathbf{0}$ *in eq.(28). Furthermore,* $D_0 = 10^{-12}\, m^2/s$*, which is a typical value for colloidal particles of about* $100\, nm$ *in diameter,* r *is given in units of* nm *and time in* ms*.*

It is easily verified that at $t = 0$ this pdf is equal to $\delta(\mathbf{r} - \mathbf{r}(0))$, where δ is the 3-dimensional delta distribution. Hence, at time $t = 0$, the Brownian particle is prescribed to have position $\mathbf{r}(0)$.

Note that the probability to find a Brownian particle at a certain position is proportional to the number concentration at that point. Equation (28) thus predicts that a region containing particles broadens due to diffusion with a rate $\sim D_0$. The probability that the particles are located within a radial shell of thickness dr and radius r, given that the particles were at the origin at time $t = 0$, is equal to $4\pi\, r^2\, P(\mathbf{r},t)\, dr$. The function $4\pi\, r^2\, P(\mathbf{r},t)$ is plotted in Fig.3, showing the broadening of an initially confined concentration profile. This diffusion process limits the separation efficiency in sedimentation and electrophoresis (of which more will be said in section 5).

3 Brownian Motion of Interacting Particles

There are two types of interactions between Brownian particles suspended in a fluid : direct interactions and hydrodynamic interactions. Direct interactions are the result of exclusion of volume by the hard core of each particle, and, for example, charges or polymers attached to the surface of the particles. Hydrodynamic interactions are interactions between the Brownian particles mediated through the fluid : when a particle moves it will induce a fluid flow which affects other particles in their motion. Both these types of interactions must be considered when analyzing Brownian motion and diffusion in concentrated systems. In the present section, we shall consider diffusion within a thermodynamic approach, which allows the inclusion of direct interactions. Hydrodynamic interactions can only be included within a microscopic approach,

where equations of motion for appropriate probability density functions are derived (the solution of which should comply with eq.(28) at low concentrations).
For concentrated systems, a distinction should be made between so-called self diffusion and collective diffusion. In addition short- and long-time regimes must be distinguished. These features will be discussed in the following subsection on an intuitive level, before embarking into more quantitative considerations.

3.1 Self Diffusion

Self diffusion is related to the dynamics of a single Brownian particle in a homogeneous system. The single particle under consideration is commonly referred to as *the tracer particle* or *the tagged particle*, while the remaining Brownian particles, with which the tracer particle can interact, are referred to as *host particles*.
As discussed before, the simplest quantity that characterizes the motion of a single Brownian particle is its mean squared displacement $W(t)$, defined in eq.(6). In section 2 on the diffusion of non-interacting Brownian particles, we have seen that on the diffusive time scale $\tau_D \gg M/\gamma$,

$$W(t) \;=\; 6\,D_0\,t\;, \tag{29}$$

with $D_0 = k_BT/\gamma$ the Einstein diffusion coefficient. Interaction of the tracer particle with surrounding Brownian particles affects the time dependence of the mean squared displacement. The most obvious way to introduce the self diffusion coefficient for interacting systems, is to replace D_0 in eq.(29) formally by the *self diffusion coefficient* D_s. This diffusion coefficient may be time dependent as a result of interactions with other Brownian particles. However, to make the connection with light scattering experiments, we shall defined the self diffusion coefficient as follows. In a light scattering experiment one measures the so-called *self dynamic structure factor* $S_s(k,t)$, which is defined as,

$$S_s(k,t) \;\equiv <\exp\{i\mathbf{k}\cdot(\mathbf{r}(t)-\mathbf{r}(t=0))\}>\;, \tag{30}$$

where *the wavevector* k is set by the scattering angle. The *generalized self-diffusion coefficient* D_s is now defined as,

$$S_s(k,t) \;\equiv\; \exp\{-D_s(k,t)k^2t\}\;. \tag{31}$$

For very dilute suspensions, where interactions between the colloidal particles are unimportant, the self dynamic structure factor is given by this expression with D_s equal to D_0. This follows from eq.(28) with $\mathbf{r}(t=0)=\mathbf{0}$ and eq.(24) with $f(\mathbf{r})=\exp\{i\mathbf{k}\cdot\mathbf{r}\}$. Hence,

$$D_s(k,t) \;=\; D_0 \;,\;\text{at infinite dilution}\;. \tag{32}$$

To make the connection with the mean squared displacement, the self dynamic structure factor is expanded in a Taylor series for small wavevectors,

$$S_s(k,t) \;=\; 1-\frac{1}{6}k^2\;<|\,\mathbf{r}(t=0)-\mathbf{r}(t)\,|^2>+\cdots\;. \tag{33}$$

On the other hand, the defining relation of D_s in eq.(31) may be expanded for small wavevectors as,

$$S_s(k,t) \;=\; 1-D_s(k=0,t)k^2t+\cdots\;. \tag{34}$$

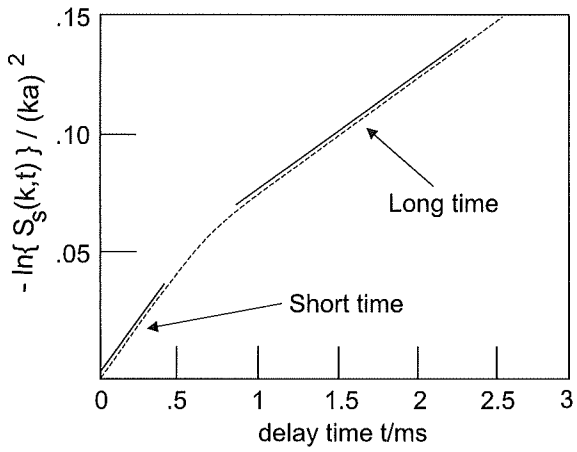


Fig. 4: *The measured mean squared displacement $W(t)$ as a function of time for a dispersion of sterically PMMA silica particles, with a volume fraction of* 0.37. *The radius of the silica particles is $a = 80\,nm$. The scattering wavevector here is equal to $0.012\,nm^{-1}$. Data (dotted line) are taken from ref.[6]. The solid lines are straight lines to indicate the linear time dependence in both the short- and long-time limit.*

Comparing the two Taylor expansions gives,

$$W(t) \;=\; 6\,D_s(k=0,t)\,t\;, \tag{35}$$

The wavevector dependent self diffusion coefficient as defined in eq.(31) at k = 0 thus generalizes the Einstein diffusion coefficient to systems of high concentrations where inter-Brownian particle interactions are important.

The above equations suggest the following experimental route for obtaining the time dependent mean squared displacement. According to eq.(34), a plot of $\ln\{S_s(k,t)\}/k^2$ as a function of the wavevector for a given time may be extrapolated to $k = 0$ to obtain $D_s(k=0,t)$. Since $S_s(k,t)$ is an even function in k, this can best be done by plotting versus k^2, which should yield a straight line for small enough wavevectors. The mean squared displacement then follows immediately from eq.(35). Such an experimental result for a concentrated dispersion of PMMA colloids is given in Fig.4 (here, PMMA stands for poly(methylmethacrylate). A Brownian particle consists of a compact, spherically symmetric polymer network). As can be seen, there are two diffusive regimes where the mean squared displacement varies linear with time.

For such highly concentrated suspensions one must distinguish between *short-time diffusion* and *long-time diffusion*. At small times, the tracer particle, on average, resides at the minima in the energy landscape created through interactions with the host particles. Short time self-diffusion thus relates to Brownian motion of the tracer sphere around the minima in the energy landscape, as sketched in Fig.5. Note that this landscape itself is fluctuating, since it is set up by interactions with the host Brownian particles which themselves exhibit Brownian motion. For later times, the tracer particle "climbes" free energy barriers, which changes the time dependence of the mean squared displacement. The self diffusion coefficient may then become time dependent. The mean squared displacement is then no longer a linear function of time. Long-

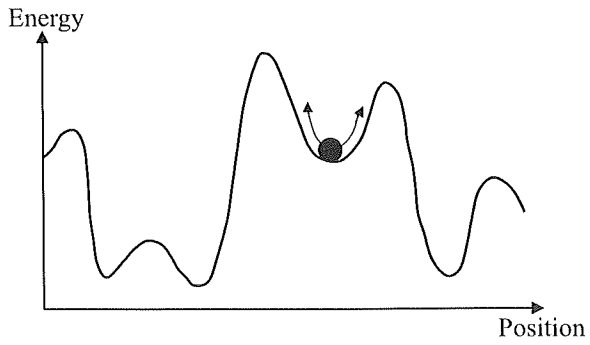


Fig. 5: *Self diffusion of a tracer particle through the energy landscape set up by the host particles.*

time diffusion is related to the escape from energy minima, and the fluctuations of the landscape itself. Since for long times the tracer particle experienced many independent displacements, the mean squared displacement is again a linear function of time, the slope of which defines the long-time self diffusion coefficient. The time at which short time regime crosses over to the long time regime is commonly referred to as the *interaction time scale*, and is of the order of magnitude of the structural relaxation times of the assembly of Brownian particles.

3.2 Collective Diffusion

Imagine a suspension where the density of Brownian particles, at some instant in time, varies sinusoidally with position (such a sinusoidal density profile is referred to as a *density wave*). That is, at time $t = 0$ say, the macroscopic density $\rho(\mathbf{r}, t = 0)$ at position $\mathbf{r}$ is equal to,

$$\rho(\mathbf{r}, t = 0) \; = \; \bar{\rho} + \rho(\mathbf{k}, t = 0) \; \sin\{\mathbf{k} \cdot \mathbf{r}\} \, , \tag{36}$$

with $\bar{\rho} = N/V$ the average number density of Brownian particles, and $\rho(\mathbf{k}, t = 0)$ the amplitude of the density wave. This density profile is sketched in Fig.6. The *wavevector* $\mathbf{k}$ determines both the direction and the wavelength of the sinusoidal density variation. For changes of the position $\mathbf{r}$ perpendicular to $\mathbf{k}$, the phase of the sine function does not change, so that the direction of $\mathbf{k}$ is in the "propagation direction" of the sinusoidal variation. A change $\Delta\mathbf{r}$ of the position $\mathbf{r}$ parallel to $\mathbf{k}$ leaves the sine function unchanged when for this displacement $\mid \Delta\mathbf{r} \mid = n \times 2\pi/k$, with n an arbitrary integer. Hence, the wavelength of the density variation is,

$$\Lambda \; = \; 2\pi/k \, . \tag{37}$$

The sinusoidal density variation may be thought of as being the result of an external field. Now suppose that this field is turned off at time $t = 0$. In a thermodynamically stable system, the amplitude of the density wave decreases with time due to thermal motion of the Brownian particles. In the initial stage of the decay, the sinusoidal shape of the density wave will be retained. At a later stage, different wavevectors, or equivalently, different wavelengths come into play as a result of interactions between the Brownian particles (see Fig.7). The strength of these interactions varies with the distance between the Brownian particles, leading to a distribution of

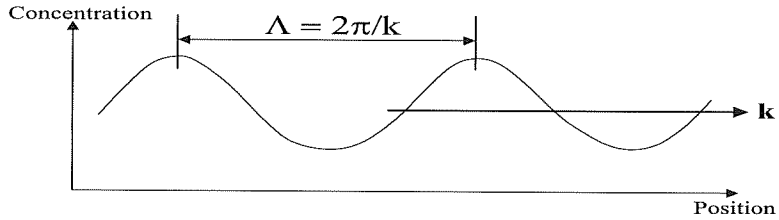


Fig. 6: *A density wave. The "propagation direction" is along the wavevector* k *and the wavelength is* $\Lambda = 2\pi/k$.

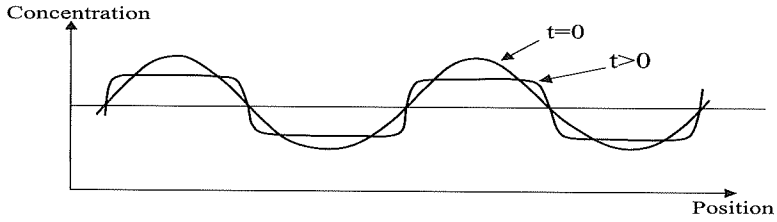


Fig. 7: *A sketch of the decay of a sinusoidal density profile. Initially the density profile will be more or less sinusoidal. At later times, however, the density profile is generally no longer purely sinusoidal, but involves many Fourier components.*

relaxation times. Spatial inhomogeneities extending over various distances relax to equilibrium with different relaxation times. The shape of the density variation is then no longer sinusoidal, but involves other "Fourier components" (other wavevectors) in addition.
The decay of such a sinusoidal density variation is a collective phenomenon, since a relaxation of inhomogeneities involves the cooperative motion of many Brownian particles simultaneously. Notice that $\rho(\mathbf{k}, t = 0)$ is the amplitude of the sinusoidal variation with wavevector k, which is just one of the many sinusoidal density variations that constitute an arbitrary spatially varying density. In addition to this particular wavevector, there are generally many more wavevectors contributing to the actual spatial variation of the density.
Let $\rho(\mathbf{r}, t)$ denote the space and time dependent macroscopic density, and $\mathbf{J}(\mathbf{r}, t)$ the flux (or current density) of Brownian particles, which is the number of Brownian particles which move across a surface perpendicular to J per unit area and unit time. The continuity equation, which expresses conservation of the number of Brownian particles, reads,

$$\frac{\partial}{\partial t}\rho(\mathbf{r}, t) \;=\; -\nabla \cdot \mathbf{J}(\mathbf{r}, t)\;. \tag{38}$$

Translational motion is driven by gradients in the density of Brownian particles. For small gradients, the flux is a linear function of these gradients. The flux at a certain position r may depend, through interactions with surrounding particles, on gradients at neighbouring positions. Furthermore, the flux at a certain time t may depend on states of the system at preceding times. The flux can thus formally be written as,

$$\mathbf{J}(\mathbf{r}, t) \;=\; -\int d\mathbf{r}' \int dt'\, D(\mathbf{r} - \mathbf{r}', t - t')\nabla' \rho(\mathbf{r}', t')\;. \tag{39}$$

The integral kernel $D(\mathbf{r},t)$ will be referred to simply as "the diffusion coefficient", which is 0 for $t < 0$, since the temporal evolution of the density cannot depend on future events. To leading order in gradients in the density, and for otherwise translationally invariant systems, the diffusion coefficient is a function of the difference vector $\mathbf{r} - \mathbf{r}'$ only. When the current density at a point $\mathbf{r}$ is fully determined by the instantaneous density gradient in that same point, so that there is no coupling with gradients in neighbouring points nor with preceding states of the system, the diffusion coefficient is proportional to a delta distribution in both position and time, that is, $D(\mathbf{r}-\mathbf{r}',t-t') \equiv D(\mathbf{r},t)\delta(\mathbf{r}-\mathbf{r}')\delta(t-t')$. We shall restrict ourselves here to cases where this is a good approximation. Moreover, to leading order in gradients, the position and time dependence of $D(\mathbf{r},t)$ can be neglected, so that it follows that,

$$\mathbf{J}(\mathbf{r},t) \;=\; -D\nabla\rho(\mathbf{r},t)\;, \tag{40}$$

where D is referred to as *Fick's diffusion coefficient* or the *gradient diffusion coefficient.* Substitution into eq.(38) leads to the following equation of motion,

$$\frac{\partial}{\partial t}\rho(\mathbf{r},t) \;=\; D\,\nabla^2\,\rho(\mathbf{r},t)\;. \tag{41}$$

This is *Fick's diffusion equation.* It is valid under the assumption that gradients in the density are small (on the basis of a microscopic approach, we shall specify later how small gradients should be).
When at time $t = 0$ all Brownian particles are concentrated in a point $\mathbf{r}(0)$, say, the solution to this equation is given by the right hand side of eq.(28), with $D_0 = D$, in accordance with the interpretation of the pdf as a concentration as discussed in the beginning of subsection 2.3.

3.3 Diffusion Towards Receptors

Consider a cell, on the surface of which N receptors are located. We shall ask for the diffusion flux of the compound that is bound by the receptors (note that the reasoning in the previous subsection is equally valid for Brownian particles as for simple molecules). This flux determines the rate at which the compound is converted after being bound to the receptor. This problem is formally the same as calculating the diffusion flux of molecules towards channels in the membrane of the cell.
The cell is assumed spherical while the receptors are modelled as disks (see Fig.8a). The concentration of the component that is bound by the receptors is assumed zero, that is, after having reached the receptor, the compound is instantaneously adsorbed, converted to a different compound and released. In addition it will be assumed that the diffusion process is stationary, that is, concentrations are independent of time.
Let us first consider the two extreme cases where either (i) the cell is fully covered with receptors or (ii) just a single receptor is present on the entire cell membrane, and then proceed to (iii) the general case.
(i) When the cell is fully covered by receptors, the entire surface of the cell can be considered to adsorb the relevant compound. In that case, Fick's law (41) can be written in spherical coordinates, where the angular dependence can be ignored (see, for example, refs.[7],[8] how differential operators are expressed in new coordinates),

$$\frac{\partial}{\partial t}\rho(r,t) \;=\; D\,\frac{d}{dr}\left(r^2\,\frac{d\rho(r,t)}{dr}\right)\;, \tag{42}$$

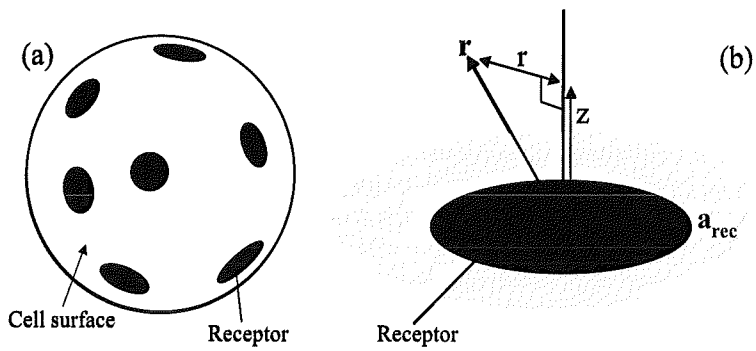


Fig. 8: *(a) A cell with receptors on its surface. The cell is assumed to be spherical. (b) The receptors are modelled as disks. The coordinates r and z are the cylindrical coordinates used in the main text.*

where r is the radial distance to the center of the cell. The general stationary solution of this equation is,

$$\rho(r) \;=\; A + B/r \;, \tag{43}$$

where A and B are constants. These constants are determined from the boundary conditions,

$$\rho(r) \;=\; \rho_0 \;\;, \;\; \text{for } r \rightarrow \infty \;, \tag{44}$$

$$\rho(r) \;=\; 0 \;\;\;, \;\; \text{for } r = a_{cell} \;, \tag{45}$$

where ρ_0 is the bulk concentration of the compounds complying with the receptor under consideration, and a_{cell} is the radius of the cell. It follows that,

$$\rho(r) \;=\; \rho_0 \left(1 - \frac{a_{cell}}{r}\right) \;. \tag{46}$$

The flux is then equal to,

$$J(r) \;=\; -D\,\frac{d\rho(r)}{dr} \;=\; D\,\rho_0\,\frac{a_{cell}}{r^2} \;. \tag{47}$$

The amount of compound that is bound in total per unit time is a measure for the *adsorption efficiency of the cell*, which will be denoted as I. According to eq.(47) we have,

$$I_{full} \;\equiv\; 4\pi\, a_{cell}^2\, J(r = a_{cell}) \;=\; 4\pi\, D\,\rho_0\, a_{cell} \;. \tag{48}$$

The subscript "full" on I indicates that this result holds for a cell that is fully covered with receptors. Note that the efficiency varies like $\sim a_{cell}$, instead of being proportional to the surface area of a receptor $\sim a_{cell}^2$, as one might except on first thought. The reason for this is that the local gradients vary on average like $\sim 1/a_{cell}$.
(ii) Let us now consider the other extreme case where only a single receptor is present on the entire cell surface. The problem is now to calculate the flux to an adsorbing disk (see Fig. 8b). When the size of a receptor is very much smaller than that of the cell, the problem

is approximately cylindrical symmetric, rather than spherical symmetric, where the angular variable is irrelevant. Fick's law (41) thus reads [7],[8],

$$\frac{\partial}{\partial t}\rho(r,z,t) \;=\; D\left(\frac{1}{r}\frac{\partial}{\partial r}\left[r\,\frac{\partial\rho(r,z,t)}{\partial r}\right]+\frac{\partial^2\rho(r,z,t)}{\partial z^2}\right)\,, \tag{49}$$

where r is now the distance to the line of cylindrical symmetry and z is the distance to the surface (see Fig.8b). The stationary flux $J_\perp$ in the z-direction is simply equal to,

$$J_\perp(r,z) \;=\; -D\,\frac{\partial\rho(r,z)}{\partial z}\,. \tag{50}$$

The stationary differential equation (49), where $\partial\rho(r,z,t)/\partial t \;=\; 0$, can be solved using the Ansatz $\rho(r,z) \equiv f(r)\,g(z)$. Substitution into eq.(49) leads to,

$$0 \;=\; \frac{1}{f(r)}\,\frac{1}{r}\,\frac{d}{dr}\left(r\,\frac{df(r)}{dr}\right)+\frac{1}{g(z)}\,\frac{d^2g(z)}{dz^2}\,. \tag{51}$$

Since this is the sum of a z-independent and r-independent part, and r and z are independent variables, we must have that,

$$\begin{aligned}\frac{1}{f(r)}\,\frac{1}{r}\,\frac{d}{dr}\left[r\,\frac{df(r)}{dr}\right] \;&=\; -\alpha^2\,,\\ \frac{1}{g(z)}\,\frac{d^2g(z)}{dz^2} \;&=\; \alpha^2\,,\end{aligned} \tag{52}$$

where α is a constant (which, for reasons to become clear in a moment, appears here squared). The boundary conditions are,

$$\begin{aligned}\rho(r\geq 0, z\to\infty) \;&=\; \rho_0\,,\\ \rho(r\to\infty, z\geq 0) \;&=\; \rho_0\\ \rho(r\leq a_{rec}, z=0) \;&=\; 0\,,\\ \frac{\partial}{\partial z}\,\rho(r>a_{rec}, z=0) \;&=\; 0\,,\end{aligned} \tag{53}$$

where a_{rec} is the radius of the receptor and, as before, ρ_0 is the bulk concentration of the compound complying with the receptor. The last boundary condition states that the flux (50) of the compound through the membrane of the cell is zero. The solution of eq.(52) for which $g(z\to\infty)=0$ is,

$$g_\alpha(z) \;\sim\; \exp\{-\alpha\,z\}\,. \tag{54}$$

where $\alpha > 0$. The parametric dependence of g on α is denoted here explicitly. The solution for f with $f(r\to\infty)=0$ can be found by substitution of a power series expansion in r. The coefficients in this expansion are then determined by equating equal powers of r. It is thus found that,

$$\begin{aligned}f_\alpha(z) \;&=\; J_0(\alpha\,r)\,,\\ \text{where}\;,\; J_0(x) \;&=\; \sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\left(\frac{x}{2}\right)^{2n}\,.\end{aligned} \tag{55}$$

The function $J_0(x)$ is referred to as a Bessel function (of order 0). It can be shown that this function tends to 0 as $x \rightarrow \infty$. Since Fick's law is a linear equation, any sum (read : integral) over α of the solution $f_\alpha(r)\, g_\alpha(z)$ is also a solution. Hence,

$$\rho(r,z) \;=\; \rho_0 + \int_0^\infty d\alpha\, C(\alpha)\, f_\alpha(r)\, g_\alpha(z) \;=\; \rho_0 + \int_0^\infty d\alpha\, C(\alpha)\, J_0(\alpha\, r)\, \exp\{-\alpha\, z\}\;, \qquad (56)$$

is a solution of Fick's stationary equation (49) that satisfies the first two boundary conditions in eq.(53). The as yet unknown function C of α must be determined such that the two last boundary conditions in (53) are fulfilled. This is only possible from detailed knowledge of properties of the Bessel function J_0, a treatment of which is beyond the scope of this lecture (see, for example, ref.[8]). It can be shown that the last two boundary conditions in eq.(53) are satisfied for,

$$C(\alpha) \;=\; -\frac{2}{\pi}\,\rho_0\,\frac{\sin\{\alpha\, a_{rec}\}}{\alpha}\;. \qquad (57)$$

Hence, the number density of the compound adsorbed to the receptor is equal to,

$$\rho(r,z) \;=\; \rho_0 - \frac{2}{\pi}\,\rho_0 \int_0^\infty d\alpha\, \frac{\sin\{\alpha\, a_{rec}\}}{\alpha}\, J_0(\alpha\, r)\, \exp\{-\alpha\, z\}\;, \qquad (58)$$

The flux at the receptor surface, according to eq.(50), is therefore equal to,

$$J(r, z=0) = -D\frac{2}{\pi}\rho_0 \int_0^\infty d\alpha\, \sin\{\alpha\, a_{rec}\} J_0(\alpha\, r) = \frac{2}{\pi}\,\frac{D\,\rho_0}{\sqrt{a_{rec}^2 - r^2}}\;,\; r \leq a_{rec}\;. \qquad (59)$$

The proof of the last equality again relies on detailed knowledge of Bessel functions and the residue theorem, a treatment of which is beyond the scope of this lecture [8]. Integration of the flux in eq.(59) over the receptor surface gives the adsorption efficiency,

$$I_{single} \;=\; 4\, D\, a_{rec}\, \rho_0\;, \qquad (60)$$

where the index refers to the single receptor being considered here. The reason that $I \sim a_{cell}$ and not $\sim a_{cell}^2$ is the same as for the fully covered cell, namely that local gradients vary like $\sim 1/a_{rec}$.

(iii) Consider now the more general case that N receptors are embedded in the cell membrane [9]. This problem can not be solved exactly. Instead, we shall resort to an analogue from electrodynamics. The stationary Fick's equation $\nabla^2\rho = 0$ is equivalent to Poisson's equation $\nabla^2 V = 0$ for the electrical potential V, which is valid in the absence of free charges and for stationary states. The potential is thus the electrical analogue of the concentration. The electric field strength $\mathbf{E}$ is equal to $-\nabla V$, and is therefore the electrical analogue of the flux $\mathbf{J} = -D\nabla\rho$. We can now define the analogue of the electric resistance as $R \equiv \rho_0/I$. For the two extreme cases considered above, these resistances are (see eqs.(48,60)),

$$\begin{aligned} R_{full} \;&=\; 1/4\pi\, D\, a_{cell}\;, \\ R_{single} \;&=\; 1/4\, D\, a_{rec}\;. \end{aligned} \qquad (61)$$

The general case considered here can approximately be mapped onto the problem of calculating the electric resistance, where current flows through a medium of finite resistance (the fluid

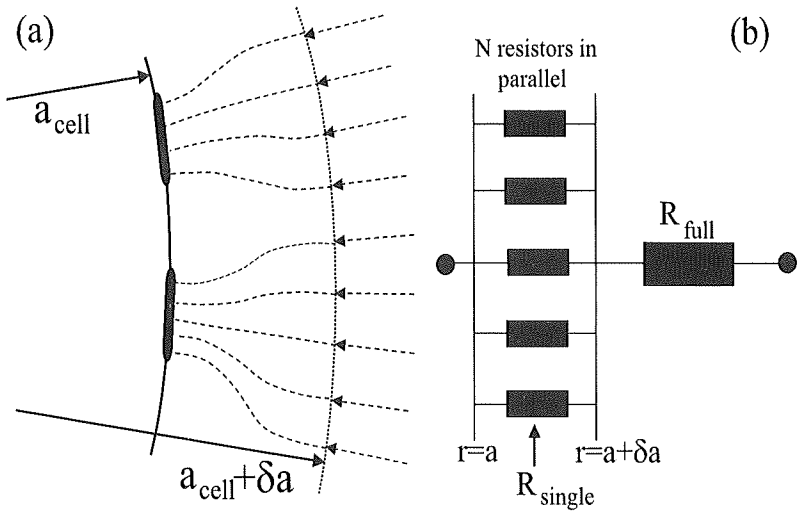


Fig. 9: *(a) Stream lines for a cell partly covered with receptors. (b) The electrical resistance circuit complying with figure (a).*

around the cell) to N conducting patches (the receptors) on a sphere (the cell), as depicted in Fig.9a. The field lines are approximately radial beyond a distance δa form the sphere's surface, which is of the order of the distance between the receptors. The electrical model is given in Fig.9b. There are N resistances in parallel, the total resistance R_{par} of which is equal to $1/R_{par} = N/R_{single} = N\,4Da_{rec}$, according to the second result in eq.(61). If the concentration at $r = a_{cell} + \delta a$ is small in comparison to ρ_0, the resistance R_+ of the material in between $a_{cell} + \delta a$ and $r \to \infty$ is equal to $R_+ = 1/(4\pi D[a_{cell} + \delta a])$, according to the first result in eq.(61). The total resistance of the circuit is thus equal to,

$$R = R_{par} + R_+ = \frac{1}{N\,4\,D\,a_{rec}} + \frac{1}{4\,\pi\,D\,[a_{cell} + \delta a]} \approx \frac{1}{4\,\pi\,D\,a_{cell}} \left(1 + \frac{\pi\,a_{cell}}{N\,a_{rec}}\right) , \quad (62)$$

where in the last equation it is used that $\delta a \ll a_{cell}$. Since δa scales approximately with the distance between the receptors, this is a good approximation when the distance between receptors is small as compared to the cell size. As compared to the adsorption efficiency I_{full} of a fully covered cell surface, see eq.(61), the adsorption efficiency I of a partly covered cell is thus smaller by a factor,

$$\frac{I}{I_{full}} = \frac{1}{1 + (\pi\,a_{cell}/N\,a_{rec})} . \quad (63)$$

The adsorption efficiency is half that of a fully covered cell when $N = \pi\,a_{cell}/a_{rec}$. This is a surprisingly small number. A typical radius of a cell is, say, $5\,\mu m$, while the size of a receptor is $1\,nm$. The number of receptors required to achieve half the adsorption efficiency as compared to a fully covered cell is thus $N = 16000$. The fraction of the cell's surface that is covered is easily calculated to be about 10^{-4}, corresponding to a mean distance between the receptors of about $140\,nm$, that is, about 140 times the receptor radius. The cell can thus accommodate hundreds

of different types of receptors, each of them working at half the efficiency as compared to a cell that is fully covered with one species of receptor.

3.4 Fick's Diffusion Coefficient (without Hydrodynamic Interactions)

In this subsection we shall derive an expression for Fick's diffusion coefficient on the basis of thermodynamic and statistical mechanics arguments, together with force balance on the Brownian time scale. This approach does not allow for the inclusion of hydrodynamic interactions, which will be discussed later.
Noting that the particle number current density is equal to $\mathbf{J} = \rho\,\mathbf{v}$, with $\mathbf{v}$ the local average velocity of Brownian particles, the equation of motion (38) can be written as,

$$\frac{\partial}{\partial t}\,\rho(\mathbf{r},t) \;=\; -\nabla\cdot[\mathbf{v}(\mathbf{r},t)\,\rho(\mathbf{r},t)]\;, \tag{64}$$

As discussed in section 2, momentum coordinates are relaxed to equilibrium with the heat bath of solvent molecules on the Brownian time scale. This implies that inertial forces are very small on the Brownian time scale as compared to other, non-inertial forces. There is a thus balance of all non-inertial forces acting on a spherical Brownian particle. There are two essentially different kinds of such forces : the hydrodynamic force $\mathbf{F}^h$ that a Brownian sphere experiences due to friction with the solvent once it attained a non-zero translational velocity, and the remaining force $\mathbf{F}$ due to direct interactions with other Brownian particles. Force balance implies that,

$$\mathbf{0} \;=\; \mathbf{F}(\mathbf{r},t) + \mathbf{F}^h(\mathbf{r},t)\;. \tag{65}$$

With the neglect of hydrodynamic interactions between the Brownian particles, the hydrodynamic force is equal to (see also eq.(1)),

$$\mathbf{F}^h(\mathbf{r},t) \;=\; -\gamma\,\mathbf{v}(\mathbf{r},t)\;, \tag{66}$$

where $\gamma = 6\pi\eta_0 a$ is the friction coefficient of the sphere with radius a in a solvent with shear viscosity η_0. Hence, once an expression for $\mathbf{F}$ is found in terms of gradients in number density of Brownian spheres, the equation of motion (64) is formulated entirely in terms of the density.
In order to obtain an expression for the interaction force $\mathbf{F}$, it is essential to be able to assume local equilibrium. The dispersion is divided into small volume elements, here after referred to as "boxes". Each box can be treated as a homogeneous, thermodynamic system in equilibrium if,
(i) intensive variables vary on a length scale that is very much larger than the linear dimension of each box, and,
(ii) the linear dimension of the boxes is much larger than the range R_D of direct interactions between the Brownian particles.
The first condition implies that mass transport is very slow in comparison to relaxation times of fluctuations within a box and that within each box intensive variables are constant, independent of position. Hence the spatial variations of the density should be very small on the length scale set by R_D. The second condition then implies that each box may be regarded as a thermodynamic system in internal equilibrium.
Consider two neighbouring boxes '1' and '2', with prescribed equal volumes V and temperature T. The number of Brownian spheres in box 1 and 2 are $N_{c,1}$ and $N_{c,2}$ (where the index "c" stands

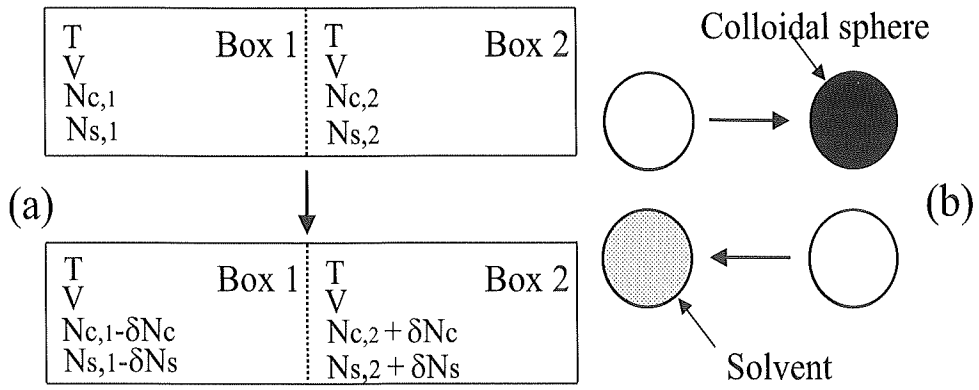


Fig. 10: *(a) Two neighbouring boxes, between which δN_c Brownian particles (the index "c" stands for "colloidal") and δN_s solvent molecules are exchanged under the influence of reversible work that acts on the Brownian particles. (b) Transportation of a Brownian sphere is accompanied by transportation of an equal volume of solvent in the opposite direction.*

for "colloidal"), and the number of solvent molecules $N_{s,1}$ and $N_{s,2}$, respectively. We shall ask for the reversible work δw^{rev} necessary to move δN_c Brownian spheres and δN_s solvent molecules from box 1 to box 2, at prescribed temperatures of the two boxes (see Fig.10a). This process is assumed reversible, complying with the local equilibrium assumption during mass transport as discussed above. The reversible work δw^{rev} necessary to achieve the above described mass transport from box 1 to box 2 is equal to $\delta F + S_1\,\delta T_1 + S_2\,\delta T_2$, where δF is the accompanied change in Helmholtz free energy, S_j is the entropy and δT_j the change of temperature of box j. Since the temperature of both boxes remains constant during mass transport, the latter, entropic contributions are absent. The work is thus equal to the change δF of the free energy. The Helmholtz free energy of each of the two boxes is a function of the number of solvent molecules and Brownian particles in the box, its volume and temperature. The reversible work necessary for the above described process is equal to (see Fig.10a),

$$\begin{aligned}\delta w^{rev} \;=\; \delta F \;=\; & F(N_{c,1}-\delta N_c, N_{s,1}-\delta N_s, V, T) + F(N_{c,2}+\delta N_c, N_{s,2}+\delta N_s, V, T) \\ & -F(N_{c,1}, N_{s,1}, V, T) - F(N_{c,2}, N_{s,2}, V, T)\,. \end{aligned} \tag{67}$$

Since the dispersion is space filling, we have in each of the two boxes (with $j = 1$ or 2),

$$v_s^0\, N_{s,j} + v_c^0\, N_{c,j} \;=\; V\;, \tag{68}$$

where v_s^0 and v_c^0 are the volumes occupied by a solvent molecule and a colloidal sphere, respectively. Due to the very large size of a colloidal sphere as compared to that of the solvent molecules, virtually all solvent molecules are surrounded by other solvent molecules, as in the pure solvent. For incompressible solvents, the specific molar volume of solvent will therefore be assumed independent of the concentration of Brownian particles and pressure. In order to indicate the independence of the molar volume v_s of solvent on intensive variables, it is denoted with a superscript "0". The specific molecular volume of a Brownian sphere is to very good approximation equal to its bare volume, and is denoted as v_c^0. It is also assumed

to be independent of concentration and pressure. A prescribed volume V thus implies that, $\delta N_s = -(v_c^0/v_s^0)\,\delta N_c$. Expansion of the free energies with respect to δN_s and δN_c, using that $\partial F(N_c, N_s, V, T)/\partial N_c = \mu_c(N_c, N_s, V, T)$, with μ_c the chemical potential of the Brownian particles, and similarly for the solvent, it is thus found that,

$$\begin{aligned} \delta w^{rev}/\delta N_c &= \mu_c(N_{c,2}, N_{s,2}, V, T) - \mu_c(N_{c,1}, N_{s,1}, V, T) \\ &\quad -\frac{v_c^0}{v_s^0}\left\{\mu_s(N_{c,2}, N_{s,2}, V, T) - \mu_s(N_{c,1}, N_{s,1}, V, T)\right\} . \end{aligned} \tag{69}$$

The continuum form of eq.(69) reads,

$$\delta w^{rev}/\delta N_c = \mathbf{L}\cdot\nabla\left[\mu_c(\mathbf{r},t) - \frac{v_c^0}{v_s^0}\,\mu_s(\mathbf{r},t)\right] , \tag{70}$$

with $\mathbf{L}$ the distance over which a colloidal sphere is moved on average on transport from box 1 to 2, that is, $\mathbf{L}$ is equal to $\mathbf{R}_2 - \mathbf{R}_1$, with $\mathbf{R}_j$ the center-position of box j. Since $\delta w^{rev}/\delta N_c = -\mathbf{F}\cdot\mathbf{L}$, with $\mathbf{F}$ the force with which surrounding material acts on a Brownian sphere (which has to be counter balanced quasi-statically by the force needed to transport a sphere from box 1 to box 2), it is found from eq.(70) that,

$$\mathbf{F}(\mathbf{r},t) = -\nabla\,\nu(\mathbf{r},t) , \tag{71}$$

where the "effective chemical potential" ν is defined as,

$$\nu \equiv \mu_c(\mathbf{r},t) - \frac{v_c^0}{v_s^0}\,\mu_s(\mathbf{r},t) . \tag{72}$$

When a colloidal particle is displaced, colloidal material is actually interchanged with an equal volume of solvent (see Fig.10b). This is the reason why in the effective chemical potential in eq.(72), the difference between the chemical potential of a Brownian particle and the chemical potential of an equal volume of solvent appears.
From the Gibbs-Duhem relation (with p the mechanical pressure),

$$0 = V\,\delta p - N_c\,\delta\mu_c - N_s\,\delta\mu_s \qquad (T = \text{constant}) , \tag{73}$$

it is easily shown that,

$$\delta\nu = \frac{V}{N_c}\,\delta\left\{p - \frac{1}{v_s^0}\,\mu_s\right\} . \tag{74}$$

This Gibbs-Duhem relation relates changes of intensive variables, under the condition that the change is established by means of exchanging only reversibel work, but no heat. Since p and μ_s are functions of $\{N_c, N_s, V, T\}$, eq.(74) can be written as,

$$\delta\nu = \frac{V}{N_c}\left\{\delta N_c\frac{\partial}{\partial N_c} + \delta N_s\frac{\partial}{\partial N_s} + \delta V\frac{\partial}{\partial V} + \delta T\frac{\partial}{\partial T}\right\}\left[p - \frac{1}{v_s^0}\,\mu_s\right] . \tag{75}$$

In all partial differentiations, p and μ_s are regarded as functions of the independent variables $\{N_c, N_s, V, T\}$. Since the addition of δN_c colloidal particles to a box is accompanied by a

decrease of the number of solvent molecules by $\delta N_s = -(v_c^0/v_s^0)\,\delta N_c$, it follows immediately that, at constant volume and temperature,

$$\delta\nu \;=\; \frac{V}{N_c}\,\delta N_c \left\{ \frac{\partial}{\partial N_c} - \frac{v_c^0}{v_s^0}\frac{\partial}{\partial N_s} \right\} \left[p - \frac{1}{v_s^0}\,\mu_s \right] . \tag{76}$$

Due to the fast adaption of the degrees of freedom of the solvent molecules to the instantaneous distribution of colloidal material and local temperature, one can think of each box as being in osmotic equilibrium with a reservoir of pure solvent. The Gibbs-Duhem relation for the pure solvent, under the restriction that only reversible work is exchanged, reads (with p_s the mechanical pressure within the pure solvent),

$$\delta p_s \;=\; \frac{1}{v_s^0}\,\delta\mu_s \;, \tag{77}$$

where it is used that for the pure solvent $V_s^r/N_s^r = v_s^0$, where V_s^r is the volume and N_s^r the number of solvent molecules in the reservoir of pure solvent (the superscript " r" stands for "reservoir"). A change of the osmotic pressure $\Pi \equiv p - p_s$ (the difference between the mechanical pressure of the suspension and the pure solvent) is therefore equal to,

$$\delta\Pi \;=\; \delta\left[p - \frac{1}{v_s^0}\,\mu_s \right] . \tag{78}$$

The operator within the curly brackets in eq.(76) is nothing but $\partial(\dots)/\partial N_c$ for the osmotic system, since one Brownian sphere is always exchanged with $-v_c^0/v_s^0$ solvent molecules. Hence, eq.(76) can be written simply as,

$$\delta\nu \;=\; \frac{V}{N_c}\,\delta N_c \frac{\partial\,\Pi}{\partial N_c} \;=\; \delta\rho\,\frac{1}{\rho}\,\frac{\partial\,\Pi(\rho,T,\mu_s)}{\partial\rho} . \tag{79}$$

In the last equation it is used that Π, being an intensive variable, depends on the extensive variables N_c and V only through the combination $\rho = N_c/V$. Replacing δ's by gradients, using that $\delta(\cdots) = \mathbf{L}\cdot\nabla(\cdots)$, as before, is thus found from eqs.(71,79) that,

$$\mathbf{F}(\mathbf{r},t) \;=\; -\frac{1}{\rho}\,\frac{\partial\Pi(\rho,T,\mu_s)}{\partial\rho}\,\nabla\rho(\mathbf{r},t) . \tag{80}$$

Note that this is equivalent to,

$$\rho\,\mathbf{F}(\mathbf{r},t) \;=\; -\nabla\Pi(\mathbf{r},t) . \tag{81}$$

The body force $\rho\,\mathbf{F}$ is thus the result of gradients in the osmotic pressure, similar to molecular systems in the absence of flow, where the body force is equal to the gradient of the mechanical pressure. In osmotic systems, work is done against gradients in the osmotic pressure. From eq.(80) for the force and eqs.(64-66), it is thus found that to leading order in gradients,

$$\frac{\partial}{\partial t}\,\rho(\mathbf{r},t) \;=\; D\,\nabla^2\,\rho(\mathbf{r},t) \;, \tag{82}$$

where Fick's diffusion coefficient D is defined as,

$$D \;=\; D_0\,\beta\,\frac{\partial\Pi(\rho,T,\mu_s)}{\partial\rho} . \tag{83}$$

where D_0 is the Einstein diffusion coefficient (16) for non-interacting particles. In deriving eq.(82), terms like $\sim \nabla\rho\cdot\nabla\rho$ have been neglected. These terms are of second order in deviations of the density from its mean value. Hence, eq.(82) is valid only to leading order in gradients *and* small deviations from mean values.
The expression (82) has been derived along different ways by Batchelor [11] (see also refs.[2],[3]). For very dilute systems, where interactions between Brownian particles are unimportant, the osmotic pressure is given by the ideal gas law (in this context commonly referred to as "van 't Hoff's law"), $\Pi = \rho\, k_B T$, and hence, $D = D_0$, as it should. For higher concentrations, the osmotic pressure can in principle be calculated from (see, for example, ref.[10]),

$$\Pi(\rho, T, \mu_s) \;=\; \rho\, k_B T - \frac{2\pi}{3}\,\rho^2 \int_0^\infty dR\, R^3\, g(R)\, \frac{dV(R)}{dR}\;, \tag{84}$$

where R is the distance between two Brownian particles and g is the pair-correlation function, which is the two-particle probability density function, re-normalized such that $g(R) \rightarrow 1$ for $R \rightarrow \infty$. To leading order in concentration, the pair-correlation function is equal to,

$$g(R) \;=\; \exp\{-\beta\, V(R)\}\;. \tag{85}$$

For a hard-core potential ($V(R) = 0$ for $R > 2a$ and $V(R) = \infty$ for $R \leq 2a$, with a the radius of the cores of the Brownian particles), eqs.(84,85) lead to,

$$D \;=\; D_0\,\left[1 + 8\,\varphi \,+\, \mathcal{O}(\varphi^2)\right]\;. \tag{86}$$

where $\varphi = \frac{4\pi}{3}\, a^3\, \rho$ is *the volume fraction of Brownian particles*. We shall see later, on the basis of a microscopic approach, that the pre-factor 8 is reduced to 1.45 due to hydrodynamic interactions between the Brownian particles, which can not be taken into account within the present thermodynamic treatment.
One might ask what the nature of the driving force for diffusion is when Brownian particles do not interact with each other. To answer this question, note that this force follows from eq.(80) with $\Pi = \rho\, k_B T$ as,

$$\mathbf{F}(\mathbf{r}, t) \;=\; -\,k_B T\, \nabla\, \ln\{\rho(\mathbf{r}, t)\}\;, \tag{87}$$

again to leading order in gradients. The logarithm indicates that this force is of an entropic nature. Even if the Brownian particles do not interact, there is a force that tends to render the system homogeneous, which maximizes the entropy. This force is usually referred to as "the Brownian force" $\mathbf{F}^{Br}$, which we will encounter again in section 4 in connection to the derivation of microscopic diffusion equations.

3.5 Diffusion in a Inhomogeneous Solvent : Brownian Motion of a Cell Towards Food-Rich Regions

Consider a fluid in which a molecular compound is dissolved in an inhomogeneous manner. We shall discuss the diffusive motion of a Brownian particle in this inhomogeneous solvent. The Brownian particle may be thought of as a cell, and the molecular compound as "food" for the cell. The concentration gradient in food will be shown to result in a systematic motion of the cell towards regions of higher food concentration.

Within the inhomogeneous environment, a change of the chemical potential on moving a cell from box 1 to box 2 is now also due to the accompanied interchange of dissolved food molecules between box 2 and box 1 (see Fig.10a in the previous subsection). Let us repeat the steps in the previous subsection in the presence of a food gradient. Let $N_{f,j}$ denote the number of food molecules in box j. Equation (69) generalizes to,

$$\begin{aligned} \delta w^{rev} &= \delta N_c \left[\mu_c(N_{c,2}, N_{s,2}, N_{f,2}, V, T) - \mu_c(N_{c,1}, N_{s,1}, N_{f,1}, V, T) \right] \\ &+ \delta N_s \left[\mu_s(N_{c,2}, N_{s,2}, N_{f,2}, V, T) - \mu_s(N_{c,1}, N_{s,1}, N_{f,1}, V, T) \right] \\ &+ \delta N_f \left[\mu_f(N_{c,2}, N_{s,2}, N_{f,2}, V, T) - \mu_f(N_{c,1}, N_{s,1}, N_{f,1}, V, T) \right] , \end{aligned} \tag{88}$$

where δN_f is the number of food molecules that is exchanged between box 2 and box 1 on moving δN_c cells from box 1 to box 2. For small concentrations of food, we still have to good approximation, $\delta N_s = -(v_c^0/v_s^0)\, \delta N_c$. The number of food molecules contained within a volume equal to that of the cell is $v_c^0\, \rho_f^0$. Hence, the number of food molecules that is exchanged is equal to $\delta N_f = -v_c^0\, \rho_f^0\, \delta N_c$, where ρ_f^0 is the local bulk concentration of food. Hence, the analogue of eq.(70) for this exchange of matter becomes,

$$\delta w^{rev}/\delta N_c = \mathbf{L} \cdot \nabla \left[\mu_c(\mathbf{r}, t) - \frac{v_c^0}{v_s^0}\, \mu_s(\mathbf{r}, t) - v_c^0\, \rho_f^0\, \mu_f(\mathbf{r}, t) \right] , \tag{89}$$

The exchange of matter described above changes the food concentration gradient. When we impose a prescribed gradient in food concentration, food has to be exchanged with box 1, to restore its original concentration. Since diffusion of the very small food molecules is much faster than that of the large cells, a constant food gradient is maintained by diffusion from food from the surroundings into box 1. The number of food molecules $\delta N_{f,1,sfg}$ that should be added to box 1 to guarantee a constant gradient in food concentration, after moving δN_c cells from box 1 to box 2, is (the index "sfg" refers to "stationary food gradient", and $\rho_{f,j}$ is the prescribed food number concentration in box j),

$$\delta N_{f,1,sfg} = \delta N_c\, v_c^0 \left[\rho_{f,1} - \rho_{f,2} \right] . \tag{90}$$

The accompanied reversible work to add this amount of food molecules is equal to,

$$\delta w^{rev}/\delta N_c = \mu_f\, \delta N_{f,1,sfg} = -\mu_f\, v_c^0\, \mathbf{L} \cdot \nabla \rho_f(\mathbf{r}) , \tag{91}$$

The total force is thus equal to, ,

$$\mathbf{F}(\mathbf{r}, t) = -\nabla\, \nu_f(\mathbf{r}, t) + \mu_f\, v_c^0\, \nabla \rho_f(\mathbf{r}) , \tag{92}$$

where the "effective chemical potential" ν_f, including food, is defined as,

$$\nu_f \equiv \mu_c(\mathbf{r}, t) - \frac{v_c^0}{v_s^0}\, \mu_s(\mathbf{r}, t) - v_c^0\, \rho_f^0\, \mu_f(\mathbf{r}, t) . \tag{93}$$

Similarly as in the previous subsection, it follows from the Gibbs-Duhem relation (with p the mechanical pressure),

$$0 = V\, \delta p - N_c\, \delta \mu_c - N_s\, \delta \mu_s - N_f\, \delta \mu_f \qquad (T = \text{constant}) , \tag{94}$$

that,

$$\delta \nu_f = \frac{V}{N_c}\, \delta \left\{ p - \frac{1}{v_s^0}\, \mu_s - \rho_f^0 \left[1 + \varphi \right] \mu_f \right\} , \tag{95}$$

where $\varphi = v_c^0 N_c/V$ is the volume fraction of cells. We shall henceforth assume that the cell concentration is very small, that is, $\varphi \ll 1$. For such low volume fractions, and at constant V and T, the analogue of eqs.(75,76) reads,

$$\begin{aligned}\delta\nu_f &= \frac{V}{N_c}\left\{\delta N_c\frac{\partial}{\partial N_c}+\delta N_s\frac{\partial}{\partial N_s}+\delta N_f\frac{\partial}{\partial N_f}\right\}\left[p-\frac{1}{v_s^0}\mu_s-\rho_f^0\mu_f\right]\\ &= \frac{V}{N_c}\delta N_c\left\{\frac{\partial}{\partial N_c}-\frac{v_c^0}{v_s^0}\frac{\partial}{\partial N_s}-v_c^0\rho_f^0\frac{\partial}{\partial N_f}\right\}\left[p-\frac{1}{v_s^0}\mu_s-\rho_f^0\mu_f\right]\ . \end{aligned} \qquad (96)$$

The Gibbs-Duhem relation (94) for the solution of food, without the cells being present, reads,

$$\delta p_s = \delta\left[\frac{1}{v_s^0}\mu_s+\rho_f^0\mu_f\right] \qquad (T=\text{constant})\ . \qquad (97)$$

where p_s is the mechanical pressure of the food solution in the absence of the cells. Hence, the first term in the square brackets in eq.(96) is equal to $\delta\Pi_f$, with Π_f the osmotic pressure of the cell solution against the food solution. Furthermore, as in the previous subsection, the operator in the curly brackets is the operator $\partial(\cdots)/\partial N_c$ within the osmotic system under consideration. We thus find the following expression for the force,

$$\mathbf{F}(\mathbf{r},t) = -\frac{1}{\rho}\frac{\partial\Pi_f(\rho,T,\mu_s,\mu_f)}{\partial\rho}\nabla\rho(\mathbf{r},t)+\mu_f v_c^0\nabla\rho_f(\mathbf{r})\ , \qquad (98)$$

where ρ is the number concentration of cells, and ρ_f that of food molecules. In the absence of a concentration gradient of cells, the force is entirely determined by gradients in food concentration. The mean velocity $\mathbf{v}_{cell}$ that the cell attains in a prescribed food gradient thus follows from eqs.(65,66) as,

$$\mathbf{v}_{cell} = \frac{v_c^0\mu_f}{\gamma}\nabla\rho_f(\mathbf{r})\ . \qquad (99)$$

Note that this finite velocity originates from the second term in eq.(98), which in turn is entirely due to the fact that a *stationary* food gradient is maintained.

The result in eq.(99) can be interpreted as follows. First we can rewrite, $\mathbf{v}_{cell} = C\,\nabla(\rho_f(\mathbf{r})/\rho_f^0)$, with $C = v_c^0\rho_f^0\mu_f/\gamma$. In the prefactor C, μ_f is the chemical potential per food molecule, so that $v_c^0\rho_f^0\mu_f$ is the chemical potential associated with the total number of food molecules that are contained in a volume equal to that of the cell. This is the change in chemical potential on moving the cell, and thus sets the driving force for motion of the cell. The inverse proportionality on the friction coefficient γ expresses the increase of the friction force on increasing the cell's velocity.

According to eq.(99) the cell moves towards regions with higher food concentration, as it should, since $\mu_f > 0$. A typical velocity of a cell can be estimated as follows. Consider a cell with a size of $5\,\mu m$. The volume of the cell is then equal to $5.10^{-16}\,m^3$. A typical value for the chemical potential is a few k_BT, say, $10^{-20}\,Nm$. The number concentration of food is typically 10^{25} molecules per m^3. A reasonable value for the gradient in food concentration is $10^{24}\,m^{-4}$. Hence, in water, $v_{cell} \approx 50\,\mu m/s$. The cell thus traverses a distances equal to ten times its own size in a second, which seems a reasonable velocity to prevent starvation.

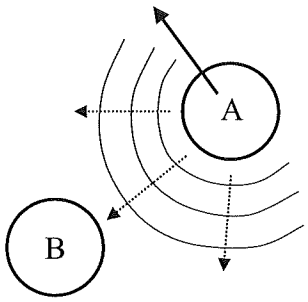


Fig. 11: *Hydrodynamic interaction. Particle A induces a fluid flow in the solvent which affects particle B in its motion.*

4 The Generalized Diffusion Equation

In previous sections we derived Fick's equation of motion for the number density of Brownian particles, which is proportional to the single particle probability density function (pdf). Fick's equation of motion is valid only for very small gradients of the density, in the short- and long-time limit. Here, we shall discuss a more general diffusion equation for the pdf of the position coordinates of N, possibly interacting, spherical Brownian particles. Under appropriate assumptions this diffusion equation reduces to Fick's law.
Before going into the derivation of the generalized diffusion equation, which is commonly referred to as the Smoluchowski equation, we shall discuss hydrodynamic interactions on an intuitive level, followed by a more quantitative treatment to obtain leading order expressions for hydrodynamic interaction function for large distances between two Brownian spheres. The calculation of more accurate expression, valid for smaller distances between the Brownian spheres, is beyond the scope of the present lecture.

4.1 A Primer on Hydrodynamic Interactions (HI)

Besides direct interactions, which are also present in molecular systems, there are so-called hydrodynamic interactions between Brownian particles immersed in a fluid. The origin of these interactions is not difficult to understand. A Brownian particle that attained a velocity at a certain time induces a fluid flow in the solvent. This fluid flow propagates through the solvent and encounters other Brownian particles (see Fig.11), which are thus affected in their motion.
Hydrodynamic interactions can be described on the basis of the Navier-Stokes equation for the solvent, provided that the Brownian particles are very large in comparison to the solvent molecules, so that they may be treated as macroscopic bodies. In the present section, generic features of hydrodynamic interactions are discussed, mostly on an intuitive level.
It can be shown that the propagation of fluid flow disturbances, created by the movement of a Brownian particle, is so fast, that the phase space coordinates of the Brownian particles hardly change during the time interval that a disturbance takes to reach other Brownian particles. It is then a good approximation to assume that hydrodynamic interactions are instantaneous. The fluid flow resulting from the movement of a Brownian particle can thus be thought of as existing, without any time delay, throughout the entire fluid. In this approximation the fluid flow at a given time is a function of the velocities and positions of all the Brownian particles at that

particular instant only. As a result, the hydrodynamic interaction forces are functions of the instantaneous momentum and position coordinates. Thus, the force $\mathbf{F}_i^h$ that the solvent exerts on the i^{th} Brownian particle is set by the instantaneous values of the momentum and position coordinates of all the Brownian particles,

$$\mathbf{F}_i^h \equiv \mathbf{F}_i^h\left(\mathbf{p}_1(t), \cdots, \mathbf{p}_N(t), \mathbf{r}_1(t), \cdots, \mathbf{r}_N(t)\right).$$

Here, the superscript "h" stands for "hydrodynamic".
It turns out that the Navier-Stokes equation, for the typical sizes and velocities of the Brownian particles and the typical shear viscosity of the solvent, can be linearized with respect to the fluid flow velocity. More precisely, the Reynolds number for the hydrodynamic problem considered here is small, which allows the neglect of the inertial terms in the Navier-Stokes equation. As a result, the hydrodynamic forces are linear functions of the velocities $\mathbf{v}_j$ of the Brownian particles. Hence, the above expression reduces to a linear form,

$$\mathbf{F}_i^h = -\sum_{j=1}^{N} \boldsymbol{\Upsilon}_{ij}(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N) \cdot \mathbf{v}_j \,. \tag{100}$$

The 3×3-dimensional matrices $\boldsymbol{\Upsilon}_{ij}$ are referred to as *microscopic friction matrices*. They are functions of all the position coordinates, but independent of the velocities. The mathematical problem concerned with hydrodynamic interactions is the explicit calculation of these matrices. A minus sign is added to the right hand-side in the above expression because the hydrodynamic force on a particle tends to be directed in the opposite direction to its velocity.
In very dilute suspensions, where distances between (the majority of) Brownian particles are very large, hydrodynamic interactions are unimportant. In that case eq.(100) reduces to,

$$\mathbf{F}_i^h \;=\; -\gamma\,\mathbf{v}_i \,, \tag{101}$$

with γ the friction coefficient of an isolated sphere. According to eq.(100) thus amounts to,

$$\boldsymbol{\Upsilon}_{ij}(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N) \;=\; \gamma\,\hat{\mathbf{I}}\,\delta_{ij} \quad , \quad \text{no HI} \,, \tag{102}$$

where $\hat{\mathbf{I}}$ is the identity matrix and δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$).
The "off-diagonal" friction matrices, $\boldsymbol{\Upsilon}_{ij}$ with $i \neq j$, describe the hydrodynamic interaction of particle j with particle i. The "diagonal" friction matrices $\boldsymbol{\Upsilon}_{ii}$ also depend on hydrodynamic interaction, and are not just equal to the friction coefficient of an isolated sphere, as one might think on first sight. The fluid flow that is the result of movement of the i^{th} sphere is "reflected" from the other Brownian particles back to the i^{th} sphere, thus exerting a force on that particle in addition to the friction force of an isolated particle.

4.2 The Leading Order Expansion of Hydrodynamic Interaction Functions : The Oseen Matrix

In the present subsection we shall calculate the asymptotic form of hydrodynamic interaction functions for large separations between Brownian particles.
It can be shown that on the Brownian length- and time scale, inertial forces in the equation of motion of Newton for the solvent flow (the so-called Navier-Stokes equation) can be neglected.

There are three non-inertial forces : forces due to gradients in the pressure, forces due friction between sliding layers of solvent and possibly external forces. These three forces add up to zero on the Brownian length- and time scale. The Navier-Stokes equation then reduces to,

$$\nabla p(\mathbf{r},t) - \eta_0 \nabla^2 \mathbf{u}(\mathbf{r},t) \;=\; \mathbf{f}^{ext}(\mathbf{r}) \;. \tag{103}$$

where p is the mechanical pressure, $\hat{\mathbf{u}}$ is the solvent flow velocity, η_0 is the shear-viscosity of the solvent and $\mathbf{f}^{ext}$ is the external force acting on the fluid. Together with incompressibility, implying that,

$$\nabla \cdot \mathbf{u}(\mathbf{r},t) \;=\; 0 \;, \tag{104}$$

the pressure and velocity can be calculated once the external force is specified. Equations (103,104) are commonly referred to as *the creeping flow equations*. These equations are the starting point for the calculation of hydrodynamic interaction functions in eq.(100).
As we shall see in subsection 4.3, the inverse of the friction tensor occurs in the generalized diffusion equation of interest. We shall thus calculate here the so-called *microscopic diffusion matrix*, which is defined as,

$$\begin{pmatrix} \mathbf{D}_{11} & \cdots & \mathbf{D}_{1N} \\ \vdots & \cdots & \vdots \\ \mathbf{D}_{N1} & \cdots & \mathbf{D}_{NN} \end{pmatrix} \equiv k_B T \begin{pmatrix} \boldsymbol{\Upsilon}_{11} & \cdots & \boldsymbol{\Upsilon}_{1N} \\ \vdots & \cdots & \vdots \\ \boldsymbol{\Upsilon}_{N1} & \cdots & \boldsymbol{\Upsilon}_{NN} \end{pmatrix}^{-1} \;. \tag{105}$$

Without hydrodynamic interactions it follows from eq.(102) that,

$$\mathbf{D}_{ij} \;=\; D_0 \hat{\mathbf{I}} \delta_{ij} \;\;,\;\; \text{no HI} \;, \tag{106}$$

where, as before, $D_0 = k_B T/\gamma$ is Einstein's diffusion coefficient of a non-interacting Brownian sphere (which is the reason that the pre-factor $k_B T$ is introduced in the definition (105)).
An external force acting only in a single point $\mathbf{r}'$ on the fluid is mathematically described by a delta distribution,

$$\mathbf{f}^{ext}(\mathbf{r}) \;=\; \mathbf{f}_0 \, \delta(\mathbf{r} - \mathbf{r}') \;. \tag{107}$$

The prefactor $\mathbf{f}_0$ is the total force $\int d\mathbf{r}' \, \mathbf{f}^{ext}(\mathbf{r}')$ acting on the fluid. Since the creeping flow equations are linear, the fluid flow velocity at some point $\mathbf{r}$ in the fluid, due to the point force in $\mathbf{r}'$, is directly proportional to that point force. Hence,

$$\mathbf{u}(\mathbf{r}) \;=\; \mathbf{T}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_0 \;. \tag{108}$$

The matrix $\mathbf{T}$ is commonly referred to as *the Oseen matrix*. This matrix connects the point force at a point $\mathbf{r}'$ to the resulting fluid flow velocity at a point $\mathbf{r}$. That $\mathbf{T}$ is only a function of the difference coordinate $\mathbf{r} - \mathbf{r}'$ follows from translational invariance, or to put it in other words, from the fact that the choice of the position of the origin is of no significance. Similarly, the pressure at a point $\mathbf{r}$ is linearly related to the point force,

$$p(\mathbf{r}) \;=\; \mathbf{g}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_0 \;. \tag{109}$$

The vector $\mathbf{g}$ is referred to here as the *pressure vector*.

Consider now an external force which is continuously distributed over the entire fluid. Due to the linearity of the creeping flow equations, the fluid flow velocity at some point $\mathbf{r}$ is simply the superposition of the fluid flow velocities resulting from the forces acting in each point on the fluid. Hence, from eq.(108),

$$\mathbf{u}(\mathbf{r}) \;=\; \int d\mathbf{r}'\, \mathbf{T}(\mathbf{r}-\mathbf{r}')\cdot \mathbf{f}^{ext}(\mathbf{r}')\;. \tag{110}$$

The same holds for the pressure,

$$p(\mathbf{r}) \;=\; \int d\mathbf{r}'\, \mathbf{g}(\mathbf{r}-\mathbf{r}')\cdot \mathbf{f}^{ext}(\mathbf{r}')\;. \tag{111}$$

In mathematical language, the Oseen matrix and the pressure vector are the Green's functions of the creeping flow equations for the fluid flow velocity and pressure, respectively. Once these Green's functions are known and the external force is specified, the resulting fluid velocity and pressure can be calculated through the evaluation of the above integrals. The calculation of the Green's functions is thus equivalent to solving the creeping flow equations, provided that the external forces are known.
Let us calculate the Oseen matrix and pressure vector. To this end, substitute eqs.(110,111) into the creeping flow equations (103,104). This leads to,

$$\begin{aligned}
\int d\mathbf{r}'\, \left[\nabla\cdot\mathbf{T}(\mathbf{r}-\mathbf{r}')\right]\cdot\mathbf{f}^{ext}(\mathbf{r}') \;&=\; 0\,,\\
\int d\mathbf{r}'\, \left[\nabla\mathbf{g}(\mathbf{r}-\mathbf{r}') - \eta_0\nabla^2\mathbf{T}(\mathbf{r}-\mathbf{r}') - \hat{\mathbf{I}}\,\delta(\mathbf{r}-\mathbf{r}')\right]\cdot\mathbf{f}^{ext}(\mathbf{r}') \;&=\; \mathbf{0}\,,
\end{aligned}$$

where $\hat{\mathbf{I}}$ is the 3×3-dimensional unit matrix. Since the external force is arbitrary, the expressions in the square brackets must be equal to zero, so that the Green's functions satisfy the following differential equations,

$$\begin{aligned}
\nabla\cdot\mathbf{T}(\mathbf{r}) \;&=\; \mathbf{0}\,,\\
\nabla\mathbf{g}(\mathbf{r}) - \eta_0\nabla^2\mathbf{T}(\mathbf{r}) \;&=\; \hat{\mathbf{I}}\,\delta(\mathbf{r})\,.
\end{aligned} \tag{112}$$

A single equation for the pressure vector is obtained by taking the divergence of the second equation, with the use of the first equation,

$$\nabla^2\mathbf{g}(\mathbf{r}) \;=\; \nabla\cdot\hat{\mathbf{I}}\delta(\mathbf{r}) \;=\; \nabla\delta(\mathbf{r})\,.$$

Now using the standard representation,

$$\frac{1}{4\pi}\nabla^2\frac{1}{r} \;=\; -\delta(\mathbf{r})\,, \tag{113}$$

it follows that,

$$\mathbf{g}(\mathbf{r}) \;=\; -\frac{1}{4\pi}\nabla\frac{1}{r} + \mathbf{G}(\mathbf{r})\,,$$

where $\mathbf{G}$ is a vector for which $\nabla^2\mathbf{G}=\mathbf{0}$. It can be shown that the condition $\mathbf{G}\rightarrow\mathbf{0}$ as $r\rightarrow\infty$ implies that $\mathbf{G}\equiv\mathbf{0}$. Hence,

$$\mathbf{g}(\mathbf{r}) \;=\; -\frac{1}{4\pi}\nabla\frac{1}{r} \;=\; \frac{1}{4\pi}\frac{\mathbf{r}}{r^3}\,. \tag{114}$$

The differential equation to be satisfied by the Green's function for the fluid flow velocity (the Oseen matrix), is found by substitution of eq.(113) into eq.(112), and using eq.(114),

$$\nabla^2 \left[\frac{1}{4\pi} \frac{1}{r} \hat{\mathbf{I}} - \eta_0 \mathbf{T}(\mathbf{r}) \right] = \frac{1}{4\pi} \left[3 \frac{\mathbf{rr}}{r^5} - \frac{1}{r^3} \hat{\mathbf{I}} \right] .$$

An obvious choice for the term between the square brackets on the left hand-side of the above expression is of the form,

$$\frac{1}{4\pi} \frac{1}{r} \hat{\mathbf{I}} - \eta_0 \mathbf{T}(\mathbf{r}) = \alpha_0 \frac{1}{r^n} \hat{\mathbf{I}} + \alpha_1 \frac{1}{r^m} \frac{\mathbf{rr}}{r^2} ,$$

with $\alpha_{0,1}$, n and m constants. These constants can indeed be chosen such that this Ansatz is the solution of the differential equation (with the boundary condition that $\mathbf{T}(\mathbf{r}) \to 0$ as $r \to \infty$). A somewhat lengthy, but straightforward calculation yields,

$$\mathbf{T}(\mathbf{r}) = \frac{1}{8\pi\eta_0} \frac{1}{r} \left[\hat{\mathbf{I}} + \frac{\mathbf{rr}}{r^2} \right] . \tag{115}$$

This concludes the determination of the Green's functions for the creeping flow equations. These functions, the Oseen matrix in particular, play a central role in the calculation of microscopic diffusion matrices.

Note that the Oseen tensor decays like $\sim 1/r$, and is thus very long ranged. Hydrodynamic interactions therefore extend over long distances.

The starting point for the calculation of the microscopic diffusion matrices is the Green's function representation (110) of the creeping flow equations. In the present situation, the external force $\mathbf{f}^{ext}$ is due to forces that surface elements of the Brownian spheres exert on the fluid. For the multi sphere problem considered here, the integral in eq.(110) is now a sum of integrals ranging over the surfaces ∂V_j , $j = 1, \cdots, N$ of the N spherical Brownian particles,

$$\mathbf{u}(\mathbf{r}) = \sum_{j=1}^{N} \oint_{\partial V_j} dS' \, \mathbf{T}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_j(\mathbf{r}') , \tag{116}$$

$$p(\mathbf{r}) = \sum_{j=1}^{N} \oint_{\partial V_j} dS' \, \mathbf{g}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_j(\mathbf{r}') , \tag{117}$$

where $\mathbf{f}_j$ is the force per unit area that a surface element with position $\mathbf{r}'$ of Brownian particle j exerts on the fluid.

We shall assume stick boundary conditions, that is, the velocity of the fluid at the spheres surface is assumed to be equal to the velocity of the corresponding surface element. Hence,

$$\mathbf{v}_i + \boldsymbol{\Omega}_i \times (\mathbf{r} - \mathbf{r}_i) = \sum_{j=1}^{N} \oint_{\partial V_j} dS' \, \mathbf{T}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_j(\mathbf{r}') \ , \ \mathbf{r} \in \partial V_i , \tag{118}$$

where $\mathbf{v}_i$ and $\boldsymbol{\Omega}_i$ are the translational- and rotational velocity of the i^{th} sphere, respectively. Since this equation is valid for *any* position $\mathbf{r}$ on the surface ∂V_i of particle i, both sides can be integrated over that surface. Due to symmetry, the rotational component on the left hand-side drops out, and we have,

$$\begin{aligned} \mathbf{v}_i = & \frac{1}{4\pi a^2} \oint_{\partial V_i} dS \oint_{\partial V_i} dS' \, \mathbf{T}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_i(\mathbf{r}') \\ & + \frac{1}{4\pi a^2} \sum_{j \neq i}^{N} \oint_{\partial V_i} dS \oint_{\partial V_j} dS' \, \mathbf{T}(\mathbf{r} - \mathbf{r}') \cdot \mathbf{f}_j(\mathbf{r}') . \end{aligned} \tag{119}$$

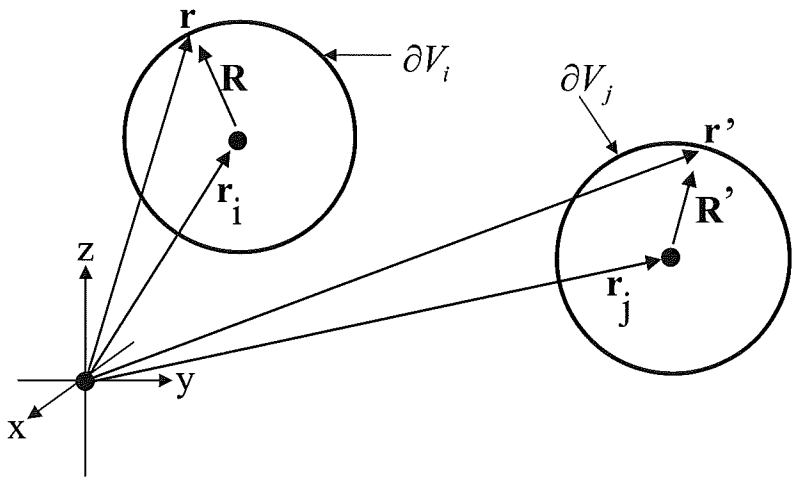


Fig. 12: *Definition of the positions* $\mathbf{R}$ *and* $\mathbf{R}'$ *on the surface of Brownian particles relative to their position coordinates* $\mathbf{r}_i$ *and* $\mathbf{r}_j$*, respectively.*

Since,

$$\oint_{\partial V_i} dS\, \mathbf{T}(\mathbf{r}-\mathbf{r}') \;=\; \hat{\mathbf{I}}\,\frac{2a}{3\eta_0}\;,\;\; \text{for } \mathbf{r}'\in\partial V_i\;, \tag{120}$$

the first term on the right hand-side of eq.(119) is thus equal to,

$$\frac{1}{4\pi a^2}\oint_{\partial V_i} dS \oint_{\partial V_i} dS'\,\mathbf{T}(\mathbf{r}-\mathbf{r}')\cdot\mathbf{f}_i(\mathbf{r}') = -\frac{1}{6\pi\eta_0 a}\,\mathbf{F}_i^h\;,$$

where the total force that the fluid exerts on the i^{th} Brownian particle is equal to,

$$\mathbf{F}_i^h(t) \;=\; -\oint_{\partial V_i} dS'\,\mathbf{f}_i(\mathbf{r}')\;. \tag{121}$$

The double surface integrals in the second line on the right hand-side of eq.(119) can be approximated, in case the distance between the Brownian particles is large, as follows. First, the integrations are performed with respect to the translated coordinates $\mathbf{R}=\mathbf{r}-\mathbf{r}_i$ and $\mathbf{R}'=\mathbf{r}'-\mathbf{r}_j$ (see Fig.12).
Let ∂V^0 denote the spherical surface ∂V_i with its center at the origin. The integrals on the right hand-side of eq.(119) are written as,

$$\frac{1}{4\pi a^2}\oint_{\partial V^0} dS \oint_{\partial V^0} dS'\,\mathbf{T}(\mathbf{R}-\mathbf{R}'+\mathbf{r}_i-\mathbf{r}_j)\cdot\mathbf{f}_j(\mathbf{R}'+\mathbf{r}_j)\;.$$

Now suppose that the distance $\mid \mathbf{r}_i-\mathbf{r}_j \mid$ between the Brownian particles i and j is much larger than $\mid \mathbf{R}-\mathbf{R}' \mid \leq 2a$, with a the radius of the sphere. The Oseen matrix $\mathbf{T}(\mathbf{R}-\mathbf{R}'+\mathbf{r}_i-\mathbf{r}_j)$ may then be replaced, to a good approximation, by $\mathbf{T}(\mathbf{r}_i-\mathbf{r}_j)$. With eq.(121) it then follows that,

$$\frac{1}{4\pi a^2}\oint_{\partial V_i} dS \oint_{\partial V_j} dS'\,\mathbf{T}(\mathbf{r}-\mathbf{r}')\cdot\mathbf{f}_j(\mathbf{r}') \approx -\mathbf{T}(\mathbf{r}_i-\mathbf{r}_j)\cdot\mathbf{F}_j^h\;.$$

For these large separations between the Brownian particles, eq.(5.51) can thus be approximated as,

$$\mathbf{v}_i = -\frac{1}{6\pi\eta_0 a}\mathbf{F}_i^h - \sum_{j\neq i}^{N} \mathbf{T}(\mathbf{r}_i - \mathbf{r}_j)\cdot\mathbf{F}_j^h \ . \tag{122}$$

Comparison with the definition (100,105) of the microscopic diffusion matrices thus leads to,

$$\mathbf{D}_{ii} = D_0\hat{\mathbf{I}} \ , \tag{123}$$

$$\mathbf{D}_{ij} = k_B T\,\mathbf{T}(\mathbf{r}_i - \mathbf{r}_j) = \frac{3}{4}D_0\frac{a}{r_{ij}}\left[\hat{\mathbf{I}} + \hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij}\right] \ , \ i \neq j \ , \tag{124}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ is the distance between the spheres i and j, and $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/r_{ij}$ is a unit vector. Furthermore, $D_0 = k_B T/6\pi\eta_0 a$ is the same Stokes-Einstein diffusion coefficient that we encountered before.

The above expressions, *the Oseen approximation for the microscopic diffusion matrices*, are valid for large distances between the Brownian particles, that is, for small values of a/r_{ij}. These results are the leading terms in an expansion with respect to a/r_{ij}.

On the two-particle level, where only two particles interact simultaneously, the general form of the diffusion matrices reads,

$$\mathbf{D}_{ii} = D_0\hat{\mathbf{I}} + D_0\sum_{j=1\,,\,j\neq i}^{N}\left\{A_s(r_{ij})\,\hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij} + B_s(r_{ij})\left[\hat{\mathbf{I}} - \hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij}\right]\right\} \ ,$$

$$\mathbf{D}_{ij} = D_0\left\{A_c(r_{ij})\,\hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij} + B_c(r_{ij})\left[\hat{\mathbf{I}} - \hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij}\right]\right\} \ , \ i \neq j \ . \tag{125}$$

The summation in the expression for the "self" diffusion matrix $\mathbf{D}_{ii}$ accounts for the fact that *all* particles in suspension reflect the field of the i^{th} particle back to that particle. The scalar functions A_s , B_s , A_c and B_c are referred to as *mobility functions*. These functions depend only on the scalar distance r_{ij} between the two spheres i and j. The mobility functions with $i = j$ are sometimes called *self-mobility functions*, and those with $i \neq j$, *distinct-* or *cross-mobility functions*. The subscripts s and c refer to "*self*" and "*cross*", respectively. The mobility functions can be calculated as a Taylor expansion with respect to the inverse distance between two Brownian particles. The first few terms in such an expansion read,

$$A_s = -\frac{15}{4}\left(\frac{a}{r_{ij}}\right)^4 + \frac{11}{2}\left(\frac{a}{r_{ij}}\right)^6 + O\left((a/r_{ij})^8\right) \ ,$$

$$B_s = -\frac{17}{16}\left(\frac{a}{r_{ij}}\right)^6 + O\left((a/r_{ij})^8\right) \ ,$$

$$A_c = \frac{3}{2}\frac{a}{r_{ij}} - \left(\frac{a}{r_{ij}}\right)^3 + \frac{75}{4}\left(\frac{a}{r_{ij}}\right)^7 + O\left((a/r_{ij})^9\right) \ ,$$

$$B_c = \frac{3}{4}\frac{a}{r_{ij}} + \frac{1}{2}\left(\frac{a}{r_{ij}}\right)^3 + O\left((a/r_{ij})^9\right) \ . \tag{126}$$

Nowadays, the numerical values of many hundreds of coefficients are known [12]. Note that the leading order terms here follow from our Oseen approximation (123). These expressions allow the calculation of transport coefficients up to leading order in concentration. For higher concentrations, when hydrodynamic interactions between more than two particles simultaneously are

important, one should derive expressions involving three and more particles. So far, only the leading order expressions for three particle hydrodynamic interactions are known. Alternatively one can perform simulations to take hydrodynamic interactions into account.

4.3 The Smoluchowski Equation for Spheres

Consider an ensemble of $\mathcal{N}$ identical containers filled with solvent, each containing just a single Brownian sphere. The "microstate" of the ensemble is characterized by $\mathcal{N}$ points in 3-dimensional space. Let $\mathcal{W}$ denote an arbitrary volume. The density of points at a certain position is proportional to the probability of finding a sphere at that position. Since the number of systems within the ensemble is fixed, the equation of motion for the probability density function (pdf) $P(\mathbf{r},t)$ of the position coordinate of the single Brownian sphere satisfies the conservation equation,

$$\frac{\partial}{\partial t}P(\mathbf{r},t) \;=\; -\nabla\cdot\left[\mathbf{v}\,P(\mathbf{r},t)\right]\;, \tag{127}$$

where $\mathbf{v} = d\mathbf{r}(t)/dt$ is the local velocity of Brownian spheres. If each container in the ensemble contains N spheres, the relevant phase space coordinates are all position coordinates $\{\mathbf{r}_1,\mathbf{r}_2,\cdots,\mathbf{r}_N\}$. The equation of motion for the probability density function P of these phase space coordinates is found from eq.(127) by simply adding the in- and out-fluxes over all spheres,

$$\frac{\partial}{\partial t}P(\mathbf{r}_1,\cdots,\mathbf{r}_N,t) = -\sum_{j=1}^{N}\nabla_j\cdot\left[\,\mathbf{v}_j\,P\,\right]\;, \tag{128}$$

where $\mathbf{v}_j$ is the velocity of the j^{th} Brownian sphere. This is an exact result, since it merely expresses conservation of the number of Brownian particles.
The translational velocities have to be expressed in terms of functions of the phase space coordinates. The key relation to express the velocities in terms of phase functions is the force balance equation, which is discussed in section SEC. There are in general four forces working on each sphere : the hydrodynamic force $\mathbf{F}_j^h$ that the solvent exerts on the sphere, the direct interaction force $\mathbf{F}_j^I$, a possible external force $\mathbf{F}_j^{ext}$ that acts on the Brownian particles (like an electric or gravitational field), and the Brownian force $\mathbf{F}_j^{Br}$ which will be discussed and specified later. Hence,

$$\textit{Total force} \;=\; \mathbf{F}_j^h+\mathbf{F}_j^I+\mathbf{F}_j^{ext}+\mathbf{F}_j^{Br} \;\approx\; \mathbf{0}\;, \tag{129}$$

The direct force is minus the gradient of the total potential energy Φ of the assembly of Brownian particles, so that, using eq.(100),

$$-\sum_{j=1}^{N}\boldsymbol{\Upsilon}_{ij}(\mathbf{r}_1,\mathbf{r}_2,\cdots,\mathbf{r}_N)\cdot\mathbf{v}_j-\nabla_j\,\Phi(\mathbf{r}_1,\cdots,\mathbf{r}_N)+\mathbf{F}_j^{ext}+\mathbf{F}_j^{Br} \;=\; \mathbf{0}\;. \tag{130}$$

Inversion of this vector equation this leads to the following expression for the velocity,

$$\mathbf{v}_j \;=\; -\sum_{i=1}^{N}\mathbf{D}_{ij}\cdot\left\{\,\nabla_j\,\Phi-\mathbf{F}_j^{ext}-\mathbf{F}_j^{Br}\,\right\}\;, \tag{131}$$

where the so-called *microscopic diffusion matrices* are defined in eq.(105).
We still have to express the Brownian force in terms of phase functions. This is achieved as follows. Since for infinite time $t \to \infty$, in the absence of an external force, the suspension attains equilibrium, the probability density function P is proportional to the Boltzmann exponential $\exp\{-\beta\,\Phi\}$, and $\partial P/\partial t = 0$. It thus follows from eqs.(128,131) that,

$$\mathbf{F}_j^{Br} \;=\; -k_B T \nabla_j \ln\{P\}\;. \tag{132}$$

This expression for the Brownian force is in agreement with eq.(87) that was found on the basis of thermodynamic reasoning. The Brownian force, which is an entropic force, arises from interactions of the colloidal particle with the solvent molecules, averaged with respect to the equilibrium probability density function for the phase space coordinates of the fluid molecules in the external field imposed by the colloidal particles, with prescribed positions. Even in a very dilute system of colloidal particles (an "ideal gas"), where interactions between the colloidal particles can be neglected, the equilibrium state is one where the macroscopic density is constant, independent of position. The forces that drive such an ideal gas to the homogeneous state are the Brownian forces.
The Smoluchowski equation for spherical particles thus follows from eqs.(128,131,132) as,

$$\frac{\partial}{\partial t} P(\mathbf{r}_1, \cdots, \mathbf{r}_N, t) \;=\; \sum_{i,j=1}^{N} \nabla_i \cdot \mathbf{D}_{ij} \cdot \left[\, \nabla_j P + \beta\, P\, \nabla_j \Phi - \beta\, P\, \mathbf{F}_j^{ext} \right]\;. \tag{133}$$

The leading order expressions (126) for the hydrodynamic interaction functions will be used in the next section to re-derive Fick's law from eq.(133), up to leading order in concentration, including the effect of hydrodynamic interactions.

4.4 Fick's Law Revisited

In this subsection we shall re-derive Fick's law up to leading order in concentration from the Smoluchowki equation (133) for spherical particles, including the hydrodynamic interactions (but without an external field). With the neglect of hydrodynamic interactions, eqs.(84,86) are recovered to leading order in concentration.
The equation of motion for the macroscopic density $\rho(\mathbf{r}_1, t)$ is obtained from the Smoluchowski equation by integration with respect to the position coordinates $\mathbf{r}_2$, $\mathbf{r}_3, \cdots$, $\mathbf{r}_N$, using the following relation between the N-particle pdf $P(\mathbf{r}_1, \cdots, \mathbf{r}_N, t)$ and the density,

$$\rho(\mathbf{r}_1, t) = N\, P_1(\mathbf{r}_1, t) = N \int d\mathbf{r}_2 \int d\mathbf{r}_3 \cdots \int d\mathbf{r}_N P(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N, t)\,, \tag{134}$$

where P_1 is the one-particle pdf that we encountered before. The problem that arises is that the microscopic diffusion matrices depend on the position coordinates of all the Brownian particles in the system under consideration. Upon integration, this leaves integrals with respect to $\mathbf{r}_2, \cdots, \mathbf{r}_N$, involving products of $\mathbf{D}_{ij}$ with $P(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N, t)$, which cannot be reduced further. We restrict ourselves here to concentrations which are so small, that events where more than two Brownian particles interact simultaneously hardly occur. For such small concentrations, the two-particle expressions for the microscopic diffusion matrices in eq.(126) may be used, which functions depend only on the difference of two particle positions. This leaves integrals involving only the lowest order pdf's, which are amenable to explicit evaluation.

To make any progress, it is necessary to assume that the total potential energy Φ of the assembly of N Brownian particles is pair-wise additive, that is,

$$\Phi(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N) \;=\; \sum_{i,j\,=\,1\;\;i\,<\,j}^{N} V(r_{ij}) \;, \tag{135}$$

which is a very good approximation for most systems. For low concentrations, where only binary interactions are considered, the expression (135) is exact. For spherically symmetric Brownian particles, the *pair-interaction potential* V is a function of the absolute distance $r_{ij} = \mid \mathbf{r}_i - \mathbf{r}_j \mid$ between two particles.
For identical Brownian particles each term in the summation yields upon integration an identical result, where i and $j = 1$ are special, since we do not integrate with respect to $\mathbf{r}_1$. Integration of the Smoluchowski equation (133) thus gives, with $\mathbf{r}_1$ replaced by $\mathbf{r}$,

$$\begin{aligned} \frac{1}{N}\frac{\partial}{\partial t}\rho(\mathbf{r},t) \;&=\; \nabla\cdot\int d\mathbf{r}_2\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\,\mathbf{D}_{11}\cdot[\nabla P + \beta P\nabla\Phi] \\ &+ (N-1)\nabla\cdot\int d\mathbf{r}_2\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\,\mathbf{D}_{12}\cdot[\nabla_2 P + \beta P\nabla_2\Phi] \;, \end{aligned} \tag{136}$$

where ∇_2 is the gradient operator with respect to $\mathbf{r}_2$. The problem is to reduce the integrals to expressions containing the density as the single variable. These integrals need to be evaluated to leading order with respect to concentration and gradients in concentration. A typical example of an integral that needs to be evaluated is discussed in the appendix. The final result is Fick's law,

$$\frac{\partial}{\partial t}\rho(\mathbf{r},t) \;=\; D\,\nabla^2\rho(\mathbf{r},t) \;, \tag{137}$$

where Fick's diffusion coefficient D is found to be equal to,

$$D \;=\; D_0\left[1 + \alpha\,\varphi + \mathcal{O}(\varphi^2)\right] \;, \tag{138}$$

with the first order in volume fraction coefficient being equal to,

$$\begin{aligned} \alpha \;=\; &-\;\beta\int_0^\infty dx\,x^3 g(ax)\frac{dV(ax)}{dx} \\ &+\;\int_0^\infty dx\,x^2\left\{A_s(ax) + 2B_s(ax)\right\}g(ax) \\ &-\;\int_0^\infty dx\,x^3\left[f_c(ax)g(ax) + A_c(ax)\frac{dg(ax)}{dx}\right] \;, \end{aligned} \tag{139}$$

where g is the pair-correlation function, and,

$$f_c(ax) \equiv x^2\frac{d}{dx}\left(\frac{A_c(ax)-B_c(ax)}{x^2}\right) + 4\,\frac{A_c(ax)-B_c(ax)}{x} + \frac{dB_c(ax)}{dx} \;. \tag{140}$$

In the evaluation of the integrals in eq.(136) (see the appendix), it turns out that "small gradients in concentration" means that concentration should vary, to a good approximation, in a linear fashion over distances equal to the range of the pair-interaction potential, which is of the order of a few nanometers to a few microns.

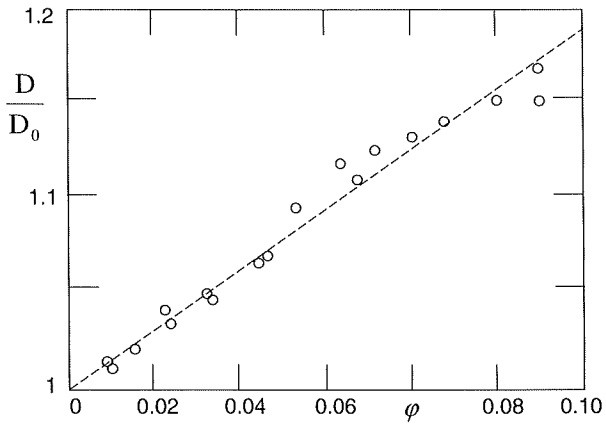


Fig. 13: *Comparison of the theoretical prediction in eqs.(138,142) with Fick's diffusion coefficient as obtained from light scattering experiments.*

The above integrals are easily evaluated for hard-sphere interactions, using eq.(85) for the pair-correlation function (where $R = |\mathbf{r} - \mathbf{r}_2|$), noting that,

$$\begin{aligned} \frac{dg(R)}{dR} &= \delta(R-2a)\,, \\ g(R)\,\frac{dV(r)}{dr} &= -\beta^{-1}\,\frac{dg(R)}{dR} = -\beta^{-1}\,\delta(R-2a)\,, \end{aligned} \tag{141}$$

where δ is the 1-dimensional delta distribution.
The first integral on the right hand-side of eq.(139) is easily calculated with the help of the above relations, and is found be equal to 8. This would be the value for α when hydrodynamic interactions are neglected, and is in agreement with eq.(86). The second integral is also easily evaluated using the approximate expressions (126) for the self-mobility functions. Its numerical value is $-111/64 = -1.734\cdots$. The function $f_c(ax)$ in eq.(140) is found from the eq.(126) to be equal to $-\frac{375}{4}x^{-8}$. The third integral thus turns out to be equal to $-5 + \frac{75}{256} = -4.707\cdots$. Hence,

$$\alpha = 1.559\,. \tag{142}$$

Using more accurate hydrodynamic interaction functions yields a value of 1.45 [11].
A comparison of this prediction with dynamic light scattering data on a hard-sphere silica dispersion is given in Fig.13. As can be seen from these data, the diffusion coefficient indeed increases with increasing concentration according to eqs.(138,142) up to volume fractions of about 10 %, but decreases at higher volume fractions.
It is evident that hydrodynamic interactions can be important for diffusion (they change the prefactor α from 8 to 1.45). Howver, when there are repulsive interactions (for example due to charges on the surfaces of the Brownian particles) with a range that is very long as compared to the core size, hydrodynamic interactions are of minor importance. Such systems may be "thermodynamically concentrated" but at the same time "hydrodynamically dilute".

5 Brownian Motion in External Fields : Sedimentation

Separation and subsequent characterization of components in a mixture (of proteins, amino acids or polymers, for example) can be achieved by applying an external field. Examples are sedimentation, where a gravitational field acts on each of the components, and electrophoresis, where an electric field is active. Due to the different force with which these fields act on the varies components in the mixture, they attain different average velocities, leading to separation. Brownian motion, however, counter acts separation. We shall analyze the separation process in case of sedimentation on the basis of the Smoluchowski equation as derived in the previous section.

Consider a small volume of a solution of a mixture of varies components that is carefully pipetted on top of pure solvent, and subsequently covered with pure solvent. After pipetting, the mixture spreads as a thin layer over the solvent below. The solvent is contained in a cylindrical vessel, with its long axis along the z-direction. The top of the container where the mixture is pipetted is located at $z = z_0$. The bottom of the container is chosen as $z = 0$ (see Fig.14). For all components in the mixture, the concentration at time $t = 0$, say, is then given by,

$$\rho(z, t = 0) \;=\; \rho_0\, \delta(z - z_0)\;, \tag{143}$$

where δ is the one-dimensional delta function, and ρ_0 is the number of particles of the corresponding species per unit area of the thin layer. The cylindrical container is then placed in a centrifuge, where the centrifugal force F acts differently on each of the species (depending on their mass corrected for buoyancy) in the $-z$ direction. The question is how the concentration profiles of each of components develop during time.

In analyzing the concentration profile, we shall assume that the solution is so dilute that interactions can be neglected. In that case the interaction potential in the Smoluchowski equation (133) can be neglected, while the microscopic diffusion matrices reduce to the diagonal form in eq.(106). Furthermore, $N = 1$ is this case, so that,

$$\frac{\partial}{\partial t}\varphi(z,t) \;=\; D_0 \frac{\partial}{\partial z}\left[\frac{\partial}{\partial z}\varphi(z,t) + \beta\,\varphi(z,t)\,F^{ext}\right] \;\;,\;\; z > 0\;. \tag{144}$$

where F^{ext} is the magnitude of the external force, which acts in the $-z$-direction, and $\varphi = \frac{4\pi}{3}a^3\rho(z,t)$ is the volume fraction of the corresponding species (a is the radius of a Brownian sphere).

Since the particles cannot cross the bottom of the container, the flux at $z = 0$ must be zero,

$$\frac{\partial}{\partial z}\varphi(z,t) + \beta\,\varphi(z,t)\,F^{ext} \;=\; 0\;. \tag{145}$$

The differential equation (144) thus has to be solved under the initial condition (143) and the boundary condition (145).

For a container of infinite extent, the sedimentation velocity of species j is equal to,

$$v_s^0 \;=\; F^{ext}/\gamma \;=\; D_0\,\beta\,F^{ext}\;. \tag{146}$$

Now define the function,

$$\varphi(z,t) \;\equiv\; u(z,t)\exp\left\{-\frac{v_s^0}{2D_0}(z - z_0) - \frac{(v_s^0)^2}{4D_0}t\right\}\;. \tag{147}$$

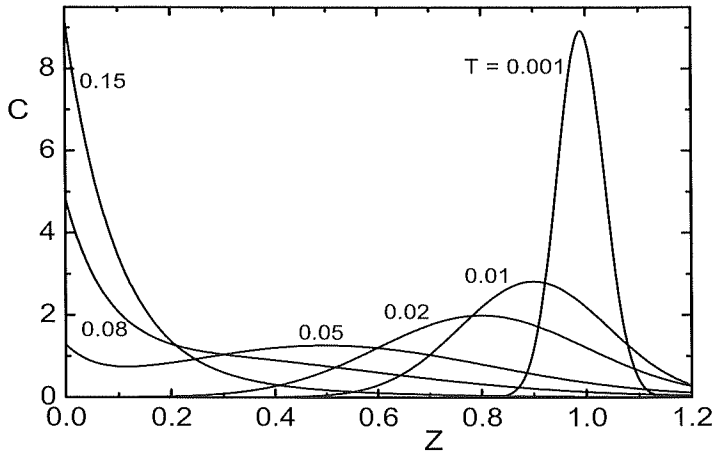


Fig. 14: *Density profiles for non-interacting particles, initially concentrated in a thin layer at* $z = z_0$. *Plotted is* $C \equiv \varphi(z,t)\, z_0/C_0$ *versus* $Z \equiv z/z_0$ *for various values of* $T \equiv D_0 t/z_0^2$, *which are indicated in the figure. The value of* $v_s^0\, z_0/D_0 = z_0\, F^{ext}/k_B T$ *is chosen equal to* 10.

It is easily shown, by substitution into eq.(144) that the new function $u(z,t)$ satisfies the Smoluchowski equation of a free particle without an external field,

$$\frac{\partial}{\partial t} u(z,t) \;=\; D_0 \frac{\partial^2}{\partial z^2} u(z,t)\;,$$

with the initial and boundary condition,

$$\begin{aligned} u(z, t=0) \;&=\; C_0\, \delta(z - z_0)\;, \\ D_0 \frac{\partial}{\partial z} u(z,t) + \tfrac{1}{2} v_s^0\, u(z,t) \;&=\; 0 \quad , \quad \text{for } z = 0\;, \end{aligned}$$

where $C_0 = \frac{4\pi}{3} a^3 \rho_0$ of the corresponding species 1 or 2. The solution of the Smoluchowski equation of a free particle without an external field, P_0, subject to the initial condition $P_0(z, t = 0) = C_0 \delta(z - z_0)$ is (see eq.(28), where the power $3/2$ in the exponential prefactor in one dimension is equal to $1/2$),

$$P_0(z,t) \;=\; \frac{C_0}{(4\pi D_0 t)^{1/2}} \exp\left\{ -\frac{(z - z_0)^2}{4 D_0 t} \right\}\;.$$

By partial integration it is easily verified that $u(z,t)$ can be expressed in terms of P_0 as follows,

$$\begin{aligned} u(z,t) \;=\; & P_0(z - z_0, t) + P_0(z + z_0, t) \\ & + \frac{v_s^0}{D_0} \int_{z_0}^{\infty} dz'\, P_0(z + z', t) \exp\left\{ \frac{v_s^0}{2 D_0} (z' - z_0) \right\}\;, \end{aligned}$$

from which it follows, using eq.(147), that,

$$\varphi(z,t) = \frac{C_0}{(4\pi D_0 t)^{1/2}}\left[\exp\left\{-\frac{(z-z_0)^2}{4D_0 t}\right\}+\exp\left\{-\frac{(z+z_0)^2}{4D_0 t}\right\}\right] \qquad (148)$$
$$\times\exp\left\{-\frac{v_s^0}{2D_0}(z-z_0)-\frac{(v_s^0)^2}{4D_0}t\right\}+\frac{C_0\,v_s^0}{\pi^{1/2}D_0}\exp\left\{-\frac{v_s^0}{D_0}z\right\}\int_{\frac{z+z_0-v_s^0 t}{\sqrt{4D_0 t}}}^{\infty} dx\,\exp\{-x^2\}\,.$$

For $D_0 t/z_0^2$ and $v_s^0 t/D_0 \ll 1$, this density profile is equal to the pdf of the free particle (in one dimension) in an external field, while for $t\rightarrow\infty$ this solution reproduces the barometric height distribution. The transition from the approximately Gaussian peak with a downward velocity v_s^0 to the barometric height distribution is clearly seen in Fig.14, where the above solution is plotted for various times.
Since the external force on the varies species in the mixture is different, the approximately Gaussian concentration profiles for intermediate times move downward with different velocities, leading to separation. The (maximum) effectiveness of the separation process is set by the diffusion coefficients of the varies species, since these determine how fast the concentration profiles broaden during sedimentation, and hence to what extent concentration profiles overlap.

A Appendix : Evaluation of integrals in the calculation of Fick's diffusion coefficient.

In this appendix we shall discuss the derivation of eqs.(138,139) from eq.(136). Let us consider, as an example, one of the integrals that need to be evaluated in eq.(136). The analysis of the remaining integrals proceeds along similar lines. Substitution of the expressions (135,125) into the second term in the first integral on the right hand-side of the integrated Smoluchowski equation gives rise to the following integral,

$$I(\mathbf{r}) = \beta\nabla\cdot\int d\mathbf{r}_2\cdots\int d\mathbf{r}_N\sum_{l=2}^{N}\Delta\mathbf{D}_s(|\,\mathbf{r}-\mathbf{r}_l\,|)$$
$$\times\,P(\mathbf{r},\mathbf{r}_2,\cdots,\mathbf{r}_N,t)\cdot\sum_{j=2}^{N}\nabla V(|\,\mathbf{r}-\mathbf{r}_j\,|)\,,$$

where,

$$\Delta\mathbf{D}_s(|\mathbf{r}-\mathbf{r}_l|) = D_0\sum_{j=1\,,\,j\neq i}^{N}\left\{A_s(|\mathbf{r}-\mathbf{r}_l|)\,\hat{\mathbf{r}}_l\hat{\mathbf{r}}_l+B_s(|\mathbf{r}-\mathbf{r}_l|)\left[\hat{\mathbf{I}}-\hat{\mathbf{r}}_l\hat{\mathbf{r}}_l\right]\right\}\,.$$

with $\hat{\mathbf{r}}_l=(\mathbf{r}-\mathbf{r}_l)/\,|\,\mathbf{r}-\mathbf{r}_l\,|$. There are $N-1$ terms with $l=j$, which all yield the same result for identical Brownian particles. We can therefore choose $l=j=2$, and the integration with respect to $\mathbf{r}_3,\,\mathbf{r}_4,\cdots,\,\mathbf{r}_N$ is only over the pdf, resulting in the pdf P_2, which is defined as,

$$P_2(\mathbf{r},\mathbf{r}_2,t) = \int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\,P(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,\cdots,\mathbf{r}_N,t)\,.$$

There are $(N-1)(N-2)$ terms with $l\neq j$, which again yield identical results upon integration. We choose $l=2$ and $j=3$. The integration with respect to $\mathbf{r}_4,\,\mathbf{r}_5,\cdots,\,\mathbf{r}_N$ is only over the pdf,

resulting in the pdf P_3, which is defined as,

$$P_3(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,t) \;=\; \int d\mathbf{r}_4 \cdots \int d\mathbf{r}_N \, P(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,\mathbf{r}_4,\cdots,\mathbf{r}_N,t) \,.$$

The above integral thus reduces to,

$$\begin{aligned} I(\mathbf{r}) \;&=\; (N-1)\beta\nabla\cdot\int d\mathbf{r}_2\,\Delta\mathbf{D}_s(\mathbf{r}-\mathbf{r}_2)P_2(\mathbf{r},\mathbf{r}_2,t)\cdot\nabla V(|\,\mathbf{r}-\mathbf{r}_2\,|) \\ &+(N-1)(N-2)\beta\nabla\cdot\int d\mathbf{r}_2\int d\mathbf{r}_3\Delta\mathbf{D}_s(\mathbf{r}-\mathbf{r}_2)P_3(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,t)\cdot\nabla V(|\,\mathbf{r}-\mathbf{r}_3\,|)\,. \end{aligned}$$

The pair- and three-particle correlation function g and g_3 for an inhomogeneous suspension are now defined as,

$$\begin{aligned} P_2(\mathbf{r},\mathbf{r}_2,t) \;&=\; \frac{1}{N^2}\rho(\mathbf{r},t)\rho(\mathbf{r}_2,t)g(\mathbf{r},\mathbf{r}_2,t)\,, \\ P_3(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,t) \;&=\; \frac{1}{N^3}\rho(\mathbf{r},t)\rho(\mathbf{r}_2,t)\rho(\mathbf{r}_3,t)g_3(\mathbf{r},\mathbf{r}_2,\mathbf{r}_3,t)\,. \end{aligned}$$

These correlation functions account for the interactions between the Brownian particles, and are simply equal to 1 for non-interacting particles. Disregarding dynamical correlations, the pair-correlation function on the pair level considered here is equal to the Boltzmann exponential of the pair-interaction potential,

$$g(\mathbf{r},\mathbf{r}_2,t) \;=\; \exp\{-\beta V(|\,\mathbf{r}-\mathbf{r}_2\,|\}\,.$$

As the next step towards the derivation of the equation of motion for the density, the inhomogeneities in density are assumed to be small. Thus, we rewrite the density as,

$$\rho(\mathbf{r},t) \;=\; \bar{\rho}+\Delta\rho(\mathbf{r},t)\,,$$

with $\Delta\rho$ a small deviation from the mean density $\bar{\rho}=N/V$. The integrals may be linearized with respect to this small deviation. Furthermore, on the pair level, the integrals may also linearized with respect to the mean density. The integral involving the three-particle pdf turns out to be of higher than linear order in either $\bar{\rho}$ or $\Delta\rho$, and is therefore omitted. After linearization the integral reduces to,

$$I(\mathbf{r}) = \frac{N-1}{N^2}\bar{\rho}\,\beta\,\nabla\cdot\int d\mathbf{r}_2\,\Delta\mathbf{D}_s(\mathbf{r}-\mathbf{r}_2)\,\Delta\rho(\mathbf{r}_2)\,g(|\,\mathbf{r}-\mathbf{r}_2\,|)\cdot\nabla V(|\,\mathbf{r}-\mathbf{r}_2\,|)\,.$$

One integral is omitted here : this is an integral over an odd function in $\mathbf{r}-\mathbf{r}_2$, which integral is zero (note that $\Delta\mathbf{D}_s$, g and V are all even functions, so that their spatial derivatives are odd functions).
The density profiles considered here are smooth on the length scale of the range of interactions between the Brownian particles. Since $\Delta\rho$ in the above integral is multiplied by ∇V, the density may be expanded to leading order in a Taylor expansion as follows,

$$\Delta\rho(\mathbf{r}_2,t) \;=\; \Delta\rho(\mathbf{r},t)+(\mathbf{r}_2-\mathbf{r})\cdot\nabla\Delta\rho(\mathbf{r},t)\,.$$

Upon substitution into the integral, only the second term survives, since the first term gives rise to an integral over an odd function, which is zero. Hence,

$$I = \frac{N-1}{N^2}\bar{\rho}\beta\nabla\nabla\Delta\rho(\mathbf{r},t) : \int d\mathbf{r}_2\,(\mathbf{r}-\mathbf{r}_2)\Delta\mathbf{D}_s(\mathbf{r}-\mathbf{r}_2) \times g(|\,\mathbf{r}-\mathbf{r}_2\,|)\cdot\nabla V(|\,\mathbf{r}-\mathbf{r}_2\,|)\,.$$

The next step is to transform to the new integration variable $\mathbf{r}' = \mathbf{r} - \mathbf{r}_2$. Using that $\nabla' V(r') = \hat{\mathbf{r}}' dV(r')/dr'$, integrating over the directions of $\mathbf{r}'$ and using that $\int d\hat{\mathbf{r}}'\hat{\mathbf{r}}'\hat{\mathbf{r}}' = \frac{4\pi}{3}\hat{\mathbf{I}}$, finally gives,

$$I = \frac{N-1}{N^2}\,\frac{4\pi}{3}\,\bar{\rho}\,\beta\,\nabla^2\Delta\rho(\mathbf{r},t)\int_0^\infty dr'\,r'^3 A_s(r')\,g(r')\frac{dV(r')}{dr'}\,.$$

The remaining integrals in eq.(136) are treated similarly.

References

[1] J. K. G. Dhont, G. Gompper, D. Richter (Eds.), *Soft Matter : Complex Materials on Mesoscopic Scale* (Lecture manuscripts of the 33th IFF winter school, Schriften des Forschungszentrums Jülich, Volume 10 (ISSN 1433-5506, ISBN 3-89336-297-5, 2002).

[2] W. B. Russel, D. A. Saville, W. R. Schowalter, *Colloidal Dispersions* (Cambridge University Press, Cambridge, 1991).

[3] J. K. G. Dhont, *An Introduction to Dynamics of Colloids* (Elsevier, Amsterdam, 1996).

[4] M. Doi, S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, Oxford, 1986).

[5] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1992).

[6] A. van Veluwen, Dissertation *Dynamic Light Scattering Studies of Diffusive Motion of Interacting Colloidal Particles* (Utrecht University, 1987).

[7] Ll. G. Chambers, *A Course in Vector Analysis* (Chapman and Hall, London, 1969).

[8] G. Arfken, *Mathematical Methods for Physicists* (Academic Press, New York, 1970).

[9] H. C. Berg, *Random Walks in Biology* (Princeton University Press, New Jersey, 1983).

[10] D. A. McQuarrie, *Statistical Mechanics* (Harper and Row Publishers, New York, 1976).

[11] G. K. Batchelor, J. Fluid Mech. **74**, 1 (1976).

[12] J. Vogel, B. U. Felderhof, J. Chem. Phys. **97**, 6744 (1992).

A 3 Self-Assembly

W. F. C. Sager
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Self-Assembly in Biological Systems 3
2.1 Cellular membranes 3
2.2 Mixed micelle mediated digestion of fats 11

3 Amphiphile Self-Assembly: Driving Force and Shape 11
3.1 Surface activity and formation of micellar aggregates 12
3.2 Shape of micellar aggregates 16
3.3 Micelle and bilayer dynamics 17

4 Thermodynamics of self-assembly 18
4.1 Micelle formation 19
4.2 Micellar growth 21

5 Amphiphilic Phase Behavior 22
5.1 Curvature energy model 22
5.2 Amphiphilic phase behavior 24

1 Introduction

The self-assembly of amphiphiles, small organic molecules that consist of a covalently linked hydrophilic and hydrophobic part, plays an important role in (cell) biology and is ubiquitous in everyday life. The hydrophobic or non-polar part (tail) of these molecules is mostly build-up by one or two hydrocarbon chains, while the hydrophilic or polar part (headgroup) can differ quite significantly in its chemical structure. Depending on the nature of the headgroup, amphiphiles can either be negatively (*anionic* headgroup) or positively charged (*cationic* head group), display no net charge (*zwitterionic* or amphoteric headgroup) or bear no charges (*nonionic* headgroup). Notwithstanding the bewildering complexity of chemical structures of biological amphiphiles, amphiphile self-assembly shows striking similarities for simple amphiphiles such as soap molecules on the one hand and complex membrane lipids on the other hand.

Based on their dual character, amphiphilic molecules adsorb at the water–air interface and can spontaneously self-associate or self-assemble in aqueous solution above the critical micelle concentration (cmc) into a variety of different structures of well defined geometry, whereby the driving force lies in the elimination of unfavorable contacts between water molecules and the hydrophobic tails. The simplest structure is the spherical micelle, in which the hydrophobic core is shielded by the headgroups from the aqueous surroundings. With increasing concentration and/or temperature a rich phase behavior is generally observed. The structural polymorphism ranges from spherical and cylindrical micelles to bilayer phases (e.g., lamellar sheets and vesicles) and more complex lyotropic crystalline phases (e.g., normal and inverse hexagonal and cubic phases). In biology, the bilayer structures play a dominant role, since they form the biological membranes (lipid bilayer), the most common cellular structures in both animals and plants. Amphiphile self-assemblies, which are sometimes referred to as association colloids or complex fluids, are soft and flexible, e.g. fluid-like, because the forces that hold the amphiphilic molecules together are not due to strong covalent or ionic bonds but arise from weaker van der Waals, hydrophobic, hydrogen-bonding and screened electrostatic interactions. The morphology of the amphiphile aggregates depends on the nature of the amphiphile, i.e. the size of the headgroup and the hydrophobic tail, but also on factors that influence the conformation of the hydrocarbon chain, the headgroup solvation and the interaction between neighboring head groups. Thus, changing the amphiphile or salt concentration, pH or temperature affects not only the interactions between the aggregates but also the intermolecular forces within each aggregate, thereby modifying the self-assembled structures themselves.

Over the last two to three decades research on amphiphile self-assembled systems has increased dramatically, both fundamentally and with respect to more advanced applications. The latter includes for example the employment of surface modified (multilamellar) vesicles in targeted drug delivery and the direct application of liquid crystalline phases as template for inorganic nanostructured materials such as mesoporous silica in catalysis. Since this contribution can only be seen as an introduction into the basic principles of self-assembly, the reader is referred for further information to more elaborated text books on this subject, such as [1],[2],[3],[4],[5].

In this Chapter we introduce the basics of amphiphile self-assembly with respect to its relevance in (cell) biology. *Section* 2, deals with the self-assembly of lipids in biological membranes and the role of mixed lipid/bile salt micelles in the digestion of fats in the intestine. Emphasis is thereby laid on demonstrating the flexibility introduced by the lipid self-assemblies to respond to changes in the biological system and to allow for structural adjustments

necessary for the fulfillment of biological functions. Biophysical aspects of (biological) membranes and membrane based processes are discussed in later chapters of this book (see D1-9). In *Section* 3, we study the thermodynamic driving force for the self-assembly of amphiphilic molecules in aqueous solution and discuss amphiphile self-assembly at the water-air surface and the formation and characteristics of micellar aggregates in detail. The thermodynamics of the formation and growth of micelles are discussed in *Section* 4, while phase diagrams and bilayer phases are introduced in *Section* 5. We also introduce the main concepts developed to describe the size and shape of micellar aggregates and the structural evolution of phases.

2 Self-Assembly in Biological Systems

In this section we discuss lipid self-assembly in biological systems. The most dominant role lipids play in biology lies in their ability to form liquid-like bilayers that can host a large number of proteins and built up an enormous variety of different membranes with complex shape and high functionality. Within the framework of this lecture we can only briefly introduce membrane-lipid systems and highlight the tricks used by nature that can be understood in terms of concepts of self-assembly developed for simple amphiphiles. Besides bilayer phases we also look at a mixed micellar lipid/bile salt system that forms in the process of fat digestion in the intestine.

2.1 Cellular membranes

Biological membranes are rather complex and diversely organized two dimensional assemblies of proteins and lipids that permit the compartmentalization of living matter. They make up the most common cellular structures in both animals and plants where they are involved in almost all aspects of cellular activity ranging from simple mechanical functions such as motility, food entrapment and transport to highly specific biochemical processes (for general books on biomembranes see [6],[7],[8],[9]). The simplest example is the cytoplasmic membrane (plasma membrane) that separates the content of a single cell from its environment and represents the only membrane system in prokaryotic cells (e.g., bacteria and similar organisms). The plasma membrane, of which a cartoon is shown in Fig. 1A, consists of a lamellar lipid bilayer into which proteins can be incorporated in different ways.

Amphiphilic proteins are, if the geometry is right, integrated by their hydrophobic parts (hydrophobic cylindrical α-helix or globule with hydrophobic surface), while hydrophilic proteins are attached through hydrophobic anchors or by linkage to membrane bound proteins. This results in a well-defined orientation of proteins. The composition of proteins differs on both leaflets of the bilayer. The membrane proteins are responsible for functional activities of the cell membrane, such as binding of hormones, transfer of signals into the cell interior, transport of ions or nutrients and segregation of products of the cell like digestive enzymes. The lipid bilayer that is built up by a large variety of lipids, functions mainly as a (hydrophobic) barrier for small polar solutes (e.g., ions and sugar) and macromolecules (e.g., peptides, proteins and nucleic acids) and serves as unique two-dimensional (dynamic) solvent for the membrane proteins. Lipid bilayer thicknesses of 60-100 Å determined by electron microscopy and X-ray diffraction, indicate that their hydrocarbon tails are in a more or less extended state. The existence of the lipid bilayer was first proposed by Gorter and Grendel [11] in 1925 on the basis of their observation that lipids extracted from erythrocytes covered twice the area when spread as a monolayer at the air water–interface than in the erythrocyte plasma membrane (the erythrocytes only membrane).

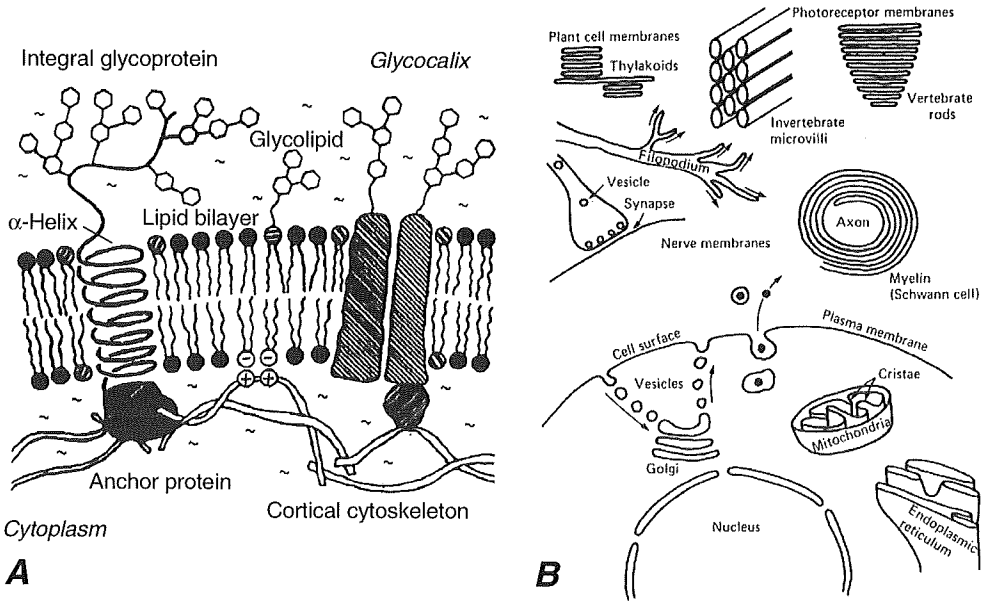


Fig. 1: *A: Schematic cartoon displaying a cross-sectional view of a plasma membrane of an animal cell showing integral proteins embedded within the 60-100 Å thick lipid bilayer, adapted from [10]. The proteins are integrated via their hydrophobic parts (e.g. α-helix) and can either span the whole membrane (transmembrane proteins with two hydrophilic groups on both sides of the bilayer) or be rooted in only one side. At the cytoplasm side (cell interior side) the plasma membrane is connected to the cytoskeleton, a polymeric actin network that forms a cortical structure close to the membrane with a thickness of 100-500 nm for enhanced mechanical strength. The outside of the plasma membrane is covered by a ~50 nm thick layer of oligosaccharides, the glycocalix, that i) prevents unspecific cell adhesion via steric stabilization and electrostatic repulsion, ii) protects the cell from attacks by exogenous enzymes and iii) provides target sides for specific cell adhesion. B: Diagram featuring different types of specialized (intracellular) membranes that offer little resistance to bending, taken from [4]. The bottom part displays a typical eukaryotic cell with different organelles showing the inner and outer membranes of the mitochondrion, and the cell nucleus. The endoplasmic reticulum represents a labyrinth of flattened membrane tubes. It is the location where all transmembrane molecules and nearly all secretory membranes are synthesized directly at the membrane. The synthesis of the lipids in the endoplasmic reticulum is vectorial, i.e. it occurs exclusively on one side of the membrane. Further transport of these biomolecules from the endoplasmic reticulum to the other organelles and the plasma membrane occurs via lipid vesicles and includes topological changes of the lipid membrane during budding off and fusion of the vesicles. Examples for nerve, photoreceptor and plant cell membranes are displayed in the upper part.*

The basic requirement for performing biological activity and for cell survival is that both the proteins and the lipids can laterally diffuse freely within the lipid membrane plane (fluid mosaic model of Singer and Nicholson) [12]. The dynamic structure allows for a rapid turnover of the membrane constituents and the formation of heterogeneous domains and local clustering of lipids and proteins that are vital for the normal functioning of the membrane and

its components. Incorporation of the –compared to membrane lipids– more rigid proteins into the liquid bilayer normally induces stresses on the lipids in their vicinity. These stresses arise if the hydrophobic regions of the protein and bilayer are of different length, for then the lipids become stretched or compressed in order to accommodate the protein. In this case the neighboring lipids adopt packings that deviate from the optimal value. Furthermore, specific electrostatic or hydrogen-bonding interactions between the hydrophilic head group and the exposed amino acid residues of the proteins can lead to additional lipid–protein interactions that often result in a preferential clustering of specific lipids around a protein [13].

Since the plasma membrane represents the only membrane system in prokaryotes (organisms which lack a cell nucleus), these organisms can be employed to study how a cell maintains and regulates the structural integrity of its membranes. If the growth conditions for bacteria such as temperature, salt content in the medium or the availability of nutrients (e.g., essential fatty acids) is varied, the cell responds to the changes by adjusting its lipid membrane composition. Lipids are then synthesized to maintain the bilayer state of the lipid layer and its fluidity. This illuminates the role of the chemical diversity of lipids and the dynamic character of their self-assembling.

In eukaryotic cells of higher organisms, the plasma membrane is only a very minor part of the –in some case quite extended– membraneous system. Eukaryotic cells of higher organisms contain a cell nucleus and several organelles and thus possess intracellular membranes that define compartments within the cell and allow to establish the necessary conditions (e.g. pH, ionic strength, digestive milieu) within the organelles (see Fig. 1B). Furthermore, each cell plays a specific role in the organism and higher organisms contain a large number of special-

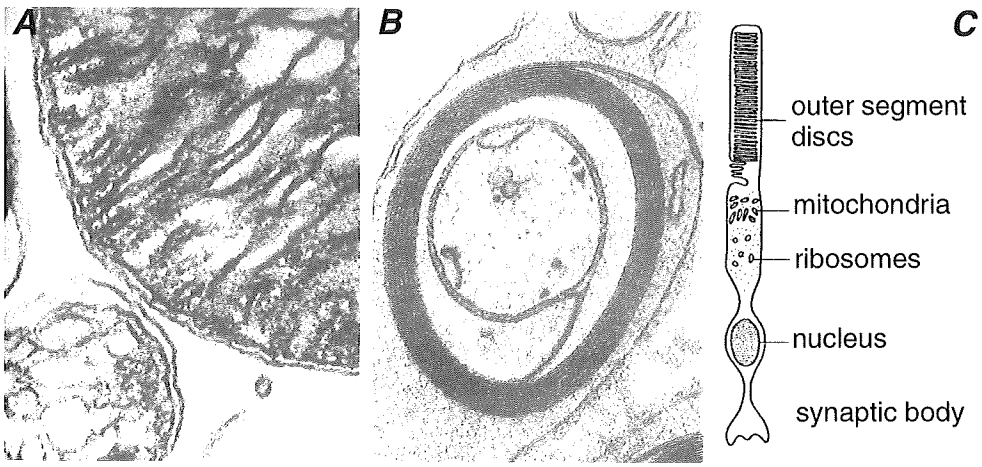


Fig. 2: *Examples of multilayered and highly folded membranes. A: Transmission electron microscopy (TEM) micrograph of a cross section through a mitochondrion, displaying a smooth outer membrane and a separate inner membrane with many folds into the interior of the organelle. B: TEM micrograph of a myelin sheath around a nerve axon. The innermost membrane is the axonal membrane, and the surrounding sheath is formed by extension of the cytoplasmatic membrane of a Schann cell, which wraps itself in several layers around the axon membrane. C: Schematic drawing of a rod photoreceptor cell. The outer segment discs are formed by inward folding of the cytoplasmatic membrane, followed by a disconnection from it. All pictures are taken from [1], for B and C see also Fig. 1B).*

ized membranes that exhibit a rich variation of membrane (key) functions. Membrane folding is often employed to increase the membrane area, and/or to obtain a structural organization relevant for specific purposes and thus allow for fundamental biochemical processes to take place. For example, electron transport and oxidative phosphorylation, processes that oxidize nutrients with the concomitant generation of ATP, are mediated by an organized battery of enzymes that are components of the highly folded inner mitochondrial membrane (see Fig. 1B and Fig. 2A). Likewise, photosynthesis, in which light energy powers the chemical combination of H_2O and CO_2 to form carbohydrates, occurs in the inner membranes of chloroplasts in the leaves of green plants. The thylakoids (see Fig. 1B) within the chloroplasts contain sections of stacked membranes that unfold upon addition of salt into a single giant vesicle, revealing that the folded state is made up by a single compartment and that the membrane folding is promoted by electrostatic attractive interbilayer forces. The two relevant photosystems are not located at the same position within the thylakoid membrane. Photosystem II prefers the folded regions and photosystem I the freely exposed membrane patches. Thus membrane folding results not only in a high area to volume ratio but can also be used to trigger lateral organization of the membrane protein systems. The processing of information, such as sensory stimuli or intercellular communications, is generally a membrane based phenomenon. Fig. 2B shows the myelin sheath that wraps in several layers around the axon of a nerve cell. The myelin membrane serves mainly as electrical isolator and therefore displays one of the lowest concentrations of proteins. The nerve impulses are mediated by nerve cell membranes and the presence of certain substances such as hormones and nutrients is detected by specific membrane-bound receptors. Fig. 2C shows isolated stacked discs of the outer segments of retinal rod cells that are formed by inward folding from the plasma membrane. A similar folding of the outer membrane occurs in retinal cone cells, but there the folds do not disconnect into discrete discs.

The protein-to-lipid ratio is for most membranes about 50%, but is of course correlated to the biological function of the membrane. Myelin sheath are with ~25% on the low end, while maximal values of ~300% are found in biochemically very active membranes such as the inner mitochondrical or bacterial membranes. For many cells the structure of the lipid bilayer is that of the fluid lamellar phase (see the myelin sheath in Fig. 2B). Besides the (flat) lamellae also two-dimensionally ordered cubic bilayer phases (see Section 5) occur based on the gyroid (G), double diamond (D) and primitive (P) periodic minimal surfaces. They have been evidenced to be present in the plasma membrane of the bacterium *Streptomyces hygroscopicus* and in highly convoluted intercellular membranes such as the smooth endoplasmic reticulum, the inner mitochondrical membrane and the nucleus of certain cells [10],[14]. Their main function is that of subcellular space organizers. In the cubic phases, the area-to-volume ratio is proportional to the inverse repeat distance in the network and thus independent of the system size facilitating a scale-up from a small to a large cell. Within the network of the cubic phase, solutes on the outside, on the inside, and in the membrane all have free transport through space.

Membrane lipids.

Membrane lipids (Greek: lipos, fat) are in water only sparingly soluble amphiphilic organic molecules of biological origin that display a large variety of chemical structures. They are carefully designed by nature to self-assemble into bilayer phases that can cope with the enormous number of biological functions of cellular and intracellular membranes. Biological membranes contain generally in the order of 100 different lipids and their lipid fractions vary between organisms, cells, different organelles within the cell, and even between the two sides of the bilayer. Myelin, for example, contains 1500 different lipid species of which 30 in a sizable amount.

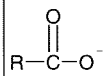
Most lipids are phospholipids or glycolipids with various headgroups and numbers of chains, chain lengths and degrees of unsaturation (see **Tab.1** for chemical structure and commonly used acronyms). Their chemical structure can be seen as variations on a common theme. The apolar group typically consists of two hydrocarbon chains originating from various saturated and unsaturated fatty acids. These chains are attached via ester, ether or amide covalent bonds to a linker group (e.g. glycerol and alkylacyl or alkenylacyl linkages), which also binds to the polar head group. The apolar parts originate from the fat metabolism. Most fatty acids thus have an even number of carbon atoms because they are normally biosynthesized by the concatenation of C_2 units. The hydrocarbon chain length typically ranges from C_{12} to C_{24}, with the degree of unsaturation increasing with the chain length. Over half of the fatty acid residues of plant and animal lipids are unsaturated (contain double bonds, usually in the *cis*-con figuration) and are often polyunsaturated (contain two or more double bonds that tend not to be conjugated). Bacterial fatty acids are rarely polyunsaturated but are commonly branched, hydroxylated or contain cyclopropane rings. Individual double-chained lipids contain quite often a saturated and an unsaturated chain. The chemical diversity of the polar groups is similar to that found for synthetic surfactants. We find ionic, zwitterionic and sugar head groups. Almost all ionic lipids carry a negative charge, a fact that has considerable significance for the electrostatic interactions. Typically 10-20% of the lipids of a membrane are charged. The nature of the head group depends on the organism. Glycolipids with mono- or disugar groups dominate in plants. In animals, phospholipids are most abundant, while in bacteria both types of lipids are common. Lipids with a single hydrocarbon chain such as the free fatty anions and lysophospholipids occur only in small quantities. Another important lipid is cholesterol that consists of a rigid, relatively planar steroid skeleton with a single OH as polar group. Cholesterol is not present in eukaryotic membranes, and thus played an important role in the development of intracellular membranes during evolution.

Besides the enormous complexity that biological membranes display, there are some important common features with respect to their self-assembly. Most lipid bilayer surfactants are double-tailed and posses rather low monomer solubilities (low cmc-values, see **Tab.1**). As we will see in the next sections, the shape of the amphiphile, i.e. the balance between its hydrophilic and hydrophobic moieties, plays a dominant role in determining which type of self-assembled aggregates form. For bilayer phases (eg., lamellar sheets as well as bicontinuous cubic phases), both moieties should be (nearly) balanced and the lipids thus contain strong apolar and polar groups. Low cmc-values in the range of 10^{-5} to 10^{-10} are necessary to assure that the membranes remain intact even when the bathing medium is grossly depleted of lipids. A vital point is the fluidity of the membranes. Fluid lipid chains are a prerequisite for i) deforming, bending and folding of the membrane, ii) to allow the mutual accommodation of different lipids as well as other molecules within the (curved) bilayer configuration, iii) to allow various solute molecules to pass through and proteins to diffuse along them. The (local) curvature can be regulated by altering the ratio of its constituent lipids. To adjust the curvature also lipids can be incorporated that do not form bilayers on their own, such as single-chained lysophospholipids on the one and cholesterol on the other hand.

Bilayer fluidity, order-disorder transition, and membrane asymmetry.
Lipid bilayers display two distinct diffusion modes. The lipids are highly mobile in the in-plane of the bilayer (lateral diffusion). X-ray diffraction patterns of bilayers at physiological temperatures display a diffuse peak, centered at a spacing of 4.6 Å, whose width is a measure of the distribution of lateral spacings between the hydrocarbon chains in the bilayer plane. This band resembles that obtained for liquid paraffins, indicating that the hydrocarbon chains undergo rapid fluxional motions involving rotations about their C–C bonds. By diffusion in

Tab.1: *Structure and properties of common biological amphiphiles with zwitterionic [a], anionic [b] or neutral head group: Fatty acids and phospholipids. The ionic states of the headgroup are given at pH 7. PC and PE are the two main lipids found in animal membranes.*

FATTY ACID



Saturated fatty acids[b]

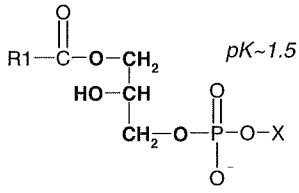
C_{12} Lauric *(44)* C_{18} Stearic *(70)*
C_{14} Myristic *(54)* C_{20} Arachidic *(77)*
C_{16} Palmitic *(63)* C_{22} Behenic

Unsaturated fatty acids[b]

9-octadecenoic *(13)* Oleic
9,12-octadecadienoic *(-5)* Linoleic
6,9,12-octadecatrienoic *(-11)* γ-Linolenic
5,8,11,14 eicosatetraenoic *(-50)* Arachidonic

numbers in brackets give the melting point of the acid in ºC

LYSO-PHOSPHOLIPID

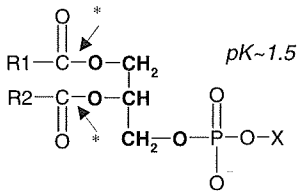


X

$-CH_2CH_2N(CH_3)_3^+$ Lysophosphatidyl*choline*[a] (lysolecithin)

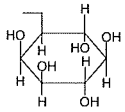
	cmc (M)	*Increment of cmc per CH_2 group*	*Average energy per CH_2 group*
C_{10}-PC	7×10^{-3}	3.2	2.8 kJ mol^{-1}
C_{12}-PC	7×10^{-4}	3.2	
C_{14}-PC	7×10^{-5}	3.2	
C_{16}-PC	7×10^{-6}		

GLYCEROPHOSPHOLIPID



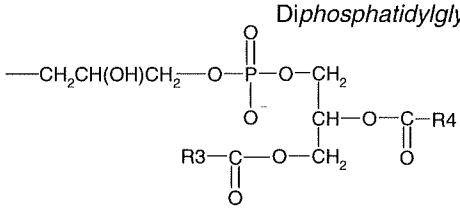
X

$-CH_2CH_2NH_3^+$ Phosphatidyl*ethanolamine*[a] PE *pK~11*
$-CH_2CH_2N(CH_3)_3^+$ Phospatidyl*choline*[a] PC (lecithin)
$-CH_2CH(NH_3^+)COO^-$ Phosphatidyl*serine*[b] PS
Phosphatidyl*inositol*[b] PI



$-CH_2CH(OH)CH_2OH$ Phosphatidyl*glycerol*[b] PG

Di*phosphatidylglycerol*[b] DPG (cardiolipin)



(in inner mitochondrical membranes) 4 alkyl chains & 2 negative charges

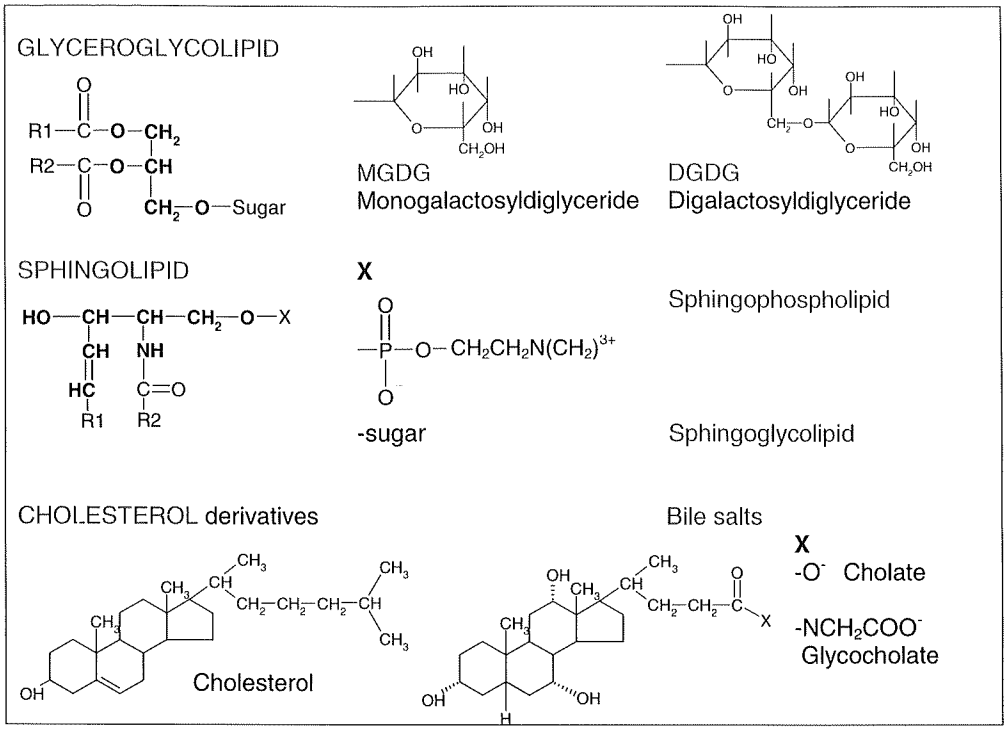
	cmc (M)	*Increment of cmc per CH_2 group*	*Average energy per CH_2 group*
C_6C_6-PC	1.5×10^{-3}	2.7	2.1 kJ mol^{-1}
C_8C_8-PC	3×10^{-4}	2.7	
$C_{10}C_{10}$-PC	5×10^{-5}	2.7	
$C_{16}C_{16}$-PC	5×10^{-6}		

*bonds that can be hydrolysed by phospholipases to form lysophospholipids.

Normally R1 corresponds to the saturated (C_{16}-C_{18}) and R2 to the unsaturated hydrocarbon chains (1-3 cis) of fatty acids, but also phospholipids with two saturated or two unsaturated chains exist. About 50% of membrane lipids have an unsaturated chain, thereby increasing the fluidity and hydrophilicity of bilayers.

R denotes the hydrocarbon chain of fatty acids

Tab.2: *Structure and properties of common biological amphiphiles: Glycolipids, sphingolipids and cholesterol (derivatives). The uncharged galactolipids DGDG and MGDG are the major constituents of plant tylakoid membranes.*



the bilayer lipids pass the 1μm length of a bacterial cell in ~1 s. In contrast, the transverse diffusion or flip-flop rates of lipids across the bilayer are very low, since they involve the rotation of the polar head groups through the center of the oil bilayer. Transverse diffusion is characterized by half-times that are minimally several days, which allows to maintain the asymmetry of biological membranes imposed during their synthesis.

Below a critical transition temperature, the lipid bilayer undergoes order-disorder transition and looses its fluidity by forming a gel-like solid. The X-ray diffraction pattern exhibits a sharp peak at 4.2 Å similar to that obtained for crystalline paraffins, which indicates that the hydrocarbon chains become fully extended and packed into a hexagonal array. The characteristic feature of the gel phase is that the alkyl chains are crystallized while liquid like solvent still exists between the bilayers. Mismatches between the cross-sectional areas of chains and head groups lead to tilted or interdigitated chains and rippled surfaces (see Fig. 3). The transition temperature of a lipid bilayer increases with the chain length and the degree of saturation of its fatty acid components in the same way as the melting temperature increases (for values see Tab.1). The transition temperatures of most biological membranes are in the range of 10 to 40 °C. Cholesterol decreases the membrane fluidity because its ridged steroid ring system interferes with the motion of the fatty acid chains thus inhibiting crystallization of the fatty acid side chains. It also broadens the temperature range of the order–disorder transition and in high concentration totally abolishes it.

The fluidity of biological membranes is one of their important physical attributes since it permits their embedded proteins to interact. The transition temperatures of mammalian membranes are well below body temperatures. Bacteria and poikilothermic (cold-blooded animals) such as fish modify –through lipid biosynthesis and degradation– the fatty acid compositions of their membrane lipids with the ambient temperature so as to maintain membrane fluidity. For example, the membrane viscosity for E. *coli* at its growth temperature stays constant as the temperature is varied from 15 to 43 °C.

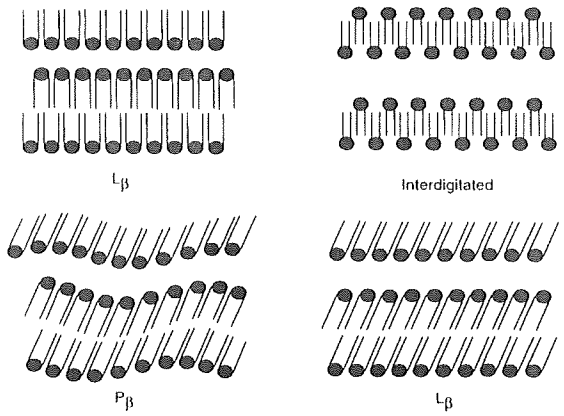


Fig. 3: *Crystalline alkyl chains have cross-sectional areas of ≈20 Å^2 per chain perpendicular to the direction of the plane. If the areas of the head group(s) and the chain(s) match, a L_β structure is adopted. In the more likely case of mismatch between them, a tilted L_β' or a rippled P_β' structure can appear. For single chain-amphiphiles, chains may even interdigitate resulting in a bilayer thickness that corresponds to the length of a single chain [3].*

Membrane curvature.

The curvature of the lipid bilayer can be regulated by altering the ratio of its constituent lipids. Local regions that are more curved to the cell interior need lipids with more bulky headgroups (cone-shaped lipids such as lysolecithin, phosphatidylcholine (PC) or glycolipids digalactosyldiglyceride (DGDG)) on the outer side of the bilayer and lipids with a more bulky tail region (wedged-shaped lipids such as cholesterol, phosphatidylethanolamine (PE) or monogalactosyl diglyceride (MGDG)) on the inner side. A vice versa situation is necessary for membrane patches that are more curved towards the outside of the cell. Only small amounts of single-chained lysolipids can be incorporated into the bilayers before they break up into, e.g., vesicles and micelles. A similar argument holds for the addition of PE and cholesterol. It is probably for this reason that natural biological membranes never contain high amounts of both cholesterol and phosphatidylethanolamine (the most common wedge-shaped lipids in animal membranes). As might be expected, cholesterol and lysolecithin –neither of which can form a stable bilayer by itself– can combine in certain proportions to form bilayers (for more information on physical properties of bilayers see [15],[16],[17]).

The common lipids in animal cells are the two phospholipids PC and PE, while in plant cells they are the glycolipids DGDG and monogalactosyl diglyceride MGDG. Depending on the ratio of these lipid types in a bilayer, they can pack together into planar bilayers or into bilayers of varying curvature and flexibility, a facility that is made use of in the lipid synthesis machinery of cells. Each of these four lipids has an uncharged head group. Their headgroup interactions are therefore entirely due to steric hydration forces that are fairly insensitive to changes in the ionic environment of the cytoplasm. The invariant packing properties of PC, PE, DGDG and MGDG make these the ideal structural building blocks for stable membrane organization.

2.2 Mixed micelle mediated digestion of fats

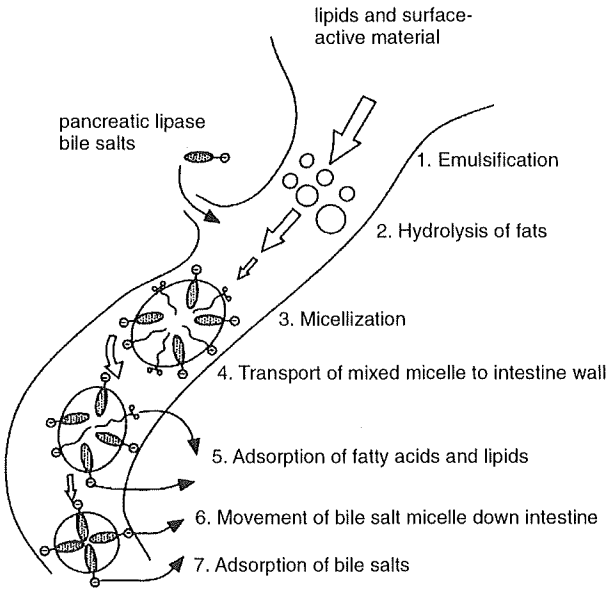


Fig. 4: *Schematic drawing showing the solubilization and digestion of fats in the stomach and upper intestine, taken from [2],[3].*

A good example for the biological relevance of micelles is the formation of mixed lipid/bile salt micelles during digestion of fats in the animal intestine sketched in Fig. 4. Bile salts that are derivatives of cholesterol (see Tab.2) are excreted by the liver into the common bile duct as mixed lecithin/bile salt micelles and act as weak amphiphiles in the stomach and the intestine. Ingested fats pass into the duodenum (the upper part of the small intestine) and are emulsified by mechanical agitation. Pancreatic enzymes attack the ester groups and brake the triglyceride fats into fatty acids and 2-monoglycerides, while bile salts are simultaneously released from the gall bladder. Bile salts are seldom absent, since without them the body cannot digest fats with chain length longer than C_8. At low pH fatty acids are insoluble in water (non-dissociated acid group) and combine with the bile salts and 2-monoglyceride lipids into mixed micelles. The mixed micelles are transported to the intestine walls, where the fatty acids are adsorbed into the bilayers of epithelial cells and diffuse across the intestinal wall, where they are catalytically re-esterified by enzymes to form triglycerides. The more polar bile salts are adsorbed by active transport in the ileum (lower part of the intestine), where they flow into the liver via the portal vein prior to being recycled.

3 Amphiphile Self-Assembly: Driving Force and Shape

In this section we introduce the basics of surfactant self-assembly mainly for simple amphiphiles but emphasize differences –if they occur– to lipids. One should keep in mind, that the self-assembly of amphiphiles is a physicochemical process by which the individual molecules are only physically associated. The size and shape of the self-assembled structures can thus change in response to small changes in concentration, salt, temperature, pH and pressure, leading to a quite complicated phase behavior. It is therefore necessary to begin by considering the factors that determine how and why certain molecules associate into various well-defined structures.

3.1 Surface activity and formation of micellar aggregates

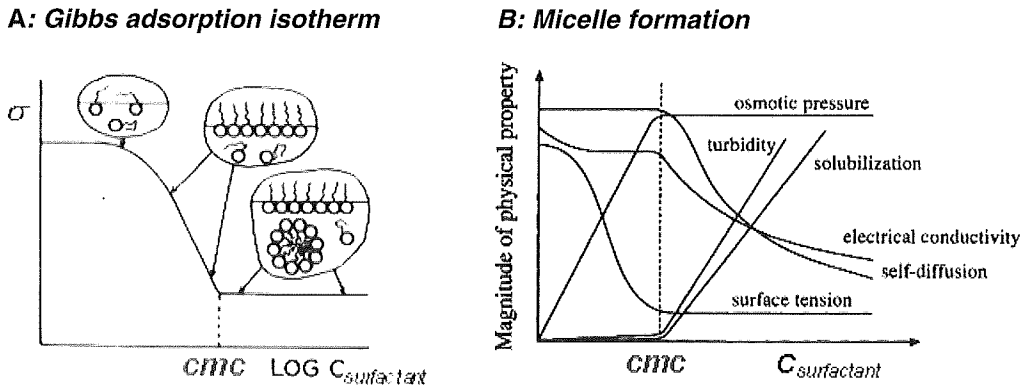


Fig. 5: *A: Typical adsorption isotherm (schematically) for an aqueous surfactant solution: Surface tension σ plotted against the logarithm of the surfactant concentration c_s* [18]. *The sketched insets show the state of the surfactant molecules in the water phase (bulk) and at the water/air surface. B: Effect of surfactant concentration on various physical properties [3]. Over a fairly narrow surfactant concentration range, all physicochemical properties that depend on the number of particles (e.g., osmotic pressure) or the particle size (e.g. turbidity) show a discontinuity, that can be used to determine the cmc.*

Surface activity and micelle formation.
Amphiphiles are surface active molecules. If surfactant is added to an aqueous solution, it will preferentially adsorb at the water/air surface. Its solubility in the aqueous phase determines the distribution coefficient between surfactant molecules in the aqueous phase and at the surface. The presence of surfactant molecules induces a surface pressure $\Pi = \sigma_0 - \sigma$, with σ the surface tension with surfactant present and σ_0 the surface tension of the bare surface without added surfactant. Furthermore, upon addition of surfactant, the surface tension decreases. In Fig. 5A a schematic plot (the *Gibbs adsorption isotherm)* of this decrease in surface tension is shown as a function of the bulk surfactant concentration c_s. For low concentrations, the decrease in surface tension is related to the adsorption Γ, which is the excess concentration of surfactants at the interface, through the Gibbs adsorption equation

$$\Gamma = -\frac{1}{kT}\left(\frac{\partial\sigma}{\partial \ln c_s}\right)_T, \qquad (1)$$

where k is Boltzmann´s constant and T the absolute temperature. The Gibbs adsorption equation supplies means to determine the average area occupied by a surfactant molecule a at the saturated macroscopic, flat interface $(a=1/\Gamma)$ [19]. At a certain concentration, the so-called **critical micelle concentration** (cmc), it becomes energetically more favourable for the surfactant molecules to aggregate into micelles. The self-assembly of amphiphiles eliminates unfavourable contacts between water and hydrophobic tails of the amphiphiles and yet permits the solvation of the polar head groups. Micelle formation is a cooperative process, since an assembly of just a few amphiphiles cannot shield its apolar tails from contact with water. Consequently, dilute aqueous solutions of amphiphiles do not form micelles until their concentration surpasses the critical micelle concentration. Above the cmc ($c_s > c_{cmc}$) the surface tension stays constant, since all additionally added surfactant molecules form micelles in the bulk water phase and the concentration of not aggregated surfactant molecules remains nearly

the same. For conventional surfactants the plateau values are for the surface tension typically in the range of 25-35 mNm^{-1}.

Experimentally, a number of physical properties that depend on the number (colligative properties) and/or size of particles present in solution exhibit a discontinuity near the critical micelle concentration (see Fig. 5B). In this way, micelle formation can be followed by measurements of these properties in dilute surfactant solutions as a function of surfactant concentration. A further possibility is to follow the uptake of hydrophobic dye molecules in the aqueous phase that become solubilized in the micellar core at surfactant concentrations above the cmc. The most widely used techniques, however, are surface tension measurement and ionic conductivity. The concept of micelle as an aggregate of surfactant molecules arose in the early part of the last century when Mc Bain in 1913 first sought to understand the anomalous concentration dependence of many physical properties of aqueous soap solutions [20].

Driving force for self-aggregation and the hydrophobic effect.
The thermodynamic driving force behind *micellization* has been the subject of intensive research for quite some time, but it is nowadays agreed upon that it is mainly of an entropic origin connected to the gain in entropy of the surrounding water molecules [1],[3]. If a hydrophobic molecule (e.g. a hydrogen molecule) is present in an aqueous environment the water molecules surrounding the hydrophobic molecule will become more ordered, resulting into a loss of entropy. Accordingly, formation of micellar aggregates, which reduce the contact between water molecules and hydrophobic chains, is entropically favored. The ordering of water molecules, responsible for the segregation of hydrophobic and hydrophilic regions in aqueous amphiphilic systems, supplies the thermodynamic explanation for the immiscibility of hydrocarbon and water. It is called **the hydrophobic effect** (or Tanford effect) and is responsible for a variety of important physico-chemical effects, such as self-aggregation, surface activity, folding of biomolecules to superstructures, etc.

In principle the hydrophobic effect is also the driving force for self-aggregation in other polar solvents, such as short chain alcohols, formamide, glycerol, and hydrazine. However, the solvophobic effect might be less pronounced. Self-aggregation has also been found in apolar solvents, such as hydrocarbons (oil). In this case small amounts of water are a necessary prerequisite for the formation of reversed micelles (in particular for ionic surfactants that need water for the hydration of the ionic group).

Factors affecting the cmc.
The value of the cmc depends on the identity of the amphiphile and the solution conditions. For amphiphiles with relatively small single tails, such as sodium dodecyl sulfate ($CH_3(CH_2)_{11}OSO_3^-Na^+$, SDS), the cmc is ~ 10^{-3} M. Those of biological lipids with two large hydrophobic tails are generally < 10^{-6} M (see **Tab.1**). It is of major relevance to know how the cmc is influenced by the molecular structure of the surfactant and formulation variables, such as temperature, ionic strength and addition of alcohols. In general one can expect that the monomeric solubility of an amphiphilic molecule is determined by two factors: i) the free energy of the transfer of the hydrophobic part of the molecule from the aqueous environment into the micellar interior and ii) the solvation energy of the hydrophilic head group.
Since the driving force for the aggregation lies in the entropy loss due to the ordering of water molecules next to the apolar tail, the cmc shows for all surfactants, ionic as well as non-ionic, a pronounced dependence on the surface area of the apolar part that is exposed to water. For linear hydrocarbons the cmc decreases exponentially with the number n of C-atoms. The gain in free energy for the transfer of one CH_2 group from the aqueous environment into the

micellar interior deduced from $\mathrm{d}\ln(cmc)/\mathrm{d}n$ lies in the range of 1.2-1.4 kT (see also **Tab.1**), and is thus only little affected by the nature of the polar head. Increasing the alkyl chain by one CH_2 group reduces the cmc by a factor of about 3. If double bonds or aromatic rings are present in the apolar chain, the cmc is slightly less reduced, indicating the more polar character induced by the π-electrons.

For ionic surfactants the value of the cmc is reduced by the addition of salt. Since the cmc is higher for stronger electrostatic interactions between the head groups, the cmc depends not only on the electrolyte concentration but also on the type of counterion. In general the solubility of the surfactant and therefore the cmc is higher when the counterion is less strongly bound to the ionic group of the amphiphile. For SDS the following sequence has been observed for the cmc:

$$Li^+ > Na^+ > K^+ > Cs^+ > N(CH_3)^{4+} > N(C_2H_5)^{4+} > Ca^{2+} > Mg^{2+}.$$

Not all amphiphiles exhibit a cmc and form micelles. Medium and long-chain alcohols are, e.g., surface active and lead to a substantial reduction of the surface tension. However, before they reach the break point in the Gibbs adsorption isotherm, they exceed their solubility in water. Instead of forming micelles, macroscopic phase separation occurs. These so called **cosurfactants** play an important role in the self-assembly of amphiphilic systems, since they form with surfactants mixed aggregates and can therefore strongly influence the properties of the amphiphilic system. Since they posses only a weak headgroup (OH group) they are often added to single-chain surfactants such as SDS to obtain bilayer phases. In this sense also cholesterol acts as cosurfactant in lipid bilayers. For these mixed surfactant systems, the situation is rather complex and the micelle formation depends on the interactions between the different types of surfactants.

Insoluble monolayers.
The hydrophobic effect is most impressively demonstrated by the observation of Benjamin Franklin in 1774, who dropped not more than a teaspoonful of "oil" onto the rough surface of a pond in London to investigate its calming effect on waves –an effect that was known since Roman times [21],[22]. The oil he used was probably similar to olive oil and contained thus lipid molecules. The "oil" spread spontaneously to cover an area of half an acre, which gives a final film thickness of about 10 Å and could have been the first reasonable measure of molecular dimensions, but Franklin did not perform the calculation. A film thickness of 17 Å was calculated by Rayleigh in 1890 from a similar experiment on a smaller scale [23].

Amphiphiles that display very low cmcs such as lipids form insoluble monolayers at the water–air interface. In this case the Langmuir balance can be used to measure the surface pressure Π, which represents the surface pressure needed to prevent the film from spreading. Plots of Π versus surface area per molecule yield two-dimensional isotherms, as depicted in Fig. 6. Insoluble monolayers are two-dimensional analogs of the bulk systems that display a number of gaslike, liquid, and solid phases. At very low surface pressures, all monolayers exhibit gaseous behavior (G). As Π approaches 0 the isotherm satisfies the two-dimensional analog of the ideal gas equation, whereby $\Pi a = k\,T$.

At the other end of the isotherm, we observe high Π, low a, and solid like behavior (S). In this case the hydrocarbon chains are oriented vertically and packed together. The film is relatively incompressible. For long-chain alcohols and esters, $a \approx 0.19\,\mathrm{nm}^2$, which corresponds to the closed packed cross section of solid hydrocarbons obtained by X-ray scattering. A monolayer transition from gas to solid is more complex than for a corresponding bulk phase. As the film

is more compressed two liquid phases, a liquid expanded (L_1) and a liquid condensed phase (L_2) can be observed. The transition from G to L_1 occurs like a gas-liquid transition with significant compression at constant pressure (first order transition). In the liquid expanded state the molecules have self-assembled into a state analogous to the one found in liquid crystalline bilayers.

A: *Surface pressure-area isotherm* **B: *Langmuir trough***

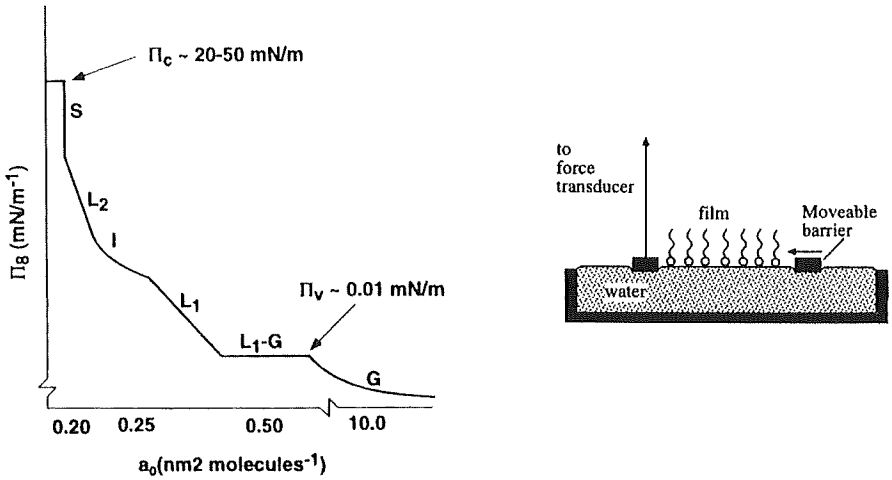


Fig. 6: *A: Two-dimensional pressure (Π_s)-area (a) isotherm typically obtained for insoluble monolayers, displaying a rich phase behavior that includes gaslike, liquid and solid phases. B: Schematic drawing of a Langmuir trough. The insoluble amphiphile is spread from a volatile solvent to ensure a uniform monolayer. The area of the monolayer is controlled through a movable barrier. The horizontal force necessary to maintain the float at a fixed position is measured using a torsion balance that gives the surface tension. The Langmuir balance can be used to measure directly the surface pressure as a function of the vapor–liquid surface area.*

They diffuse rapidly along the surface and there is a substantial conformational disorder in the alkyl chains. Further compression of the monolayers leads to the condensed liquid phase. Both the L_2 and solid regions show lower compressibility than the L_1 region due to strong intermolecular forces.

The actual shape of the surface pressure-area isotherm depends on the factors such as the size of the headgroup, the presence of polar groups, the number of hydrocarbon chains and their conformation (straight or bent) that determine the lateral interactions between molecules. The Π-a curve of, e.g., n-hexadecanoic acid shows a first order transition from the G to the L_1 phase, the next transition to the condensed phase when a is approximately equal to a closed packed fatty acid chain. Increasing the chain length of the fatty acid enhances the tendency for the formation of condensed phases. Indeed, n-docosanoic acid undergoes a direct transition from the from G to L_2 and several condensed phases (S) form at lower film areas that display different tilt and packing of the chains. The latter depend not only on surface pressure but also on temperature (as we have seen for lipid bilayers in Section 2.1) Condensed phases can

display a number of structures that are two-dimensional liquid crystal or crystal phases of hexagonal or rectangular packings. Although amphiphiles form lyotropic liquid crystal phases in the bulk, these two-dimensional phases are in fact analogues of thermotropic smectic liquid crystals. This is caused by the fact that it is the packing of the molecules that determines the symmetry of the crystal structure and not that of the molecular aggregates such as micelles and lamellae.

3.2 Shape of micellar aggregates

In the case of micelle formation the operating aggregation forces due to the hydrophobic effect are opposed by the repulsion between head groups. The result is that only finite (self-assembled) micellar aggregates are formed with a well-defined number of surfactant molecules N_p per micelle (N_p~50-80). The size and shape of the micellar aggregates depend on the geometry of the surfactants. This dependence can well be described by a geometric model that is based on the (critical) **packing parameter** v/al, whereby l and v correspond to the length and volume of the aliphatic tail and a corresponds to the effective head group area. The number of surfactant molecules N_p is determined by dividing the total volume of a given micellar aggregate by the volume of a single molecule or by dividing the total surface area by the head group area. For example, for spherical micelles this means:

$$N_p = \frac{\frac{4}{3}\pi R_{micelle}^3}{v} \quad , \quad N_p = \frac{4\pi R_{micelle}^2}{a} \quad , \quad \left(\frac{v}{al} \leq \frac{1}{3}\right) . \tag{2}$$

The condition that the packing parameter $v/al \leq 1/3$ for spherical micelles results from the assumption that the radius of the micelle $R_{micelle}$ cannot exceed l. While for a saturated hydrocarbon chain, v and l can easily be calculated from the van der Waals radii, a depends not only on the size of the head group but also on pH, surfactant and electrolyte concentration for ionic surfactants [3]. Head group areas are usually determined from measured values of the size of the micelle and the number of surfactant molecules per micelle. The head group area is typically in the range of 55-70 Å^2. For ionic surfactants, the effective head group area decreases with the addition of salt due to screening of the electrostatic interactions. For commonly used non-ionic surfactants that are composed of a hydrocarbon chain and a water-soluble ethyleneoxide chain (C_iE_j, whereby the subscripts denote the number of methyl C_i and oxyethylene E_j repeats), the head group area increases with chain length j and decreases with increasing temperature. For increasing temperature the hydrogen bonds between the ethyleneoxide oxygen atoms and the surrounding water molecules are broken resulting in a smaller effective head group area. Membrane lipids show a wide variety of headgroups that ranges from relatively simple groups such as carboxylic acids, amine and quaternary ammonium salts to very complex and bulky polysaccharides (see **Tab.1**).

The different possible packing geometries of surfactant molecules are listed in Tab.3 together with the corresponding range of the packing parameter. Generally, surfactants shaped in the way that the effective head group takes more space than the tail region (v/al<1/3) form spherical micelles. If v/al approaches 1 first cylindrical micelles and finally (almost) flat films form as they occur in vesicles (bilayers) and lamellar crystals. If v/al>1, the surfactants are not anymore soluble in water. They will form inverted crystalline structures such as inverted cubic or inverted hexagonal phases at high surfactant concentrations. In this case water is enclosed within the structures that are continuous in surfactant. Dissolved in an apolar solvent (oil) they will form inverse micelles. In this case the head groups form the interior of the reverse micelle and are shielded by the apolar moiety from the bulk oil.

Packing parameters

	*v/al*<1/3	spherical micelles
	*1/3*<*v/al*	cylindrical micelles
	*1/2*<*v/al*<1	vesicles, flexible bilayers
	*v/al*≈1	lamellae, planar bilayers
	*v/al*>1	inverse cylindrical and spherical micelles

Tab.3: *Surfactant parameter range for different surfactant aggregates* [2].

Turning from strongly curved micelles to flat surfactant sheets and even to oppositely curved aggregates (inverse micelles), can generally be achieved by either increasing the bulkiness (volume) of the hydrophobic moiety (e.g. double tailed instead of single tailed surfactants) and/or by decreasing the head group area. The latter can be realized by the addition of salt in the case of ionic surfactants, decrease in chain length j and increase in temperature in the case of nonionic surfactants, and/or addition of long chain alcohols (cosurfactants) in multi-component mixed systems. The two hydrocarbon tails of glyceroposphoipids and sphingolipids give these amphiphiles more or less rectangular shapes and allow them to form bilayers.

3.3 Micelle and bilayer dynamics

Amphiphile self-assemblies are highly dynamic systems. Surfactant molecules that are dissolved in the aqueous phase exchange rapidly with those in the self-assemblies. For micellar aggregates two time scales for dynamic processes have been found by fast relaxation kinetics measurements (e.g., temperature jump, pressure jump or ultrasonic adsorption). In these experiments, a system that is originally at equilibrium is subjected to a small but rapid change in an intensive parameter (e.g. T or p), and its relaxation to the new equilibrium state followed. The first process detected has been associated with the exchange of monomer between micelles and is characterized by a relaxation time τ_1 of 10^{-6} to 10^{-3} s. In this case, monomers leave some micelles and associate to other micelles, restoring an optimal size distribution. However, such processes leave the number of micelles unchanged. For the most probable aggregation numbers, N-1 and N, involved in this process, we have at equilibrium $k^+[S_1][S_{N-1}]=k^-[S_N]$. Since the monomer concentration equals the cmc, this reduces to $k^+c_{cmc}\simeq k^-$. k^+ has been found to be close to the diffusion limited rate. Because the micelle is essentially a liquid like drop, the monomer association involves no slow structural reorganization or bond formation. The sole barrier felt by a surfactant entering the micelle arises from long range electrostatic repulsions due to the micellar charge –if present. For surfactants with longer alkyl chain length, the cmc is lower (see Section 3.1) and thus the electrolyte concentration (monomer surfactant plus counter-ion) decreased. The charge of the micelle is therefore less screened and slows down the association for long-chain ionic surfactants compared to short-chain ones. In this case, a weakly decreasing k^+ and a strong decreasing cmc with increasing hydrocarbon chain length causes k^- to decrease even more strongly.

The slower process that takes place in micellar solutions, characterized be a relaxation time τ_2 of $10^{-3}$to 1 s, corresponds to the rate of micelle formation and disintegration and involves thus a change in the number of micelles. If a particular micelle starts to dissolve it has to pass

through a number of unfavorable aggregation states. As we will see in the next Section, it is thus the cooperative character of micelle formation that causes micelles to pass through a kinetic bottle neck as they disintegrate, therefore $\tau_2 \gg \tau_1$.
Bilayer phases are typically formed by surfactants with two hydrocarbon chains, whose cmcs are much lower. This results in a considerably slower monomer exchange. Formation and disintegration of bilayers involves rearranging of the entire bilayer structure. Bilayers respond therefore much more sluggish to changes in their environment. It can require hours, days, or even years for a bilayer (in e.g. a lamellar phase) to achieve an equilibrium state. Tab.4 summarizes the main differences between micellar and bilayer aggregates.

Tab.4: *Comparison between properties of micellar and bilayer aggregates, taken from [3].* The Kraft temperature is the temperature at which the solubility equals the cmc, which is often observed for ionic surfactant solutions. At lower temperature the surfactants form hydrated crystal phases. Above the Kraft temperature the solubility increases rapidly and a single homogeneous micellar phase forms over a large composition range.*

Property	*MIcelle*	*Bilayers*
Monomer solubility	~10^{-2} M	10^{-5}-10^{-10} M
τ for monomer exchange	10^{-5}-10^{-10} s	10^{-5}-10^{-10}
τ for aggregate life time	10^{-5}-10^{-10} s	days to years
Characteristic temperature	Krafft temperature*	Chain-melting temperature
Structural dimensionality	all directions equivalent	Lateral diffusion rapid, flip-flop slow
Aggregation pattern	Forms well defined aggregates at well-defined cmc	Basic structural unit appears in the a variety of global structures

4 Thermodynamics of self-assembly

The formation of micellar aggregates is not simply a first-order phase transition but thermodynamically a much more complicated process.

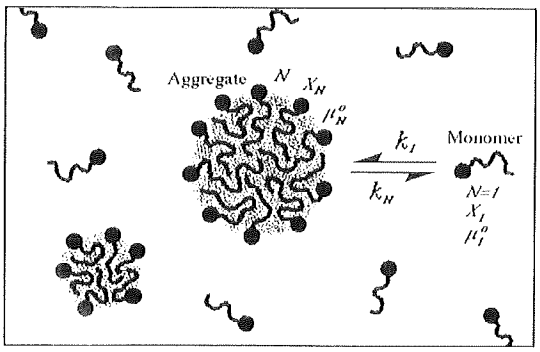


Fig. 7: *Association of N monomers into an aggregate (e.g. a micelle).*

The main purpose of this chapter is to present a comprehensive, thermodynamic framework for treating the different sizes and shapes of micellar aggregates in dilute aqueous surfactant solutions.

4.1 Micelle formation

The thermodynamics of the situation at hand is sketched in Fig. 7. Surfactant molecules are either present as un-aggregated monomers or are part of an aggregate consisting of N surfactant molecules (N=2,3,...). In equilibrium the chemical potential of a surfactant molecule in any of the aggregates should be the same, i.e.

$$\mu = \mu_1^0 + kT \log X_1 = \mu_2^0 + \frac{1}{2}kT \log \frac{1}{2} X_2 = \mu_3^0 + \frac{1}{3}kT \log \frac{1}{3} X_3 = \ldots$$

monomers dimers trimers

or

$$\mu_N = \mu_N^0 + \frac{1}{N}kT \log\left(\frac{X_N}{N}\right) = \text{constant}, \qquad N = 1,2,3,\ldots, \tag{3}$$

where μ_N is the chemical potential of a molecule in an aggregate with aggregation number N. The chemical potential μ_N^0 is the standard part of the chemical potential in an aggregate of aggregation number N and X_N is the total concentration (more strictly the mole fraction) of all surfactant molecules in aggregates of size N (the situation $N = 1$ corresponds to isolated molecules, or monomers, in solution). Eq. (**3**) assumes ideal mixing and is restricted to dilute systems where inter-aggregate interactions can be ignored. If we consider now as shown in *Fig.*6 the formation of an aggregate by the association of N monomers, using the equilibrium condition we now get that

$$\mu_N^0 + \frac{1}{N}kT \log\left(\frac{X_N}{N}\right) = \mu_1^0 + kT \log X_1, \tag{4}$$

and thus

$$X_N = N\left\{X_1 \exp\left(\frac{\mu_1^0 - \mu_N^0}{kT}\right)\right\}^N = NX_1^N \exp\left(-\frac{N\Delta E_N}{kT}\right), \tag{5}$$

where $\Delta E_N = \mu_N^0 - \mu_1^0$ is the free energy difference for a single surfactant molecule in an N-aggregate compared to being single. Eq.(**5**) together with the conservation law which states that the sum of all surfactants in all aggregates is given by the total surfactant concentration C

$$C = X_1 + X_2 + X_3 + \ldots = \sum_{N=1}^{\infty} X_N \tag{6}$$

completely determines the solution to these equations. It is good to realize, however, that all the relevant physics lies in ΔE_N and how it depends on N. Aggregates will only form when there is an energetic benefit in the cohesive energies between the molecules in the aggregate compared to the monomeric states. Suppose we assume that there is no such energetic benefit ($\Delta E_N = 0$) and the surfactant molecules in all different sized aggregates (including monomers) experience the same interaction with their surroundings. The result is that μ_N^0 is the same in all aggregates (with different N) and Eq.(**5**) then reduces to

$$X_N = NX_1^N \tag{7}$$

Since $X_1 << 1$, we must have $X_1 > X_2 > X_3 > X_4 > \ldots$ so that most of the molecules will be in the monomer state ($N = 1$) and the presence of larger aggregates becomes progressively less probable. This model therefore does not lead to the formation of micelles. The necessary condition for the formation of large stable aggregates is that $\mu_N^0 < \mu_1^0$ for at least some value of N.

In the **closed association model** it is assumed that only micellar aggregates of a given aggregation number m are in equilibrium with the monomers. In this case we model ΔE_N in the following way:

$$\Delta E_N = \begin{cases} 0 & \text{for } N = 1 \\ -\Delta G_{mic} & \text{for } N = m \\ \infty & \text{for } N \neq 1, m \end{cases}, \quad X_N = \begin{cases} X_1 & \text{for } N = 1 \\ X_m & \text{for } N = m \\ 0 & \text{for } N \neq 1, m\,. \end{cases} \tag{8}$$

Eq.(**5**) then becomes

$$X_m = mX_1^m \exp\left(-\frac{m\Delta G_{mic}}{kT}\right)$$

and (**9**)

$$X_m = m\left(\frac{X_1}{X_0}\right)^m \qquad \text{with} \qquad X_0 = \exp\left(-\frac{\Delta G_{mic}}{kT}\right)$$

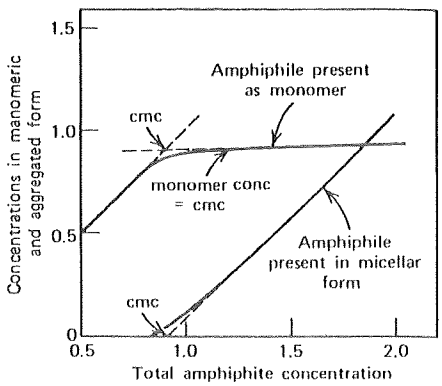


Fig. 8: *Monomer and aggregate concentration as a function of total concentration (schematic).*

The monomer and micelle concentrations are plotted schematically as a function of the total concentration C in Fig. **8**. At low surfactant concentrations when $X_1 < X_0$, we have that $(X_1/X_0)^m \ll 1$ and only monomers form ($X_m \approx 0$). Since $C = X_1 + X_m$, we have that X_1 increases linearly with C in the interval $0 < C < X_0$. When X_1 approaches X_0, we have $(X_1/X_0)^m \approx 1$ and $X_m \gg X_1$. Therefore, above X_0 all additionally added amphiphile goes into the micelles and X_1 stays roughly constant. As a result, we may interpret X_0 as the critical micelle concentration. Recalling the definition of X_0 we thus have that

$$\Delta G_{mic} = kT \ln(X_0) = kT \ln(cmc / 55.5 \text{ mol/l}) \tag{10}$$

Here we have made the usual assumption to express all concentrations in mole fraction units ((mol dm^{-3})/55.5 or M/55.5 for aqueous solutions). With this equation one can directly relate the measured cmc values to the free energy of micellization ΔG_{mic} and deduce the increment values for the transfer of CH_2 into the micelle (see **Tab.1**).

4.2 Micellar growth

We have now considered a simple model for micelle formation in which only aggregates of a specific size are energetically allowed. In reality there will of course be a distribution of sizes determined by the complete function μ_N^0. Furthermore, we have basically assumed, although not explicitly used, that the micelles are more or less spherical objects. Again, that may not always be experimentally appropriate and one might imagine the formation of cylindrical or disc-like aggregates. Let us now consider for instance the formation of cylindrically shaped micelles and consider the dependence on N of μ_N^0 for large aggregate size. In this case, one should make a distinction between surfactant molecules that are situated at the cylindrical part of the surface and those surfactant molecules in either of the two end-caps. Since the relevance of the chemical potential difference between the surfactant molecules at the end-caps and the interior of the cylinder, describes effects that are of order $1/N$, we might suppose the following form for the chemical potential μ_N^0

$$\mu_N^0 = \mu_\infty^0 + \frac{\alpha kT}{N}, \tag{11}$$

where μ_∞^0 and α are constants that we will assign thermodynamic meaning to below. One may extend this form for μ_N^0 also for the two other geometries of spherical and disc-like micellar aggregates. In general one may show that

$$\mu_N^0 = \mu_\infty^0 + \frac{\alpha kT}{N^p}, \tag{12}$$

Where $p=1$ corresponds to cylindrically shaped micelles, $p=1/2$ corresponds to discs, and $p=1/3$ to spherically shaped micelles. The value of the exponent p for spheres and discs can be calculated from a thermodynamic analysis in which the two leading contributions to the chemical potential are derived from a consideration of the bulk free energy per molecule with end-effects. For the spherical micelles the bulk contribution to the free energy is simply a pressure times volume term with the leading correction determined by a surface tension times surface area term:

$$\begin{aligned} &F = PV_S + \sigma A_S, \quad \text{with } N = \frac{4\pi R^3}{3v}, V_S = \frac{4}{3}\pi R^3, \text{ and } A_S = \pi R^2 \\ &\frac{F}{N} = Pv + \frac{\sigma v}{R} = \frac{\sigma}{N^{1/3}}\left(\frac{4}{3}\pi v^2\right)^{1/3} \\ &\mu_N^0 = \mu_\infty^0 + \frac{\alpha kT}{N^{1/3}} \quad \Rightarrow \quad \alpha = \frac{\sigma}{kT}\left(\frac{4}{3}\pi v^2\right)^{1/3} \end{aligned} \tag{13}$$

Similarly for the disc-like micelles, the main contribution to the free energy is a surface tension times surface area term with the leading correction given by the line tension times the circumference of the disc:

$$F = \sigma A_D + \tau L_D, \quad \text{with} \quad N = \frac{\pi R^2}{a}, A_D = \pi R^2, \text{ and } L_D = 2\pi R$$
$$\frac{F}{N} = \sigma a + \frac{2\tau a}{R} = \sigma a + \frac{2\tau(\pi a)^{1/2}}{N^{1/2}} \tag{14}$$
$$\mu_N^0 = \mu_\infty^0 + \frac{\alpha kT}{N^{1/2}} \quad \Rightarrow \quad \alpha = \frac{2\tau(\pi a)^{1/2}}{kT}$$

Apart from being able to determine the exponent p, one gets explicit expressions for the constants μ_∞^0 and α in this way.

With the general expression for μ_N^0 in Eq.(**12**), one may now determine the whole distribution X_N. Now, it turns out that such a description is very useful to describe cylinders and especially the growth of cylinders as it is determined by the end-cap energy α, however for discs and spheres it turns out that the average size of the aggregates diverges above the critical micelle concentration. This means that for discs and spheres there is a phase transition to a separate phase, strictly to an aggregate of infinite size. For discs this would correspond to a transition to a lamellar phase. The consequence is that one very rarely encounters self-assembled discs and spheres of a certain size, whereas long cylinders of a certain length are much more common.

5 Amphiphilic Phase Behavior

5.1 Curvature energy model

The global features of the phase behavior of self-assembling systems, including the formation of crystalline phases, have been attributed to changes in the interfacial surfactant layer in particular to its elastic properties. These elastic properties have successfully been described using an, at present, well-established approach based on the statistical mechanics of curved interfaces, that was first introduced by Helfrich in 1973 [24],[25]. The **Helfrich free energy** F_H describes the bending energy of an arbitrarily shaped surface in terms of the radius of the spontaneous curvature, H_o, the rigidity constant associated with bending, κ; and the rigidity constant associated with Gaussian curvature, $\bar{\kappa}$,

$$F_H = \int dA\left[f_o + 2\kappa(H - H_o)^2 + \bar{\kappa}K\right]. \tag{15}$$

The above form for the free energy is an expansion up to second order in curvature and features an integral over the whole surface area, A, of the mean curvature, $H = ½\,(1/R_1+1/R_2)$, and Gaussian curvature, $K = (1/R_1R_2)$ with R_1 and R_2 the principal radii of curvature at a certain point on the surface A. In terms of R_1 and R_2, F_H can be written as

$$F_H = \int dA \left[\sigma - \frac{2\kappa}{R_o}\left(\frac{1}{R_1}+\frac{1}{R_2}\right) + \frac{\kappa}{2}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)^2 + \frac{\bar{\kappa}}{R_1 R_2} \right], \quad \textbf{(16)}$$

where $R_o = 1/H_o$ is the radius of spontaneous curvature and σ is the surface tension of the planar surface. The constant f_o is defined such that σ comprises all terms that are independent of curvature: $\sigma = f_o + 2\kappa H_o^2$.

The four parameters σ, R_o, κ and $\bar{\kappa}$ can, until now, not be obtained from a molecular model allowing the direct input of the properties of the surfactant molecules and the interactions of the hydrophilic and hydrophobic moieties with their solvent. They can experimentally be determined or simulated for (simple) model systems with well defined interaction potentials (e.g. a "surfactant" with Lennard-Jones interactions). A global insight into the general phase behavior of self-assembling systems can be gained by minimization of the Helfrich free energy for the different phases. Phase boundaries and coexistence regions can then be obtained in terms of R_o and, e.g. $\bar{\kappa}/\kappa$. It is therefore important to know how these curvature parameters are related to experimentally accessible variables, such as temperature or salt concentration.

The *spontaneous curvature*, H_o, is defined as the curvature H a surfactant film will adopt when the film is totally unconstrained. In this case the film is in its lowest free energy state and thus $\partial F_H / \partial H = 0$ for $H = H_o$. The spontaneous curvature H_o corresponds to the energetically most favorable packing configuration of the surfactant molecules in the interface. In principal, it should be closely related to the packing parameter introduced previously and it thus depends on the molecular geometry of the surfactant and the solubilization of the hydrophilic part in the aqueous phase. This depends in turn on formulation parameters such as temperature, (pressure), salt concentration in the aqueous phase, and the presence of cosurfactants (e.g., alcohol). A large value for H_o favors the formation of spherical or cylindrical aggregates, whereas when H_o is zero (balanced state), the film preferably forms flat planar bilayers or bicontinuous structures with saddle shaped curvature for which $R_1 = -R_2$. When H_o is negative inverse structures are formed in which the surfactant layer is curved against water.

For a given type of surfactant, the spontaneous curvature can be tuned and adjusted by the aforementioned formulation variables. For non-ionic surfactants of the C_iE_j type the strongest effect on H_o is due to a change in temperature. Upon increasing the temperature the hydrogen bonds between the hydrating water molecules and the ether groups of the ethyleneoxide moiety E_j are broken with the result that the surfactant molecules become increasingly less soluble in the aqueous phase. The interfacial surfactant film changes its spontaneous curvature from $H_o > 0$ via its balanced state at $H_o \cong 0$ to $H_o < 0$. The actual transition temperatures depend on the length of the ethylene C_i and ethylene oxide E_j blocks of the surfactants. For ionic surfactants the strongest influence is due to the salt concentration in the aqueous phase. If the salt concentration is increased the effective charge of the head group will be screened and the surfactant heads will repel each other less which can lead to a change in sign of H_0. For both kinds of surfactants, the addition of alcohols with more than four C-atoms stabilizes structures which are curved toward the aqueous phase. Ionic surfactants with a strong head group (e.g. SO_3^- or SO_4^-), such as sodium dodecylsulfat (SDS), can only be brought into their balanced state (or even curved toward water) if salt and a long chain alcohol are added, while films of a double tailed ionic surfactant, such sodium di(2-ethylhexyl)sulfosuccinate (AOT), and lipids are close to their balanced state.

The *(bending) rigidity constant* or *bending elastic modulus*, κ, is, in principle, a measure of how easily an interfacial surfactant film can be bent and influences both the static and dynamic properties of surfactant self-assemblies. The dimension of the bending rigidity κ is energy with values typically between 1-20 kT. Experimentally, κ can be obtained from ellipsometric measurements of fluctuating surfactant interfaces [26]. If the bending elastic constant κ is of the order of kT the interfacial film is flexible and a bicontinuous structure will exist if the spontaneous curvature of the surfactant film is adjusted to be close to zero ($H_o \cong 0$) by, e.g., variation of T, salt and/or alcohol concentration. If κ is much larger than kT, the film is rigid and spherical or cylindrical micelles exist up to very high volume fractions with phase inversion taking place preferably via a lamellar crystalline phase. The flexibility is increased when i) short chain hydrocarbon chains or cosurfactants are added, ii) double chain surfactants with unequal chains are used and iii) the salt concentration of the aqueous phase is increased. For the *rigidity constant associated with Gaussian curvature* or *the saddle splay modulus*, $\bar{\kappa}$, only few reliable measurements are available. The integral of the Gaussian curvature depends solely on the topology of the surface (Gauss-Bonnet theorem). Deformations of the film leave the integral over K thus unchanged. This value only changes if the film breaks due to the formation of holes or if two separated domains join. The typical trend is that $\bar{\kappa}$ is negative. The magnitude of it is usually smaller than that of κ in the same system. Experiments have shown, that κ and $\bar{\kappa}$ are approximately constant over the temperature range encountered, but depend for ionic surfactant layers strongly on the salt concentration of the aqueous phase [27]. Generally, the interfacial layer is stiffer when no screening occurs at no or little added salt.

We should mention here that the Helfrich energy is not only widely used to predict the phase behavior of amphiphilic systems, but it is has also been proven its use when describing fluctuations of planar surfactant layers, or the undulations of membrane bilayers. This description is an extension of the capillary wave theory of simple surfaces where the driving force against fluctuations is surface tension. Since in amphiphilic systems the surface tension is small, the bending rigidity becomes the dominant quantity describing the interfacial fluctuations down to a microscopic scale (this scale is set by the ratio of the bending rigidity and the surface tension). In fact, the relevance of the bending rigidity for the description of the fluctuations in amphiphilic layers, has been used to determine its value from measurements of the spectrum of fluctuations either by reflectivity or in computer simulations.

5.2 Amphiphilic phase behavior

The phase behavior of aqueous surfactant solutions is generally quite complicated but follows certain trends. At low concentrations micelles are formed and are arranged in a liquid-like structure without long-range order. At higher concentrations cylindrical micelles form that can pack into ordered hexagonal arrays. At high concentrations, amphiphiles tend to self-assemble into a number of fluid lyotropic crystalline phases that develop according to their average interfacial mean curvature. The nature of the lyotropic liquid crystal phase formed by amphiphiles in solution is described at a molecular level by the Helfrich energy and the surfactant packing parameter. The sequence hexagonal H_1 phase $\rightarrow$ lamellar L_α phase $\rightarrow$ inverted hexagonal H_2 phase is quite generally realized. Phases on the low surfactant side of the L_α phase have increasingly positive mean curvature (the interfacial film is curved towards the hydrocarbon region), while the phases towards the surfactant rich side of the L_α phase have increasingly negative mean curvature (curvature towards the aqueous region). In the regions between the hexagonal and the lamellar phase and the hexagonal and the micellar phase micelles can fill space more efficiently by ordering themselves in different cubic arrays.

This occurs on either side of the region of zero mean curvature.

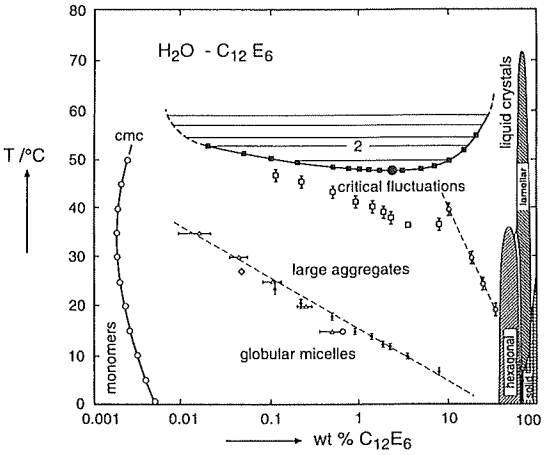
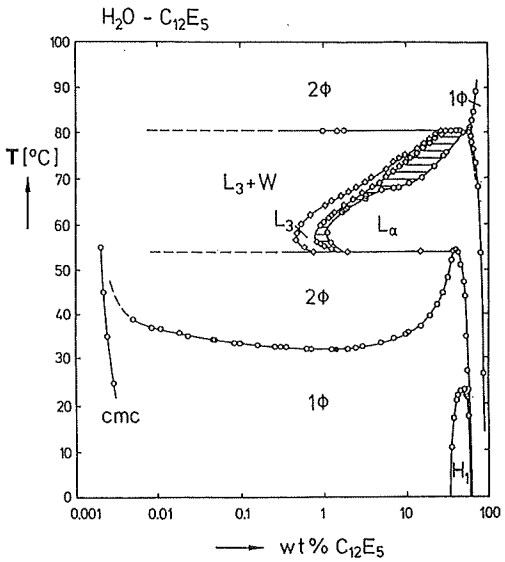
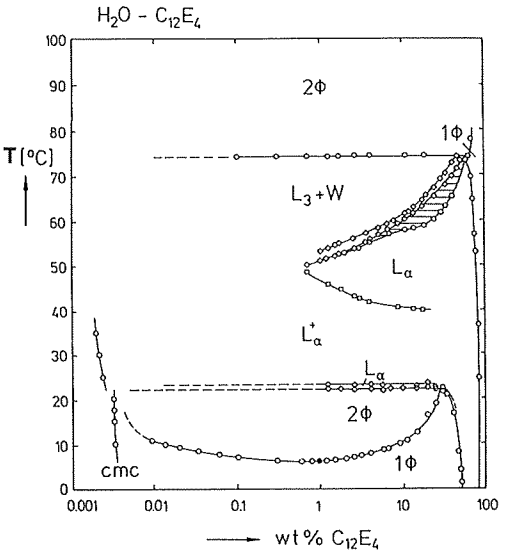


Fig. 9: *Phase diagrams for different water-C_iE_j systems* [29]. *With decreasing number of ethyleneoxide groups the headgroup becomes smaller and the tendency to form bilayer phases increases. The L_α region extends to lower surfactant concentrations and can also form at even lower temperatures. At high temperature a miscibility gap with a lower critical point forms in which the micellar solution 1ϕ separates into a micelle-lean and a micelle-rich region. For $C_{12}E_6$ the region is also shown in which structural changes occur from spherical micelles to elongated structures*





The free energy differences between a cubic phase or a hexagonal phase is quite small. The structures that are based on saddle-splay surfaces are bicontinuous cubic phases. They are characterized by non-zero mean curvature and negative Gaussian curvature. Normal bicontinuous structures consist of two continuous channels of water separated by a bilayer of surfactant molecules. In a gyroid bicontinuous phase, the two continuous channels are formed from labyrinths with threefold connecting nodes.

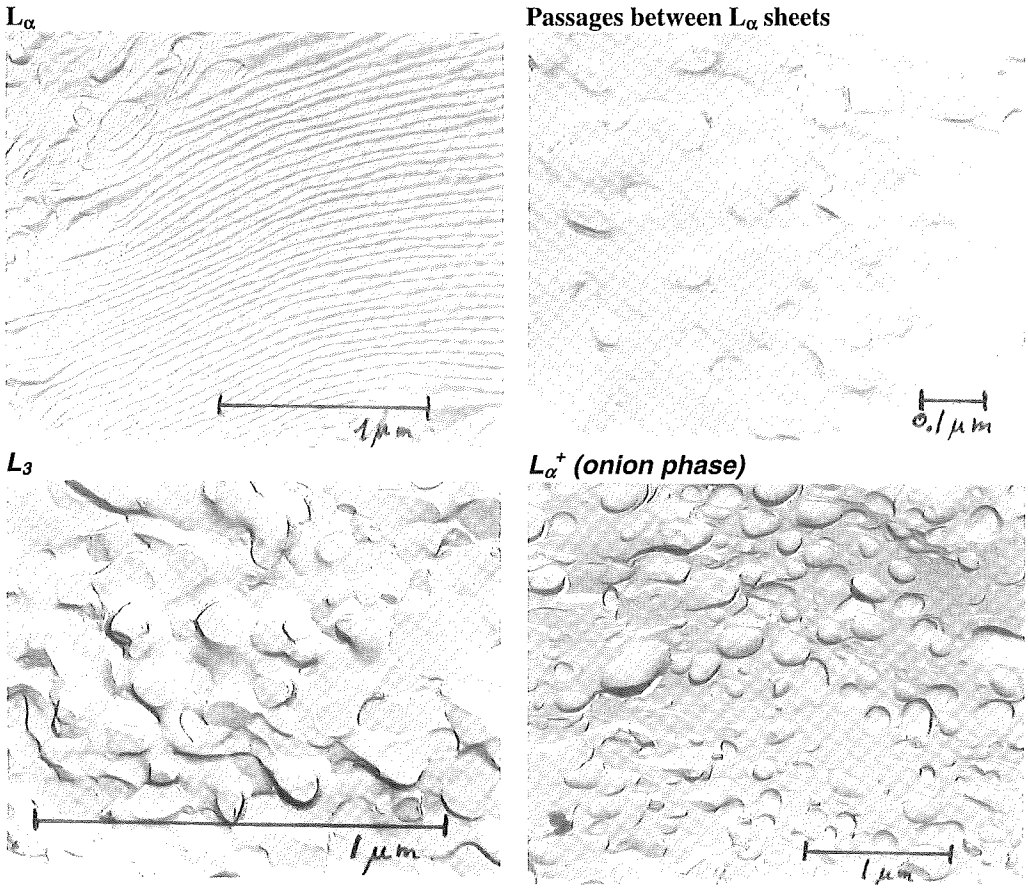


Fig. 10: *Freeze fracture transmission electron micrographs of dilute bilayer phases formed by the nonionic $C_{12}E_5$ [29]. The L_α phase consists of a stack of flat bilayer sheets bilayer that are separated by the aqueous phase. The L_3 or sponge phase is a distorted cubic phase that borders the L_α phase at the water rich side of the phase diagram. The bilayer extends in an unlimited distance in the dimensions. It is isotropic and displays birefringence under shearing. The L_3 phase forms from the L_α by creating passages between the lamellar sheets. The L_α^+ phase consists of multi-lamellar vesicles (onions) that form under shear in certain regions of the L_α phase.*

To emphasize the transition from micellar systems to bilayer phases, Fig. 9 shows phase diagrams for different nonionic surfactants of the C_iE_j series. With decreasing size of the headgroup, the tendency to form bilayer phases increases. $C_{12}E_6$ displays the L_α phase only at very high surfactant concentrations. By lowering the number of ethylene oxide groups from 5 to 4 the L_α phase extends to lower surfactant concentrations and also forms already at lower temperatures. Temperature decreases the hydration of the ethylene oxide groups and leads therefore also to a lowering of the head group efficiency and curves thus the interfacial layer towards zero mean curvature. The L_α phase is surrounded on its low surfactant side by a bicontinuous isotropic sponge phase (L_3) that can be seen as a distorted cubic phase. In certain

regions of the La phase multi-lamellar vesicles can form under shear. TEM micrographs of the different bilayer phases are displayed in Fig.10.

Biological amphiphiles such as phospholipids and glycolipids exhibit a rich lyotropic phase behavior, forming liquid crystalline structures with 1-D, 2-D, or 3-D periodicities. The 1-D fluid lamellar phase forms the basic bilayer structure of biological membranes. More complex structures such as 2-D and 3-D structures, such as the inverse hexagonal (H_2) and inverse cubic phases, may form in certain specialized membrane structures, and might be important in dynamical processes in membrane processes which involve topological changes in the membrane structure. In order to understand the possible role of these complex-crystalline phases in biology, it is necessary to characterize the physico-chemistry which underlies their phase behavior. In the L_α phases of phospholipids, the molecules typically have lateral diffusion constants of 10^{-12}–10^{-11} m^2s^{-1}, and have long-axis rotational correlation times of 10^{-9}–10^{-8} s. In terms of the bilayer elastic-mechanical properties, the area compressibility modulus is typically in the range 0.1-1 Nm^{-1}, and the bending modulus (mean curvature elastic modulus) is typically $4x10^{-20}$ J (10 *kT*), so thermal fluctuations (in conjunction with hydration forces) are usually too weak to swell the water layers to more than approximately 30 Å (however, if electrical charges are present on the lipid headgroups, indefinite swelling can occur).

References

[1] C. Tanford, *The Hydrophobic Effect: Formation of Micelles and Biological membranes,* 2nd ed. (Krieger Publishing Company, Malabar, Florida, 1991).

[2] I. W. Hamley, *Introduction to Soft Matter: Polymers, Colloids, Amphiphiles and Liquid Crystals* (Wiley, Chichester, West Sussex, 2000).

[3] D. F. Evans, H. Wennerström, *The Colloidal Domain: Where Physics, Chemistry, Biology, and Technology meet*, 2nd ed. (Wiley-VCH, New York, 1999).

[4] J. N. Israelachvili, *Intermolecular and Surface Forces*, 2nd ed (Academic Press, London, 1992).

[5] *Micelles, Membranes, Microemulsions, and Monolayers*, edited by W. M. Gelbart, A. Ben-Shaul, and D. Roux (Springer-Verlag, Berlin, 1994).

[6] E. E. Bittar, *Membrane Structure and Function* (Wiley-Interscience, New York, 1980).

[7] R. Harrison and G. G. Lunt, *Biological Membranes*, 2nd ed. (Halsted Press, Wiley, New York, 1980).

[8] M. Jain, *Introduction to Biological Membranes* (Wiley, New York, 1988).

[9] *Structure and Dynamics of Membranes*, edited by R. Lipowsky and E. Sackmann (Elsevier, Amsterdam, 1995).

[10] J. M. Seddon, Ber. Bunsenges. Phys. Chem. **100**, 380 (1996).

[11] E. Gorter an F. Grendel, J. Exp. Med. **41**, 439 (1925).

[12] S. J. Singer and G. L. Nicholson, Science **175**, 720 (1972).

[13] G. Benga and R. P. Holmes, Prog. Biophys. Molec. Biol. **43**, 195 (1984).

[14] T. Landh, FEBS Lett. 369, 13 (1995).

[15] S. Carnie, J. N. Israelachvili, and B. A. Pailthorpe, Biochim. Biophys. Acta **554**, 340 (1979).

[16] D. J. Murphy, FEBS Lett. **150**, 19 (1982).

[17] E. W. Kaler, A. K. Murthy, B. E. Rodriguez, and J. A. N. Zasadzinski, Science **245**, 1371 (1989).

[18] J. Th. G. Overbeek, K. Ned. Acad. Wet. Ser. B **89**, 61 (1986).

[19] R. J. Hunter, *Introduction to Modern Colloid Science* (Oxford University Press, Oxford, 1993).

[20] J. W. McBain, Trans. Faraday Soc. **9**, 99 (1913).

[21] B. Franklin, Phil. Trans. Roy. Soc. **64**, 445 (1774).

[22] C. H. Giles, Chem. Ind., 1616 (1969).

[23] Rayleigh, Lord, Proc. Roy. Soc. **47**, 364 (1980).

[24] W. Helfrich, Z. Naturforsch. C **28**, 693 (1973).

[25] S. A. Safran, *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* (Perseus Books, Cambridge, MA, 1994).

[26] H. Kelly, B. P. Binks, Y. Hendrikx, L. T. Lee, J. Meunier, Adv. Colloid Interface Sci. **49**, 85 (1994).

[27] D.J. Mitchell, B. Ninham, Langmuir **5**, 1121 (1989).

[28] *Physics of Amphiphiles: Micelles, Vesicles and Microemulsions*, edited by V. Degiorgio and M. Corti (North Holland, Amsterdam, 1985).

[29] R. Strey, Ber. Bunsenges. Phys. Chem. **100**, 182 (1996).

A 4 Molecular Components of the Cell

A. Baumann
Institut für Biologische Informationsverarbeitung 1
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 The Living World **2**

2.1 Structural features of eukaryotic cells 2

2.2 Molecular Composition of Cellular Membranes 5

2.3 The Nucleus 7

2.3.1 Nucleic Acids 10

2.3.2 Synthesis of proteins 13

2.4 Mitochondria produce cellular energy 15

3 Summary **17**

A Appendix: Glossary **19**

1 Introduction

All living organisms are made up of small elementary units: the cells. Although cells are remarkably diverse in structure and function within a single organism or between different organisms, they share various features. In the following, I will introduce the main cellular organelles, explain their biochemical function and describe the properties of essential molecules of the living cell. This is a tutorial for physicists interested in biochemical and molecular cell function. There are several superb textbooks available that comprehensively cover this research area. The text and figures of this review are, in part, based on the text books "Biochemistry" by Berg et al. [1], "Lehninger, Principles of Biochemistry" by Nelson and Cox [2], "Molecular Cell Biology" by Lodish et al. [3], and "Molecular Biology of the Cell" by Alberts et al. [4]. The interested reader is also referred to other textbooks cited at the end of the text [5],[6].

2 The Living World

The diverse organisms of the modern world can be divided into three fundamental groups: *Eukarya* (eukaryotes), *Bacteria*, and *Archaea*. Eukarya comprise all macroscopic organisms. To name just a few this group includes human beings, apes, cats, birds, fish, insects, worms, and plants. In addition to macroscopic organisms many microscopic, unicellular organisms such as the yeast also classify as eukaryotes. The common and name-giving characteristic of all eukaryotes is the presence of a well-defined *nucleus* within each cell.
Unicellular organisms, which lack a nucleus, are referred to as *prokaryotes*. There are two groups of prokaroytes: the bacteria and the archaea. This differentiation was introduced when it was realised that certain bacteria-like organisms (archaea) are biochemically distinct from better-characterised bacterial species. Modern bacteria inhabit almost every ecological niche in the biosphere. The best studied prokaryotic cell is the bacterium *Escherichia coli*. It is a usually harmless habitant of the human intestinal tract. With a length of ~2 μm and a diameter of ~1 μm it is rather small. For the rest of this review, however, I will focus on the properties of eukaryotic cells.

2.1 Structural features of eukaryotic cells

Multicellular organisms contain many different types of cells, which vary in size, shape, and function. The human body, for example, contains at least 10^{14} cells. Typically eukaryotic cells are 5 to 100 μm in diameter and thus are much larger than prokaryotic cells. Some cells of the nervous system (*motor neurons*) possess very long cellular extensions. These *axons* can be ~1 m long. They originate in the spinal cord and pass signals to muscle fibers in legs or toes.
Despite their morphological and functional differences, all eukaryotic cells share important features (Fig.1). The plasma membrane defines the surface of the cell, separating its contents from the surroundings. The external surface of the cell is in contact with other cells and the extracellular fluid containing solutes and nutrient molecules. The plasma membrane is composed of enormous numbers of lipid and protein molecules, held together primarily by noncovalent hydrophobic interactions, forming a thin, hydrophobic layer around the cell. The membrane is a barrier to the free passage of inorganic ions and most other charged or polar compounds. Nevertheless, cells continuously have to exchange "substances" with their environment. They have to take up nutrients and to get rid of metabolized or degraded cellular components. To manage this logistic problem, specific transport proteins in the plasma membrane allow the passage of certain ions and molecules. Other membrane proteins not participating in 'cargo' processes, include receptors that transmit signals from the outside to

the inside of the cell and enzymes that participate in membrane-associated reaction pathways. Whatever the mode of signal transduction, surface receptors characteristically act as signal amplifiers: a single molecule activating a single receptor causes the flux of thousands of ions through an opened channel or the synthesis of thousands of molecules of intracellular messenger molecules. In addition to the plasma membrane, eukaryotic cells also contain internal membranes that form the boundaries of organelles such as the endoplasmic reticulum, Golgi complex, lysosomes, mitochondria, and nucleus.

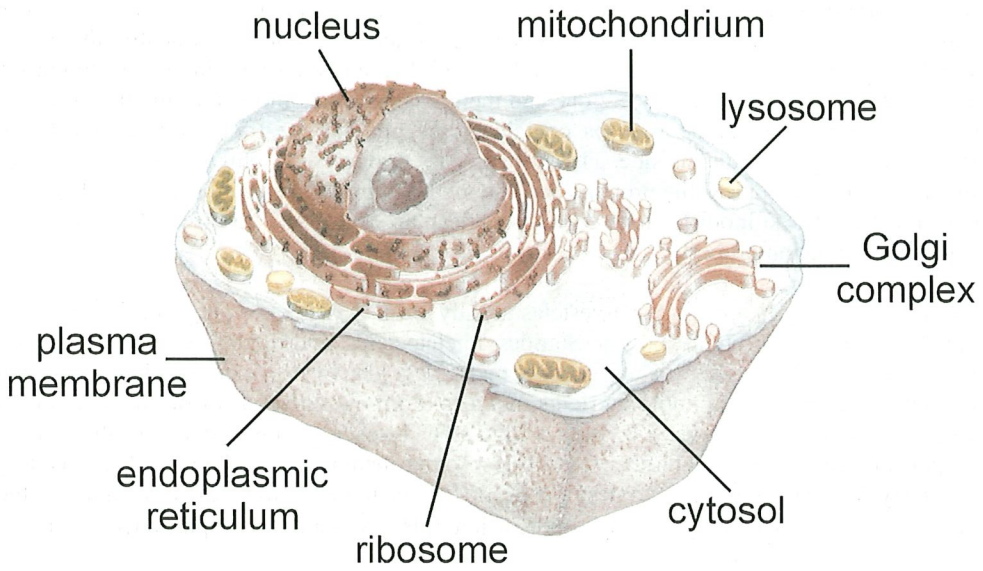


Fig. 1: *Schematic drawing of a typical animal cell. Animal cells are usually 2 to 30 µm in diameter. Structures and subcellular components are labelled and explained in the text. Modified from [2].*

The internal volume bounded by the plasma membrane, the *cytoplasm*, is composed of an aqueous solution, the cytosol, and a variety of insoluble, suspended particles. Initially, the cytosol was thought to be a fairly homogeneous "soup" in which all of the organelles floated. However, it is now known that the cytosol is a highly concentrated aqueous solution with complex composition and gel-like consistency. A major constituent is the *cytoskeleton*, a meshwork-like structure composed of polymeric protein fibres, called microtubules and filaments, that participate in maintaining cell shape and mobility and also provide anchoring points for other cellular structures. Dissolved in the cytosol are many proteins, small organic molecules (metabolites), and intermediates of biosynthetic and degradative pathways. Among the particles operating in the cytosol are the *ribosomes*. They are the sites at which protein synthesis occurs. Ribosomes engaged in protein synthesis often come in clusters called polysomes and are located on the outer surface of the *endoplasmic reticulum*.

The endoplasmic reticulum (ER) is a highly convoluted, three dimensional network of membrane-enclosed spaces extending throughout the cytoplasm and enclosing a subcellular compartment separate from the cytoplasm. The flattened branches of this compartment are continuous with each other and with the nuclear envelope. In cells specialized for the

secretion of proteins, the ER is particularly prominent. Ribosomes that synthesize proteins destined for export attach to the surface of the ER, and the secretory proteins are passed through the membrane into the lumen as they are synthesized. By contrast, proteins required in the cytosol are synthesized on ribosomes unassociated with the ER. An area where thousands of ribosomes are attached to the ER has a granular appearance and therefore is called *rough endoplasmic reticulum*. In some regions of the cell, the ER is free of ribosomes. This *smooth endoplasmic reticulum* is physically continuous with the rough ER. The smooth ER is the site of lipid biosynthesis.

Most eukaryotic cells have *Golgi complexes* (see Fig.1). The Golgi complex is a system of membranous sacs arranged as flattened stacks. Structurally and functionally the Golgi complex is asymmetric. The cis side faces the rough ER, and the trans side faces the plasma membrane. Newly synthesized secretory and membrane proteins move into the lumen of the rough ER. Small membrane vesicles containing the proteins are transported to the Golgi complex and fuse with the cis side. The proteins pass through the Golgi complex to the trans side. In the Golgi complex, the proteins become modified by sulfate groups, carbohydrates, or lipid moieties. These modifications serve as molecular "addresses" to direct the proteins to their final destination properly.

Lysosomes (see Fig.1) are spherical vesicles usually 1 μm in diameter. They contain enzymes capable of digesting proteins, polysaccharides, nucleic acids, and lipids. Lysosomes function as cellular recycling centers. They contain many degradative enzymes which break down complex molecules, for example nutrients, but also worn-out organelles from the cell's own cytoplasm. The degradative enzymes within the lysosome could act on all cellular components were they not confined by the lysosomal membrane. Another strategy keeping these enzymes inactive in the cytoplasm, is the difference in pH between the lysosome and the cytosol. The lysosomal compartment is more acidic ($pH \leq 5$) than the cytoplasm ($pH \approx 7$).

Another important organelle of the cell is the *mitochondrion*. Mitochondria typically have a diameter of 1 μm, similar to that of bacterial cells. Most animal cells contain several hundred to a thousand of these membrane-bounded organelles. Mitochondrial enzymes catalyze the oxidation of organic nutrients by molecular oxygen. The chemical energy released in mitochondrial oxidations is used to generate ATP, the major energy-carrying molecule of cells. The ATP diffuses to all parts of the cell and provides the energy for cellular work.

The largest subcellular structure depicted in Fig.1 is the *nucleus*. The nucleus contains nearly all of the cell's DNA. The nucleus is surrounded by a nuclear envelope, composed of two membrane bilayers separated by a narrow space and continuous with the rough ER. At intervals, the inner and outer nuclear membranes are pinched together and form nuclear pores. Associated with the pores are specific transporters, protein complexes that allow certain macromolecules to pass between the cytoplasm and the aqueous phase of the nucleus. Traffic into the nucleus includes enzymes required for DNA replication, transcription, and RNA processing. Passing out through the nuclear pores are messenger RNA molecules which are translated into protein on ribosomes in the cytoplasm.

2.2 Molecular Composition of Cellular Membranes

In the following, the basic features of biological membranes of eukaryotic cells will be described. More information on biological membranes can also be found in chapters D.1 and D.5.

All biological membranes share certain fundamental properties. They are impermeable to most polar or charged solutes, but permeable to nonpolar compounds. Membranes are sheetlike structures. They are 6 to 10 nm thick. Biological membranes consist of a lipid bilayer and proteins. Membrane lipids are relatively small molecules that have both hydrophilic and hydrophobic moieties. Among the most biologically significant properties of lipids are their hydrophobic properties. These properties are mainly due to a particular component of lipids: fatty acids. Fatty acids are hydrocarbon chains of various length. In most biological membranes fatty acids typically consist of 14 to 24 carbon atoms. Fatty acids not only differ in chain length. They can also be saturated or unsaturated which means that carbon atoms are connected by single or double bonds, respectively.
There are three major kinds of lipids found in a biological membrane: phsopholipids, glycolipids, and cholesterol. *Phospholipids* are abundant in all biological membranes. A phosopholipid molecule is constructed from four components: fatty acids, glycerol or sphingosine, a phosphate, and an alcohol attached to the phosphate. In phosphoglycerides (Fig.2), the carboxyl groups of two fatty acid chains are esterified to the hydroxyl groups at C-1 and C-2 of glycerol. The phosphoric acid is esterified to the hydroxyl group at C-3 of glycerol. The phosphate can form another ester bond to the hydroxyl group of one of several alcohols. Some common alcohol moieties of phosphoglycerides are choline, ethanolamine, the amino acid serine, and the sugar inositol.

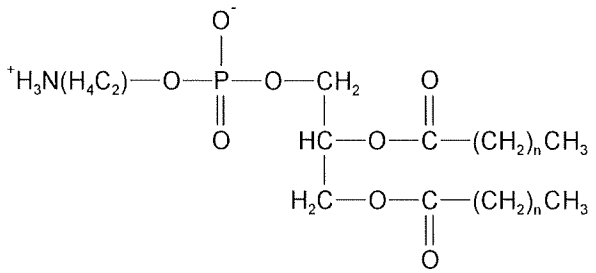


Fig. 2: *Molecular structure of a phosphoglyceride (here: phosphatidylethanolamine) found in cell membranes. Two fatty acid molecules ($CH_3(CH_2)_nCOOH$) are linked by ester bonds to the central alcohol, glycerol (CH_2OH-$CHOH$-CH_2OH). The third hydroxyl group of glycerol is esterified to phosphoric acid (PO_4) which also is connected to one of several alcohols, in this case ethanolamine ($^+H_3N(C_2H_4)OH$).*

Glycolipids, as their name implies, are sugar-containing lipids. The glycolipids in animal cells are derived from sphingosine, an amino alcohol that contains a long, unsaturated hydrocarbon chain. One or more sugars are attached to the primary hydroxyl group of the sphingosine backbone. Glycolipids are oriented in a completely asymmetric fashion with the sugar residues always on the extracellular side of the membrane.
Cholesterol is another important constituent of biological membranes. It is a steroid. In membranes, cholesterol is oriented parallel to the fatty acid chains of the phospholipids. Cholesterol is found to varying degrees in all animal membranes. A cholesterol content of up

to 25% has been determined in some neurons whereas the steroid is essentially absent from intracellular membranes.

What properties enable phospholipids to form membranes? Membrane formation is a consequence of the amphipathic character of the molecules. The polar head groups, formed by the alcohol and phosphate (see Fig.2), favour contact with water. Their hydrocarbon chains interact with one another rather than preferring water. These strongly opposed preferences of the hydrophilic and hydrophobic moieties of membrane lipids can be satisfied by forming a lipid bilayer, composed of two lipid sheets (Fig.3). The hydrophobic tails (= fatty acid chains) of each sheet interact with one another, forming a hydrophobic interior that acts as a permeability barrier. The hydrophilic head groups interact with the aqueous medium on each side of the bilayer. The hydrophobic interactions between the hydrocarbon chains are the major driving force for the formation of lipid bilayers. Van der Waals attractive forces between the hydrocarbon tails favour close packing of the tails. In addition, there are electrostatic and hydrogen-bonding attractions between the polar head groups and water molecules. Thus, lipid bilayers are stabilized by a full array of forces that mediate molecular interaction in biological systems.

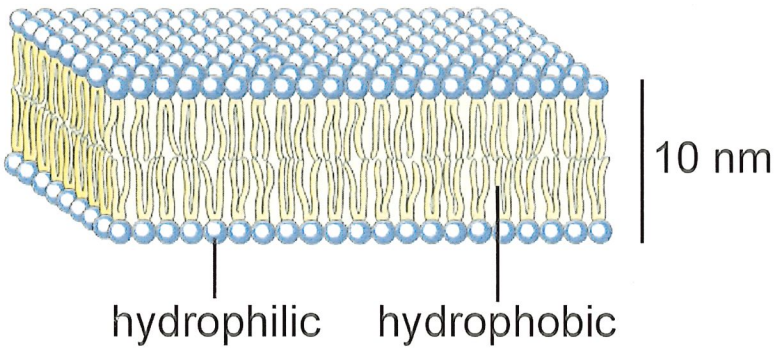


Fig. 3: *Biological membranes consist of a two-dimensional sheet of lipid molecules arranged in a lipid bilayer. The hydrophilic head groups (blue) of the lipids are in contact with water at each surface of the bilayer. The hydrophobic chains of the fatty acids (yellow) are protected from water and form the core of the bilayer. Modified from [2].*

Although the lipid bilayer structure itself is stable, the individual phospholipid molecules have great freedom of motion within the plane of the membrane. The interior of the bilayer is fluid. Individual hydrocarbon chains of fatty acids are in constant motion produced by rotation about the carbon-carbon bonds. The degree of fluidity depends on lipid composition and temperature. At low temperature, lipid motion is restricted and the bilayer exists as a nearly crystalline array. Above a certain temperature (= transition temperature), the crystalline solid changes to a more fluid phase; the transition temperature depends on the lipid composition of the membrane. High amounts of cholesterol and unsaturated fatty acid chains in phospholipids result in lower transition temperatures.
Another type of lipid motion involves the movement of an entire lipid molecule relative to its neighbours. A molecule in one monolayer of the lipid bilayer can diffuse laterally so fast that it 'circumnavigates' a cell within seconds. This rapid lateral diffusion within the plane of the

bilayer randomizes the positions of individual molecules in a few seconds. In recent years, however, experimental evidence has been accumulated that some areas of a cell's membrane possess a very fixed lipid and cholesterol composition. Such areas are referred to as *lipid rafts*. It is assumed that certain membrane proteins preferentially assemble in these areas ('microdomains') to ensure a specific and rapid response to external stimuli.
In addition to lateral movements of lipid molecules one might also expect diffusion from one face of the bilayer sheet to the other. This "flip-flop" transition requires a polar head group to leave its aqueous environment and move into the hydrophobic interior of the bilayer. This process is energetically unfavourable due to a large, positive change in the free energy. Sometimes, however, a cell necessitates to move lipids from one side of the bilayer to the other. For this purpose, specific proteins in the plasma membrane, called *flippases*, provide a transmembrane path to circumvent the energetically unfavourable flip-flop diffusion.

As for the flip-flop mechanism of single lipid molecules, the properties of the lipid bilayer also impair the movement of charged or polar molecules across a biological membrane. A measure often used to describe and to compare the abilities of substances to pass the membrane is the permeability coefficient (P [cm s^{-1}]). The coefficient's dimension is given in terms of velocity, though one should keep in mind that there is a barrier, the membrane, that hinders the free movement of the molecules. In general terms, the more polar the substance the slower the movement across the lipid bilayer. What are the steps necessary for a given molecule to traverse the membrane? First, it has to shed its solvation shell of water. Second, it becomes dissolved in the hydrocarbon core of the membrane. Thirdly, the molecule diffuses through the hydrocarbon core to the other side of the membrane, where it becomes resolvated by water. Ions such as Na^+ or K^+ traverse the membranes very slowly because the replacement of its coordination shell of polar water molecules by nonpolar interactions with the membrane interior is energetically highly unfavourable. Similarly as described for the flip-flop mechanism of individual lipid molecules, there exist large families of membrane proteins that mediate the transport of polar molecules across the lipid bilayer: ion channels and pumps. These proteins are introduced in chapters D.7 and D.8.

2.3 The Nucleus

The nucleus of a non-dividing cell is filled with a material called *chromatin*. Chromatin consists of DNA and proteins bound tightly together. Chromatin is the substance of the *chromosomes*. The cells that make up most of the body of a multicellular organism, have two copies of each chromosome. These somatic cells are diploid. Gametes, e.g. egg and sperm, are having only one copy of each chromosome. They are haploid. As depicted in Fig.4 a human somatic cell has 23 pairs of chromosomes. The DNA, wound up in the chromosomes, contains all the information to encode every protein of the cell. Each chromosome consists of two identical *chromatids*. When a cell divides, the two chromatids separate, one moving to each pole of the cell, where they become part of the newly formed nucleus of each daughter cell.

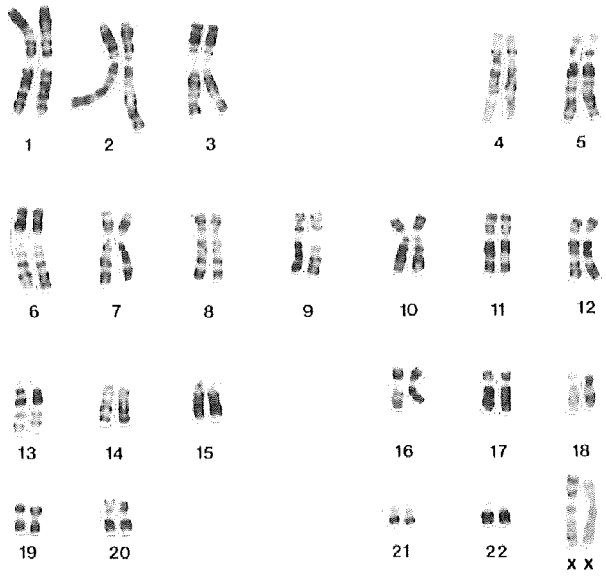


Fig. 4: *Karyotype of human female metaphase chromosomes. Taken from [5].*

In the nucleus, the DNA is tightly bound to a family of positively charged proteins, the *histones*. The histones and DNA associate in complexes called *nucleosomes*, in which the DNA strand winds around a core of histone molecules (see Fig.5). The DNA of a single human chromosome forms about a million nucleosomes. The nucleosomes associate to form very regular and compact supramolecular complexes. The resulting chromatin fibers, about 30 nm in diameter, condense further by forming a series of looped regions and cluster with adjacent looped regions. This tight packing of DNA into nucleosomes achieves a remarkable condensation of the DNA molecules. The DNA in the chromosomes of a single somatic human cell would have a combined length of about 2 m if fully extended, whereas the combined length of all 46 chromosomes is about 200 nm (= 2×10^{-7} m). In the following chapter, the molecular properties of the nucleic acids DNA and RNA will be described.

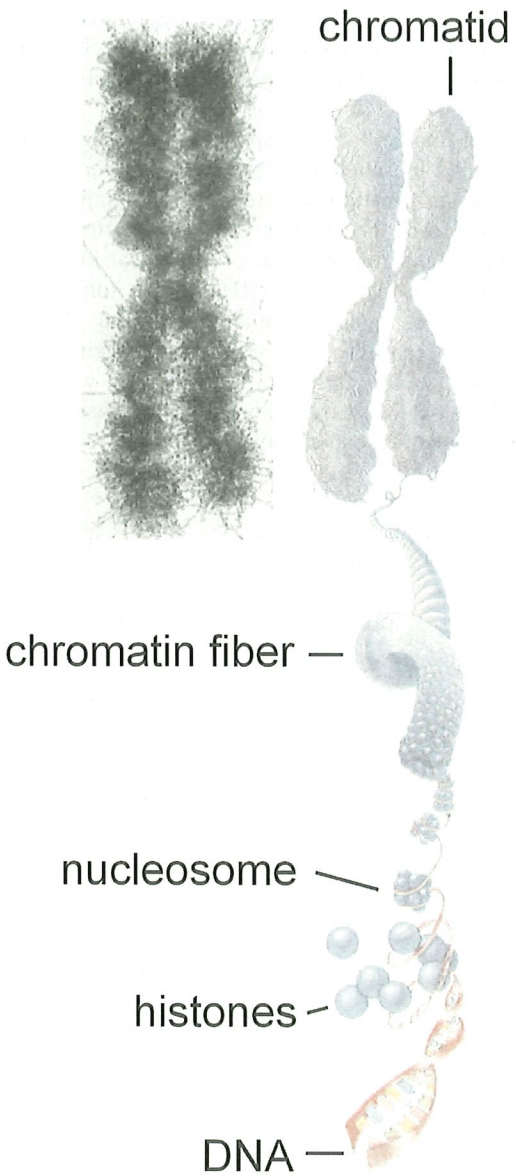


Fig. 5: *Shown is an electron micrograph of one human chromosome (left). Every chromosome is composed of two chromatids, each consisting of tightly folded chromatin fibres. Each chromatin fiber is formed by packaging of a DNA molecule wrapped around histone proteins to form nucleosomes. Modified from [2].*

2.3.1 Nucleic Acids

In modern organisms, deoxyribonucleic acids (DNA) and ribonucleic acids (RNA) are the molecular repositories of genetic information. The structure of every protein is a product of information programmed into the nucleotide sequence of a cell's nucleic acids. The ability to store and transmit genetic information from one generation to the next is a fundamental condition of life.

The nucleotide sequence in the cell's DNA specifies the amino acid sequence (see 2.3.2) of every protein in the cell, and the nucleotide sequence of every RNA. A segment of a DNA molecule that contains the information required for the synthesis of a functional biological product, is referred to as a *gene*. The human *genome* is estimated to contain up to 100,000 different genes. The storage and transmission of biological information are the only known functions of DNA. In comparison to DNA, RNAs have some broader range of functions. Several classes of RNAs are found in cells. Ribosomal RNAs (rRNA) are structural components of ribosomes, the particles that synthesize proteins. Messenger RNAs (mRNA) carry genetic information from a gene to a ribosome, where the encoded protein is synthesized. Transfer RNAs (tRNA) are adapter molecules that translate the information from the messenger RNA into a specific sequence of a protein.

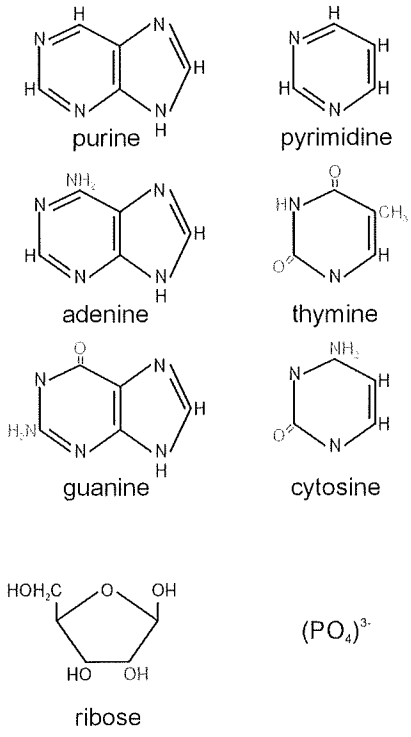


Fig. 6: *Structure of nucleotides. Nucleotides consist of three components: a nitrogen-containing base, a pentose, and a phosphate. The bases are derivatives of either purine (adenine or guanine) or pyrimidine (thymine or cytosine). In DNA, the hydroxyl group on the pentose ring (depicted in red) is replaced by a hydrogen.*

The nucleic acids of a cell consist of nucleotides (Fig.6). Nucleotides have three characteristic components: a nitrogen-containing base, a pentose, and a phosphate. The bases are derivatives of two parent compounds, purine and pyrimidine. The base of a nucleotide is covalently linked by an N-β-glycosyl bond to the pentose, and the phosphate is esterified to the 5' carbon of the pentose. Both DNA and RNA contain two purine bases, adenine (A) and guanine (G), and two pyrimidines. In both nucleic acids one of the pyrimidines is cytosine (C), but the second is thymine (T) in DNA and uracil (U) in RNA. Nucleic acids have two kinds of pentoses. The nucleotides of DNA contain 2'-deoxy-D-ribose, and the nucleotides of RNA contain D-ribose.
The successive nucleotides of both DNA and RNA are covalently linked through phosphate-group bridges, in which the 5'-hydroxyl group of one nucleotide is joined to the 3'-hydroxyl group of the next nucleotide by a phosphodiester linkage (Fig.7).

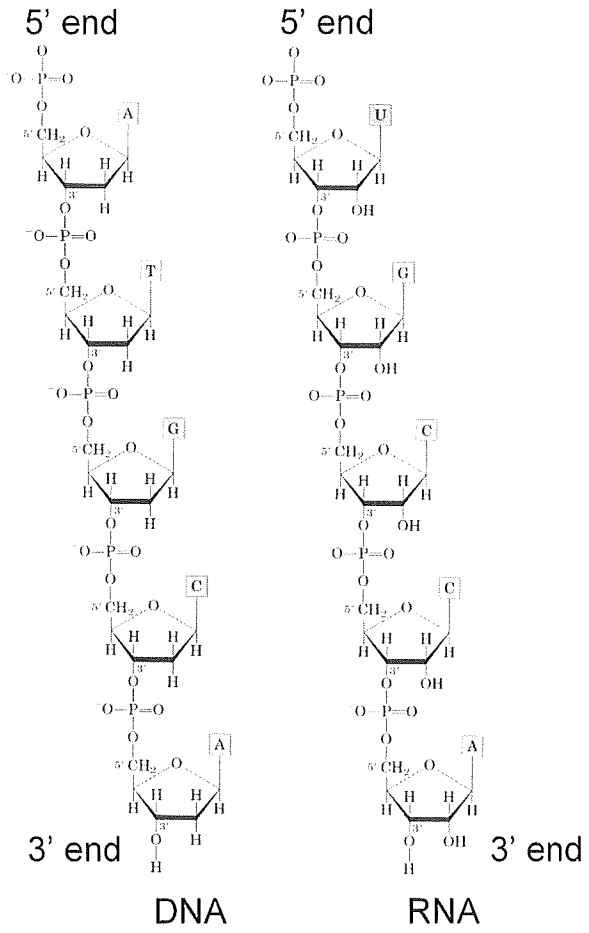


Fig. 7: *Covalent bonds in the backbone of DNA and RNA. The phosphodiester bonds link successive nucleotide units. The backbone of alternating pentose and phosphate groups in nucleic acids is highly polar. Nucleic acids are also referred to as poly-anions. Modified from [2].*

Thus the covalent backbones of nucleic acids consist of alternating phosphate and pentose residues, with the bases joined as side groups to the backbone at regular intervals. The backbones of both DNA and RNA are hydrophilic. The hydroxyl groups of the sugar residues form hydrogen bonds with water. The phosphate groups are completely ionized and negatively charged at pH 7. These negative charges are compensated by ionic interactions with positive charges on proteins, e.g. histones, on metal ions, and on polyamines.

The covalent structure of nucleic acids accounts for their ability to carry information in form of the linear sequence of bases along a nucleic acid chain. Another important feature is the ability of the bases to form specific base pairs resulting in a helical structure consisting of two strands of nucleic acids. The existence of specific base-pairing interactions was discovered when the three-dimensional structure of DNA was solved by X-ray diffraction [7]. The diffraction patterns indicated that DNA was formed of two chains that wound in a regular helical structure (Fig.8). The two helical DNA chains are coiled around a common axis. The sugar-phosphate backbones are on the outside of the helix, and the purine and pyrimidine bases are facing the inside of the helix. The bases are nearly perpendicular to the helix axis. The helical structure repeats every 3.4 nm. There are 10 bases per helix turn. The diameter of the helix is 2.0 nm. Altogether these are the essential features of the Watson-Crick model of DNA.

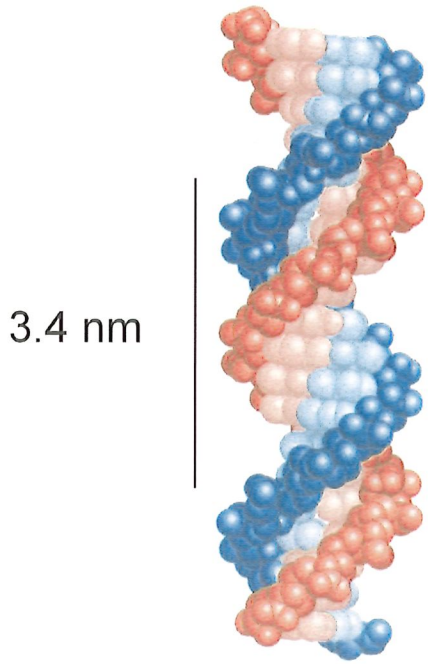


Fig. 8: *Model of double-helical DNA. One DNA strand is shown in blue, the other in red. The structure repeats along the helical axis at intervals of 3.4 nm, which corresponds to 10 nucleotides on each chain. Modified from [1].*

The purine and pyrimidine bases of both strands are stacked inside the double helix, with their hydrophobic ring structures coming very close together. Each nucleotide base of one strand is paired in the same plane with a base of the other strand. In DNA helices, adenine is paired with thymine (A-T) and guanine is paired with cytosine (G-C). The base pairs are held together by hydrogen bonds.

The flow of genetic information in normal cells is from DNA to RNA to protein. All cellular RNAs are synthesized by specific enzymes (RNA polymerases) from DNA as a template. This process is called *transcription*. In contrast to DNA molecules, RNA molecules are single stranded. RNAs, however, may also contain double-helical regions that arise from the folding of the chain into hairpin loop-like structures.
RNA polymerase must detect and transcribe discrete genes from within large stretches of DNA. Remember that the morphological and functional properties of cells in an organism differ considerably. Beyond housekeeping genes that are expressed in every cell and which are necessary for metabolic processes, cellular differentiation relies on the synthesis of specific proteins encoded by genes that are active in one but not the other cell type. How is differentiation achieved molecularly? Genes do not cover the entire length of the DNA harboured in the nucleus. They are separated by intervening non-coding stretches of DNA. Regulatory elements in the DNA called *promoters* specifically bind RNA polymerase and determine where transcription begins. Promoter sites are also recognized by sets of specific proteins called transcription factors. It is the *transcription factors* in combination with RNA polymerase deciding whether a gene is kept silent or transcribed into messenger RNA. When mRNA is synthesized, it finally is translated into a protein. It is the different repertoire of proteins that determine the cell's functional and morphological properties.

2.3.2 Synthesis of proteins

Proteins are the working molecules of the cell. Proteins are constructed from 20 different *amino acids*. All amino acids share a basic design. A central carbon atom (C_α) is bonded to an amino group, to a carboxyl group, to a hydrogen atom, and to one variable group (R), called a side chain (Fig.9). The side chains give the amino acids their individuality. The amino acids represent the alphabet in which linear proteins are "written". As already mentioned, the flow of biological information is from: DNA → RNA → protein.

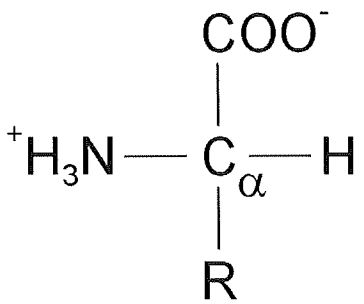


Fig. 9: *Structural formula of amino acids.*

The mRNA is the template for protein biosynthesis. How does it direct amino acids to become joined in the correct sequence to form a protein? The adapter molecule in protein synthesis is transfer RNA. The tRNA contains an amino acid-attachment site and a template (= mRNA) recognition site. A tRNA molecule carries a specific amino acid to the site of protein synthesis. The carboxyl group of this amino acid is esterified to the 3'- or 2'-hydroxyl group of the ribose unit at the end of the tRNA chain (= amino-acyl-tRNA). This reaction is catalyzed by a specific enzyme (= synthetase). There is at least one specific synthetase for each of the 20 amino acids. Because four nucleotides cannot specify the linear arrangement of 20 possible amino acids in a one-to-one manner, a group of nucleotides is required to symbolize each amino acid. The code employed must be capable of specifying at least 20 "words". If two nucleotides were used to code for one amino acid, then only 4^2 different code words could be formed. That would be an insufficient number. If a group of three nucleotides are used for each code word, then 4^3 code words can be formed. Therefore, a code using three nucleotides will have more than enough units to encode 20 amino acids. In fact, some amino acids are encoded by more than one codon. This situation is also referred to as *'the genetic code is degenerate'*. An advantage of the degeneracy is that the system becomes less vulnerable to mutations in the nucleic acids. In addition, the DNA base composition may vary without altering the amino acid sequence of the proteins encoded by the DNA. Thus, it is not surprising that the template-recognition site on tRNA is a sequence of three nucleotides called *anticodon*. The anticodon on tRNA recognizes a complementary sequence on the mRNA by base pairing.

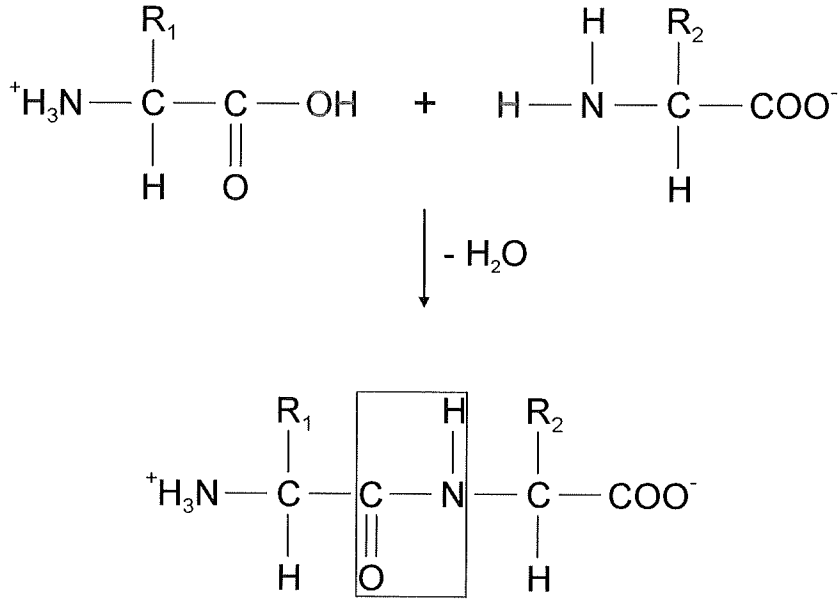


Fig. 10: *Schematic drawing of the formation of a peptide bond. Proteins are formed by linking the α-carboxyl group of one amino acid (R1) to the α-amino group of another amino acid (R2) with a peptide bond. The formation of a dipeptide from two amino acids is accompanied by the loss of a water molecule.*

Ribosomes are the molecular machines that coordinate the interplay of tRNAs, mRNA, and proteins that leads to protein synthesis. A eukaryotic ribosome consists of a large and a small subunit. Both subunits come together forming the functional ribosome with a relative molecular mass of 4,200 kDalton. As a whole, the ribosome is a complex composed of individual RNA molecules (rRNA) and more than 50 proteins [8].

Protein biosynthesis takes place in three phases: initiation, elongation, and termination. In eukaryotes, the first (initiating) amino acid is methionine. The initiating codon on the mRNA strand and thus the start site of protein synthesis is AUG. In most cases, eukaryotic mRNA has only one start site and hence is the template for a single protein. A small subunit of the ribosome attaches to the mRNA molecule and searches for the AUG codon by moving step-by-step in the 3' direction. A group of proteins called *initiation factors* assist the ribosome finding the initiation site. Pairing of the anticodon of the amino-acyl-tRNA loaded with methionine to the AUG codon of mRNA signals that the target has been found. Only after the initiation site is located does the large subunit join the complex. For the protein chain to begin to grow, a second amino-acyl-tRNA must be properly positioned on the ribosome. This is the job of another group of proteins called *elongation factors*. Now, two ribosomal sites are occupied by tRNA molecules: the *A site* accomodates the incoming amino-acyl-tRNA that is to contribute a new amino acid to the growing chain; the *P site* contains the tRNA still linked to all the amino acids added to the chain so far. A peptide bond (see Fig.10) is formed between the adjacent amino acids. In this process the last bound amino-acyl-tRNA moves from the A site to the P site, displacing the initially bound tRNA and giving way for positioning another amino-acyl-tRNA to the A site to continue elongation of the protein chain. The average rate of synthesis is about five amino acids per second. The process of *translocation* continues step-by-step until all amino acids encoded by the mRNA have been added. Unique codons in the mRNA signal the end of protein biosynthesis. When these stop codons are recognized by *termination factors*, the complex falls apart and releases the newly synthesized protein.

At the end of biosynthesis, the protein is a linear stretch of covalently linked amino acids. To become a functional protein within the cell it has to adopt a unique three-dimensional structure. A protein adopts a stable, folded conformation mainly through noncovalent ionic, hydrogen, van der Waals, and hydrophobic interactions. Its stability is further enhanced by disulfide bonds formed between amino acid residues harbouring reactive sulfhydryl (SH) groups. Detailed information on protein folding can be found in chapters C.5, C.6, and C.7.

2.4 Mitochondria produce cellular energy

All processes involved in the growth and metabolism of cells require an input of energy. In most cases, free energy is supplied by the hydrolysis of the high-energy phosphoanhydride bonds in adenosine triphosphat (ATP):

$$ATP^{4-} + 2H_2O \rightarrow AMP^{2-} + 2(PO_4)^{2-} + 2H^+ \qquad (1)$$

Per mol of ATP hydrolised, 14.6 kcal of energy are produced. Energy released by the hydrolysis is used directly to power many otherwise energetically unfavourable events, such as the transport of molecules against a concentration gradient, exemplified by the Na^+-K^+ ATPase, beating of cilia, and contraction of muscle. The energetic biochemical reactivity of ATP is used in many pathways, including synthesis of nucleic acids and proteins from nucleotides and amino acids, respectively.

Since ATP is heavily consumed by every cell, it has to be regenerated efficiently and rapidly. The cellular place of ATP generation are the mitochondria (Fig.11).

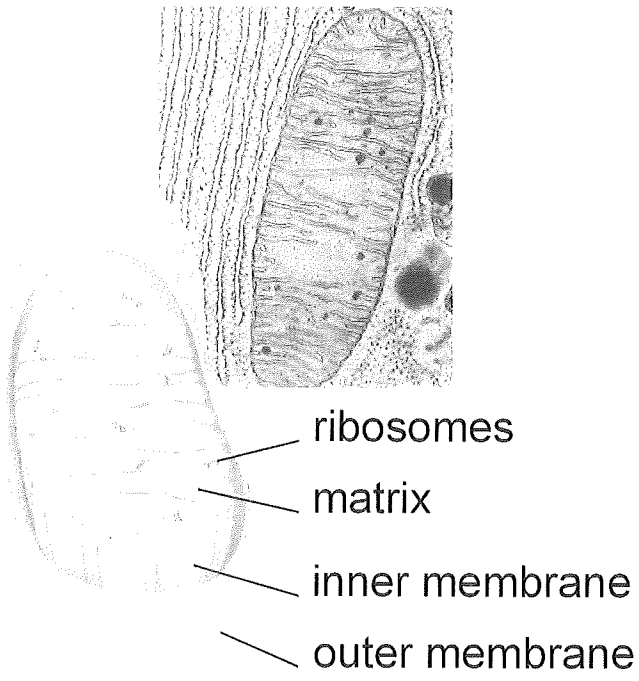


Fig. 11: *Structure of a mitochondrion. The electron micrograph shows the extensive folding of the inner membrane. Modified from [1].*

Each mitochondrion has two membranes. The outer membrane is unwrinkled and completely surrounds the organelle. The inner membrane has foldings called *cristae*, which give rise to a large surface area. Enclosed by the inner membrane is the *matrix*, a very concentrated aqueous solution of enzymes and chemical intermediates involved in energy-yielding metabolism. Unlike other membranous structures in the cell, mitochondria are produced only by division of previously existing mitochondria. They contain their own DNA, RNA, and ribosomes.

How do mitochondria produce ATP? In most aerobic organisms, oxidative phosphorylation is the process in which ATP is formed as a result of the transfer of electrons from reducing substances (NADH, $FADH_2$) to O_2 by a series of electron carriers. When glucose is completely oxidized to CO_2 and H_2O, 26 molecules of ATP are formed. Oxidative phosphorylation is mechanistically complex. The elucidation of the mechanism of oxidative phosphorylation has been one of the most challenging problems in biochemistry.

The flow of electrons from NADH or $FADH_2$ to O_2 through protein complexes located in the mitochondrial inner membrane leads to the pumping of protons (H^+) out of the mitochondrial matrix.

The resulting uneven distribution of protons generates a pH gradient and a transmembrane electrical potential that creates a *proton-motive force*. ATP is synthesized when protons flow back to the mitochondrial matrix through an enzyme complex. Thus, the oxidation of carbon fuels and the phosphorylation of ADP (adenosine diphosphate) to ATP are coupled by a proton gradient across the inner mitochondrial membrane (Fig.12).

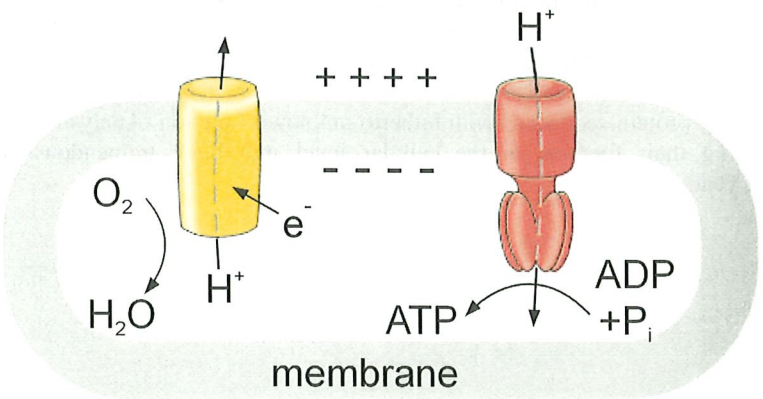


Fig. 12: *Production of ATP. Proton (H^+) fluxes across the inner membrane of mitochondria couple oxidative phosphorylation of carbon fuels (sugar, fatty acid) to ATP synthesis. Modified from [1].*

Oxidative phosphorylation is the culmination of a series of energy transformations that are called *cellular respiration*. First, carbon fuels are oxidized to yield electrons with high transfer potential. Then, this electron-motive force is converted into a proton-motive force and, finally, the proton-motive force is converted into phosphoryl transfer potential. The conversion of electron-motive force into proton-motive force is carried out by three electron-driven proton-pumps. These are large protein complexes spanning the mitochondrial membrane. The complexes contain multiple oxidation-reduction centers, including quinones, flavins, ion-sulfur clusters, and copper ions. The final phase of oxidative phosphorylation is carried out by *ATP synthase*. The enzyme complex is a molecular motor made of two operational units: a rotating component and a stationary component. The flow of protons back to the mitochondrial matrix through ATP synthase provides the force for the rotation. The rotation induces structural changes in the enzyme complex that result in the synthesis and release of ATP from the enzyme.

3 Summary

This chapter aimed to introduce the reader to some of the morphological and molecular properties of a living cell. Subcellular structures like the nucleus, the cellular membranes, mitochondria, or lysosomes have been described. Pivotal to all cellular functions is the presence and activity of specific proteins.

Proteins are synthesized from rather simple monomeric units, the amino acids. Proteins acquire their functional activity after folding in a characteristic three-dimensional structure. To a certain extent, the structural features of a protein are already determined by the genetic

information encoding the protein. The genetic information is stored in the DNA of the cell. The DNA is wound up as a nucleic acid/protein complex in the chromosomes. Transcription of the DNA produces molecular messengers required to "translate" the genetic information into a protein sequence. As a general rule, the flow of genetic information is from DNA $\rightarrow$ RNA $\rightarrow$ protein, but proteins in turn synthesize both DNA and RNA. All cellular processes need a continuous supply of energy. In biological systems, most of the energy is produced by hydrolysis of ATP. In mitochondria, the aerobic degradation of sugar or fatty acids to CO_2 and H_2O is coupled to the resynthesis of ATP. Though the biochemical and molecular mechanisms described in this section are known for reasonable time, biochemists and cell biologists are facing novel challenges. The advent of completely sequenced genomes led to the prediction of protein sequences with hitherto unknown function. Analysing these proteins and unravelling their function at the cellular level will be a tremendous challenge in forthcoming years.

A Appendix: Glossary

amino acid	monomeric unit of a protein
anticodon	three-base group on a tRNA molecule
ATP	adenosine triphosphate
ATP synthase	enzyme complex that forms ATP from ADP and phosphate
axon	long process of a neuron that conducts nerve impulses (= action potentials)
cholesterol	major sterol (= structural lipid) in animal tissues
christae	folding of the inner membrane of a mitochondrium
chromatid	one of two daughter strands of a chromosome
chromatin	complex of DNA and proteins constituting the eukaryotic chromosome
chromosome	DNA molecule and its associated proteins; stores and transmits genetic information
cytoplasm	portion of a cell's content outside the nucleus
cytoskeleton	filamentous network providing structure and organization to the cytoplasm
DNA	deoxyribonucleic acid
degenerate codons	two or more codons code for the same amino acid
elongation factor	specific proteins required in the elongation process of protein synthesis by ribosomes
endoplasmic reticulum	system of double membranes in the cytoplasm of eukaryotic cells
eukaryote	organism having a nucleus, chromosomes and internal organelles
flippase	lipid transporter catalysing the movement of lipid molecules between the two sides of a membrane bilayer
gene	chromosomal segment that codes for a functional protein or RNA
genome	all the genetic information encoded in a cell
glycolipid	lipid containing polar, hydrophilic carbohydrate head group
Golgi complex	complex membranous organelle of eukaryotic cells; functions in protein modification and secretion
histones	proteins rich in basic amino acids found in chromosomes
initiation factor	specific proteins required in the initiation process of protein synthesis by ribosomes
lipid raft	specialized membrane domains composed mainly of cholesterol and sphingolipids
lysosome	cellular organelle containing a large variety of hydrolytic enzymes
matrix	aqueous content of a mitochondrion with dissolved solutes
mitochondrion	organelle of eukaryotic cells; site of ATP production
motor neuron	neuron conducting electrical impulses from the spinal cord to muscles
nucleosome	structural unit for packaging chromatin
nucleus	cellular organelle containing the chromosomes
phospholipid	lipids containing charged, hydrophilic phosphate head groups
plasma membrane	physical barrier that surrounds the cell surface and encloses the cytoplasm
prokaryote	unicellular organism without a nucleus and intracellular organelles
promoter	DNA sequence at which RNA polymerase binds and initiates transcription
proton-motive	electrochemical potential inherent in a transmembrane gradient of H^+

force	concentration
replication	synthesis of daughter DNA strands identical to parental DNA strands
ribosome	small particles that are the site of protein synthesis
RNA	ribonucleic acid
termination factor	specific proteins required in the termination process of protein synthesis by ribosomes
transcription	enzymatic process whereby the genetic information contained in DNA is used to specify a messenger RNA molecule
transcription factor	specific proteins required in the initiation process of messenger RNA synthesis
translocation	movement of the ribosome along an messenger RNA molecule

References

[1] Berg, J.M., Tymoczko, J.L., and Stryer, L. "*Biochemistry*" 5th Ed. (W.H. Freeman & Co., New York, 2001). See especially chapters 5, 12, 14, and 29.
[2] Nelson, D.L. and Cox, M.M. "*Lehninger Principles of Biochemistry*" 5th Ed. (Worth Pub., New York, 2000). See especially chapters 2, 10, 11, and 19.
[3] Lodish, H.F., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D., and Darnell, J. "*Molecular Cell Biology*" 4th Ed. (W.H. Freeman & Co., New York, 1999).
[4] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. "*Molecular Biology of the Cell*" 4th Ed. (Garland Science Pub. Inc., New York, 2002).
[5] Kleinig, H and Sitte, P. "*Zellbiologie*" (Spektrum Wissenschaftlicher Verlag, Heidelberg, 1999).
[6] Gilbert, S.F. "*Developmental Biology*" (Sinauer Assoc. Inc., 2000).
[7] Watson, J.D. "*The Double Helix: A Personal Account to the Discovery of the Structure of DNA*" (Norton & Co. Inc., New York, 1981)
[8] Moore, P.B. (1998) The three-dimensional structure of the ribosome and its components. Annu. Rev. Biophys. Biomol. Struct. 27, 35-58.

TECHNIQUES AND METHODS

B

B 1 Scattering Experiments

W. Schweika
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Scattering experiments with neutrons and photons **3**
2.1 Principle of a scattering experiment 3
2.2 Energies and momentum 5
2.3 Interaction with matter - scattering amplitudes 7
2.4 Radiation Sources 9
2.5 Instrumentation 10

3 Scattering cross-section and correlation functions **12**
3.1 Pair-correlation functions 13
3.2 Integral scattering 14
3.3 Different sorts of scatterers 15
3.4 Coherent and incoherent scattering 15
3.5 Different probes 16

4 Examples and applications **17**
4.1 Structural correlations 17
4.2 Relaxations 21
4.3 Dynamic excitations 25

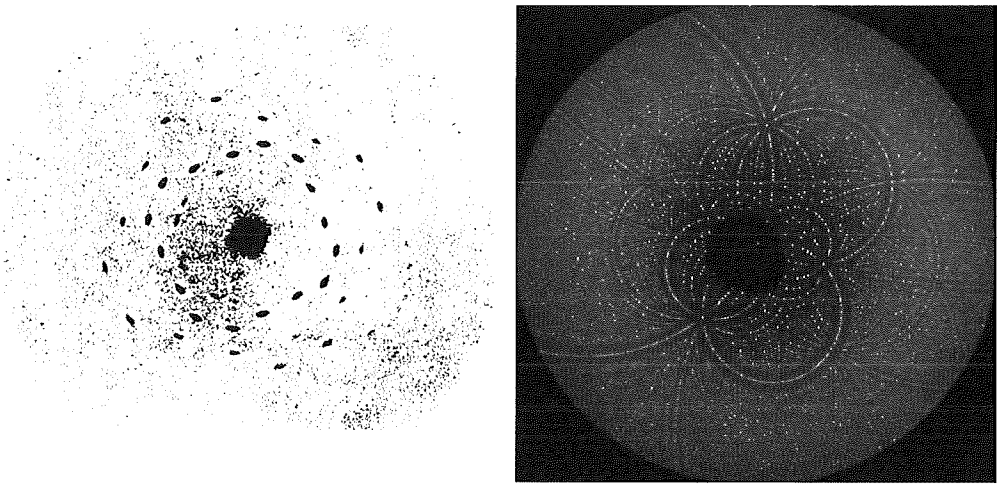


Fig. 1: *Left: one of the first X-ray interference patterns (film) of crystals observed by Friedrich, Knipping and von Laue (1912)[1]; right: a nowadays typical Laue diffraction image (CCD camera) of a much more complex system (cubic 2Mn catalase [[2]]).*

1 Introduction

The direct microscopic observation of smallest structures is limited by the wave length of visible light, whose spectrum ranging from the violet-blue 380 nm to red of about 760 nm wavelength and this is by far too large to explore details on finally atomic length scales. It was a major step in science early in the last century, when M. von Laue, W. Friedrich and P. Knipping proved that the newly discovered x-ray radiation can produce interference images from the atomic structure of crystals.

Meanwhile new sources, synchrotrons provide a much superior quality and quantity of x-ray radiation. The source brilliance has increased by 18 orders of magnitude compared to the former x-ray tubes. With the new Free Electron Laser for x-rays, (XFEL) project at DESY/Hamburg there will be a further increase of ten orders of magnitudes for the peak brilliance, which will permit to record such exposures as shown in Fig. 1 with a single bunch of 100 fs for time-resolved measurements of conformational transitions. The new oppurtunities will, however, also be accompanied with the challenging problem that the intense and hot keV x-ray beam can easily burn a sample.

The information content of scattering experiments, however, is different from real space images obtained in microscopy. Obviously, scattering patterns cannot be immediately understood without at least some education. We shall see that scattering intensities do not provide the structure itself but different quantities, *i.e.* the pair correlation functions in space and time. They tell us how particles or atoms "behave" with respect to another. The determination of the structure itself is the task of crystallography, and will be discussed in detail in the next lecture (B2). The loss of information about the phases of scattering amplitudes remains a central problem of structure determination. However, in many cases, the information about *typical* correlations in a system is more interesting than its *individual* structure. The reason is that correlations as obtained by scattering reflect the intrinsic interactions among the particles of the observed sys-

tem. The dynamic properties and characteristic time-correlations on atomic scales can also be accessed, and this has been for long times a particular domain of neutron scattering.
Neutrons are another important probe to study matter on atomic length scales. They are produced by research reactors or modern spallation sources. Neutrons are sensitive to different isotopes. In particular contrast variation techniques with selective H/D substitution are used for investigations of soft matter. Furthermore, neutrons are particularly gentle probe, non-invasive and non-destructive. In contrast to electrons, both neutrons as well as x-rays interact only weakly with matter. Therefore, the first Born approximation usually applies, and scattering intensities can be quantitatively analyzed using kinematic scattering theory. X-rays and neutrons differ in their properties in many respects, which is a reason for being complementary research tools.
Light scattering is an important laboratory method, which, – for transparent samples –, covers larger length scales, such as of colloids in solution. The underlying concepts of scattering as set theoretically by van Hove are very general. In principle, they can also be applied to elastic and inelastic scattering of other probes, like electrons or light. There is a large variety in experimental designs and methods, often highly specialized to interesting scales in space and time, and only an introduction to scattering and some overview can be given.

2 Scattering experiments with neutrons and photons

2.1 Principle of a scattering experiment

The scheme of a typical scattering experiment can be described as follows, see Fig. 2. A monochromatic beam, defined by a plane wave with wavelength λ, momentum $\hbar k$, and wave number $k = 2\pi/\lambda$, hits the sample. According to Huygens principle each particle of the sample may give rise to scattering of the incident wave. However, we shall assume that the scattering probability is so weak that we do not need to include multiple scattering processes. This assumption is also known as first Born approximation, which is the basis for kinematic scattering theory. In the simplest case of point-like scatterers (size $\ll \lambda$), the scattered waves can be described by spherical waves. Scattering intensities are measured by a (far distant) detector which covers a certain solid angle $r^2 d\Omega$ that is defined by experimental slits. The intensity is the modulus squared of the sum of all scattered waves. Therefore, the intensity is given by the sum of all pair-products of individual scattering amplitudes, *i.e.* the information on the pair-correlation of scattering centers. According to the law of conservation of momentum and energy, see Fig. 3, the energy transfer $\hbar\omega$ and momentum transfer $\hbar\vec{Q}$ between radiation probe and sample in the scattering process is

$$E_0 - E_1 = E_f - E_i = \hbar\omega \tag{1}$$

and

$$\hbar\vec{k}_0 - \hbar\vec{k}_1 = \hbar\vec{Q}. \tag{2}$$

Depending on the arrangement of the scattering centers in the sample, one may see interferences in the scattering patterns. Consider for example the Bragg scattering of an ideal crystal (see Fig. 4), in which constructive interference only occurs under the Bragg-condition

$$n\lambda = 2d\ \sin\ \theta \qquad \text{(Bragg's law)} \qquad Q = 2k\ \sin\ \theta \tag{3}$$

where the phase differences of scattered waves from different atomic layers separated by a distance d equal to λ (or a multiple of it). In a real experiment, the incoming beam will be

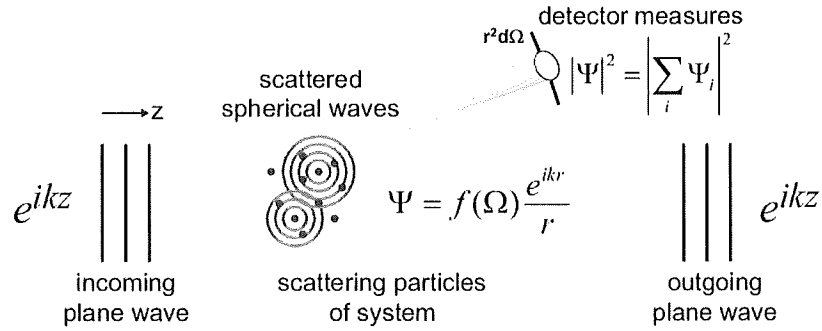


Fig. 2: *Scheme of a scattering experiment.*

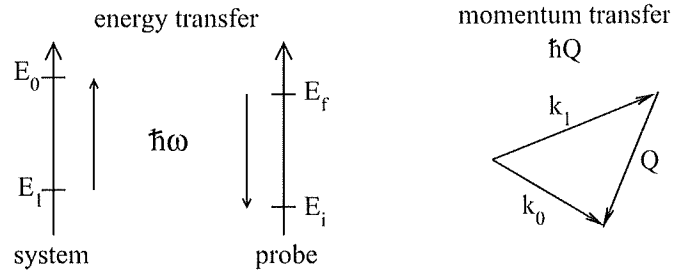


Fig. 3: *General case of inelastic scattering.*

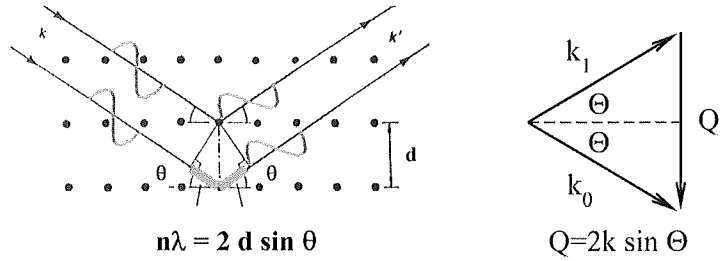


Fig. 4: *Bragg equation for constructive interference of elastic scattering from an ideal crystal. For a given wavelength λ, the larger the characteristic distance d in a sample, the smaller will be the scattering angle $\vartheta = 2\theta$ or (the modulus of the) scattering vector $\vec{Q}$.*

defined with a finite resolution $\Delta\lambda$. Backscattering techniques in x-ray and neutron scattering are used to achieve particularly high energy resolution. Differentiating the Bragg equation yields

$$\frac{\Delta E}{E} \propto \frac{\Delta\lambda}{\lambda} = \frac{\Delta d}{d} + \Delta\theta \cot\theta \qquad \text{or} \qquad \frac{\Delta Q}{Q} = \frac{\Delta k}{k} + \Delta\theta \cot\theta \tag{4}$$

for an ideal lattice neglecting a variation in spacings d. It shows how the resolution can be varied by the Bragg reflection of a crystal monochromator. A typical wavelength resolution that is also easy to achieve will be in the order of 10^{-2}. A much better resolution ($\approx 10^{-5}$) can be obtained in backscattering, for $\theta \rightarrow 90$. In this case, essentially the number of reflecting layers determines the resolution limit. Nowadays resolutions better than 10^{-7} have been achieved with synchrotron x-ray radiation using higher order Bragg peaks.

Coherence. Intensity is related to the square of scattering amplitudes. However, in order to observe constructive interference from many scattering particles, a sufficient coherence of the incoming beam is required. For a wavelength distribution $\lambda \pm \Delta\lambda$ the wave will be out of phase after a path of $n\lambda = (\lambda - \Delta\lambda)(n + 1/2) = l_\parallel$, where n is a positive integer number,

$$l_\parallel = \frac{\lambda^2}{2\Delta\lambda} \tag{5}$$

which provides an upper limit of the longitudinal coherence length $l_\parallel$. The transverse coherence length is reversely proportional to the angular divergence of the incident beam $\gamma \approx \sin\gamma = \Delta k_\perp / k$.

$$l_\perp = \frac{\lambda}{2\gamma} \tag{6}$$

This estimate is found from considering the Fraunhofer diffraction at a single slit of size $d = l_\perp$. In usual scattering experiments the coherence volume $V_{coh} = l_{\perp,x} l_{\perp,y} l_{\parallel,z}$ is much smaller than the sample size.

2.2 Energies and momentum

	x-rays (Photon) electromagnetic waves	neutron	electron particle waves	helium
rest mass	0	m_n	$m_e \approx m_n/1839$	$m_{He} = 4m_n$
charge	0	0	1	0
spin	1	1/2	1/2	0
energy	$E = c\,h/\lambda$		$E = h^2/\,2m\lambda^2$	
at $\lambda = 1$Å	12.39875 keV	81.81 meV	150 eV	20.45 meV

with
$\mathrm{c} = 2.9979 * 10^8 \mathrm{m/s}$, $\hbar = \mathrm{h}/2\pi = 1.0545 * 10^{-34}\mathrm{Js}$
$\mathrm{m_e} = 9.109389710^{-31}\mathrm{kg}$, $\mathrm{m_n} = 1.674928610^{-27}\mathrm{kg}$

The relations between energy and wave number or wave length are

$$E = \hbar kc \quad \text{in case of photons and } E = (\hbar k)^2/(2m) \quad \text{for a particle of mass } m. \tag{7}$$

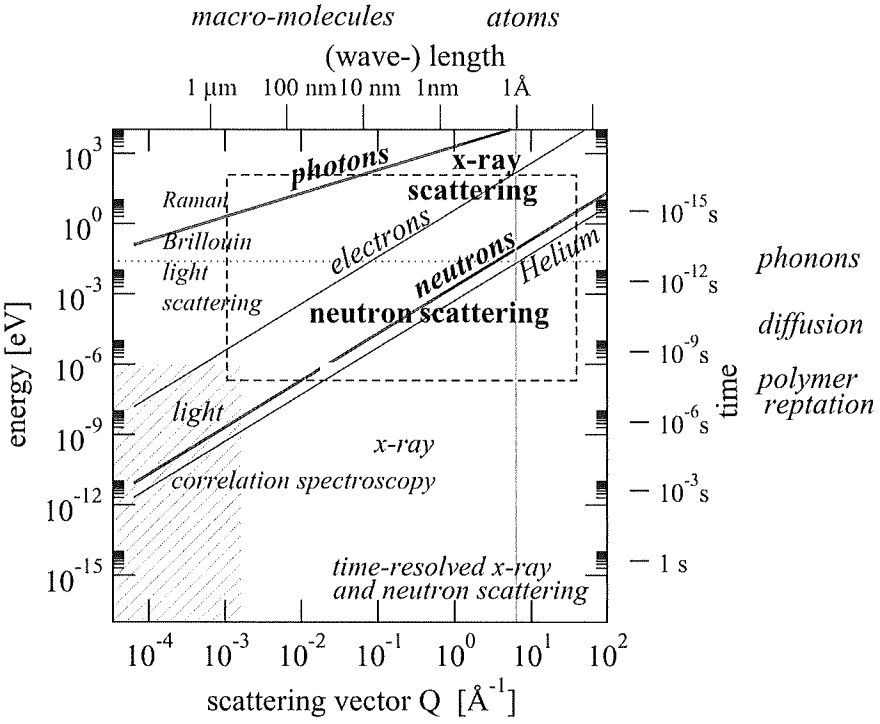


Fig. 5: *Photon and neutron dispersion and range in structures and dynamics.*

It expresses that electromagnetic waves can be described by massless photons which have particle properties a momentum $\vec{p} = \hbar\vec{k}$, and second, that particles like neutrons with a momentum $\vec{p} = mv = \hbar k$ have wave properties with a *de Broglie wavelength* $\lambda = 2\pi/k = h/(mv)$. The dispersion relation between energy and momentum $E(k)$ (see Fig. 5) is one key parameter to discuss the suitability of a probe for scattering experiments. We see that both neutrons with thermal energies and x-rays of $\approx 10\ keV$ have wavelengths in the order of Å ($= 10^{-10}\,m = 0.1\,nm$) which is required to resolve structures with atomic resolution. Of course, larger structures can also be explored with sufficiently good resolution at small scattering vectors Q (small scattering angles). (See Bragg equation and resolution consideration.) The accessible dynamic range of a probe is determined by its own energy and momentum and the achievable energy and momentum resolution. Compared to x-rays the energy of neutrons itself is much better adapted for investigations of thermal excitations like phonon vibrations or diffusion processes. While a typical experimental energy resolution set by a monochromator crystal may be in the order of 0.01, with backscattering techniques a resolution of 10^{-4} can be achieved, which yields about μeV resolution for neutrons. Using higher order reflections from perfect silicon at 3rd generation synchrotron sources a resolution of 10^{-7} sets a practical limit to about meV resolution for x-rays.

Scattering of charged particles, like electrons, is relatively strong. Therefore, the penetration depth is small. Since $m_e = 9.1091 * 10^{-31}\ kg$, the kinetic energy is about 150 eV at $\lambda = 1$Å. This is the reason why electron diffraction with such a wavelength is limited to surfaces (LEED).

Although with higher energies (MeV) and high resolution (10^{-5}) it is possible to measure bulk properties, as done in transmission electron microscopy.
Helium atoms with suitable wavelength and energy resolution have very weak penetrating power because of Coulomb interaction. This probe can be advantageously used for studies of surface properties, *e. g.* inelastic scattering studies of phonon modes at surfaces.
X-rays and in particular neutrons have a much larger penetration depth and they are important bulk probes. There are other techniques such as time-resolved scattering experiments, photon correlation spectroscopy and neutron spin-echo spectroscopy to explore slower dynamics, as characteristic for polymers, macromolecules, colloids, membranes *etc.* For transparent samples light scattering completes the accessible structural and dynamical range at small Q.

2.3 Interaction with matter - scattering amplitudes

The interaction of electromagnetic radiation with matter is quite complex. A free electron in the electric field of an electromagnetic wave will be forced by Coulomb interaction to an oscillatory movement in the oscillating electric wave field. Like in an antenna, the accelerated charge will become itself the source of radiation. The scattering cross-section for x-rays is usually given in electron units ($e.u.$) and refers to the *Thomson* cross-section for a free electron

$$\frac{d\sigma}{d\Omega}_{Thomson} = r_e^2 \, (\vec{e}_0 \cdot \vec{e}_1)^2, \tag{8}$$

where $r_e = 0.282 \cdot 10^{-12}\, cm$ is the *classical electron radius* that can be regarded as the scattering amplitude of the electron. Furthermore, the cross-section depends on the polarization $\vec{e}_0$ of the electric field of the incident wave with respect to the polarization $\vec{e}_1$ of the scattered wave. The atomic scattering amplitude, – in the limit of high photon energies compared to electronic binding energies –, is given by the Fourier transform of the electron density distribution $\rho(\vec{r})$,

$$b(\vec{Q}) = \frac{r_e}{e} \int_V \rho(\vec{r}) e^{i\vec{Q}\cdot r} d\vec{r}. \tag{9}$$

For small Q, the ratio b/r_e equals the atomic number Z. The absorption cross-section is essentially determined by the photo effect. Large steps can be seen near to the absorption edges, where resonant effects become important for both imaginary and real part of the scattering amplitude

$$b(Q, E) \approx b(Q) + b'(E) + ib''(E). \tag{10}$$

Although the change due to $b'(E)$ may be only of a few electron units, it provides the possibility to change and to highlight specific atomic scattering amplitudes, for most of the elements of the periodic table. With respect to soft matter and biological matter investigations, unfortunately, the binding energies of the light elements are very small (H: 16 eV, C: 284.2 eV, O: 543.1 eV).
Neutrons are elementary particles that decay with a half life time of about 15 minutes, $n \rightarrow p + e^- + \bar{\nu}$. Details of its internal structure, which is formed by three valence quarks and gluons, are not of relevance in the present context. There is the strong nuclear interaction between neutron and the nuclei of a sample, which is empirically described by the *Fermi pseudo potential*

$$V(\vec{r}) = \frac{2\pi\hbar^2}{m} \, b \, \delta(\vec{r}). \tag{11}$$

The delta-function applies since the size of the nuclei and of the neutron is negligible small (fm $= 10^{-15}$ m) compared to the wavelength λ (in the order of Å $= 10^{-10}$ m). Therefore, the

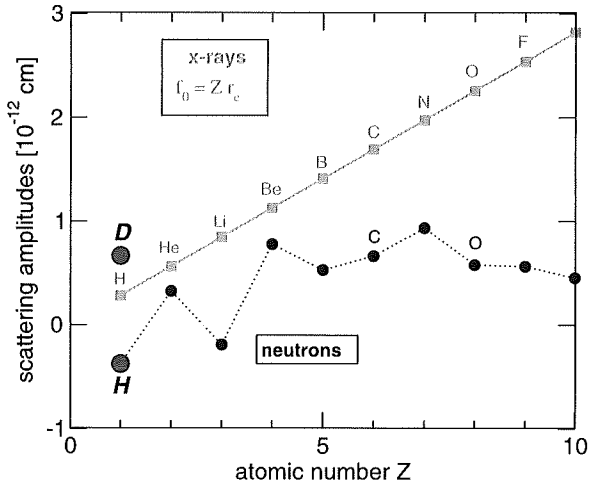


Fig. 6: *X-ray and neutron scattering amplitudes of light elements. Note the large isotope effect (H/D) of hydrogen for neutrons.*

scattered waves are spherical waves ($s-waves$) and the scattering amplitude are simple constants, named scattering lengths b, independent on $\vec{Q}$. The scattering lengths b vary in a not very regular manner for different elements, and some of them are even negative. In contrast to x-rays, neutrons are relatively well sensitive to light elements. Because of the nuclear nature of interaction, different isotopes of the same element will have different scattering lengths. Particularly, for hydrogen there is a huge contrast between the scattering lengths of the proton ($b_H = -0.3739 * 10^{-12}$cm) and the deuteron ($b_D = 0.6671 * 10^{-12}$cm), which provides an elegant mean for contrast variation techniques (exemplified further below). Since b_H is negative, it is even possible to create mixtures of solvents, where the average scattering length density disappears.

The neutron also has a spin, which gives rise to the neutron's magnetic moment. The nuclear interaction depends on the actual spin of the compound of neutron and nucleus of an atom. Since the spins in the nuclei of a sample are usually unpolarized, there will be a randomly fluctuating part of the scattering amplitudes, which gives rise to the so-called incoherent cross-section of neutrons. A further peculiarity of hydrogen is that its incoherent cross-section is the largest of all elements and isotopes. One should note that with polarized neutrons and polarization analysis of the scattered intensities, it is possible to separate coherent and incoherent cross-sections. Another aspect of relevance for research on magnetic properties, is, of course, that the magnetic moment of the neutron will interact with the magnetic moments of the electrons in the atomic orbitals, which results in significant magnetic scattering amplitudes.

In light scattering [7] the wavelengths are large compared to atomic sizes. Therefore, what matters is the dielectric response and the polarisibility of the medium at this length scale. The scattering amplitudes will be specific to the actual material under study rather than to elementary properties.

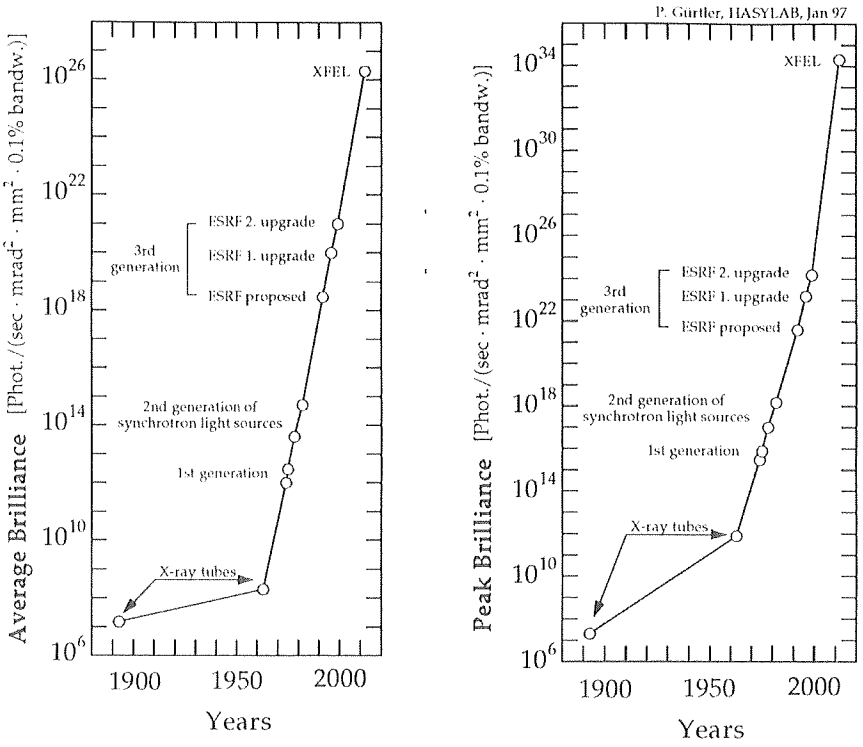


Fig. 7: *Brightness development of x-ray sources since the discovery of X-rays by W.C. Röntgen 1895.*

2.4 Radiation Sources

The development in x-ray flux both in quantity and quality is indeed remarkable, when comparing the x-ray tubes with modern *synchrotron* storage rings, see Fig. 7. These new sources offer a continuous spectrum of high brilliant and intense flux that is much superior to conventional laboratory x-ray tubes. The radiation is produced by the circular and thus accelerated movement of very fast and highly relativistic electrons (or positrons) in synchrotron storage rings. Relativistic effects are the high collimation of this radiation in the direction the circumventing electrons, and the polarization in the plane of the electron orbit, and elliptically out of plane. There is also an inherent time structure (in the 100 ps range), which is related to the ring properties and its cavities, the accelerating devices. There are further devices, wigglers and undulators, to increase the intensity and brilliance of the radiation. They consist of periodic magnetic arrays that force the electrons onto a sinusoidal path. Each curvature N contributes to the emitted radiation intensities (wiggler device). In undulators the superposition of these waves is in phase and shows constructive interference for certain wavelengths $\hat{\lambda}$ and an increase of $I(\hat{\lambda})$ proportional to N^2. The same principle will be also the basis for the Free Electron Laser project at DESY (Hamburg).

The neutrons that we use as probes in condensed matter research are produced in research

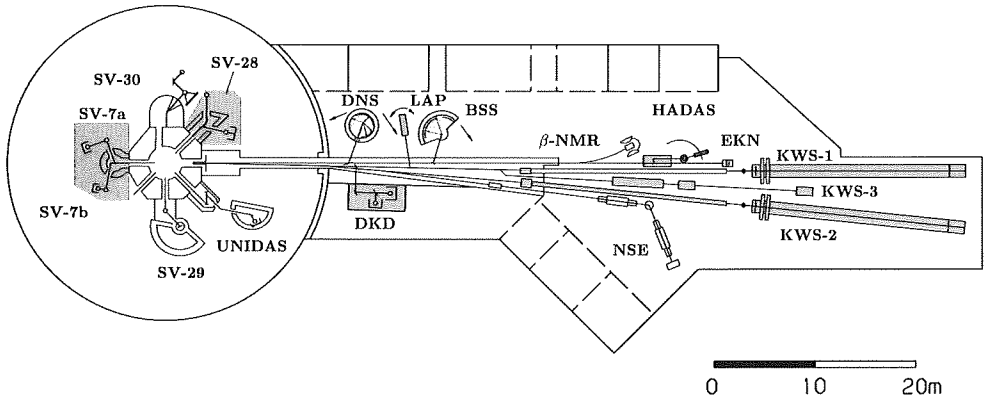


Fig. 8: *Neutron scattering experiments at the Jülich research reactor.*

reactors, for instance at the Institut Laue Langevin (ILL) or here at Jülich. *Neutron research reactors* are optimized for high neutron flux densities (thermal isotropic flux $10^{15}\, n/(cm^2\, s)$ and $2.5 * 10^{14}\, n/(cm^2 s)$ respectively). However, with the more than 30-years-old ILL-reactor, the reactor technology has already reached a practical limit for neutron fluxes. Nowadays, further progress is made by the new spallation sources that are currently built in Japan and USA. Unfortunately, the even more ambitious ESS (European Spallation Source) is pending for a positive political decision. These spallation sources produce pulsed neutron beams which is obviously beneficial for time-of-flight instruments but also for many new types of instrumentation. In a nuclear spallation reaction high energy protons (GeV) hit targets of heavy elements (W, Pb, Hg, etc). The expected gains ($\sim 10^2$) in intensities and efficiencies may appear modest compared to the ongoing increases for X-ray sources. However, one might view it like that: it will be the last and very valuable major step forward for neutron sources, and these places will leading neutron science for the next several decades.
Since neutrons are produced in nuclear reactions, their initial energies are very high (MeV). Therefore, neutrons are cooled down to thermal energies in moderators, the cooling water (H_2O or D_2O) of the reactor core. The wavelength distribution can be further modified and adapted for the instrumentation using in special hot (graphite $\sim 1000°\, C$) or cold moderators (liquid H_2, 20K). Through beam-holes neutrons are guided to the scattering experiments. By total reflection they can be transported in *neutron guides* over longer distances to the various experimental stations. The experimental scattering facilities at the Jülich research reactor are shown in Fig. 8.

2.5 Instrumentation

Many x-ray beamlines for scattering experiments might look quite similar at first glance. Along the beam path one finds a crystal monochromator, a mirror, a sample station with several degrees of freedom to orient the sample and a detector. Mirror systems are used to select by total reflection the desired wavelength set by the monochromator but to suppress any possible contributions from higher harmonics λ/n. Beam sizes are small and the natural collimation given by the synchrotron is already very good. Lenses and mirrors may be used to produce intense beams

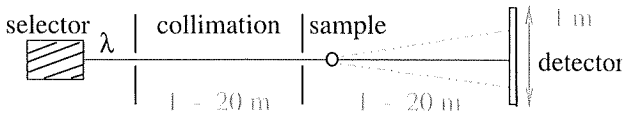


Fig. 9: *Scheme of small angle scattering instruments.*

in a micro-focus for very small samples or to view small parts of samples only. While in the early days of x-rays data were taken by films, the request for absolute or quantitatively comparable intensities has introduced single detectors (ionization chambers, diodes, semiconducting detectors). Nowadays image plates and CCD cameras become more and more important for providing quantitative data in simultaneous measurements of large sections in Q-space. This is of particular relevance, when moving detector motors is more time consuming than actual counting.

The isotropic neutron flux is comparatively weak, and any trimming to the various resolution elements has to be optimized with care not to loose too much of the precious intensities. Therefore the instrumentation for neutrons is quite diversed. There are typically very specialized instruments, as shown for example in Fig. 8: classical triple axis instruments (UNIDAS and SV-30), multi-detector instrument for powder diffraction (SV-7), (4-circle) single crystal diffraction instruments (SV-28), the long small angle scattering instruments (KWS-1, KWS-2, KWS-3), backscattering instrument (BSS), reflectometer (HADAS), thermal and cold time-of-flight instruments (SV-29 and DNS), neutron spin-echo (NSE) and others.

Small angle scattering instruments are one of the most used instruments, particularly for investigations of soft matter. With their aid we explore structural inhomogeneneities on larger than atomic length scales (10 to 1000 Å). An instrument for neutron small angle scattering experiments follows a fairly simple principle. Typically, one chooses a relative broad wavelength band with an average wavelength of several Ångstroem from the cold neutron spectrum by use of a rotating velocity selector. In the next step the neutrons are well collimated by slits before they hit the sample. The scattered neutrons are detected under small angles by a 2-dimensional position sensitive detector at appropriate large distances from the sample. Flight paths are evacuated to reduce absorption and background scattering. The set-up measures diffraction, energy integrated elastic intensity which contains only the structural information given by the instantaneous spatial correlations on length scales that are much larger than near atomic distances.

Triple axis spectrometers (TAS) consist of a monochromator preparing the incident beam by Bragg reflection (1. axis), sample (2. axis), and analyzer crystal (3. axis) to determine the final wavevector and energy.

Thermal and cold neutrons have velocities in the order of 1000 m/s. By measuring the time of flight for an initially pulsed monochromatic neutron one obtains the energy dependence of the scattered neutrons. With area detectors covering a large solid angle such time-of-flight (TOF) instruments are very efficient to measure extensively regions in (Q,ω)-space.

High energy resolution ($\sim \mu$eV) is achieved by backscattering (consider Eq. 4 for $2\Theta \rightarrow \pi$). The energy transfer at the sample is matched by the wavelength shift of the moving monochromator (Doppler effect), similar to principles of Mössbauer spectroscopy.

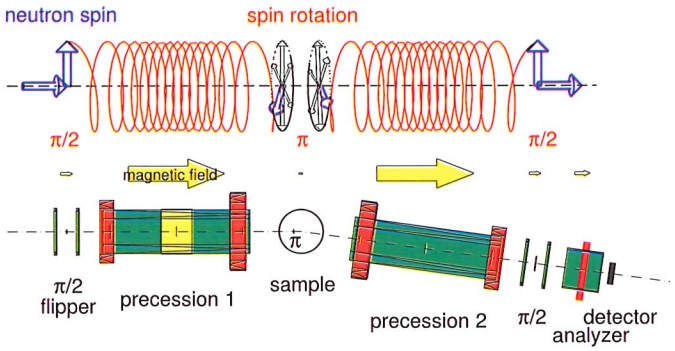


Fig. 10: *Scheme of neutron spin-echo (NSE): Larmor precessions of the neutron spins depend on the neutron velocity. For equal magnetic fields in both precession coils and elastic scattering the precession is symmetric in time using the π-flipper for time reversal (upper part); the NSE set-up at Jülich (lower part).*

Neutron spin-echo (NSE) technique achieves the best dynamic resolution ($\sim 0.1\mu eV$) to study particularly slow dynamics. The scheme of a NSE spectrometer is depicted in Fig. 10. The idea is to measure the energy transfer in the sample and velocity change of the neutron via the Larmor precessions of neutrons in a magnetic field. Considering the case of elastic scattering, the precession in the magnetic field of the first coil will be exactly compensated by the reverse field of the second coil and the final polarization will be the same as and the *echo* of the initial polarization. In case of inelastic scattering neutrons spent longer or shorter times in the secondary coil, which can be compensated for by tuning an additional magnetic field, and the echo will appear again for symmetric Larmor precessions. The clue is that spins work as individual clocks of the neutrons. This allows the use of a broad incident wavelength spectrum with high intensity *and* high resolution. Lateral field homogeneity is of importance and achieved by Fresnel correction coils. Divergence is decoupled from energy resolution and can be adapted to the Q-resolution. Note that for practical convenience, instead of opposite fields, a "π-flipper" near to the sample inverts the precession angle, which has the effect of a time-reversal.

3 Scattering cross-section and correlation functions

This chapter will describe the relation between scattering and correlation functions. In the following we shall consider only a description of scattering within the *kinematic approximation*, or *first Born approximation*, we assume that multiple scattering do not have any practical relevance. This condition is in general not fulfilled for the scattering of electrons. For neutron and x-ray scattering this condition is usually fulfilled apart from Bragg scattering of relative perfect crystals or in case of total reflection from surfaces. In such cases we would need to apply dynamic scattering theory.

Within the first Born approximation the double differential scattering cross-section per atom

can be written as

$$\frac{d^2\sigma}{d\Omega dE'} = \frac{k'}{k}\frac{1}{(2\pi\hbar)}\int \langle \frac{1}{N}\sum_{l,l'} b_l b_{l'} e^{-i\vec{Q}\cdot\vec{R}_l(0)} e^{i\vec{Q}\cdot\vec{R}_{l'}(t)}\rangle e^{-i\omega t} dt \tag{12}$$

where the double sum is taken over all scattering amplitudes b_l with their phase factors $e^{i\vec{Q}\cdot\vec{R}_l(t)}$; $\hbar\omega$ is the energy transfer between probe and system. We may rewrite Eq. 12 into a product of two factors, one that is related only to properties of the probe, and a second that characterizes only the properties of the sample. Therefore, we consider first the simplest case of identical scatterers $b = b(\vec{R}_l)$. The average defines the time dependent pair correlation function $G(\vec{r}, t)$, where $\vec{r}_{ll'} = \vec{R}_l(0) - \vec{R}_{l'}(t)$. The average includes a configurational average with respect to the coherence volume defined by experimental parameters, a time average during the data collection, and a thermal average of the unperturbed system. Replacing the double sum by an integral representation, and introducing difference vectors, one obtains

$$\frac{d^2\sigma}{d\Omega dE'} = b^2\frac{k'}{k}(2\pi\hbar)^{-1}\int\int G(\vec{r}, t)\; e^{i\vec{Q}\cdot\vec{r}}\; dr\; e^{-i\omega t} dt. \tag{13}$$

The Fourier transform of the pair correlation function $G(\vec{r}, t)$ is called the intermediate scattering function $S(\vec{Q}, t)$,

$$\frac{d^2\sigma}{d\Omega dE'} = b^2\frac{k'}{k}(2\pi\hbar)^{-1}\int S(\vec{Q}, t) e^{-i\omega t} dt. \tag{14}$$

The Fourier transform in time of $S(\vec{Q}, t)$ yields the dynamic scattering function $S(\vec{Q}, \omega)$,

$$\frac{d^2\sigma}{d\Omega dE'} = b^2\frac{k'}{k}\hbar^{-1} S(\vec{Q}, \omega). \tag{15}$$

3.1 Pair-correlation functions

There is a clear classical interpretation of pair-correlation function $G(\vec{r}, t)$.
It is the probability that there is a particle at place $\vec{r}$ and time t, when there has been a particle at place 0 and time 0.
The pair-correlation function provides a suitable description for instance of the structural properties of a crystal, the diffusion of a particle, or of lattice vibrations (phonon modes). In the first case, we shall describe the equilibrium positions by the time average of the pair correlation function $G(\vec{r}, t)$. This will be essentially $G(\vec{r}, \infty)$. The scattering contribution will be purely *elastic*, $S(\vec{Q}, \omega) = S(\vec{Q}, 0)$. The periodic motion of lattice vibrations causes *inelastic* scattering, $S(\vec{Q}, \omega)$ with $\omega \neq 0$. If we observe relaxation processes, for instance, damping of a vibration due to friction, their signature will be an intrinsic finite line width in ω. Similarly, in case of diffusing particles there will be *quasi-elastic* broadening of the scattering $S(\vec{Q}, \omega \approx 0)$ instead of a delta function $\delta(\omega)$ ("Doppler effect").
Detailed balance. At low temperatures it is still possible by scattering to create an excited state in the sample, however, since the probability for excited states will be low, vice versa any energy transfer from sample to the probe will become more and more unlikely. The asymmetry of intensities in energy gain and loss can be described by the condition of *detailed balance*

$$S(\vec{Q}, \omega) = exp(\frac{\hbar\omega}{kT}) S(-\vec{Q}, -\omega). \tag{16}$$

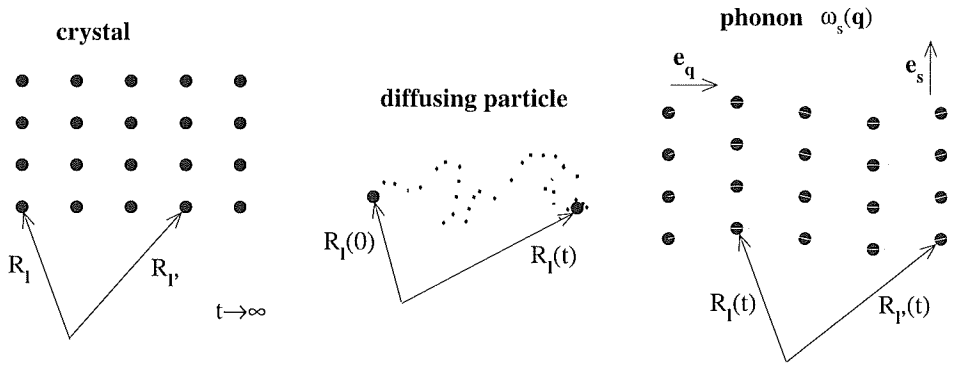


Fig. 11: *Left: Time-independent pair-correlations of crystals can be described in terms of unit cell (here one atom only) and elementary translations (d); elastic Bragg scattering occurs at* $nQ = 2\pi/d$.
Middle: Diffusive motion of a single particle can be described by the self-correlation function $G_s(\vec{r}, t)$*, and leads to quasi-elastic scattering.*
Right: Lattice vibrations (here a transverse phonon mode is shown) can be defined by time-dependent pair-correlations, and give rise to inelastic scattering. In all cases, directions of correlations are probed in scattering via the scalar product $\vec{Q} \cdot \Delta\vec{R}$.

The above classical interpretation of $G(\vec{r}, t)$ is based on the symmetry of $S(\vec{Q}, \omega)$, which means $\hbar\omega \ll kT$. Otherwise $G(\vec{r}, t)$ will not be anymore a real but complex function. Using the *detailed balance* condition as a (quantum mechanical) correction, we may still apply the classical interpretation of the pair-correlation function $G(\vec{r}, t)$.

3.2 Integral scattering

One may note that the term elastic is often defined with respect to experimental resolution. In x-ray scattering one usually integrates over all thermal excitation energies. Inelastic scattering is often related to transitions in the electronic system. With respect to the correlations among the many atoms of a structure, one measures the static structure factor

$$S(\vec{Q}) = \int_{-\infty}^{+\infty} S(\vec{Q}, \omega)\ d\omega. \tag{17}$$

$S(\vec{Q})$ is the Fourier transform of the instantaneous pair-correlation function $G(\vec{r}, 0)$, a snapshot of the correlations in the structure. $G(\vec{r}, 0)$ is also known as the Patterson function. In liquids all atoms move. Therefore, the true elastic scattering $S(\vec{Q}, 0)$ disappears but $S(\vec{Q})$ remains meaningful quantity.

3.3 Different sorts of scatterers

Scattering of different atoms or particles μ, ν can be introduced by labeling all scattering amplitudes

$$\frac{d^2\sigma}{d\Omega dE'} = \sum_{\mu\nu} b_\mu b_\nu \frac{k'}{k}(2\pi\hbar)^{-1} \int\int G_{\mu\nu}(\vec{r},t)\; e^{i\vec{Q}\cdot\vec{r}}\; dr\; e^{-i\omega t} dt. \tag{18}$$

This implicite definition of $G_{\mu\nu}(\vec{r},t)$ still includes correlations of dinctinct as well as identical types μ, ν. The part which describes the correlation between atoms or particles of *distinct* character, $G^D_{\mu\nu}(\vec{r},t)$ becomes visible in scattering accordingly to the difference in scattering amplitudes of types μ, ν. The average of the scattering amplitudes does not distinguish different particles but leads to scattering even if $b_\mu = b_\nu$. Since the concentration fluctuation is accompanied by a scattering amplitude fluctuation, $\Delta b = b_\mu - b_\nu$ the scattering weight for the scattering cross-section of different sorts of scatterers will be given by the square of the contrast of two scattering amplitudes, $|\Delta b|^2 = |b_\mu - b_\nu|^2$ and by the product of the concentrations c_μ and c_ν:

$$\frac{d^2\sigma}{d\Omega dE'_{(\mu>\nu)}} = \frac{k'}{k}\hbar^{-1} \sum_{\mu\nu} c_\mu c_\nu |b_\mu - b_\nu|^2 S^D_{\mu\nu}(\vec{Q},\omega). \tag{19}$$

In some cases there may be a feature in the scattered intensities that can be obviously related to a specific pair-correlation function. However, in general we shall encounter the problem that a single measurement is insufficient to understand a system of many components. The combined use of different probes altering the scattering contrast can be a valuable tool. Particularly important for scattering investigations of organic materials is the possibility of contrast variation for neutrons using the hydrogen isotopes H and D. An example will be given further below.

3.4 Coherent and incoherent scattering

A special case of Eq. 19 would be a completely random distribution of different scatterers b_μ and b_ν. In this case $S_{\mu\nu}(\vec{Q},\omega)$ reduces to the Fourier transform in space and time of only the *self-correlation function*. In x-ray scattering such an example is known from the Laue scattering of a random alloy.

Neutron scattering amplitudes are both specific to the nuclear isotope, and dependent on the spin of the scattering compound neutron-nucleus. Each of the two properties gives rise to randomness and to fluctuating parts of scattering amplitudes and to scattering intensities that are only related to self-correlation functions. Therefore, the scattering cross-section for neutrons is usually decomposed into two parts, the *coherent* and *incoherent* scattering

$$\frac{d^2\sigma}{d\Omega dE'} = \frac{k'}{k}\frac{1}{(4\pi\hbar)}\{\sigma_{coh} S_{coh}(\vec{Q},\omega) + \sigma_{inc} S_{inc}(\vec{Q},\omega)\}, \tag{20}$$

where $\sigma_{coh} = 4\pi\overline{b}^2$ and $\sigma_{inc} = 4\pi(\overline{b^2} - \overline{b}^2)$ denote the coherent and incoherent scattering cross-sections respectively. Note the equivalence $\sum_{\mu\nu} c_\mu c_\nu |b_\mu - b_\nu|^2 = \overline{b^2} - \overline{b}^2$. It can be seen that the *incoherent* scattering is a special case of Eq. 19.

The integral *incoherent scattering* $S_{inc}(\vec{Q})$ (snapshot) is isotropic

$$S_{inc}(\vec{Q}) = \int_{-\infty}^{+\infty} S_{inc}(\vec{Q},\omega)\; d\omega = 1. \tag{21}$$

Incoherent scattering is used to study the dynamics of individual scatterers, for example the self-diffusion of atoms or molecules in liquids or solids. For unbound translational diffusion with random walk properties, the probability to find the particle again at the same site decreases exponentially with time. Hence the Fourier transform of the self-correlation function $G_s(\vec{r},t)$ yields a *Lorentzian* form of the energy dependence. The dynamics in real systems may reflect a variety of other characteristics, ballistic behavior of atoms in liquids on very short time scales, bounded motions at interfaces, reptation of long polymer chains etc. The energy width of the (incoherent) scattering, in the asymptotic limit of small Q, can be related to the macroscopic (self) diffusion coefficient. With neutrons and isotopic substitution of H and D, it is also possible to emphasize the incoherent and coherent scattering intensities respectively. Last but not least one should mention the possibility of polarization analysis for neutron scattering. Using polarized neutrons and polarization analysis the coherent part of the scattering can be quantitatively separated from the (spin) incoherent part.

3.5 Different probes

Neutrons clearly offer a number of unique and interesting possibilities in scattering investigations of condensed matter. The importance of the van Hove concept of scattering and correlation functions is, however, that it applies in general also for other scattering probes, provided that the kinematic approximation is fulfilled. Since $\frac{d\sigma}{d\Omega}_{probe}$ is separable, within this approximation we observe the dynamic response $S(\vec{Q},\omega)$ of the unperturbed system.

$$\frac{d^2\sigma}{d\Omega dE'} = \frac{d\sigma}{d\Omega}_{probe} \frac{k'}{k}\hbar^{-1}S(\vec{Q},\omega) \tag{22}$$

where

$$\frac{d\sigma}{d\Omega}_{neutron} = b^2 \tag{23}$$

$$\frac{d\sigma}{d\Omega}_{electron} = \frac{4(Z-f(Q))^2}{Q^4}\left(\frac{m_e e^2}{\hbar^2}\right)^2 \quad \text{(Rutherford cross section)} \tag{24}$$

$$\frac{d\sigma}{d\Omega}_{x-rays} = (\vec{\varepsilon}\cdot\vec{\varepsilon})^2\frac{e^2}{m_e c^2} \quad \text{(Thomson cross section)} \tag{25}$$

$$\frac{d\sigma}{d\Omega}_{light} \propto |\alpha|^2 \quad (\alpha \text{ polarisibility}) \tag{26}$$

The *Thomson* cross-section is the differential cross-section for x-ray scattering on a single electron charge. $e^2/(m_e c^2) = r_o \approx 2.82 \times 10^{-13}$cm, the *classical electron radius*. X-ray scattering of an atom is described in terms of atomic form-factors. In the high energy limit, where the energy of the incident x-ray wave E_0 is large compared to electron binding energies, the atomic form-factor $f(Q)$ represents the Fourier transform of the electron density distribution of the atom. In forward scattering at low Q, $f(Q \to 0) = Z$, where Z is the atomic number or number of electrons.

The *Rutherford* cross-section for electrons depend on the proton number Z and also on the atomic form-factor f(Q). $(m_e e^2/\hbar^2)^{-1} = a_0 \approx 0.529$Å, the Bohr radius.

4 Examples and applications

4.1 Structural correlations

For structural investigations one often measures only the integral scattering function $S(\vec{Q})$, that is related to the structure as seen by instantaneous pair-correlations. For the discussion of liquids it is also convenient to represent the pair-correlation function in terms of the density-density correlation $\langle n_{\vec{R}} n_{\vec{R}'} \rangle$

$$S(\vec{Q}) = \frac{1}{N} \langle \sum_{l,l'} e^{i\vec{Q}\cdot(\vec{R}_l - \vec{R}_{l'})} \rangle = \frac{1}{N} \int d\vec{R} d\vec{R}' \langle n_{\vec{R}} n_{\vec{R}'} \rangle e^{i\vec{Q}\cdot(\vec{R}-\vec{R}')}. \tag{27}$$

It is further advisable to decompose the actual density $n_{\vec{R}}$ into its average $\langle n_{\vec{R}} \rangle$ and its fluctuations $n_{\vec{R}} - n$. The substitution $\langle n_{\vec{R}} n_{\vec{R}'} \rangle = n^2 + \langle n_{\vec{R}} n_{\vec{R}'} \rangle - n^2$ leads to an appropriate separation of scattering

$$S(\vec{Q}) = N\delta(\vec{Q}) + n \int d\vec{r}\; G(\vec{r}) e^{i\vec{Q}\cdot\vec{r}}. \tag{28}$$

In case of a **continuum** without any fluctuations, $\langle n_{\vec{R}} n_{\vec{R}'} \rangle = n^2$,

$$S(\vec{Q}) = N\delta(\vec{Q}). \tag{29}$$

There is no true scattering (only contributions to the transmitted beam for Q=0).
Considering ideal **crystals**, the Fourier transform of a periodic structure in real space will be a periodic structure in Q-space, where $\vec{G}_{hkl}$ denotes a reciprocal lattice vector,

$$S(\vec{Q})_{hkl} = N\delta(\vec{Q} - \vec{G}_{hkl}). \tag{30}$$

An **ideal gas** is the simplest example of matter with discrete or discontinuous nature. All distinct pair-correlations vanish and only the self-correlation function ($l = l'$ term) contributes to $G(r)$. The Fourier transform of the delta function in real space leads to a constant in Q-space,

$$S(Q) = N\delta(Q) + 1. \tag{31}$$

The scattering function for **liquid** or **real gas** includes non-vanishing contributions of distinct pair-correlations

$$S(Q) = N\delta(Q) + 1 + 4\pi n \int r^2 dr\; (G_d(r) - 1) \frac{sin\; Qr}{Qr}, \tag{32}$$

where the third term only describes the contributions from the *distinct* pair-correlations $G_d(r)$, which can be obtained by Fourier transformation. For instance typical nearest neighbor distances and characteristic coordinations of atoms can be determined this way. Another feature of the scattering of liquids or amorphous materials, which are *homogeneous* on large scales, is the asymptotic weak scattering for small Q. It can be related to the macroscopic compressibility.

Liquid and amorphous polymers.

The microscopic local conformation of polymers should be relevant for macroscopic material properties. Therefore, the understanding of structure and function of polymer elements are of

interest. Fig. 12 shows the scattering of PMMA (Polymethylmethacrylate) with two different isotopic substitutions in the methyl-groups. By contrast variation different weights are put on the correlation functions of PMMA, which explain the observed peak shifts. Polarization analysis has been used to separate coherent and incoherent scattering. The coherent scattering is purely non-spin flip scattering, while 2/3 of the incoherent scattering is spin-flip scattering. This can be important if the incoherent scattering contribution from H is large.

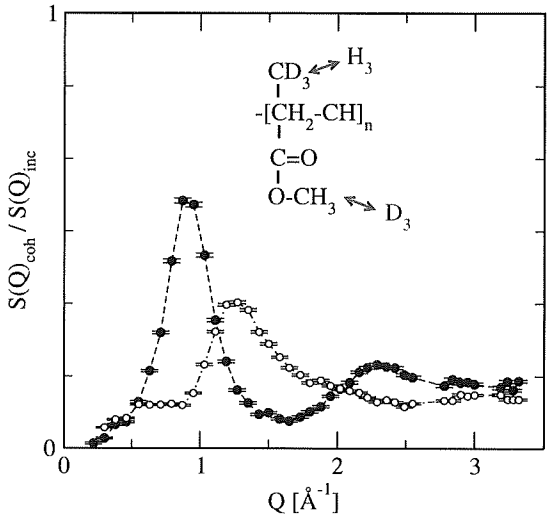


Fig. 12: *Coherent scattering of PMMA at 300 K with typical features of scattering from homogeneous liquids or amorphous solids in the range of correlations among near atoms* $Q \sim$ Å^{-1}. *The scattering at small Q is weak as it can be related to compressibility. Isotopic substitution replacing H and D (grey symbols) yield a different emphasis and information about the pair correlations in the sample. Polarization analysis has been used to separate coherent and incoherent scattering (Data of Arbe* et al. *measured on the DNS instrument).*

The more general case has to include systems that exhibit structural **inhomogeneities on larger length scales**. For example one may consider the morphology of a phase separating system like oil and water mixtures discussed further below. It will be usual and more convenient to describe the scattering in terms of scattering length (or amplitude) density $\rho = nb$, where n is the particle density.

In 3.5 we have tacitely used the convolution theorem to separate the probe and system dependent parts of the scattering cross-section. It states that the sum or integral over objects in real space which are decorated with the same shape function can be described by the product of the two Fourier transforms of object and shape. For instance the x-ray atomic form-factor is the Fourier transform of the atomic electron density distribution, which is usually an easily separable and a less interesting quantity.

Likewise for a solution one might wish to distinguish the correlation among different particles from contributions of the shape morpholoy. However, it could also be of interest to verify a specific shape model. The scattering cross-section for a random distribution of particles occuring

with a concentration of c_p is given by

$$\frac{d\sigma}{d\Omega}(\vec{Q}) = c_p(1-c_p)\Delta\rho^2 V_p^2 S_p(\vec{Q}), \tag{33}$$

where $\Delta\rho = \rho_p - \rho_{matrix}$. The scattering function of the particle $S_p(\vec{Q})$ is related to the modulus of the squared form-factor, the scattering amplitude of the particle

$$S_p(\vec{Q}) = |f_p(\vec{Q})|^2 \tag{34}$$

where $f(Q)$ is normalised to the forward scattering (*id est* the integrated scattering length density), and given by

$$f(\vec{Q}) = \frac{\int_V dr' \rho(\vec{r}') e^{i\vec{Q}\cdot\vec{r}'}}{\int_V dr' \rho(\vec{r}')}. \tag{35}$$

For a homogeneous sphere one obtains for instance

$$f(Q)_{sphere} = \frac{3(\sin\ QR - QR\ \cos\ QR)}{(QR)^3}. \tag{36}$$

The asymptotic behavior for large Q is proportional to Q^{-4}. This characteristic Q^{-4} dependence is known as the *Porod law* and it is valid in general for structures with smooth interfaces. The scattering intensity that stems from the contrast at a smooth bulk surface to air or vacuum shows the same asymptotic Q-dependence, also known as Fresnel reflectivity. Rough interfaces show scattering functions that are faster decaying with Q. Fractal structures, i.e. volume fractals $V \propto r^{d_V^f}$ and $d_V^f < 3$ or surface fractals with d_S^f show a different behavior. Scattering functions of *random chains* show a Q^{-2} asymptotic behavior (Debye model for polymers).
Furthermore, in the limit $Q \rightarrow 0$, where S(0)=1, the scattering of dilute particles can be directly related to the particle volume and the radius of gyration R_g. At small Q, for $QR_G < 1$, the so-called *Guinier approximation* $S(Q) \approx exp(-1\frac{}{3}Q^2R_G^2)$ holds. In case of non-dilute mixtures, correlations among the particles become more and more important. For instance repulsive interactions would lead to a decrease of scattering in forward direction, similar to the case of homogeneous solutions with low compressiblity.

In a recent neutron small angle scattering study the bicontinuous **structure of a micro-emulsion** of oil and water has been investigated. It was found that the interfacial membrane made of a non-ionic surfactant can be modified by small amounts of amphiphilic block copolymers with a boosting effect on emulgating efficiency of the surfactant. The microscopic picture of this system with high technological interest shown in Fig. 13 evolves from *neutron small angle scattering experiments* applying a sophisticated *contrast variation technique*, which will be discussed in further detail. Spatial correlations of objects that are large compared to the wavelength of the scattering probe will show characteristic interference features at small scattering angles as follows from Eq. 3 (*Bragg's law*). It is apparent that if there is no scattering contrast for the constituents, there will be no scattering signal. Scattering of a binary mixture is uniquely described by only a single independent pair-correlation function. (In optics known as Babinet's principle. We neglect small effects due to the inherent discrete atomic structure and compressibility.) Refering to the example of a quaternary mixture, a single measurement will yield only intensity which sums up the contributions of all partial scattering functions, where each of it is related to a specific pair-correlation function. A separation is only possible in experiments with different scattering contrasts.

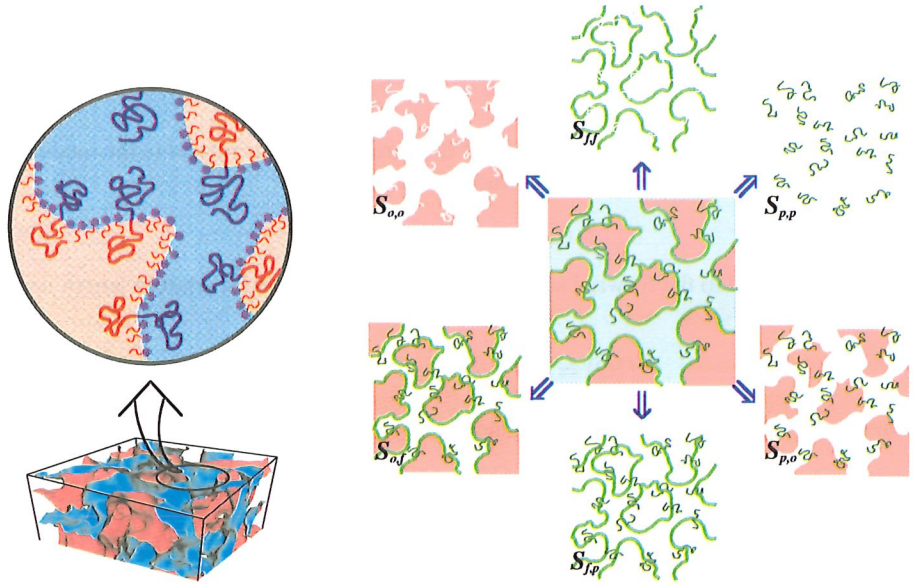


Fig. 13: *Schematic illstration of a bicontinuous microemulsion structure of oil and water. Additional amphiphilic block copolymers are distributed uniformly on the surfactant membrane. Right side illustrates the contrast variation by H/D substitution for specific pair-correlations of the constituents oil,water, film, and polymer.*

Whenever chemical synthesis allows for isotopic substitution of Hydrogen by Deuterium it provides an elegant way to highlight specific desired contrasts. Deuterium has a rather similar scattering length to C and O but differs distinctly from the negative scattering amplitude of hydrogen. Hence, it is possible to match the contrast of all but one of the constituents to reveal its characteristic correlation function. This approach has been systematically applied by in SANS experiments by Endo *et al.*, see Fig. 14.
The part of the intensity that is related to the average scattering amplitude will be negligibly small as it describes only the forward scattering and the weak scattering related to compressibility (see the homogeneous case above). For small Q, $Q < 2\pi/(\text{atomic spacings})$, the expression

$$I(Q) = \sum_{\mu,\nu} \rho_\mu \rho_\mu S_{\mu\nu}(Q) \tag{37}$$

is essentially determined by heterogeneous fluctuations (Eq. 20). Therefore,we may also add or subtract, for convenience, a constant term from the scattering amplitudes. This allow the scattering cross section to be rewritten as a function of contrast to any of its constituents:

$$I(Q) \approx \sum_{\mu\nu} (\rho_\mu - \rho_s)(\rho_\nu - \rho_s) S_{\mu\nu}(Q). \tag{38}$$

The partial scattering functions as obtained from a series of measurements with different scattering contrast are shown in Fig. 14. The scattering functions describing the oil-oil correlations S_{oo} are equivalent to the complementary water-water correlations S_{ww} (Babinet's principle). The double logarithmic plot reveals a characteristic Q^{-4} dependence, which shows that the

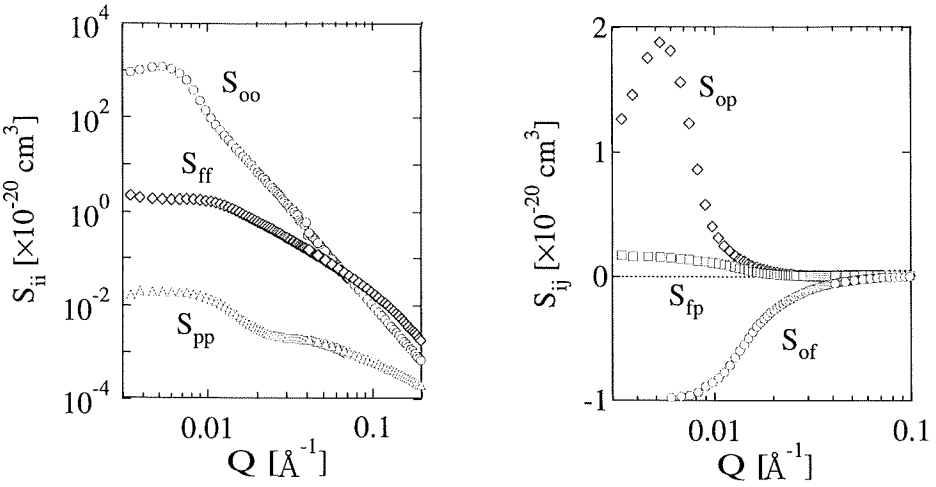


Fig. 14: *Partial scattering functions of an oil-water emulsion-surfactant mixture derived from various contrast matching experiments in neutron small angle scattering.*

interface of the oil-water emulsion is essentially very smooth. The maximum and the slight decrease of S_{oo} towards small Q is an interference that can be related to typical length scales for thickness of layers and their typical distances. This feature would be inconsistent with a picture of isolated random droplets of oil or water. The other scattering functions for film-film and polymer-polymer correlations, S_{ff} and S_{pp}, exhibit quite different characteristics. Note that the interference (cross) terms may be negative, although the total intensity is always positive. Based on the very detailed information from scattering, it was possible to achieve a quantitative description [8], which finally has led to the schematic Fig. 13. Furthermore, the theoretical models provide an understanding of the effect of polymers for the emulgating properties of the oil-water emulsion. Apparently, the di-block copolymers pinned to the interface have a strong effect on the elastic properties of the interface.

4.2 Relaxations

The dynamic of atoms that are tightly bound in molecules and crystals has essentially vibrational character, with frequencies related to their interaction potential. In disordered matter, as for instance for molecules or larger aggregates in solution, relaxation processes due to stochastic motions become important. In order to study diffusion processes at ambient temperature the relevant time scales may be in the order of nano-seconds and exceedingly larger depending on the size of the object. With usual energy-resolving scattering experiments even with neutrons and high energy resolution, there are practical limits and also intensity limits for resolutions much better than μeV. Therefore, only faster processes can be studied in such experiments, where the quasi-elastic broadening of the scattered intensity is measured using time-of-flight spectrometers or triple-axis instruments. In these experiments, the scattering cross-section is directly related to $S(\vec{Q}, \omega)$. However, there are special techniques like neutron spin-echo (NSE) and x-ray time correlation spectroscopy, in which one can explore the region of slower time domains by observing the intermediate scattering function $S(\vec{Q}, t)$. In x-ray or photon corre-

lation spectroscopy one measures the time-dependence of diffraction intensities. In order to see any characteristic temporal fluctuations on a microscopic scale, it is necessary that such fluctuations are not completely wiped out by the configurational average taken in the scattering experiment. With highly coherent beams this can be achieved, as shown in an example further below. Alternatively, one might focus the beam to the length scale of the temporal fluctuations of interest.

Relaxation processes like diffusion have been studied by neutron scattering techniques (time-of-flight spectroscopy, backscattering, and neutron spin-echo). The accessible dynamical range is given by frequencies ranging from 10^{14} to 10^{7} s^{-1} and achieves a spatial resolution from several 100 Å down to near atomic distances.

Complementary, photon correlation spectroscopy with visible light covers the low frequency dynamics ($\omega \leq 10^{6} s^{-1}$) in the regime of small Q $\leq 4 \times 10^{-3}$Å^{-1} but only in materials that are optically transparent. X-rays from 3rd generation synchrotrons open new possibilities to study low frequency dynamics (10^{6} to 10^{-3} s) in a Q-range that includes even atomic resolution (10^{-3} up to several Å^{-1}).

In X-ray photon correlation spectroscopy the time dependent intensities of speckle intensities is measured that appear in scattering using highly coherent radiation, see Fig. 15 The "noisy" speckle pattern does not stem from insufficient counting statistics, but a very precise observation of the actual correlations over a large coherence volume. Within time-averaging or by configurational averaging by a usual and less coherent beam all the speckle structures will fade out. The idea of correlation spectroscopy is to determine pixel-wise the time-dependence of a speckle pattern as shown in Fig. 15, which yields the intermediate scattering function related to $S(\vec{Q}, t)$.

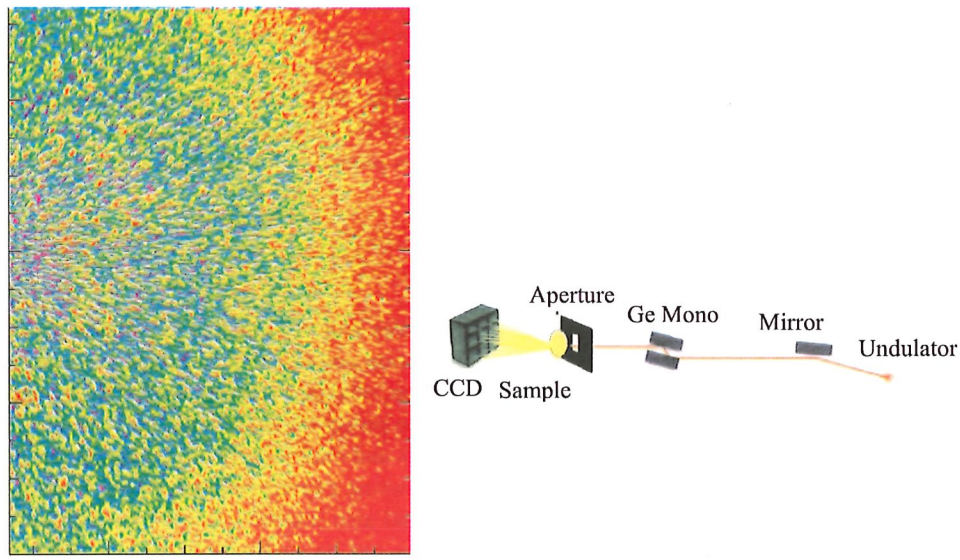


Fig. 15: *Coherent x-ray scattering of silica aerogel showing speckle pattern* (left). *XPCS experimental set-up at the APS sector 8* (right).

The temporal correlations of the intensities are quantified via the normalized time correlation

function

$$g(t) = \frac{\langle n(t')n(t+t')\rangle}{\langle n\rangle^2} \tag{39}$$

where $n(t)$ is the integrated intensity $I(Q,t)$ around a time t, and the average denotes a time average. This intensity correlation function is a second order correlation function that can be related to the first order correlation function of the field amplitudes,

$$g^{(2)}(Q,t) = \frac{\langle I(Q,0)I(Q,t)\rangle}{\langle I(Q)\rangle^2} = 1 + \beta\frac{\langle E(Q,0)E^*(Q,t)\rangle}{\langle I(Q)\rangle^2} = 1 + \beta|g^{(1)}(Q,t)|^2. \tag{40}$$

We may identify $g^{(1)}(Q,t)$ as the intermediate scattering function $S(\vec{Q},t)$
the Fouriertransform of the pair-correlation function $G(\vec{r},t)$. Depending to the amount of incoherent beam admixtures, β varies in between zero and one, and is therefore a set-up dependent quantity. In case of an ideal coherent beam, β would be equal to unity and tends to zero for loss of coherence. For a translational diffusion process of a single particle the expected functional form of the correlation function is an exponential relation

$$g^{(2)}(Q,t) = 1 + A(Q)\mathrm{e}^{-2\Gamma_Q t}, \tag{41}$$

where $\Gamma = Q^2D$ can be related to a diffusion coefficient D. In principle all correlations between interacting diffusing particles are incorporated in Eq. 41. However, in the dilute regime this will be determined only by the diffusion of a single particle:

$$D(c \to 0) = \frac{kT}{6\pi\eta R}, \tag{42}$$

where k is the Boltzmann constant, η denotes viscosity, and R is the (hydrodynamic) radius of the diffusing particle.
In a pioneering work this method has been applied to study diffusion of a colloidal palladium aggregate sol [9]. A more recent example shown in Fig. 16 is related to the *structure and dynamics of colloidal diffusion* in a charge stabilized suspension of polystyrene (PS) latex in glycerol [11]. The time integrated scattering $I(Q)$, (Fig. 16 left side), can be well described by a spherical particles of 67 nm size, whose form-factor is described by Eq. 37. Correlations among the particles increase with their volume fraction, as seen from the scattering at low Q, and stem from repulsive interactions. The time-dependent intensity correlation function allows to analyse the characteristic dynamic relaxational behavior of this material. It was measured on a time scale of seconds. There is a change from a constant at short times, which is larger than one and depends only on the coherence of the set-up, to unity at large times. This change reflects a simple exponential decay in time of $S(Q,t)$, and therefore also of the pair-correlations $G(r,t)$, which can be related to the self-diffusion of individual particles. However, the relaxations are in general more complex as described in [11]. With increasing volume two distinct relaxations appear at short and long time scales, respectively, where the particles turn from random walk to interactive motions. The next example considers reflectivy measurements from a more complex system of smetic liquid-crystal membranes[12], and particularly its time-dependent behavior. The scattering geometry is depicted in Fig.17. Without going into details about reflectivity measurements, one can understand that the anisotropy of these scattering intensities can be related to vertical correlations and lateral pair-correlations within the membrane, depending $\vec{q}_z$ and $\vec{q}_\perp$ respectively, the components of the scattering vector $\vec{Q}$ (=$\vec{q}$). There are time-dependent

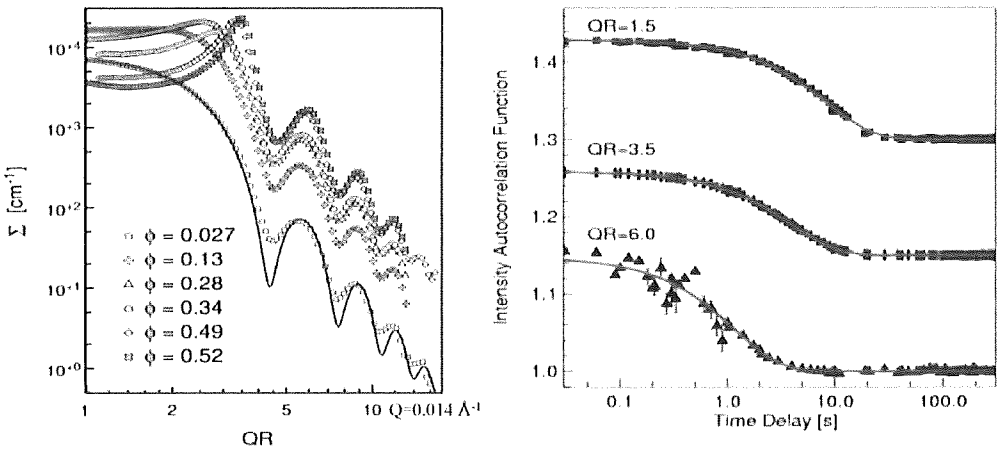


Fig. 16: *Structure and dynamics of a charge stabilized suspension of polystyrene latex in glycerol. Left: time average diffraction* $\langle I(Q)\rangle$. *Modulations of the scattering function are characteristic for spherical particles. The effects of increasing concentration seen at small* Q *are due to repulsive interactions. Right: For a volume fraction* $\Phi = 0.28$ *of PS in glycerol, the time-dependent intensity autocorrelation function is given by a simple exponential decay of* $g^{(1)}(Q,t)$ *(from [11])*

fluctuations of the membrane structure in form of thermal displacements of the layer, with specific characteristics of the anisotropic structure. Compared to the previous example, by using a fast detector and a special hardware for the auto-correlation of intensities, it has been possible to measure much faster dynamics on time scales down to hundreds of nano-seconds (Troika beamline at the ESRF). The results Fig. 17b indicate a different slope of relaxations for vertical relaxations and lateral relaxations within the membrane.

Neutron spin-echo spectroscopy also provides a direct measurement of the intermediate scattering function $S(\vec{,}t)$ und is thus particularly suited to study relaxational processes. The experimental set-up is shown in Fig. 10. A closer look at neutron spin-echo intensities reveals that the integral scattering $S(\vec{Q}) = \int d\omega\, S(\vec{Q},\omega)$is measured, however, with the weight of the precessing polarisation, $P_z = \frac{1}{2}(1 + cos\,\delta\Phi)$,

$$I \propto \int P_z\; S(\vec{Q},\omega)\; d\omega = \frac{1}{2}\{S(\vec{Q}) + \int cos\,\delta\Phi\; S(\vec{Q},\omega)\; d\omega\}. \tag{43}$$

The change of the total precession angle $\delta\phi = const\; J_0\; \omega$ is proportional to the field integral $J_0 = \int B\; dl$, and therefore proportional to the magnetic field B. Recognizing that the term $const\; J_0$ is a time variable, the integral with respect to ω turns out to be the cosine Fourier transform of $S(\vec{Q},\omega)$ and yields the intermediate scattering function $S(\vec{Q},t)$,

$$I \propto \frac{1}{2}\{S(\vec{Q}) + S(\vec{Q},t)\}. \tag{44}$$

Fig. 17a shows data measured with the neutron spin-echo spectrometer IN15 at the ILL. These experiments reveal faster dynamics of the membranes for in-plane fluctuations in the range of nano-seconds. From these data it has been concluded that the two characteristics of the thermal

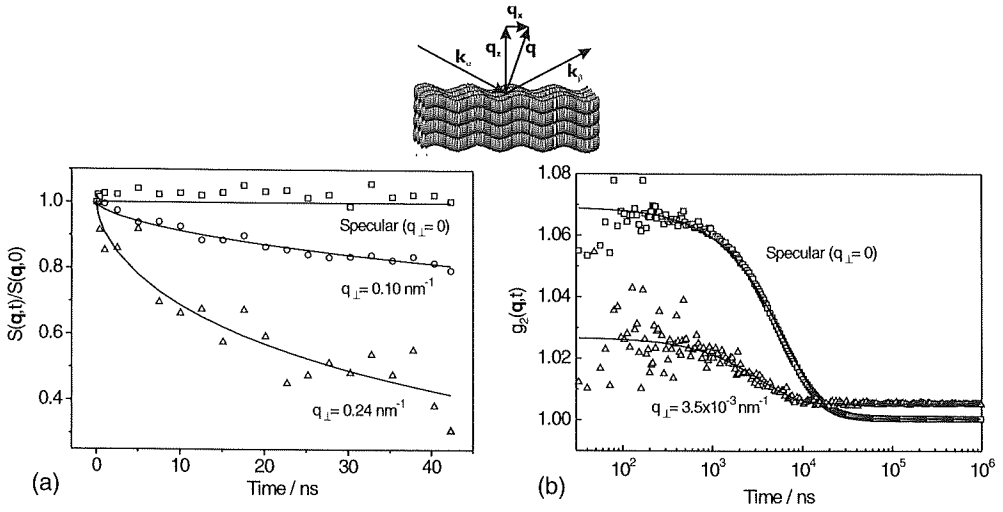


Fig. 17: *Scattering geometry in a reflectivity study of fluctuations in smectic membranes by neutron spin-echo* (left) *and XPCS* (right). *Specular reflectivity only probes the vertical displacement fluctuations perpendicular to the membrane layers. For off-specular scattering, one can see the influence of lateral fluctuations, which shows much faster relaxations.*

motion have two signatures, that are ruled by bulk elasticity at short times and surface tension at larger times [12].

4.3 Dynamic excitations

Rosalind Franklin's x-ray diffraction work on DNA has been a milestone in history of science, and enabled Crick and Watson to solve the double helix structure of DNA (1953). Nowadays, there are attempts to understand more of the DNA properties by studying the dynamic response. In particular, vibrational modes involving collective motions of groups of atoms in the helix structure of DNA have been investigated by neutron scattering [13]. The vibration spectrum is of particular interest since it offers the key to the specific bonding and restoring forces of the molecular structure.

As a prerequisite of the following discussion we shall consider the "form-factor" of a phonon excitation, which can be written as

$$f(\vec{Q})_{phonon} = \sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_l(t)} \approx \sum_l \mathrm{i}\,\vec{Q}\cdot\vec{u}(t)\,\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_0} = \mathrm{i}\,\vec{Q}\cdot\vec{u}_{\vec{Q}}(t), \tag{45}$$

where $\vec{r}(t) = \vec{r}_0 + \vec{u}(t)$ is decomposed in time average positions and time-dependent displacements; the phase factors have been expanded in terms of small displacement variables $\vec{u}(t)$. The phonon scattering function will be given by

$$S(\vec{Q},\omega) = \int dt\,\mathrm{e}^{-\mathrm{i}\omega t}|f(\vec{Q})_{phonon}|^2 = \int dt\,\mathrm{e}^{-\mathrm{i}\omega t}\langle(\vec{Q}\cdot\vec{u}_{\vec{Q}}(0))(\vec{Q}\cdot\vec{u}_{\vec{Q}}(t))\rangle. \tag{46}$$

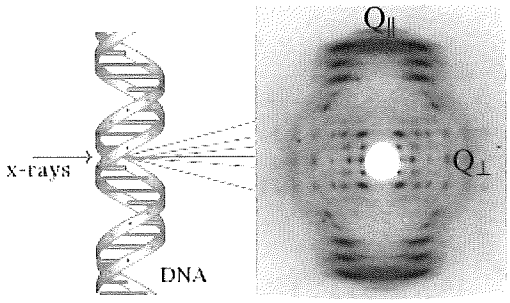


Fig. 18: *X-ray diffraction of DNA-helix.*

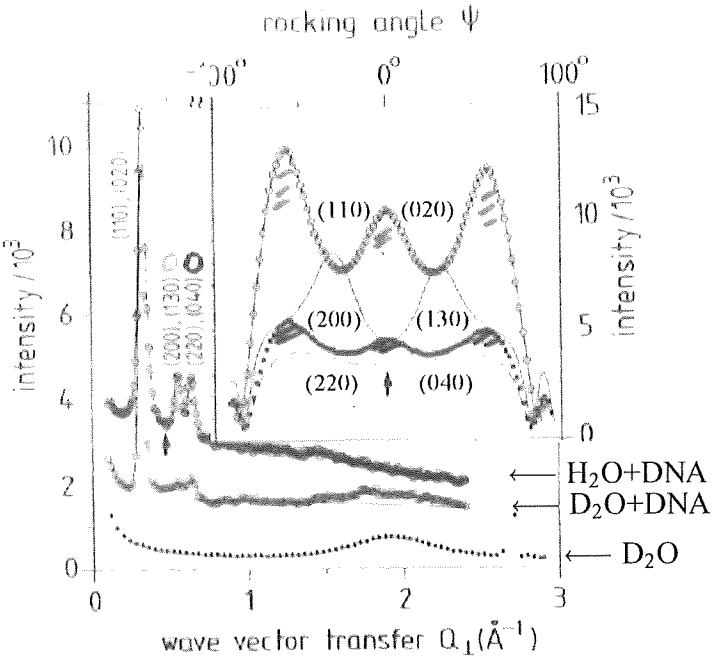


Fig. 19: *Neutron diffraction of a humid, oriented Na-DNA sample. The Q-scan perpendicular to the helical axis and the rocking scans* (inset) *on the observed three peaks reveal a hexagonal array of DNA in the crystal. The difference in the Bragg intensities due to the contrast of H_2O and D_2O shows that the water is part of the ordered structure.*

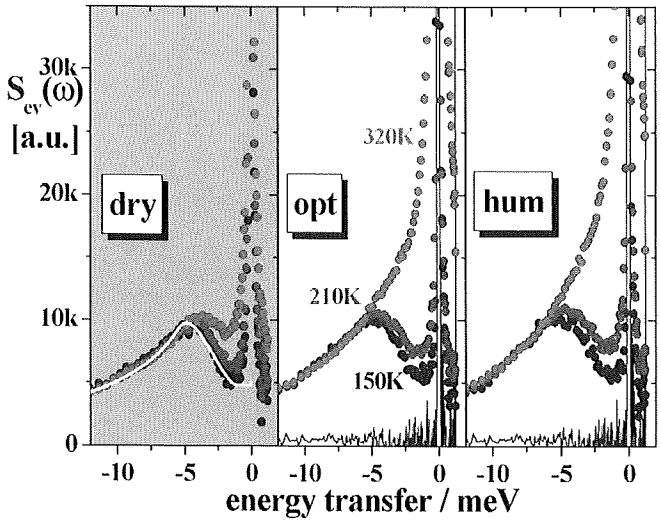


Fig. 20: *Neutron TOF spectroscopy of the DNA-hydrogen relaxations as a function of humidity and temperature. A strong increase of the incoherent quasi-elastic scattering is found with increasing D_2O for temperatures above the "glass-transition" of humid Li-DNA at $T_G \approx 210K$. Intensities have been scaled to the phonon peak near -4 meV.*

One can see first, that the intensity will increase proportional to Q^2, and secondly, that by the scalar product $\vec{Q} \cdot \vec{u}_{\vec{Q}}(t)$ scattering is specifically sensitive not only to the modulus but also to the direction of phonon displacements $\vec{u}(t)$. In general, there are at least three independent acoustic modes, a longitudinal mode with displacement $\vec{u}$ parallel to the propagation vector $\vec{q}$ (an example would be a sound wave), and two transverse modes with the displacement vector $\vec{u}$ perpendicular to the scattering vector $\vec{q}$ (related to shear forces). The scattering vector $\vec{Q}$ may be chosen such as to probe either longitudinal or transverse modes. Inelastic scattering experiments on the dispersion of phonon branches yield the complete information about the bonding properties of a crystal.

Here, we shall consider recent experiments on the dynamics of DNA. The vibration spectrum of DNA can be measured by inelastic scattering with neutrons and x-rays.

DNA is a rather delicate material. DNA samples have been prepared by intricate preparation of wet-spinning. They consist of an essentially hexagonal array of DNA helices, as can be seen from the Q-scan perpendicular to the helical axis and also from the rocking curves in the inset of Fig. 19. From structural investigations by x-ray diffraction it is known that the conformation is quite sensitive to the water and salt environment. Bragg peaks appear in layers perpendicular to the helical axis. Their scattering vectors relate to integer ratios of the number of base-pairs per turn and mark shallow minima as a function of its humid and salty enviroment, which emphasize the role of subtle interactions. The fact that H_2O participates in the order of the DNA structure itself can be infered from neutron scattering, see Fig. 19. Indeed, changing the contrast by substitution of H_2O by D_2O has a marked effect on the Bragg intensities. This effect has been used to identify the stabilizing water backbone in the minor groove of the helical screw. The background for case of the H_2O envelope stems from incoherent scattering of H. The decrease

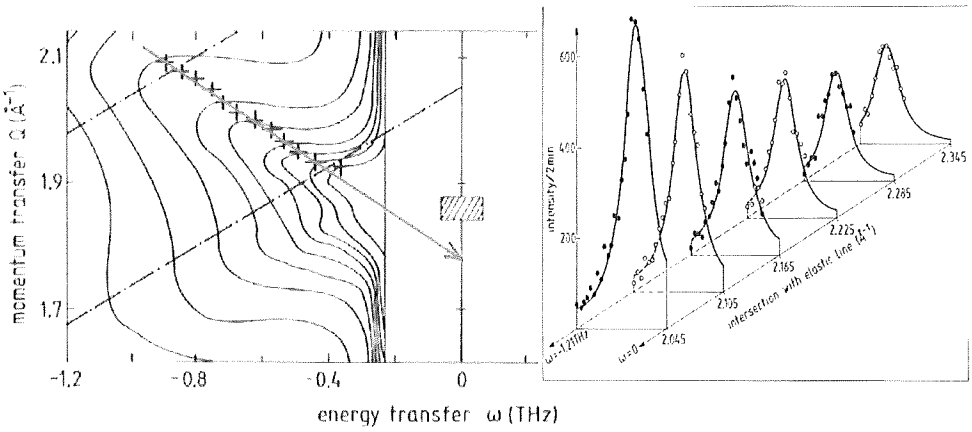


Fig. 21: *Contour map of coherent inelastic neutron scattering of B-DNA in the $(\vec{Q}_{\|}, \omega)$-plane displays LA-phonon intensity. The strong elastic scattering and Bragg peak* (hatched brick) *is not shown. Contours have been obtained from $(\vec{Q}, \omega)$ scans* (dashed-dotted lines)*, as shown in the right figure (Grimm 1987).*

with Q can be related to the mean-square fluctuations $\langle u^2 \rangle$ of the individual hydrogen atoms which reduce the elastic part of the scattering ($\sim exp(-\frac{1}{3}Q^2\langle u^2 \rangle)$ Debye-Waller factor, see also Guinier approximation).

Before discussing phonon excitations, let us consider another aspect of the envelope of humid water, which provides a certain local mobility for the DNA. Local diffusion and relaxation have been investigated by inelastic neutron scattering. Fig. 20 shows results from time-of-flight measurements on Li-DNA with various degrees of humidity and at different temperatures. Again one has made use of the scattering contrast between H and D. Because of the huge incoherent scattering cross-section of hydrogen (H),the surrounding D_2O-water does not contribute significantly to the quasielastic scattering signal. Nonetheless, as can be seen for increasing humidity above the "glass-transition" temperature $T_G \approx 210K$, the surrounding D_2O-water must be the cause for the increasing mobility of the DNA-hydrogen, consequently as well as for the whole DNA [13].

Collective phonon modes have been observed by neutron triple-axis spectroscopy. This scattering appears in the coherent scattering function $S(\vec{Q}, \omega)$. The relatively high background of incoherent inelastic scattering due to H (see previous Fig. 20.) is independent of the direction of $\vec{Q}$ and can thus be subtracted by a reference measurement. Fig.21 shows data with $\vec{Q}$ parallel to the helical axis near to a Bragg peak, which corresponds to the inverse stacking distance of the base-pairs. Near to the Bragg peak one can see two ridges of intensity which seems to spring off from the Bragg peak, although its immediate vicinity cannot be accessed. Obviously, the phonon branch can be identified as the branch of longitudinal acoustic (LA) modes along the helical axis. A more careful analysis, however, shows that the slope of about 2200 m/s does not extrapolate to the Bragg point as expected for an acoustic mode. This puzzling result might indicate that the LA modes interact with another phonon excitation of different symmetry. However, DNA is of course not a strictly periodic helix, and disorder effects may play a role as well. One may note that in a helical structure acoustic modes give rise to simultaneous torsional and compressional vibrations. Other phonon branches are of interest as well. For instance, the

radial optical modes will probe fluctuations between base-pairs along the axis.
It is worthwhile to mention further that very recently, by high-resolution x-ray scattering (meV) at the ESRF, longitudinal modes of the DNA have been studied at small scattering vectors, where one probes the restoring force of the DNA-spring by the collective motion of the whole DNA [14].

References

[1] Max von Laue, *Concerning the detection of X-ray interferences*, Nobel Lecture (1915).

[2] XFEL Design study. Eds. R. Brinkmann, G. Materlik, J. Rossbach, A. Wagner, *Conceptual design of a 500 GeV e^+e^- linear collider with integrated x-ray Laser facility.* Vol. 2, DESY 1997-048.

[3] G.L. Squires, *Introduction to the theory of thermal neutron scattering*, Cambridge University Press, Cambridge (1978).

[4] T. Brückel, *Scattering (chapter A1)* in *Soft Matter, complex materials on mesoscopic scales*, Schriften des Forschungszentrums Jülich, Matter and Materials **10** (2002).

[5] W. Schweika, *Streuexperimente zur Struktur und Dynamik kondensierter Materie (chapter 7)* in *Komplexe Systeme zwischen Atom und Festkörper* 25. Ferienkurs, Forschungszentrum Jülich, ISBN-3-89336-128-6 (1994).

[6] V. Sears, Neutron News, Vol. 3, No. 3, 1992, pp. 29-37; http://www.ncnr.nist.gov/resources/n-lengths/

[7] J. K. G. Dhont, *Light Scattering (chapter 3)* in *An Introduction to Dynamics of Colloids* (Elsevier, Amsterdam, 1996)

[8] H. Endo *et al.*, Phys. Rev. Lett. **85**, 102 (2000); H. Endo *et al.*, J. Chem. Phys. **115**, 580 (2001).

[9] T. Thurn-Albrecht, W. Steffen, A. Patkowski, G. Meier, E.W. Fischer, G. Grübel, and D. L. Abernathy, Phys. Rev. Lett. **77**, 5437 (1996).

[10] G. Grübel, *(B9) X-Ray Photon Correlation Spectroscopy* Schriften des Forschungszentrums Jülich, Matter and Materials **3** (2000).

[11] L. B. Lurio *et al.*, Phys. Rev. Lett. 84, 785 (2000).

[12] I. Sikharulidze, B. Farago, I. P. Dolbnya, A. Madsen, and W. H. de Jeu, Phys. Rev. Lett. 91, 165504 (2003).

[13] H. Grimm *et al.*, Phys. Rev. Lett. 59, 1780 (1987); H. Grimm and A. Rupprecht, Physica B **174** 291 (1991), and Physica B **234-236** 183 (1997).

[14] H. Grimm *et al.* and M. Krisch, to be published.

B 2 Protein Crystallography: X-rays

J. Labahn
Institut für biologische Strukturforschung (IBI-2)
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 The physics in crystallography 3
2.1 Generation and description of X-Rays 3
2.2 The observations and its cause 4
2.3 The scattering geometry 4
2.4 Encoding of structural information (phase difference) 5
2.5 The equation for the molecular structure-factor 6
2.6 The crystal as an signal amplifier 6

3 The informational content of the observations 9
3.1 The Patterson function (heavy atom substructures) 9
3.2 Solving structures by correlation (molecular replacement) 10
3.3 Symmetry and reciprocal space 10

4 Solving the Phase Problem 10
4.1 Molecular Replacement 10
4.2 Isomorphous replacement 11

5 Density and Model 12
5.1 Density refinement 13
5.2 Completeness and local correctness (difference-density) 13
5.3 Global correctness of the model 14

A Appendix: The atomic form factor 15

1 Introduction

Many biopolymers and their complexes can be crystallised by simply reducing their solubility by gradual change of the solvent properties [1]. The atomic structure of these compounds can be resolved by X-ray crystallography. The amount of information obtained depends on the degree of ordering in the crystals. Typical examples of low ordering are the 1-dimensional DNA-fibres and the 2-dimensional layers of membrane protein layers (e.g. Purple membrane). The dimensionality of ordering is often not caused by a principle disorder of the molecules but represents disordered packing. This kind of disorder can be overcome by changing the crystallisation conditions.
Disorder within the molecules that form the crystal may be reduced sometimes by adding compounds that stabilise the molecular structure by specific binding.
Special problems arise when time-resolved crystallography is planned [2]. Changes in the structure must appear simultaneously for all the molecules in the crystal. Such a crystal is ordered in four dimensions. If the changes do not appear simultaneously the spatial ordering is disturbed and ordering is lost.

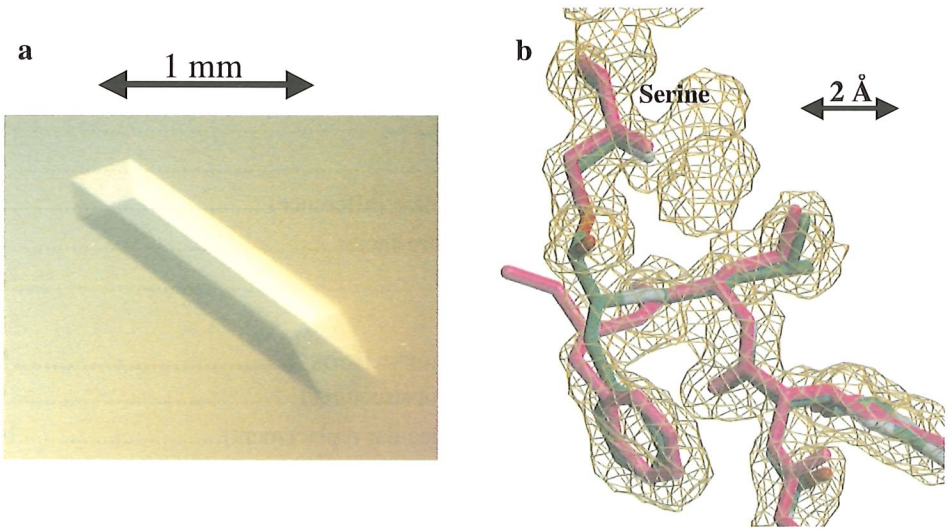


Fig. 1: *Characterisation of the mechanism of the Peptideamidase (St. maltophilia)*
***a** single crystal, that diffracts to 1.4 Å resolution*
***b** structural detail of covalently bond Serine-Chymostatin in mutant K121A with electron-density, overlayed (magenta): the structure in the non-covalent wild-type structure. The chirality at the acetal-carbon is clearly identified.*

An approach more general applicable is the structural characterisation of states along the reaction trajectory. In Fig.1b a detail of two structures is shown. Both contain a bound inhibitor molecule in their reaction centre. For the wild-type enzyme no covalent bond between the inhibitor and the serine residue is observed [3]. If a catalytic lysine residue (not visible in Fig.1b) is mutated to alanine the covalent linkage between the serine nucleophile of the protein and the aldehyde-group of the inhibitor is formed. The difference between alanine

and lysine being that alanine can not assist the hydrolysis of the acetal-bond by supplying a proton.

2 The physics in crystallography

In physics the most widely used approach to investigate the structure of composites is scattering. Obviously scattering requires to use a beam of something that actually interacts with the structure forming elements of the object.
Composites like molecules are structured clusters of charged particles, namely the nucleii and the electrons. To probe the spatial distribution of these charges, either other charges or photons may be used.
The strength of the Coloumb-interaction causes the charges in the beam to be effected in a complex manner on its passage through the object. The appearance of inelastic scattering signifies that the object is actually changed upon measurement. Consequently the electron beam is of limited utility for in-depth probing of large molecules. Photons on the other hand interact in comparison only weakly with charges. Ideally we would like to use light in scattering experiments as it is realised in the light microscope: By passing the scattered light through a lense an image of the object is obtained and can be viewed directly.
Unfortunately the wavelength employed to resolve the atoms in a molecule must be of the order of their distance ($\sim 10^{-10}$ m). For light of such a short wavelength the index of refraction is very close to unity: There are no material lenses that would allow to obtain an image of the electron distribution within a molecule by X-ray scattering. The structure of the molecule has to be calculated from the observed scattering [5].

2.1 Generation and description of X-Rays

The electromagnetic waves required in crystallography are generated either by a synchrotron source or by in-house equipment.
In a synchrotron electrons or positrons are accelerated to nearly the speed of light. When forced onto a trajectory of small diameter (wiggler), the particles will show strong radiation. Out of this continuous spectrum the desired wavelength is selected by a system of mirrors and monochromators. Therefore the wavelength is tuneable. Consequently the use of synchrotron radiation offers not only X-rays of high brilliance, but allows to measure significant data of the anomalous dispersion at the absorption edge of heavy atoms [4].
For in-house generation of X-rays the wavelength-characteristic of elements is utilised: Typically electrons are accelerated onto a copper target. The copper atoms are ionised by the impact of the electrons. The atoms react to the ionisation by a reorganisation of their electronic structure with the emission of a characteristic radiation (Fig.2a). The characteristic CuK_α-radiation of wavelength 1.54 Å obtained by a filter foil (Ni) is employed for crystallographic experiments.
At a distance from the source that is very large compared to the wavelength the X-ray is most conveniently described as a plane wave (Fig.2b):

$$\mathbf{E}(\mathbf{r},t)e^{i\varphi} \qquad (1)$$

Here the exponential describes the position of the wave-front with respect to a given co-ordinate system. For the primary beam the phase may be set to zero at the co-ordinate origin.

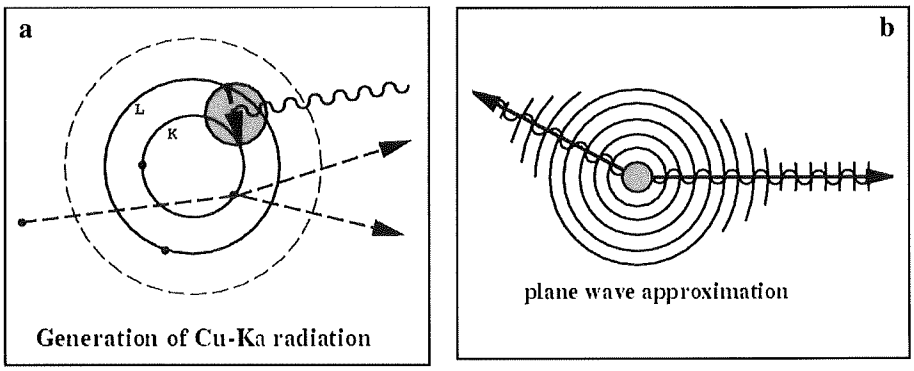


Fig.2: *Generation of X-rays in a in-house setup employing a Copper target*

2.2 The observations and its cause

We expect the charges q to be accelerated by the electromagnetic field E_0 of the primary beam.

$$a = qE_0 / m_q \tag{2}$$

The secondary field produced by the accelerated charges is independent of the mass of the charge and depends only on the charge and the accerleration. From the mass ratio of nucleii and electrons it is obvious that the acceleration of the nucleii is neglectable in comparison to the acceleration of electrons. Therefore we can consider the observed scattering to be solely due to electrons.
At the point of observation the resultant field will have a certain energy-density that moves with the speed of light. Therefore the energy per unit area and time transferred onto the detector – the intensity will be given as:

$$I = \eta \varepsilon \varepsilon_0 |E|^2 \tag{3}$$

Here η denotes the efficiency of the detector. Typically Image-Plate and CCD-detectors are in use.

2.3 The scattering geometry

The principal measurement parameters are the distance of detection and the angle with respect to the primary beam. The dependent quantity will be the observed intensity of scattering.

The direction of propagation of the field is described by an unit vector $\mathbf{S}_0$ for the primary beam and $\mathbf{S}_S$ for a scattered beam. The change of direction upon scattering is given as **S**. The angle between $\mathbf{S}_0$ and $\mathbf{S}_S$ is 2θ (Fig.3).

$$\mathbf{S} = \mathbf{S}_S - \mathbf{S}_0 \tag{4}$$

For an observed intensity I(S) the resolution is calculated from the geometry of the Ewald construction (Fig.3, right top and middle).

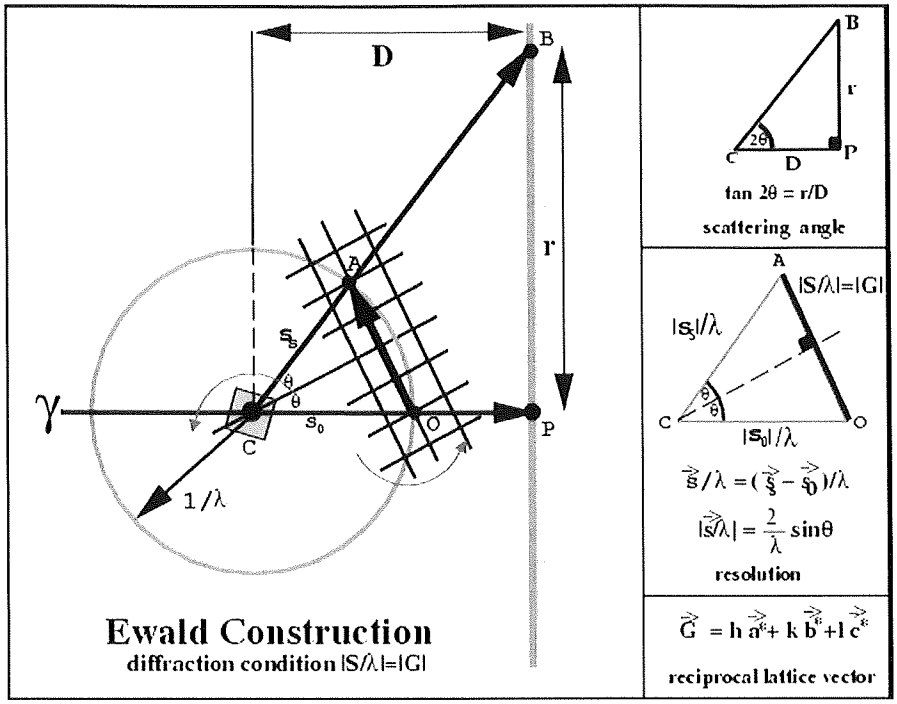


Fig.3: *The Ewald Construction connects the formulas and concepts used in crystallography. Note, in many textbooks* |S| *denotes not an unit vector, but its modulus represents the resolution (*|S/λ|*).For crystals the reciprocal space point h=0, k=4, l=0 falls onto the Ewald sphere and thereby obeys the reflection condition* **S**/λ=**G**. *The intensity I(0,4,0) is observed at* B. *The reciprocal lattice is covered in sect. 2.6 and 3.3*

In practice an area detector (typical radius r = 150 mm) at a variable distance (typically D=75–500 mm) from the object is used in protein crystallography. The maximal scattering angle observed can be calculated from tan2θ = r/D to be about 60° for the closest position of the detector. To collect complete data-sets the crystal will be rotated during a measurement (typically 1° / image, rotation axis is viewing axis in Fig.3).

2.4 Encoding of structural information (phase difference)

Any point of the electron-density in the object will generally contribute to the resultant of scattering at a detector point. The different waves will have in general different phases with respect to each other (Fig.4). These phase differences are caused by the different path length the individual waves must take because they are generated at different positions in the object. With respect to a wave through the co-ordinate origin the path difference and the corresponding phase difference is calculated for a scattering event at a position **r** observed for scattering vector **S**:

$$\varphi(\mathbf{r},\mathbf{S}) = 2\Pi\,\mathbf{r}\,\mathbf{S}/\lambda \qquad (5)$$

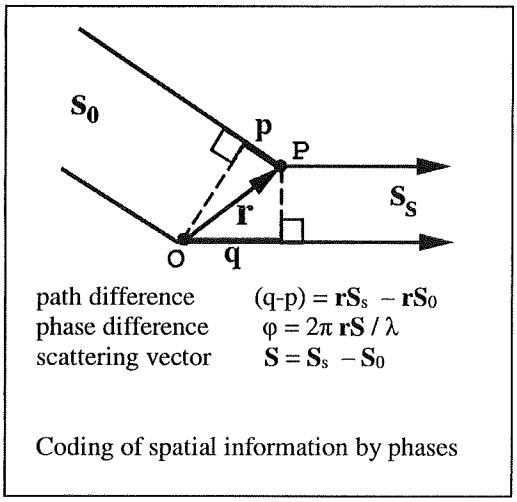


Fig. 4: *The path difference for rays scattered at different points P with respect to a common point of detection determines the observed intensity.*

2.5 The equation for the molecular structure-factor

From an Ansatz for differential scattering of the electron density ρ at a point **r** observed at a scattering angle **S**, we obtain directly the total scattering upon integration. The resultant scattered field is in crystallography denoted by F.

$$dF(\mathbf{r},\mathbf{S}) \sim \rho(\mathbf{r}) e^{i\varphi(r,S)} dV \Rightarrow F(\mathbf{S}) \sim \int \rho(\mathbf{r}) e^{i\varphi(r,S)} dV \tag{6}$$

From the structure-factor equation we obtain the electron-density equation by the application of the Fourier-transform:

$$F(\mathbf{S}) = |F(\mathbf{S})| e^{i\alpha} = \int_V \rho(\mathbf{r}) e^{2\Pi i \mathbf{r}\mathbf{S}/\lambda} dV \overset{FT}{\Leftrightarrow} \rho(\mathbf{r}) = \int_S F(\mathbf{S}) e^{-2\Pi i \mathbf{r}\mathbf{S}/\lambda} dS \tag{7}$$

From the observed intensities (eq.3) we can calculate the structure-amplitude by taking the square-root but not the structure-factor that we need to calculate the electron-density (eq.7b). The phase of the structure-factor is lost upon measurement. This is the **phase-problem**.

2.6 The crystal as an signal amplifier

X-ray scattering of a single molecule is by many orders of magnitude to weak to be observed. Scattering at a large ensemble of molecules will result in observable intensities. But this ensemble must be ordered to yield really useful data. The highest degree of ordering is found in crystalline matter. In the simplest case, a (triclinic) crystal is described by a group of molecules (the repetition unit) and three base vectors **a,b,c**. Their length and the three enclosed angles are the cell constants. Any repetition of the group within the crystal can be generated through a translation **T** by a linear combination of the base vectors (Fig.5).

$$\mathbf{T} = m\mathbf{a} + n\mathbf{b} + o\mathbf{c} \tag{8}$$

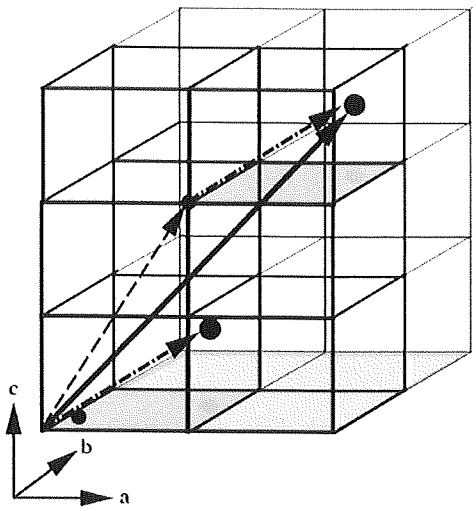


Fig. 5: *Simplification of the description of an unit cell contents e.g. a molecules co-ordinates by intro-duction of an intra-unit-cell position vector for atoms and a unit cell translation vector to describe the unit cell repeat.*

Any point $\acute{\mathbf{r}}_i$ within the crystal may be represented by a position vector within one of the unit cells $\mathbf{R}_i$ and the position of the unit cell in the crystal $\mathbf{T}_{mno}$

$$\acute{\mathbf{r}}_i = \mathbf{T}_{m,n,o} + \mathbf{R}_i \tag{9}$$

from eq.7,9

$$F(\mathbf{S}) = \sum_m \sum_n \sum_o \int \rho(\mathbf{R})^{2\Pi i(\mathbf{T}_{m,n,o}+\mathbf{R})\mathbf{S}/\lambda} dV \tag{10}$$

with eq.8 and reordering

$$F(\mathbf{S}) = \sum_m \sum_n e^{2\Pi i(ma+nb)\mathbf{S}/\lambda} \int \rho(\mathbf{R}) \sum_o e^{2\Pi i(oc+\mathbf{R})\mathbf{S}/\lambda} dV \tag{11}$$

F will be large if the sum in the integral is large:

$$\sum_o e^{2\Pi i(oc+\mathbf{R})\mathbf{S}/\lambda} = \sum_o e^{i(\alpha+o\beta)} \quad \text{with} \quad \alpha = 2\Pi i\, \mathbf{RS}/\lambda \quad \text{and} \quad \beta = 2\Pi i\, \mathbf{cS}/\lambda \tag{12}$$

In the complex plane it can be seen easily that the condition for constructive interference is $\beta = l*2\Pi$, where l is an integer (Fig.6). Analogous conditions hold along the other two axis. Therefore the Laue equations are obtained:

$$\mathbf{aS} = h\lambda \quad \mathbf{bS} = k\lambda \quad \mathbf{cS} = l\lambda \tag{13}$$

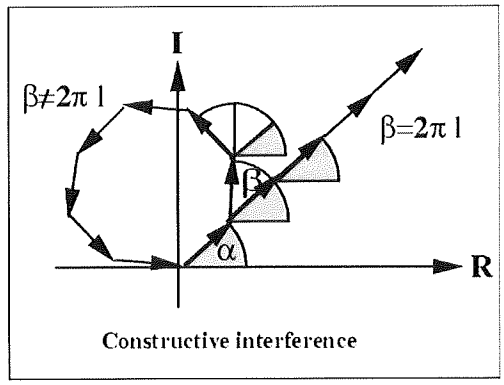


Fig. 6: *From a crystal a strong signal is only observed if special relation for the lattice-constant dependent phase-angle beta holds: The Laue equations (eq.13).*

These equation show that we will observe for a crystal diffraction. By constructive interference of the scattered waves we observe intensities only for discrete vectors **S.** Therefore the index-triple h,k,l of a reflection is used instead of **S** to denote the diffraction geometry in crystallography. We may consider the index-tripel h,k,l to mean lattice points in an abstract space - the reciprocal space - with the lattice constants **a*,b*,c***:

$$\mathbf{S}/\lambda = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \tag{14}$$

By multiplying eq.14 with the lattice constant **a** and comparing with eq.13, we find:

$$\mathbf{a}\mathbf{a}^* = 1 \quad \mathbf{a}\mathbf{b}^* = 0 \quad \mathbf{a}\mathbf{c}^* = 0$$

We conclude that a reciprocal lattice vector **a*** is at 90° with respect to **b** and **c** and has the length 1/a. The relation between crystal and reciprocal lattice is pictured by the Ewald construction (Fig.3). Upon rotation of the crystal arround C (Fig.3) the reciprocal lattice rotates arround O. Whenever a reciprocal lattice vector **G** falls onto the Ewald sphere an intensity is observed.

If we rewrite $\mathbf{R}_i$ in fractional co-ordinates as $\mathbf{R}_i = x\,\mathbf{a} + y\,\mathbf{b} + z\,\mathbf{c}$ in eq.10, we obtain for eq.11 with the Laue conditions (eq.13):

$$F(h,k,l) = \sum_m \sum_m \sum_o \int \rho(x,y,z)\, e^{2\Pi i(hm+kn+lo)} e^{2\Pi i(hx+ky+lz)} dV$$

$$\text{because } e^{2\Pi i(hm+kn+lo)} = 1 \tag{15}$$

$$F(h,k,l) = w \int \rho(x,y,z)\, e^{2\Pi i(hx+ky+lz)} dV$$

Here W = M N O is the number of unit cells in the crystal. Each cell may contain one molecule. The observed intensity is W^2 times the intensity of that of a single molecule and at least W times the intensity of an unordered ensemble of W molecules!
If we consider an average unit cell of $(60\ Å)^3$ and an average protein crystal of $(0.3\ mm)^3$ we obtain $W = 4.2 * 10^{13}$.

3 The informational content of the observations

Eq.8b shows that we need to know the phases to calculate the electron-density. But we measure only intensities that allow to calculate the amplitude. The question arises, if we can extract any structural information from the amplitudes.

3.1 The Patterson function (heavy atom substructures)

Considering the Patterson function:

$$P(\mathbf{u}) = \int \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})dV \qquad (16)$$

Denoting the tripel h,k,l by the vector **H,** it can be shown by substituting the electron-density equation into the Patterson function (eq.16) that P(u) can be calculated from the observed intensities:

$$P(\mathbf{u}) \sim \sum_H |F(\mathbf{H})|^2 \cos(2\Pi\, \mathbf{Hu}) \qquad (17)$$

If we imagine one point-atom with electron-densities ρ_1 at locations $\mathbf{r}_1$ in the unit cell there are only atom-atom vectors of the lattice type (eq.10). Only for u = $T_{m,n,o}$ the corresponding product of electron-densities will be larger than zero: The Patterson function forms a lattice with the same lattice-constants as the unit-cell.
Considering point-atoms, we may replace the electron density by the number of electrons Z. If we have two point-atoms with electron-densities (numbers) Z_1 and Z_2 at locations $\mathbf{r}_1$ and $\mathbf{r}_2$ we find for the origin of the unit cell P(0,0,0) = Z_1^2 + Z_2^2: The Patterson function is a weighted vector density function. The weight being the product of electron-densities and the ends of the vector.
Additionally we will observe signals for centro-symmetrically distributed cross-vectors at $\mathbf{u}=\mathbf{r}_2-\mathbf{r}_1$ and $-\mathbf{u}$, that are useful for the determination of co-ordinates:

$$P(\mathbf{r}_2-\mathbf{r}_1) = Z_1 Z_2 \quad P(\mathbf{r}_1-\mathbf{r}_2) = Z_1 Z_2 \qquad (18)$$

Strong peaks at **u** imply that there are many atoms that are related by **u** or there are atoms with high electron-density that are related by **u**. We may interpret a pair of peaks (eq.18) as two sites of high electron-density with co-ordinate (0,0,0) and ($\mathbf{r}_2$- $\mathbf{r}_1$) or ($\mathbf{r}_1$- $\mathbf{r}_2$), if the molecule contains heavy atoms. Naturally one of the two pairs of co-ordinates is wrong.
Note that the choice for the origin of the co-ordinate system is in principal arbitrary, though restricted if symmetry elements should have special co-ordinates (see sect.3.3). Therefore one site can be given the origin co-ordinate.
In practice heavy-atom derivatives are prepared by soaking the crystals with a solution of a reactive heavy-atom compound (typically 0.1 to 10 mM, 1-7 days). A reasonable choice for a heavy atom are elements like Hg or Pt with about ten times the number of electrons of an average protein atom.
For analysis the function is calculated with squared differences to sharpen contrast by partial removal of the cross-vectors of the light protein atoms with the heavy atoms:

$$|F_{use}| = |F_{PH}| - |F_P| \;|$$

Here F_{PH} and F_P are mean amplitudes for derivatised and native crystals, respectively.

3.2 Solving structures by correlation (molecular replacement)

If we have a priori knowledge of the structure, either in knowing the exact structure of a part of the molecule or of a similar molecule, we must place these fragments in the proper orientation at the proper place in the unit cell of the crystal under investigation.
The placing problem can be solved in two steps. First the orientation is found by calculating the correlation function between the Patterson function of the observed intensities and the function calculated for the search model in various orientation. To exclude inter-molecular vectors but include intramolecular vectors only vectors up to about 80 % of the radius of the search model are used in the correlation. In an analougous manner the translation can be found that finally place the model.

3.3 Symmetry and reciprocal space

Additional symmetry elements may appear in the unit cell. If such an element relates every molecule in the crystal to another molecule, it is a crystallographic symmetry element. Otherwise it is called a non-crystallographic or local symmetry. Crystallographic symmetry elements are used to reduce the repetition group to the asymmetric unit of the cell.
Any crystallographic symmetry element of the crystal will also be a symmetry element of the reciprocal space. This can easily be verified by writing the structure-factor equation (eq.15) as a summation of atom scattering and introducing the symmetry elements. For the example of a 2-fold rotation axis along **c** with equivalent atom positions (x,y,z) and (-x,-y,z):

$$F(h,k,l) = w\sum_{i/2} f_i \left(e^{2\Pi i(hx_i+ky_i+lz_i)} + e^{2\Pi i(-hx_i-ky_i+lz_i)} \right)$$
$$F(h,k,l) = w\sum_{i/2} f_i \left(e^{2\Pi i(-hx_i-ky_i+lz_i)} + e^{2\Pi i(hx_i+ky_i+lz_i)} \right) \quad (19)$$

$$F(h,k,l) = F(-h,-k,-l) \quad \text{there is a 2-fold axis along c*} \quad (20)$$

f_i denotes the scattering of atom i located at the origin (tabulated in [6])

This conservation of rotational symmetry in reciprocal space has the effect that the total value of the rotation angle Φ of the crystal covered during data collection is halved, if the crystal axis are properly aligned with respect to the rotation axis of measurement.

Note that the observed reciprocal space (the intensities) always has a centre of inversion, described by Friedel's Law:
This law does not hold at wavelength near the absorption edge of the atoms [4].

$$|F(h,k,l)| = |F(-h,-k,-l)| \quad (21)$$

4 Solving the Phase Problem

4.1 Molecular Replacement

If a partial structure has been found (sect.3), structure-factors for this model can be calculated by the structure-factor equation (eq.15). Often an electron-density calculated from the

observed intensities and the phases of the partial model will reveal additional details of the molecular structure (see sect.5), allowing an iterative completion of the structure solution. This requires typically a completeness of the partial model of more than 60 % with data to 2.5 Å resolution.
For one heavy-atoms as a partial structure this approach requires data to a resolution of about 1.2 Å that are rarely available.

4.2 Isomorphous replacement

From two heavy atom derivatives the missing phases can be calculated by the Harker construction, if the heavy atom substructure has been solved and the protein structure is unchanged.
The isomorphy requirement means that eq.22 must hold:

$$\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H \tag{22}$$

The position of the heavy-atoms may be found by analysing the Patterson-functions (sect.3.1). The structure-factors for the heavy-atom sub-structure can be calculated by the structure-factor equation using atomic form-factors (see Appendix A).
To minimise the phase error due to measurement errors for the intensities additional derivative data are often required.

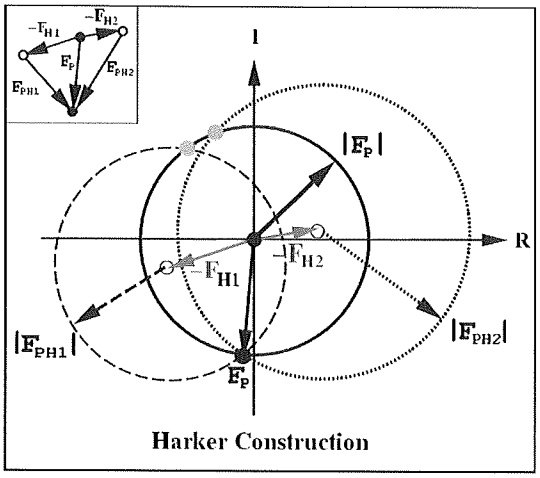


Fig. 7: *Geometric determination of Phases using two heavy-atom derivatives with known substructure by performing he addition of waves in the complex plane. Two derivatives define theoretically the sought phase.*

A useful variations of phase determination by multiple isomorphous replacement (MIR) is the use of anomalous differences. When Friedel's Law does not hold (eq.21) , the differences

$$|F_{ano}(H)| = (|F(H)| - |F(-H)|)/2 \tag{23}$$

encode positional information about the atoms that brought by this difference. Consequently these differences can be treated like an additional derivative.
For multiple wavelength anomalous dispersion (MAD), data are collected at different wavelength around the absorption edge of an atom type in the molecule. Therefore multiple sets of anomalous differences are obtained. By choosing one of the data sets to represent F_P the data for other wavelength may be treated as derivatives as described above for MIR-

phasing. The difference between the observed intensities rely on slight changes of the atomic form-factor as a function of wavelength (Appendix A). These small differences in intensity may become undetectable because of a disturbance by slight non-isomorphism introduces by derivatisation. This problem can be remedied by introducing a scatterer like Selenium that replaces the sulphur in methionine amino-acids. This replacement is accomplished by the organism that produces the protein, when fed with seleno-methionine or selenedioxide under conditions of restricted sulphur supply.

5 Density and Model

The electron-density calculated is interpreted as a chemical model. This means that the structural information is given as atomic co-ordinates and temperature-factors that describe the discrepancy of the electron-density distribution for the atoms from tabulated form-factors (Appendix A).

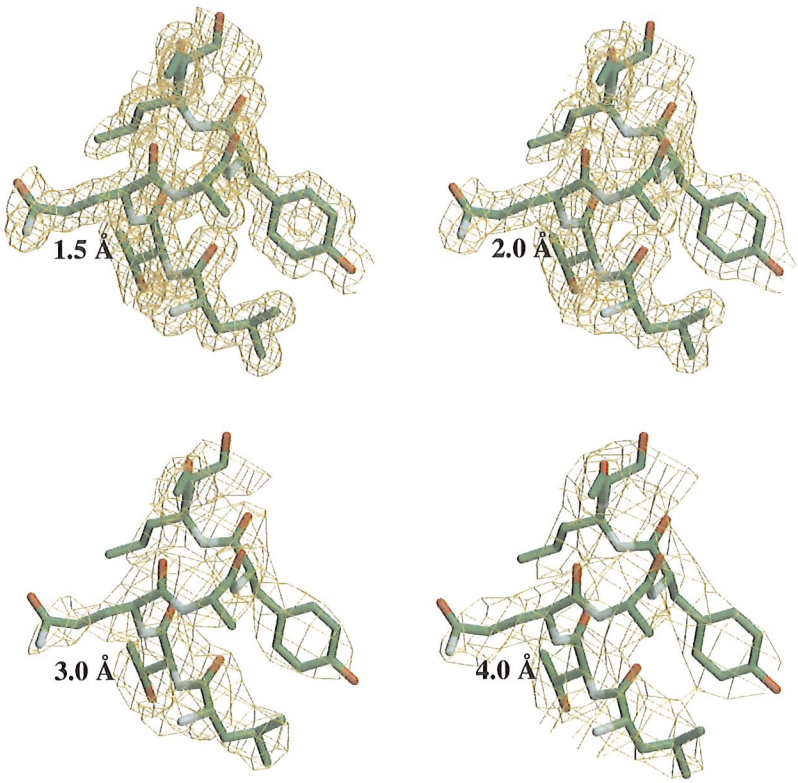


Fig. 8: *Improvement of electron-density as a function of resolution. The densities are recalculated from model co-ordinates. Therefore the densities are free of any error. At 1.5 Å all details are resolved. At 2.0 Å details like the low density region within the tyrosine ring get lost. At 3.0 Å density for terminal groups like the hydroxy-group at the tyrosine disappear. At 4.0 Å the side chain densities for groups ranging from leucyl to tyrosyl become hardly different. Smaller side chain density e.g. at threonine is hardly different from carbonyl-group density.*

Eventually the electron-density must be interpreted in terms of atom positions. Besides phase quality resolution is most important (Fig.8). If the resolution is better than 2 Å automatic procedures (algorithms) may give a complete or nearly complete atomic model. Alternatively a manual interpretation must be performed by finding the chain-trace of the polymer in the density and subsequent identification of the amino-acids.

5.1 Density refinement

An interpretable electron-density may be obtained by a density refinement, if the primary phasing is insufficient. Protein crystals contain typically about 50 % (v/v) solvent. The major part of this solvent is disordered and hardly contributes to the observed intensities. If the solvent region in the unit cell can be identified, it is possible to obtain an improved density using observed amplitudes and phases calculated from a modified version of the original density. Solvent flattening means to remove the spurious density variations in the solvent region. The recalculated density will show improved density also in the protein region, because generally any structure-factor will influence all point in the electron-density (eq. 7).
A similar approach can be used if the asymmetric unit has a local symmetry. The density can be improved by averaging corresponding density point to remove noise.

5.2 Completeness and local correctness (difference-density)

Phases and amplitudes are obtained by calculating structure-factors for a model. The comparison of model amplitudes and amplitudes calculated from observed intensities serves as an indicator for the overall correctness and completeness of the model (s.b.).

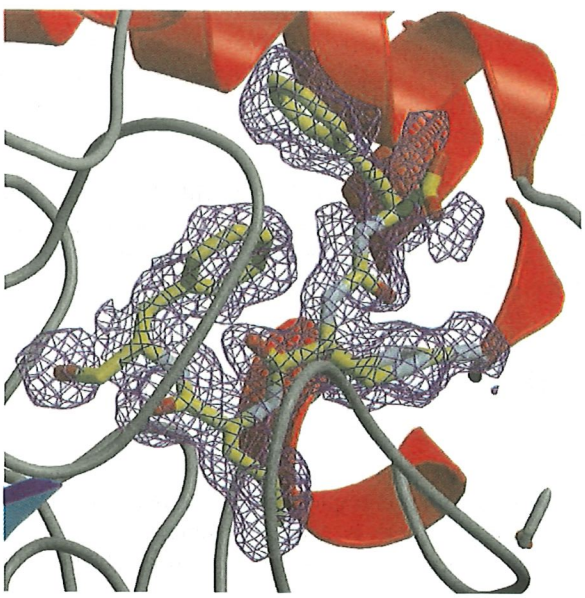


Fig. 9: *Difference-density (blue) for the inhibitor calculated from* $|F_{PI}|-|F_P|$, α_P *for peptideamidase [3].*
PI: from crystal with inhibitor
P: from crystal without inhibitor

Final structure model of inhibitor superimposed.
Carbon-atoms: yellow
Nitrogen-atoms: blue
Oxygen-atoms: red

The structure of the protein is shown as a cartoon:
sheet: blue-green
helical: red
loop: grey)

The local correctness and completeness of a model is investigated by difference-densities. Using the calculated (model) phases α_c two densities may be calculated. The model-density calculated with the (calculated) amplitudes Fc simply represents the current model in terms of electron-density instead of atomic co-ordinates. The mixed density calculated with the

amplitudes from the measurement (Fo) represents a model-biased approximation of the true density. The difference of these densities may be directly calculated employing the difference of the amplitudes (Fo-Fc) in density calculation. This difference-density will ideally be zero for atoms placed properly in the model. Wrongly placed atoms should show up as negative difference-density. A positive signal represents density (atoms) missing in the model. A variant of this type of difference-density is the usage of difference-amplitudes from two experimental data sets employing phases obtained for one of them (Fig.9).

5.3 Global correctness of the model

If the model is complete, the structure solution ends with an refinement of the atomic positions to maximise the agreement between observations and intensities calculated from model co-ordinates by minimising the difference between observed and calculated structure-amplitudes (the R-value):

$$R = \sum_H \| F_o | - | F_c \| / \sum_H | F_o | \tag{24}$$

The process of structure solution requires normally that primary phases e.g. from MIR must refined by applying density modifications (sect.5.1) or combination with information from partial models. During this process several cycles of forward and backward calculations of electron-densities and structure-factors are performed. As any density point contributes to all amplitudes and vice versa model errors are folded into the phases and phase errors and errors in the observation may be folded into the model. This effect is monitored by reserving 5-10% of the data for a free set that is only used for monitoring overfitting. Ideally the crystallographic R (R_{work}) and the R_{free} should be identical. In reality however the model is fitted to the data and therefore also to the error in the data. The discrepancy of the R-values generally indicate the quality of the data and consequently the quality of the model obtained from that data. If the difference of the R-values exceeds some percent-points the number of model parameters must be reduced by using constraints.

A good model will not only agree with the observed diffraction, it will also show a good stereo-chemistry. The observed bond length and angles will have values that are in agreement with chemical knowledge. The overall stereo-chemical agreement is reported by the rms-difference between chemically expected e.g. l_o and the actual values for the model.

$$\mathrm{rms}(\Delta l) = \left(\sum_j \left(l_j - l_{j,0} \right)^2 \right)^{1/2} \qquad 1 := \text{bond length} \tag{25}$$

Restraints on bond-length and angles are regularly employed to improve the effective observation parameter ratio o/p. As o ~ $(|S/\lambda|)^3$ there is a clear resolution boundary that separates principally over-determined problem from under-determined problems that can only be solved by applying constraints.

Appendices

A Appendix: The atomic form factor

Eq.15 describes the scattering in terms of an electron-density function (scale factors omitted).

$$F(h,k,l) = \int \rho(x,y,z)\, e^{2\Pi i(hx+ky+lz)} dV \qquad (15)$$

We may rewrite eq.15 as a summation of the contribution of atoms using:

$$F(h,k,l) = \int \rho(x,y,z)\, e^{2\Pi i(hx+ky+lz)} dV$$

$$F(h,k,l) = \int \sum_j \rho_j(x,y,z)\, e^{2\Pi i\left(h\left[x_j+X_j\right]+k\left[y_j+Y_j\right]+l\left[z_j+Z_j\right]\right)} dV$$

Here x_j denotes the x co-ordinate of an electron-density point with respect to the centre of mass of atom j and X_j denotes the x-coordinate of the centre of mass.
By reordering:

$$F(h,k,l) = \sum_j \mathbf{f}_j\, e^{2\Pi i\left(hX_j+kY_j+lZ_j\right)}$$

with the atomic form-factor (tabulated for the elements in [6]

$$f_j = \int \rho_j(x,y,z)\, e^{2\Pi i(hx+ky+lz)} dV$$

Obviously the tabulated form-factors ${}^0\mathbf{f}$ can not describe the distortions of the electron-density due to chemical bonding and thermal vibrations of the atoms. These phenomena can not be clearly separated in protein crystallography because of lacking data resolution. They are described by a temperature-factor Bj for each atom that relates to the mean square of displacement $\hat{u}^2$:

$$B := 8\Pi^2\, \hat{u}^2$$

Using temperature-factors the effective scattering of the atom is given by;

$$\mathbf{f} = {}^0\mathbf{f} \exp\left[-B\left(\sin\theta/\lambda\right)^2\right]$$

For high resolution data an anisotropic temperature-factor may be required. Note that by the temperature-factor the scattering weakens with increasing resolution.

References

[1]McPherson A., Preparation and Analysis of Protein Crystals
J.Wiley&Sons, USA (1989)
[2]Schlichting I., Physikalische Methoden: Die 4. Dimension in der Kristallographie
Chemie in unserer Zeit 29, 230-240 (1995)
[3]Labahn J. et al. An alternative mechanism for amidase signature enzymes.
J Mol Biol. (2002) 322,1053-64
[4]Helliwell J.R., Macromolecular crystallography with synchrotron radiation
Cambridge University Press (1992)
[5]Blundell, T.L, Johnson L.N., Protein Crystallography
Academic Press London (1976)
[6]International Tables for Crystallography
Kluwer Academic publisher Dordrecht (1995)

B 3 NMR Analysis of Biopolymer Structure

D. Willbold and B. W. Koenig
Institut für Biologische Informationsverarbeitung 2
Forschungszentrum Jülich GmbH
Institut für Physikalische Biologie
Heinrich-Heine-Universität Düsseldorf

Contents

1 Introduction **2**

2 NMR Basics **2**

2.1 Relations between NMR parameters and molecular structure 4

2.2 Size limit of high-resolution NMR in liquids 7

3 Multidimensional FT-NMR **7**

4 Strategy for NMR structure determination **9**

5 Protein dynamics from NMR **10**

6 Molecular complexes **10**

6.1 Slow, intermediate, and fast exchange conditions 11

7 Biological NMR has a bright future **13**

1 Introduction

Since the discovery of the phenomenon of nuclear magnetic resonance (NMR) in 1945 by two independent groups of physicists led by Felix Bloch and Edward Purcell (Nobel Prize in 1952), NMR spectroscopy has evolved into an indispensable and extremely versatile tool for determination of composition, three-dimensional (3D)-structure and dynamic properties of molecules in chemistry and finally biology. Introduction of the Fourier transform (FT) principle in the early 1970s drastically increased the amount of information accessible by NMR and paved the way for high-resolution, multidimensional NMR - a quantum jump recognized by the award of the Nobel Prize to Richard Ernst in 1991. Beginning in the 1980th Kurt Wüthrich (Nobel Prize in 2002) developed a systematic strategy to assign the NMR frequencies of proteins, to measure structural restraints, and to calculate the 3D-structure of biological macromolecules in solution on the basis of NMR data. Today NMR is one of the two principal techniques for determination of atomically resolved protein structures. In contrast to X-ray diffraction, which is the other method, no crystals are required for NMR.

The lecture will provide on overview on basic principles and strategies upon which modern biological NMR spectroscopy is based. The potential of NMR spectroscopy in life science will be illustrated using recent research results from the authors laboratory.

2 NMR Basics

NMR relies on the interaction of the magnetic moment of nuclei with an external magnetic field. The magnitude of the angular momentum, J, of a nuclei depends on its spin quantum number I, which is either integral or a half-integral:

$$\left|\vec{J}\right| = \sqrt{I(I+1)}\hbar \tag{1}$$

Here $\hbar$ is h/2π, where h is Planck's constant. In a magnetic field the angular momentum is quantized and can adopt (2I+1) different orientations. The projection of J on the magnetic field direction is $m\hbar$, where m is the direction quantum number that can assume either any integer or any half-integer number from –I to +I. The angular momentum J is associated with a magnetic moment, μ. The component of μ in the magnetic field direction (μ_z by definition) is given by:

$$\mu_z = m\gamma\hbar \tag{2}$$

where γ is the gyromagnetic ratio of the nucleus, which is an isotop specific constant. Each of the (2I+1) allowed orientations of J in a homogeneous magnetic field B_0 is associated with an energy level:

$$E_m = -m\gamma\hbar B_0 \tag{3}$$

Only nuclei with a non-zero nuclear spin give rise to NMR phenomena. Three spin ½ nuclei are almost exclusively used in high-resolution NMR on biomacromolecules, ^{1}H, ^{13}C, and ^{15}N. Unfortunately, the most abundant carbon (^{12}C with I=0) and nitrogen (^{14}N with I=1) isotopes are not suited for high-resolution NMR in solution. Isotope labelling strategies have been developed to produce biomolecules enriched with ^{13}C and/or ^{15}N isotopes for NMR studies.

Spin ½ nuclei, to which we will restrict our analysis, can occupy only two different states in a magnetic field: state α with m=+½ and state β with m=-½. The energy difference between these two states is:

$$\Delta E = E_\beta - E_\alpha = \gamma\hbar B_0 = \hbar\omega_0 \tag{4}$$

where ω_0 is called the Larmor frequency of the spin. In thermal equilibrium the occupation of the two states, N_α and N_β, respectively, follow Boltzmann statistics:

$$\frac{N_\alpha}{N_\beta} = \exp\{-\Delta E/k_B T\} \approx 1 - \frac{\Delta E}{k_B T} = 1 - \frac{\gamma\hbar B_0}{k_B T} \tag{5}$$

The energy difference ΔE is very small compared to the thermal energy $k_B T$, resulting in a tiny difference between N_α and N_β on the order of parts per million (ppm). This small ΔE causes the inherently low sensitivity of NMR in comparison with most other spectroscopic techniques.

In the classical description of NMR a spin ½ nucleus precesses with its Larmor frequency on the surface of a double cone around the B_0 direction of the NMR magnet (Fig. 1). The slightly higher number of spins in state α gives rise to a macroscopic magnetization vector, M_0, along the B_0 direction. Transitions between the two energy levels are induced by irradiation of the spin system with a radio-frequency (r.f.) that matches the Larmor frequency.

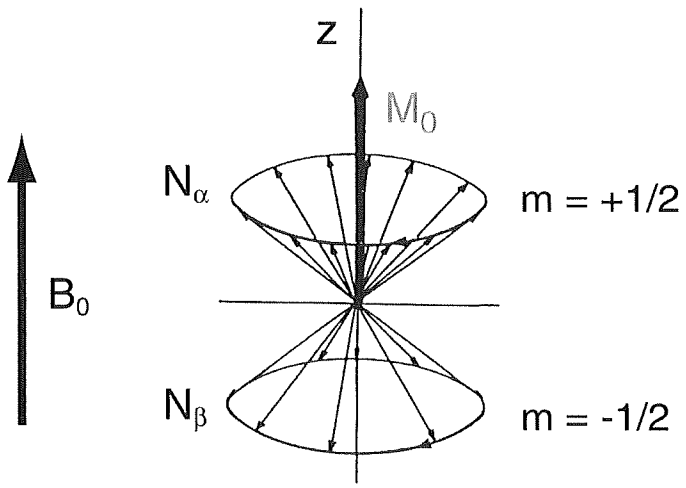


Fig. 1: *Precession of spin ½ nuclei on the surface of a double cone in the static B_0 field of the NMR magnet results in a macroscopic magnetization M_0*

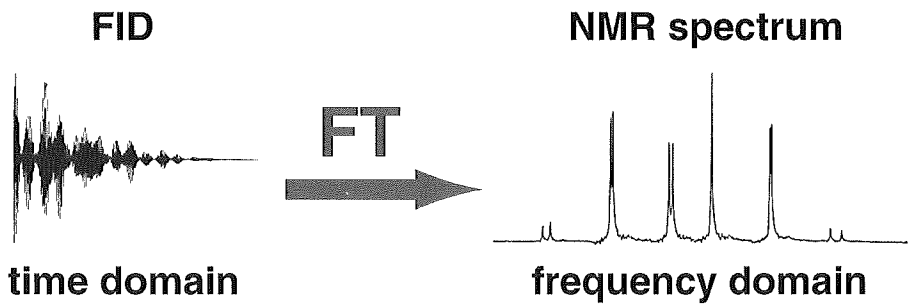


Fig. 2: *In pulsed Fourier transform (FT) NMR spectrometers transverse magnetization precessing in the receiver coil during data acquisition gives rise to an oscillating voltage with an amplitude that decays in time. This time domain signal is referred to as free induction decay (FID). The NMR spectrum in the frequency domain is derived from the FID by a Fourier transformation.*

Sample irradiation at frequency ω_0 can be applied in such a way, that it is equivalent to the presence of a constant magnetic field B_1 directed along the x-axis of a coordinate frame, which rotates around B_0 (i.e. the z-axis) at a rate corresponding to the Larmor frequency ω_0. In the classical picture, B_1 exerts a torque on the magnetization vector M_0, resulting in precession of M_0 around the x-axis. The angle covered by the precessing vector M_0 is determined by the duration of the irradiation. If irradiation is stopped once M_0 has arrived in the transverse xy-plane, the irradiation pattern is called a 90° pulse. Following a 90° pulse M_0 undergoes precession in the transverse plane. However, precession of the individual spins that make up M_0 will gradually dephase as a result of spin-spin interactions, a process called transverse or T_2 relaxation. Over time, the equilibrium spin state with M_0 aligned parallel to B_0 will recover due to spin-lattice interactions, also known as longitudinal or T_1 relaxation. Depending on the details of r.f. irradiation a pulse can excite all spins of a given isotopic species (hard pulse) or the spins within a narrow frequency band only (selective or soft pulse). During the lifetime of transverse magnetization, precession of M_0 gives rise to an electrical signal in the receiver coil, which wraps around the sample tube. The detected signal is a superposition of electrical currents oscillating with the precession frequencies of all spins in the transverse plane. The amplitude of the signal decays over time due to T_2 relaxation. The time domain signal is also referred to as free induction decay (FID). A Fourier transformation (FT) renders the time domain signal into a frequency domain NMR spectrum (see Fig.2).

2.1 Relations between NMR parameters and molecular structure

The gyromagnetic ratio γ is an isotope specific constant. Therefore, the Larmor frequencies and the spectroscopic sensitivity of ^1H, ^{13}C, and ^{15}N spins are quite different. For example, in a 14.1 Tesla magnet ^1H, ^{13}C, and ^{15}N nuclei resonate at 600, 150 and 60 MHz, respectively, and protons are by far the most sensitive nuclei in biological NMR. A closer view at the spectral regions near the three Larmor frequencies reveals a rich fine structure. The insert in Fig.3 shows a high-resolution ^1H spectrum of a protein, which typically covers a frequency range of about 10 ppm of the Larmor frequency, i.e. about 6 kHz in case of a 14.1 Tesla instrument. The signal dispersion reflects the influence of the electronic and magnetic environment of the nucleus on the local field experienced by the spin, which is referred to as nuclear shielding. The resulting frequency offset of an observed spin relative to a reference

frequency (typically the ^{1}H signal of tetramethylsilane in proton spectra) is used to define the so called ***chemical shift*** of the spin in ppm units. Protons localized in comparable nuclear environments are expected to show similar chemical shifts. Indeed, in a typical protein spectrum resonances can be grouped. Backbone amide protons show the weakest shielding resulting in rather high resonance frequencies, followed by protons in aromatic rings and side chain amide protons. H^{α} protons experience an intermediate chemical shift. The highest degree of shielding is observed for side chain methylene and methyl protons in that order (Fig.3). Moreover, the chemical shift of backbone H^N and H^{α} protons as well as of C^{α}, C^{β}, and C' carbons show a statistically significant dependence on the local secondary structure of the polypeptide chain.

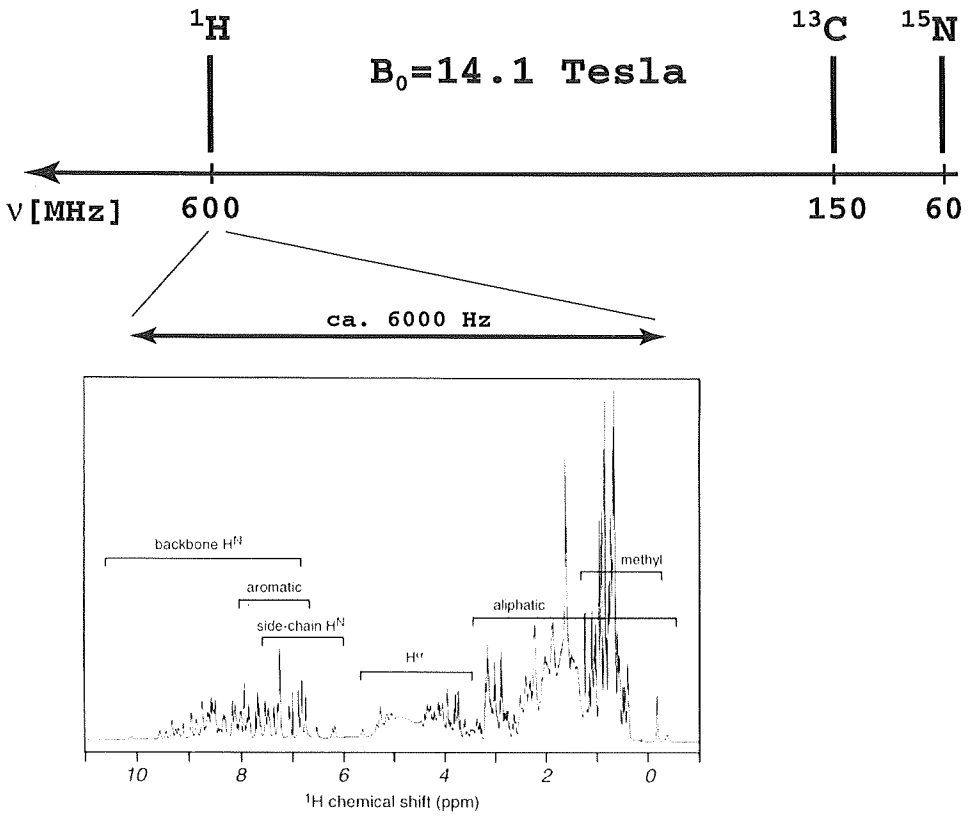


Fig. 3: *In a 14.1 Tesla magnet ^{1}H, ^{13}C, and ^{15}N nuclei resonate at 600, 150 and 60 MHz, respectively. The insert shows a typical high-resolution ^{1}H NMR spectrum of a protein with a frequency dispersion of about 6 kHz at this field strength. Frequency ranges for specific types of protons in a protein are indicated.*

Nuclei that are connected by covalent bonds do interact with each other by ***scalar or indirect spin-spin coupling***, which is mediated by bonding electrons. Scalar coupling causes slight modifications of the energy levels of the interacting spins that are reflected in typical multiplet splittings of resonance lines. The scalar coupling constant, J, specifies the strength of the interaction and is given by the multiplet splitting (in Hz). The scalar interaction is

strongest for directly bonded spins (nomenclature: $^1J_{IS}$), followed by geminal ($^2J_{IS}$) and vicinal ($^3J_{IS}$) couplings. Scalar coupling between nuclei with more than 5 intervening bonds are usually too weak for detection. Typically, only the isotropic component of scalar coupling is of concern in high-resolution NMR in liquids. Isotropic scalar coupling is independent of molecular orientation and B_0 field strength. The strength of vicinal scalar couplings within the protein backbone is sensitive to the corresponding torsion angles, i.e. ^3J contains structural information. The dependence of ^3J on a torsion angle ϕ is given by the semi-empirical Karplus relation:

$$J(\phi) = A\cos^2(\phi - 60) - B\cos(\phi - 60) + C \qquad (6)$$

where A,B, and C are empirically determined constants that have specific values for individual types of torsion angles.

Heteronuclear scalar couplings in proteins provide a rather selective mechanism by which magnetization can be transferred between coupled nuclei. This effect is widely used in multidimensional heteronuclear correlation experiments that allow efficient assignment of resonance frequencies in biopolymers. Another application of heteronuclear through-bond couplings is heteronuclear-editing of ^1H-^1H spectra (see below).

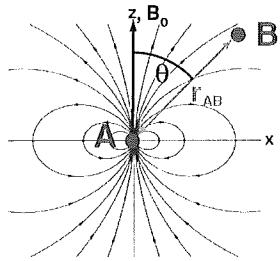


Fig. 4: *A spinning nucleus gives rise to a magnetic dipole field. For clarity the dipole field arising from spin A is drawn only. The effective magnetic field sensed by spin B is a superposition of the static field B_0 and the dipole field of spin A.*

A spinning nucleus acts as a small magnet and is thus the source of a magnetic dipole field. If an isolated two spin system of nuclei A and B is subjected to the B_0 field of the NMR spectrometer, the effective magnetic field at the position of B will be a superposition of B_0 and the local component of the dipole field originating from A (Fig.4). As a result of ***dipolar or direct coupling*** between two immobilized spins the resonance lines of nuclei A and B will both show a splitting that amounts to:

$$D_{AB} = -\frac{\mu_0}{8\pi^2}\hbar\frac{\gamma_A\gamma_B}{r_{AB}^3}\frac{3\cos^2\theta - 1}{2} \qquad (7)$$

where μ_0 is the permitivity of vacuum, r_{AB} is the internuclear distance between A and B, and θ specifies the angle between the internuclear vector AB and the B_0 field direction. D_{AB} is a rich source of structural information due to the dependence on r_{AB} and θ. For directly bound nuclei and for neighbouring protons D_{AB} can be very large and may reach tens of kHz. In the presence of molecular motion a reduced dipolar splitting is observed that derives from Eq. (7)

if the second Legendre polynomial is replaced by its time average. Small proteins with a molecular weight below 30 kDa typically undergo rapid isotropic rotational diffusion in aqueous solution at ambient temperature that is characterized by a correlation time of a few nanoseconds. Dipolar couplings are usually averaged to zero under such conditions and the geometric information is lost. However, residual dipolar couplings (RDCs) can be recovered if a small anisotropic component is imposed on the motion of the solute. An anisotropic motion of the solute may be accomplished by incubation with a dilute liquid crystalline phase. Typically, RDCs are scaled down by a factor of about 1,000 relative to the static coupling. RDCs are observed for the strongest dipolar interactions only. Due to the small size of RDCs (in the single to double digit Hz range) the high-resolution character and simplicity of the liquid-state NMR spectra is retained.

Dipole-dipole interactions between nuclear spins also provide a very efficient relaxation mechanism. In particular, direct spin-spin coupling is the physical basis of cross-relaxation, a magnetization transfer mechanism that gives rise to the ***Nuclear Overhauser enhancement*** (NOE) effect. If the magnetization of a particular spin is perturbed, the spin system tries to re-establish the equilibrium state by exchanging magnetization between the manipulated spin and spins in its surroundings. Magnetization transfer occurs at a rate that is related to r^{-6}, where r is the distance between the pair of spins. The size of signal enhancement for an interacting spin is proportional to r^{-6}, provided the NOE is observed shortly after the perturbation. As a result of the r^{-6} dependence the NOE effect decreases rapidly as the internuclear distance r increases. For protons no effect is observable for distances beyond 5 Å. Relaxation phenomena that are based on dipolar couplings are always operational, no matter if dipolar splittings are visible in the NMR spectra of the solute or if they are averaged to zero by fast isotropic motions. Internuclear distances between the protons of the biomacromolecule studied are usually the most important source of structural information available from NMR.

2.2 Size limit of high-resolution NMR in liquids

Anisotropic, tensorial interactions like dipolar couplings, the chemical shift anisotropy, and the anisotropic component of the scalar spin-spin interaction can cause extensive line broadening and an intangible superposition of multiplet pattern in the NMR spectrum. The high-resolution character of solution-NMR spectra with narrow spectral lines relies on nearly complete averaging of such anisotropic interactions by fast Brownian motion of the molecule under study. The rate of rotational diffusion of a solute depends on several factors, in particularly on the size of the molecule. This puts an upper limit on the molecular weight of the biomolecules that can be studied. For current routine structure determination at atomic resolution this limit is on the order of 30 kDa. However, by using a new class of NMR experiments, which are based on total relaxation optimised spectroscopy (TROSY), resolved spectra can be obtained from substantially larger molecules (so far up to 900 kDa) and strategies for structure determination of proteins larger than 40 kDa are under development.

3 Multidimensional FT-NMR

The FT-NMR principle allows simultaneous excitation of the entire spectrum of the studied nucleus and thus dramatically accelerates the process of data acquisition. A moderately sized protein contains no less than thousands of protons. Even under ideal conditions with a line width of the ^1H NMR signals of a few Hertz there is heavy overlap in a conventional one-dimensional NMR spectrum, which prevents unique signal assignment and collection of structurally relevant NMR parameters. The problem of signal overlap has been overcome by

an ingenious concept, that allows to spread the NMR signals in two, three or even four dimensions. What is a two-dimensional (2D) NMR experiment? In a 1D experiment the spin system is forced to evolve under the influence of a specific r.f. pulse sequence, the transverse magnetization present at the end of this sequence is read out by recording the FID s(t) as a function of time, and the time-domain spectrum s(t) is transferred to the frequency domain by a Fourier transformation. A 2D experiment consists of a series of 1D experiments, which are all identical except for one systematically incremented delay, t_1, in the pulse sequence. This variable delay constitutes a second time dimension. A typical 2D experiment can be divided into four essential building blocks (Fig. 5). First, the spin system is subjected to a well defined sequence of r.f. pulses and delays to create a desired spin state (preparation period). Second, the spin system is allowed to evolve during the variable delay time t_1 (evolution period). The state of the spin system present at end of the evolution period is highly specific and depends on the length of t_1. During the following "mixing period" another specific pulse sequence is applied to the spin system to facilitate magnetization transfer and/or to probe interactions between the active spins. Finally, during the detection period the spins precessing in the xy-plane give rise to a FID, $s(t_1, t_2)$, that is recorded as a function of the time t_2. A two-dimensional FT converts the data into a 2D frequency domain spectrum $S(\omega_1, \omega_2)$, where the precession frequencies during evolution and detection determine the coordinates of the signal peaks.

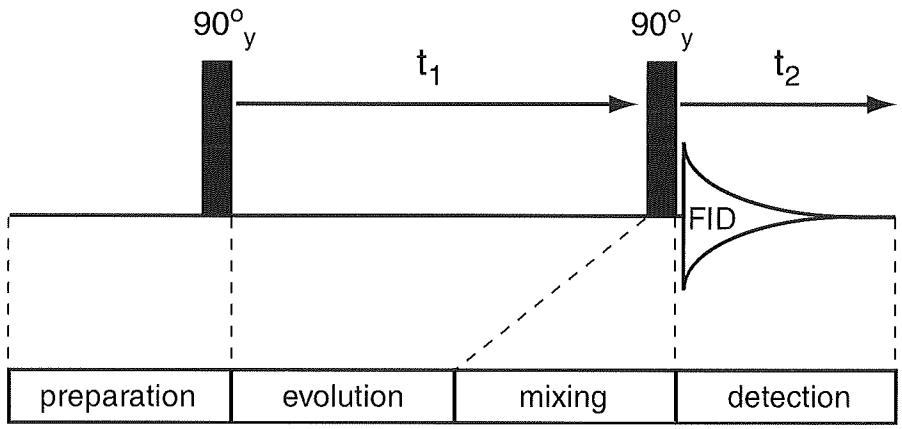


Fig. 5: *Basic scheme of a two-dimensional NMR experiment (below). Four building blocks are distinguished (cf. text). The pulse sequence of a basic COSY experiment is shown above. In the COSY experiment both the preparation and the mixing period consist of a single 90° r.f. pulse. The indirect time domain (variable delay, t_1) and the direct time domain (time passing during acquisition of the FID, t_2) are sampled during the evolution and detection periods, respectively.*

Suitable combinations of different preparation, evolution, mixing, and detection schemes allow for a wide range of homo- and heteronuclear NMR experiments. The pulse sequence applied during the mixing period can facilitate magnetization transfer by through-bond (scalar) or by through-space (dipolar) interactions. The scheme in Fig.5 can be extended by inclusion of additional sets of evolution and mixing periods, allowing for 3D and 4D experiments.

4 Strategy for NMR structure determination

Sample preparation with the biological macromolecule of interest under close-to-native conditions is the first step during 3D structure determination. In most cases, the sample needs to be labelled with stable isotopes, e.g. with ^{13}C and ^{15}N. For proteins, this is usually achieved by recombinant expression in bacteria grown on ^{13}C-enriched glucose and/or ^{15}N-enriched ammonium chloride. Next, all or at least most of the observed ^{1}H, ^{13}C, and ^{15}N resonances must be assigned to nuclei of the macromolecule. An extensive set of 3D NMR experiments is recorded, typically requiring several weeks of instrument time. Dedicated software tools are increasingly used in the assignment process. Nevertheless, resonance assignment remains the most time-consuming part of structure determination.

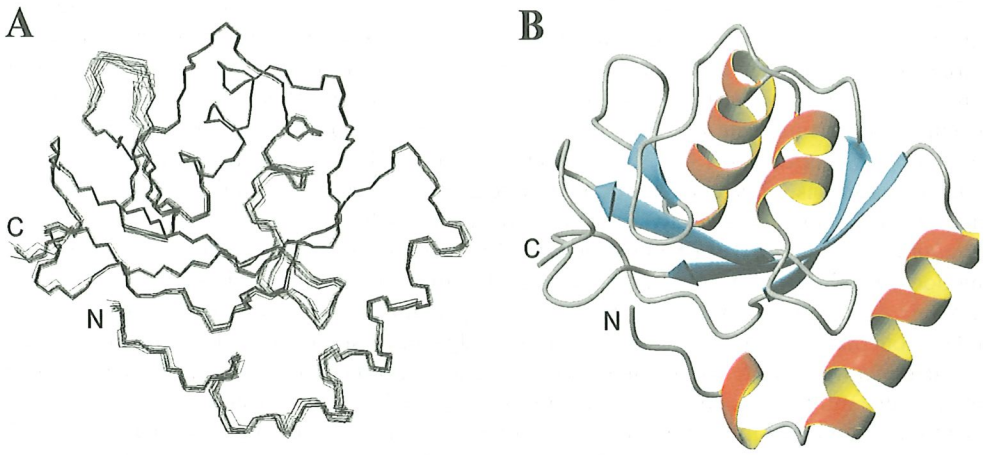


Fig. 6: *NMR structure of the human GABA receptor type A associated protein (GABARAP). A total of 4634 NOE based distance constraints (including 1387 long-range NOEs) were derived from 3D NOESY spectra. A: Superposition of the backbones of all 15 calculated structures. The rmsd between backbone atoms is 0.028 nm. B: Ribbon diagram of the averaged GABARAP structure [1].*

The third step is extraction of structure-relevant information from the NMR data. Usually, this requires acquisition of additional multidimensional NMR spectra. NOE based cross-relaxation reveals distances between pairs of protons. ^{3}J couplings contain information on torsion angles. Chemical shift of selected nuclei (e.g. H^{N}, C^{α}, C^{β}, H^{α}, C') are sensitive to local secondary structure. RDCs provide the relative orientation of bond vectors.

The final step of NMR structure determination is a computer-based search for three-dimensional model structures that are consistent with the entire set of NMR-derived structure restraints. The search is often implemented as a restrained molecular dynamics (rMD) simulation. Empirical force fields (based on typical values of bond lengths, bond angles, and dihedral angles as well as on electrostatic and van der Waals potentials) are combined with pseudoenergy terms derived from the NMR data (^{1}H-^{1}H distances, torsion angles, bond vector orientations, etc.). These potentials drive the structure toward a conformation with reduced restraint violations by following a heatup-cooldown protocol, referred to as simulated annealing (SA). The SA-protocol aims to identify atomic coordinates of the macromolecule that satisfy the experimental structure restraints in an unbiased fashion while exploring all regions of conformational space. The NMR data do not uniquely define a 3D structure,

because in general only a sparse subset of geometrical constraints is observable. Therefore, structure calculation is repeated frequently using randomly selected starting structures to produce an ensemble of low-energy structures (Fig. 6). A "good" ensemble shows minimal violations of input restraints while the root-mean-square-deviation (rmsd) between the atomic coordinates of the calculated structures closely matches the experimental uncertainty of the NMR derived restraints.

5 Protein dynamics from NMR

NMR can provide site-specific insight into protein dynamics and kinetics of exchange processes in addition to atomically resolved protein structure. Various NMR parameters like relaxation times, heteronuclear NOEs, order parameters, partially averaged tensorial quantities, or simply the NMR line-shape are sensitive to different types and modes of molecular motion. This wide array of experimental NMR parameters allows analysis of processes with time scales from picoseconds to seconds. Protein backbone dynamics occurring on a timescale of 10^{-8} s are observable by ^{15}N spin relaxation techniques, which monitor the reorientation of ^{15}N-H^N bond vectors or by ^{13}C relaxation measurement of backbone carbons. Side chain motions (timescale of 10^{-10} s) can be studied based on ^{13}C, ^{15}N, and ^{2}H relaxation data. Analysis of such data allows to differentiate between the *global motion* of the protein, described by the rate of overall molecular tumbling in case of isotropic motions or otherwise by the components of the diffusion tensor, and site specific *internal motions*, frequently characterized by means of a generalized order parameter. Combination of relaxation time measurements with molecular dynamics simulations does provide a rather detailed picture of protein dynamics. Chemical and conformational exchange processes on the micro- and millisecond time scale are probed with relaxation measurements or line-shape analysis. Rotation of membrane-anchored proteins around the membrane normal occur in the microsecond range and may get reflected in partial averaging of dipolar couplings and ^{2}H NMR quadrupolar anisotropy.

Lateral diffusion of non-aggregated proteins within a membrane is slower (milliseconds) and gives rise to partial averaging of the chemical shift anisotropy of ^{31}P nuclei. However, investigation of membrane protein dynamics requires solid-state NMR techniques, which are outside the scope of this manuscript.

6 Molecular complexes

Binding strength is of central importance for selection of appropriate NMR experiments in studies of molecular complexes. Strongly bound complexes with an affinity that exceeds the micromolar range can often be considered as a single entity and the weight limit for high-resolution NMR applies to the overall complex. Within this weight limit the structure of the entire complex is accessible. Isotope labelling of only one interaction partner allows reduction of spectral complexity and application of elaborate spectral editing schemes. Deuterium labelling might be advantageous to reduce dipolar relaxation and thus spectral linewidth.

Weakly bound complexes (affinity smaller than 10 mM) offer an opportunity to obtain at least some structural data of molecular assemblies that exceed the weight limit of high-resolution NMR. Of practical importance is the case where the interacting molecules have very different size. The smaller molecule in the complex is usually termed the "ligand". Fast exchange between free and bound ligand is crucial for a class of experiments that allow determination of the bound ligand structure. Cross-relaxation is much more efficient in the macromolecule-

bound ligand compared to the free state. Dipolar couplings are zero in case of an isotropically tumbling free ligand but can be very large if the ligand is bound to an aligned macromolecule. NOEs and RDCs detected with standard high-resolution NMR on the free form of a rapidly exchanging ligand are therefore dominated by contributions originating in the bound state and can be used to characterize of the bound ligand. This approach was recently used to determine the high-resolution structure and orientation of a fragment of the G protein transducin bound to the membrane receptor rhodopsin in its natural environment, a disk membrane vesicle prepared from rod cells of bovine retinas [2].

Changes of chemical shifts of protein spins upon ligand binding can be exploited to map the protein's ligand-binding site(s) and to quantify the interaction strength. The chemical shift reflects the immediate environment of a nuclear spin, any binding of a ligand close to the observed nucleus is likely to change it ("chemical shift perturbation"). Amino acid residues that show chemical shift changes upon addition of ligand must be somehow affected by the binding. Very likely, these residues are part of the ligand-binding site. Heteronuclear single quantum coherence (HSQC) spectra are particularly suitable for such "NMR titration" studies, usually conducted with uniformly ^{15}N-labelled protein (Fig. 7). The ^{1}H-^{15}N HSQC experiment correlates the chemical shift of the ^{15}N-nucleus of a NH_n-group with the chemical shift of the directly bound proton(s). A ^{1}H-^{15}N HSQC spectrum of a protein shows exactly one cross-peak (specified by the ^{1}H and ^{15}N NMR frequencies of the correlated spins) for each amino acid that arises from the backbone amide group (except for prolines) plus additional cross-peaks for amino acids that have NH_n-groups in their side chains (the amide groups of Asn and Gln, the amino group of Lys, the guanidinium group of Arg, and the aromatic H^N protons of Trp and His).

Provided ligand binding is sufficiently strong it is very likely that it can be localized based on chemical shift perturbations in the assigned ^{1}H-^{15}N HSQC spectrum no matter were on the protein the binding occurs. Alternatively, cross-saturation NMR experiments have been used to map ligand-binding sites.

6.1 Slow, intermediate, and fast exchange conditions

Chemical shift perturbation experiments rely on the existence of two different states of the protein studied, a ligand-bound and a ligand-free conformation. A protein resonance affected by ligand binding contains contributions from both conformations. The appearance of the resulting NMR signal depends on the time scale of the exchange between both states. If the exchange rate is "slow" compared to the chemical shift difference, $\Delta\omega$, between the resonance frequencies of an observed spin in the bound and free states, respectively, then the intensity of the free-state resonance will decrease while the bound-state NMR signal will appear and grow upon continued addition of ligand. In case of "slow" exchange, the positions of the NMR signals do not change. In contrast, if the exchange rate is "fast" compared to $\Delta\omega$, then only one NMR signal is observed for a given nuclear spin and its frequency will shift from the ligand-free to the ligand-bound position during titration. The situation is more complex in case of "intermediate" exchange where the exchange rate is similar to $\Delta\omega$. Typically, this causes appreciable line broadening and can even lead to the disappearance of the NMR signal.

An example of chemical shift titration experiments is given in figure 7. ^{1}H-^{15}N HSQC spectra of ^{15}N-labelled full-length Nef protein from HIV-1 were recorded in the presence of increasing amounts of the CD4(403-433) peptide, a fragment of the human CD4 receptor [3].

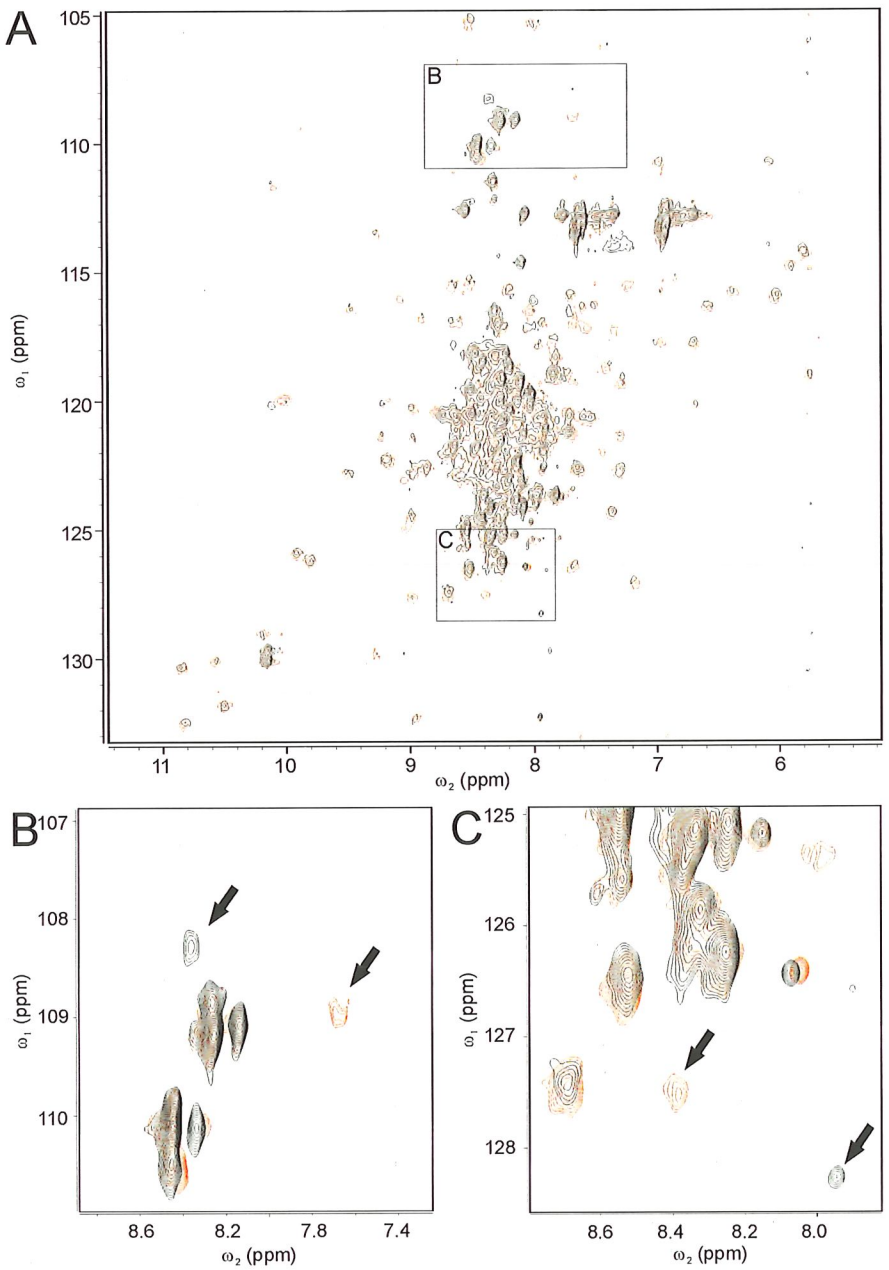


Fig. 7: *Overview (A) and two enlarged regions (B and C) of superimposed ^{1}H-^{15}N HSQC spectra of human immunodeficiency virus type 1 (HIV-1) Nef protein (180 µM, ^{15}N-labelled) in the absence (red contour lines) and presence (black contour lines) of an equimolar amount of CD4(403-433) peptide. During titration with peptide (not shown) the red coloured peaks indicated by arrows did not shift, but their intensities decreased and new peaks (black coloured, indicated by arrows) appeared. NMR spectra were recorded at 750 MHz proton frequency on a Varian Unity INOVA spectrometer at 25 °C [3].*

A dissociation rate of less than 100 Hz is expected for the complex of Nef with the CD4 peptide assuming a dissociation constant of about 1 μM or lower, even if the association rate is diffusion controlled (10^7 to 10^8 Hz M^{-1}). Thus, exchange between free and CD4-bound Nef should be slow on the NMR chemical shift time scale for at least some of the ^1H,^{15}N amide resonances. Indeed, intensities of a few amide resonances in the ^1H-^{15}N HSQC spectra decreased without shifting while new resonances appeared. The intensities of the new signals increased upon titration with CD4 peptide. The proton chemical shift differences of the two pairs of resonances labelled with arrows in figures 7B and 7C amount to 510 and 330 Hz, respectively, indicating that exchange between bound and unbound Nef is significantly slower than 300 Hz. A number of other resonances (one can be seen in Fig.7C) shifted during titration with CD4 peptide by up to 30 Hz, suggesting the dissociation of the complex is fast compared to this time scale. Taken together these observations indicate a dissociation rate of the complex of about 100 Hz. Assuming a diffusion controlled association rate of $< 10^8$ Hz M^{-1} this translates into a dissociation constant of 1 μM or less, which is in perfect agreement with results obtained with other methods.

7 Biological NMR has a bright future

Rapid progress has been made in several areas of biological NMR over the past few years. Introduction of the TROSY principle in the late nineties dramatically increased the size of supramolecular assemblies that are amenable to NMR. In combination with new isotope labelling strategies and efficient ways to measure RDCs the TROSY principle enables investigation of molecular interactions in large functionally active complexes of two or more macromolecular components. Solution NMR is now being used to study large protein-nucleic acid complexes as well as membrane proteins reconstituted and solubilised in mixed micelles with detergent and lipids.

It became feasible to employ NMR in drug discovery and optimisation projects with large receptors. Big efforts are devoted to development of automated high-throughput NMR methods for screening large compound libraries for potential drug leads. Small organic molecules that bind to proximal subsites of a drug target are identified by NMR-based mapping of binding sites, optimised, and finally linked together to produce high-affinity ligands, a strategy dubbed "SAR (structure activity relations) by NMR".

The increasing number of completely sequenced genomes including the human genome has ignited "structural genomics" initiatives in several research centres around the world. Currently there is an urgent need to develop high-throughput technology for structure determination. The ultimate goal is generation of a comprehensive atlas of protein folds. High-resolution NMR is a major player in this rapidly evolving field. Recent efforts are targeted on development of highly automated protein production, data acquisition, and data processing methodology. The focus is particularly on automated resonance assignment as well as on iterative evaluation of NOESY spectra and structure calculation. The large number of atomically resolved protein structures deposited in data banks provide a rich source of sequence-specific structure information that is increasingly used to supplement "real" data, e.g. as "data base derived potentials" in rMD driven structure determination or in structure prediction algorithms.

High-resolution NMR is successfully used to address the problem of protein folding. Besides characterising structure and/or flexibility of folding intermediates, NMR can follow the structural and dynamic changes taking place during folding in real time. Flexibility of certain protein domains is crucial for function and intermolecular recognition. For example, in protein misfolding diseases like bovine spongiform encephalopathy (BSE) it is the conformation and partial flexibility that distinguishes the benign "cellular" (correctly folded) and the disease related (misfolded) form of the same protein (e.g. the prion protein).

The current chapter can only provide a very sketchy and naturally limited summary of the basic methodology and current research areas of biological NMR. For an in depth account of the theoretical framework as well as on practical aspects of modern protein NMR superb monographs are available (e.g. [4], [5]). More recent applications and novel developments in the field of biological high-resolution NMR are discussed in recent review articles (e.g. [6], [7]) and books (e.g. [8]).

References

[1] T. Stangler, L.M. Mayr, D. Willbold, J. Biol. Chem. 277, 13363 (2002)
[2] B.W. Koenig, G. Kontaxis, D.C. Mitchell, J.M. Louis, B.J. Litman, A. Bax, J. Mol. Biol. 322, 441 (2002)
[3] A. Preusser, L. Briese, A. Baur, D. Willbold, J. Virol. 75, 3960 (2001)
[4] Protein NMR Spectroscopy: Principles and Practice, J. Cavanagh, W.J. Fairbrother, A.G. Palmer, and N.J. Skelton (Academic Press, 1996)
[5] Bioanalytik, Kapitel 15: NMR-Spektroskopie von Biomolekülen, F. Lottspeich, H. Zorbas (Hrsg.), (Spektrum Akademischer Verlag, 1998)
[6] K. Wüthrich, Bioscience Reports 23, 119 (2003) (Nobel lecture)
[7] A. Bax, Protein Science 12, 1 (2003)
[8] Nuclear Magnetic Resonance of Biological Macromolecules, Methods in Enzymology, vol. 338 & 339, T.L. James, V. Dötsch, U. Schmitz (Eds.) (Academic Press, 2001)

B 4 Atomic Force Microscopy & Laser Tweezers

A. Engel and M. Hegner
M.E. Müller Institut, Biozentrum und
Institut für Physik Universität Basel

Contents

1 Introduction 2

2 Instrumentation 2

2.1 The Atomic Force Microscope 2
2.2 The Optical Tweezers 3
2.2.1 Origin of Optical Forces 4
2.2.2 Experimental Details 5

3 Imaging 6

3.1 DLVO Forces 6
3.2 Cantilevers & Tips 6
3.3 Examples 7

4 Single Molecule Force Spectroscopy 9

5 Experiments using optical tweezers 13

6 Perspectives 14

1 Introduction

While structural biology now has tools to elucidate the static structure of molecular complexes at atomic resolution, it is still difficult to assess dynamical changes during their functional cycle. To observe biomolecules at work they must reside in their native environment. For soluble proteins this is a physiological buffer, while membrane proteins require in addition to be embedded in a lipid bilayer. The atomic force microscope (AFM) [1] is still the only instrument that provides subnanometer spatial resolution and can be operated in solution. It is therefore an important tool, in spite of the limitation that it can only provide information on the surface structure and dynamics of the biomolecule. Progress has been achieved in several laboratories by optimizing sample preparation [2-5] and image acquisition [6,7] methods, and by continuous developments of the instrumentation[8-11].

Topographs (maps of the surface topography) of biomolecules acquired with the AFM reveal the object in its most native state. The high signal-to-noise (S/N) ratio provided by this instrument allows submolecular features to be discerned on single biomolecules. Structural changes at their surfaces can be detected with a sufficient time resolution to monitor conformational changes involved in biological processes. In addition, the AFM stylus is a nanotool that allows single molecules to be manipulated. Supramolecular structures and single biomolecules can be dissected, using the sharp tip as nanoscalpel, or biomolecules can be attached to the stylus and unfolded while the stylus is retracted. The forces measured during this process give novel information on the nature of molecular interaction and the forces that stabilize the building blocks of live.

While the AFM measures forces between 20 pN and 2000 pN (1 pN = 10^{-12} N), optical tweezers are even more sensitive, allowing the detection of forces as small as 1 pN. Using such optical tools, forces generated by single molecular motors can be directly measured, and the nanomechanics of single biomolecules investigated.

2 Instrumentation

2.1 The Atomic Force Microscope

A topograph is recorded by raster scanning the sample below the stylus that is attached to a flexible cantilever while a servo displaces the sample vertically to keep the cantilever deflection constant. An optical system resolves cantilever deflections of 0.1 nm, which corresponds to a force difference of typically 10-50 pN. With modern instruments stable contact mode operation (see Fig.1) is possible at forces of some 50 pN, provided the sample is in an aqueous solution. Various ways to exploit the deflection signal yield quite different types of images, as illustrated and explained in Fig. 1a. The possibility to raster scan and manipulate biological macromolecules under physiological conditions is a major advantage of the AFM. Fig.1b displays the experimental details of an AFM, which allows scanning a sample in buffer solution, and exhibits an efficient system for changing the buffer.

The sensitivity of the cantilever deflection detector has promoted force measurements with the AFM. Unfolding forces of biomolecules that are tethered to substrate and tip can be monitored while retracting the tip. To record such force-distance curves, the tip is approached vertically towards the sample until a molecule is attached to the tip, and subsequently retracted ('fly-fishing'). Acquisition of high-resolution images before and after 'unzipping' a biomolecules allows the damage produced to be directly visualized. Only specialized AFMs can resolve forces below 20 pN, a limitation preventing the direct monitoring of molecular motors, whose forces are in the range of 1 – 10 pN.

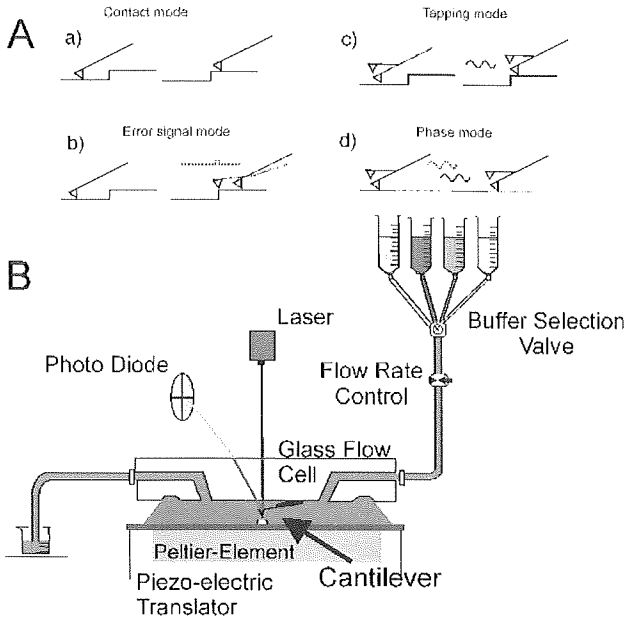


Fig. 1: ***AFM Instrumentation and modes of operation.*** *A) Imaging modes: In the contact mode (a) the servo system moves the tip up over elevations and down over depressions while the sample is raster scanned below the stylus. The servo needs a difference signal to react for keeping the cantilever deflection constant even at sharp edges (b). This error signal provides an image that reveals the edges of the surface topography. Approaching the sample with an oscillating cantilever results in tapping of the sample by the stylus (c). This reduces the oscillation amplitude providing the signal to activate the servo. Because the tip-sample contact is disrupted periodically, the friction forces are eliminated in the tapping mode. The phase difference between the measured oscillation (solid wave) and the oscillation driving the cantilever depends on the mechanical properties of the sample (d). This phase signal thus produces a sensitive material contrast. B) Key elements of an AFM are the cantilever with a pyramidal stylus that touches the sample, an optical lever consisting of a laser and a photo diode to measure the cantilever's deflection, a piezo-electric translator to displace the sample in x, y, z, and a computer to control these movements and store the surface contours. The instrument is working in buffer solution under ambient condition. The cantilever (typical dimensions ~200 x 20 x 0.5 μm, spring constant 0.1-0.01 N/m) is deflected upwards when the tip is pushed towards the sample surface (repulsive forces) or downwards when the tip is retracted from the sample surface (attractive forces). Different liquids can be injected and a Peltier-element allows precise adjustment of the temperature.*

2.2 The Optical Tweezers

Optical tweezers (OTs) measure forces as small as 0.3 pN. To this end, a laser beam is focused to a diffraction-limited spot using a microscope objective. The intense light gradient near the focal region can achieve stable three-dimensional trapping of dielectric objects, varying in size from about a hundred nanometers up to tens of micrometers (Fig.2A). As result of its high force sensitivity OTs is the technique of choice for assessing biomechanical forces. To study the mechanical properties of a single biomolecule a handle can be attached

such as a small dielectric sphere (Fig.2B). Fig.3A illustrates that the force regime in which OTs and AFM can be used overlap nicely. OTs are preferably applied in experiments on molecular motors, entropic elasticity of molecules, and conformational folding of proteins, while rupturing of bonds and molecular interactions are mainly investigated with the AFM. While some technical details of OTs instruments are given below, a more detailed description of OTs is given in [11].

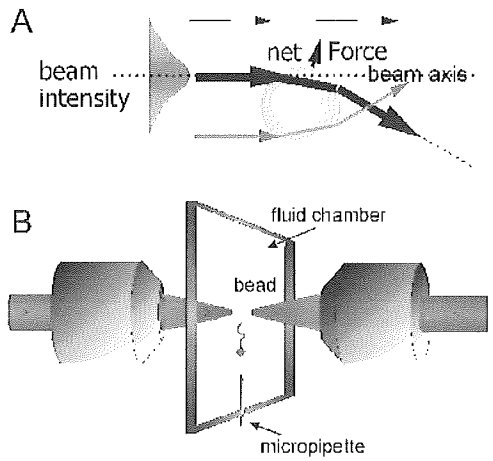


Fig. 2: ***Optical tweezers.*** *A) Optical forces: A transparent dielectric micron-sized particle with an index of refraction larger than the surrounding medium is pushed towards the largest intensity of the light. All light rays are refracted when entering the particle. As result of the second Newton's law, the change in light momentum flux (the force) causes a reaction force on the particle. Center rays contain more photons than outer rays and exert therefore more force. The resulting net force is shown. B) Schematic representation of an OTs experiment. A microscope objective lens with a high NA is used to focus the laser light to a diffraction-limited spot. At the focus, where the spatial gradient is maximal, particle such as beads can be trapped. Light is collected with a condenser lens, which converts angular deflections into transverse deflections that can be monitored on a position-sensitive detector (PSD). A single molecule can be attached between the trapped bead and a bead on a micropipette through a receptor-ligand bridge. Mechanical properties of single molecules can be therefore investigated if the micropipette is placed onto a piezoelectric element. When the trapped bead experiences the force (arrow), it moves slightly away from its stable position. As in AFM, this leads to a deflection on the detector.*

2.2.1 Origin of Optical Forces

As demonstrated in 1970 by Ashkin light can be used to trap and accelerate dielectric micron-sized particles [12]. For this experiment, a stable optical potential well was formed using two slightly divergent counter-propagating laser beams. This pioneering study established the groundwork for the OTs technique [13]. At the diffraction-limited focus of a laser beam, not only dielectric particles spheres can be trapped but also biological organisms such as cells, virus, or bacteria [14-16]. However, optical forces are very small, since 100 mW of power focused to a spot of 1 μm diameter ($\approx 10^7$ W/cm^2) produces forces of only a few tens of pN on

a micron-sized particle (Fig.2A). For biological applications, it is therefore imperative to choose a laser excitation, which (i) does not rise the temperature of the surrounding medium (i.e., water) and (ii) prevents biological damage. Near infrared excitation is best suited, but the wavelength region between 700 and 760 nm should be avoided [17,18].

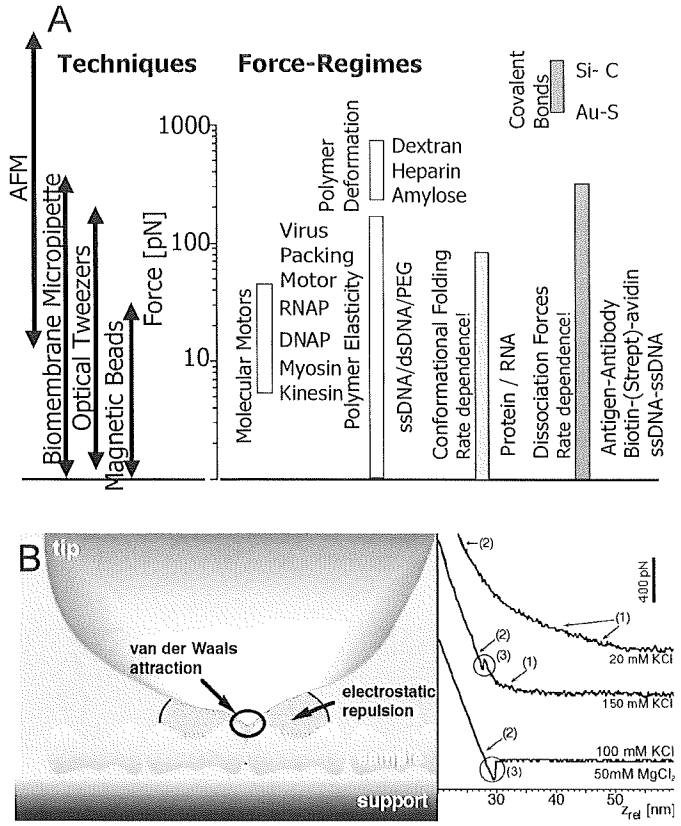


Fig. 3: ***Molecular forces.*** *A) Acquisition of biomolecular forces in aqueous environments. Experiments with motors proteins or molecular mechanics are shown on the right. Unfolding of individual proteins and dissociating biomolecular bonds depend on the rate of the applied external force.B) DLVO Forces: Force-distance-curves recorded on the extracellular purple membrane surface. Data were obtained for different electrolyte concentrations at constant pH (7.6). Force-distance curves were recorded during the approach of sample and AFM tip. The dotted lines (....) represent force-distance curves without electrostatic repulsion. Scan frequency 1.97 Hz; scan range 50 nm (512 pixel). Arrows (1) mark the onset of measurable electrostatic repulsion, whereas arrows (2) indicate the point of contact between tip and sample.*

2.2.2 Experimental Details

An optical trap for investigating biomolecules consists of (i) a beam expander to overfill the back aperture of the microscope lens, (ii) a high-quality microscope lens with a high NA to produce a steep spatial gradient, and (iii) a laser providing a wavelength in the near infrared. Some rather simple modifications of a commercial inverted microscope are sufficient to build

an OTs [19]. When the requirements such as beam steering, high mechanical stability, proper spatial filtering of the laser, and preventing mode hopping of the laser have to be considered, it is best to build an OTs on a conventional optical table with custom optics and electronics [16]. As result of the difference in index of refraction between oil and water inside the chamber oil immersion microscope lenses are not suited for OTs. Fortunately, high numerical water-immersion lenses are now available commercially. OTs setups allow complementary biological information to be gathered simultaneously (Fig.2B).

3 Imaging

3.1 DLVO Forces

To achieve high resolution, topographs are recorded in buffer solution. In a simplified model, electrostatic and van der Waals forces govern the tip sample interactions in aqueous solutions. Hydrophilic surfaces are often charged in water, leading to long-range electrostatic interactions. They can be attractive or repulsive, depending on the surface charges, which depend on the pH. Screening the surface charges with electrolytes allows the electrostatic interactions to be controlled. Since the stylus (silicon nitride, Si_3N_4) is negatively charged at neutral pH, and protein layers are often negatively charged as well, the electrostatic forces are frequently repulsive. In biological systems, van der Waals interactions do not depend on the ionic strength, decay rapidly, and are always attractive. The DLVO (Derjaguin, Landau, Verwey, Overbeek) theory describes these forces quantitatively and allows the interactions between a spherical tip and a planar sample to be modeled, providing clues to optimize the recording conditions [6].

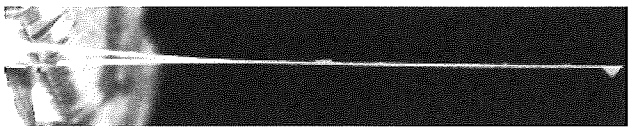


Fig. 4: ***Cantilever & tip.*** *The cantilever is a blade about 1 μm thick and 200 μm long. At ist end a pyramidal tip protrudes that has a spherical end with a radius of 10 nm.*

3.2 Cantilevers & Tips

Cantilevers are produced by microfabrication (Fig.4). The geometry, the density and the tensile strength of the material dictates their mechanical properties. Commercial cantilever suitable for biological applications exhibit a force constant of 0.5 to 0.1 N/m and resonance frequencies in the range of 10 Khz when operated in fluid. While suppliers specify tip radii of 10-50 nm, topographs of flat biological surfaces that exhibit a resolution of 1 nm have been acquired routinely [4,20-22]. Therefore, the tips employed most likely had a single nm-sized asperity that protruded sufficiently to contour the finest surface structures. Such a small asperity exerts a prohibitively high pressure on the underlying structure, inducing its deformation. However, electrolytes can be used to adjust the tip-sample interactions, provided that the electrostatic forces are repulsive. The tip then surfs on a cushion of electrostatic repulsion while the small asperity is in contact with the sample [6]. Fig.3B illustrates this situation, and it shows the changes of repulsive forces between the stylus and the sample resulting from changes of the ionic strength.

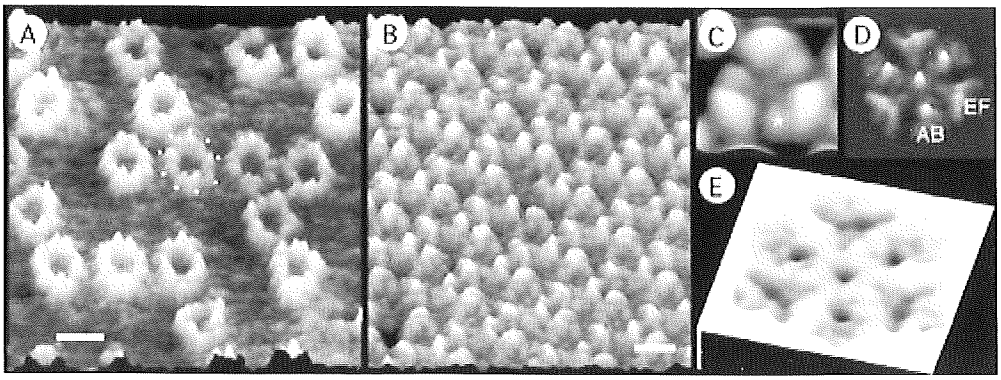


Fig. 5: ***AFM imaging.*** *A) Reconstituted light-harvesting complexes LH2 of Rubrivivax gelatinosum recorded with a commercial AFM in buffer solution. LH2 is a cylindrical structure that spans the membrane and has a diameter of 5 nm. LH2 is assembled from nine α/β heterodimers that protrude by 1.4 nm from the membrane surface [5]. The nine ends of the heterodimers are particularly distinct in the ring marked with dots. This topograph suggests that the AFM is suitable for imaging native membranes at high resolution. B-E) Sampling of the conformational space of a protein surface with the AFM [23]. Panel B) displays the cytosolic surface of a purple membrane, which consists of regularly packed bacteriorhodopsin (bR) trimers and lipids. The major protrusions at the periphery of the trimers (see C) protrude 0.8 nm out of the membrane and represent the loop connecting helices E and F [49]. Smaller protrusion reflect the shorter loop connecting helices A and B. At a closer look into panel B) significant variability of the trimer topography is seen. While the average in C) reflects the prominent features of the cytosolic bR surface, the map in panel D) displays the probabilities of finding maxima related to protruding features at a given position. The EF-loops exhibiting significant flexibility are more delocalized than the AB-loops that occupy a defined position (marked with an asterisk in D). The signal about the three-fold axis results from a lipid molecule protruding out by 0.1 nm, which is only occasionally visible in the raw data, but if present precisely localized in the center of the trimer. Therefore, a strong signal emerges in the probability map (D), whereas no signal is present in the average (C). Precise localization of a surface feature suggests the corresponding structure to be stabilized in a deeper potential well than a floppy feature, which may exhibit a pronounced thermal motion. The corresponding energy landscape is shown in E) (see text). The scale bars in A) and B) represent 10 nm and the full gray scale is 2 nm in A) and 1 nm in B). C)-E) have a side length of 5 nm.*

3.3 Examples

Operating commercial AFMs under such optimal recording conditions, the surfaces of biomolecules are contoured routinely at a lateral resolution better than 1 nm and a vertical resolution around 0.1 nm. Fig.5 shows some examples that document the amazing quality of topographs acquired with the AFM. Fig.5A displays bacterial light-harvesting complexes reconstituted in a lipid bilayer[5]. These ring-shaped complexes exist in two types, the LH2 complexes that consist of 9 α/β heterodimers, and of the LH1 complexes consisting of 16 α/β heterodimers. The latter rings are larger and accommodate the reaction center. Both rings are buried in the bilayer and contain 2 chromophores per α/β heterodimer that either collect the

light or act as couplers, funneling the photon energy to the reaction center. The example shown documents the possibility to image macromolecular complexes within a native membrane and to study their interactions. Fig.5B documents how the conformational space of a membrane protein surface can be sampled with the AFM[23].

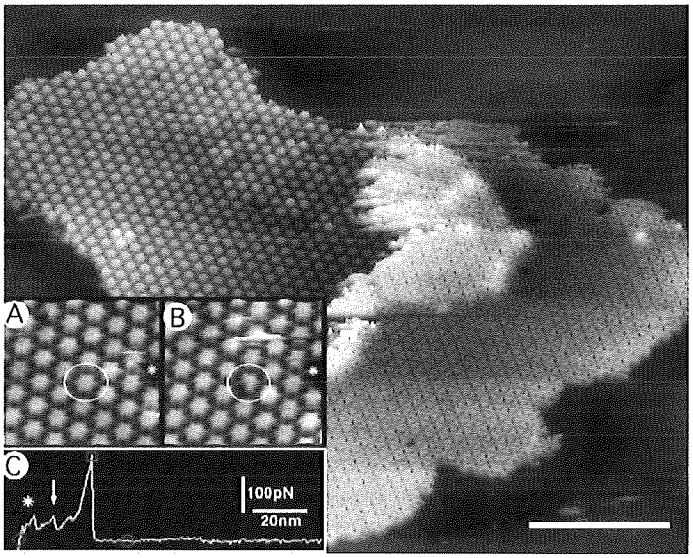


Fig. 6: ***AFM imaging.*** *Manipulating supramolecular assemblies and single molecules with the AFM. Bacterial surface layers (S-layers) are regular protein networks that protect archaea and bacteria from hostile environments. The left panel displays stacked S-layers isolated from the Corynebacterium glutanicum [25]. As result of precise hydrophobic interactions these S-layers stack in register, burying the corrugated surface (left layer). To unveil this surface, the top surface of the sandwich (here on the right) can be pushed away by the stylus. Inset: Unzipping of single and multiple proteins building this S-layer is illustrated by three cases. The surface before and the damaged surface after the molecular manipulation reveal the vacancy produced by the molecular unzipping process (A & B). Asterisks indicate features to identify the scanned area. The corresponding force-extension curves display the energetics of the unfolding process. Each event consists of a weak (arrow) and a strong force peak and corresponds to the extraction of a protein dimer from the hexameric unit building the hexagonal S-layer. The scale bar represents 200 nm and the subframes have a side length of 100 nm. The grey scale corresponds to 5 nm.*

In this case, a native 2D crystal composed of bacteriorhodopsin (bR) trimers and lipids has been studied. The topograph shows the conformation of the cytoplasmic surface of bR in its most native state. To record this image, the force applied to the stylus was approximately 50 pN, preventing a force induced conformational change of the loop connecting helices E and F. This loop is deformed when the force is increased to >100 pN, thereby changing the conformation of the bR surface. Interestingly, the variation of this region among different atomic models from x-ray crystallography is pronounced: the EF loop is involved in the contacts leading to the 3D crystals and its conformation is dictated by the 3D packing arrangement of the bR molecules. Although all the bR trimers are identical in the 2D crystal

displayed in Fig. 5B, their surface structure varies significantly. These changes are related to the flexibility of the bR surface rather than any noise introduced by the AFM. The probability p(x,y) to find a certain loop at a certain position (x,y) can be determined by mapping the corresponding peak positions of all individual bR trimers. This map is readily converted to a free energy landscape F_l using Boltzmann's law:

$$F_d(x,y) = -kT \ln\{p(x,y)\} \quad (1)$$

Forces need to be minimized to obtain high-resolution images, whether the AFM is operated in the contact mode or the tapping mode. However, the stylus may be used as nanoscalpel to disrupt supramolecular assemblies[24]. In this case, the force applied to the tip is increased to 1-10 nN, depending on the damage to be achieved. Quite small forces (typically 1 nN) are sufficient to separate stacked layers of membranes or 2D crystals [25](Fig.5A). Even smaller forces and repeated scanning at high magnification suffice to push away extrinsic proteins that are specifically complexed to an integral membrane protein[26].

4 Single Molecule Force Spectroscopy

A modern commercial AFM offers both a high spatial resolution and a force sensitivity down to the Piconewton range. Such instruments allow the forces of single molecular interactions to be measured [27-30]. For single molecule force spectroscopy using the AFM, molecules must be tethered to both the support and the tip in some ways. Whilst a variety of chemical surface activation techniques have been used, large two-dimensional (2D) assemblies such as bacterial S-layers or reconstituted membranes are simply physisorbed to mica (Fig.6 left panel). The AFM stylus is then attached to an individual protein of such an assembly by pressing the stylus down with a force of ~1 nN for approximately 1 s, which induces the denaturation of some protruding domain that is thus stuck to the silicon nitride tip. Specifically designed mutant proteins that bear a cystein on a surface exposed domain allow specific attachment of the protein to a gold-coated stylus. Alternatively, a ligand may be covalently immobilized on the stylus through a suitable spacer, while the receptor is tethered to the support. When approaching the tip to the support, a bond may form between ligand and receptor. In all cases, the change in cantilever deflection upon tip retraction is then recorded, yielding a force-extension curve (Fig.6 insert)[25]. By scanning the sample before (Fig.6A) and after (Fig.6B) such molecular 'unzipping' events, the structural damage can be directly assessed and correlated with the respective force-extension curve (Fig.6C).

In addition, the AFM offers the possibility of unbinding biomolecular bonds (e.g. ligand/receptor). Because the unbinding of biomolecular bonds is a dynamic process, the acquisition of force-extension curves at various extension rates in these experiments is also referred to as dynamic force spectroscopy, DFS. It has long been known that only molecules with an excess of energy over the average energy of the population can participate in chemical reactions. Accordingly, reactions between ligands and receptors follow pathways (in a virtual energy landscape) that involve the formation of some type of high-energy transition states whose accessibility along a reaction coordinate ultimately controls the reaction rate. But because quantitative measurements have only been possible on large molecular ensembles, only macroscopic thermodynamical quantities could be determined, e.g., the free energy of complex formation and/or dissociation. From DFS measurements, the energy landscape of a single molecular interaction can be mapped, giving a detailed insight into the reaction pathway of single molecular processes. Elaborate thermodynamical models describing the rupture of a single bond are described in [31-33].

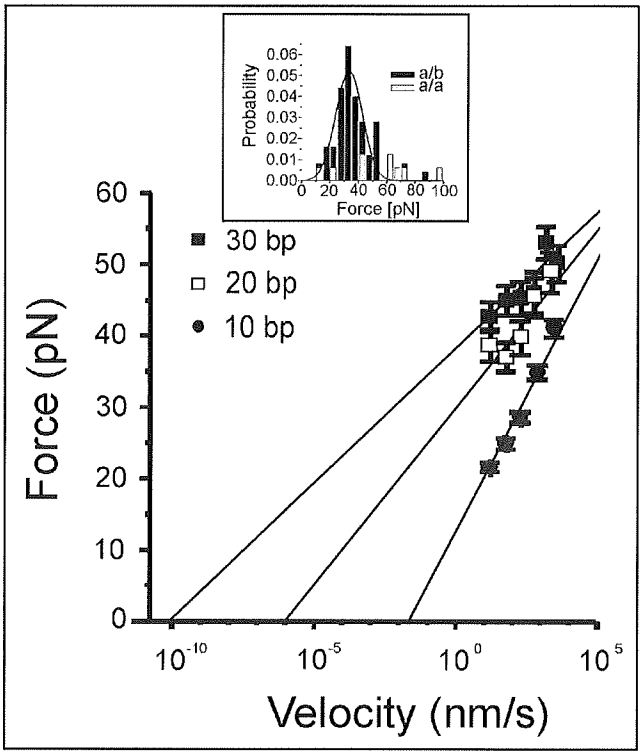


Fig. 7: ***Dynamic force spectroscopy.*** *Rupturing dsDNA molecules: Velocity dependence of the most probable unbinding force. Back squares (a-tip/b-surface, 30 bp), empty squares (a-tip/c-surface, 20 bp), circles (a-tip/d-surface, 10 bp). From a linear fit, both the force-scales* $F^0 = k_BT/\Delta x$ *and thermal off-rates can be determined. Inset specificity of the rupturing force: A typical probability distribution for the rupture force (about 500 approach/retract cycles, retract velocity 100 nm/s). For this experiment, an oligomer a (see text) was attached to the tip of the AFM-cantilever and its complement b was immobilized on the surface (complements were pulling apart at their opposite 5'-ends). Gray rectangles (a against a), black rectangles (a against b). To minimize unspecific interactions (e.g. a against a) and multiple unbinding events, 30-nm-long PEG linkers were attached to the 5'-ends. Note that the scale-force* F^0 *can be in principle determined from the width of the distribution.*

DFS measurements can be performed with an unmodified commercial AFM (see Fig.7), or using external data acquisition capabilities that enhance sensitivity and flexibility of the instrument. To achieve reproducible measurements, the spring constants of all cantilevers used must be calibrated. The thermal fluctuation method [34] is currently the most convenient method and gives an absolute uncertainty of < 20%. For the temperature measurements presented below, the temperature was controlled using a home built cell where the buffer solution that immersed both the probe surface and the AFM cantilever was in contact with a Peltier element, driven with a constant current source. Measurements at different points of the cell showed deviations of less than 2°C. Preparation and immobilization of all oligonucleotides was carried out using the protocol described in [30,35].

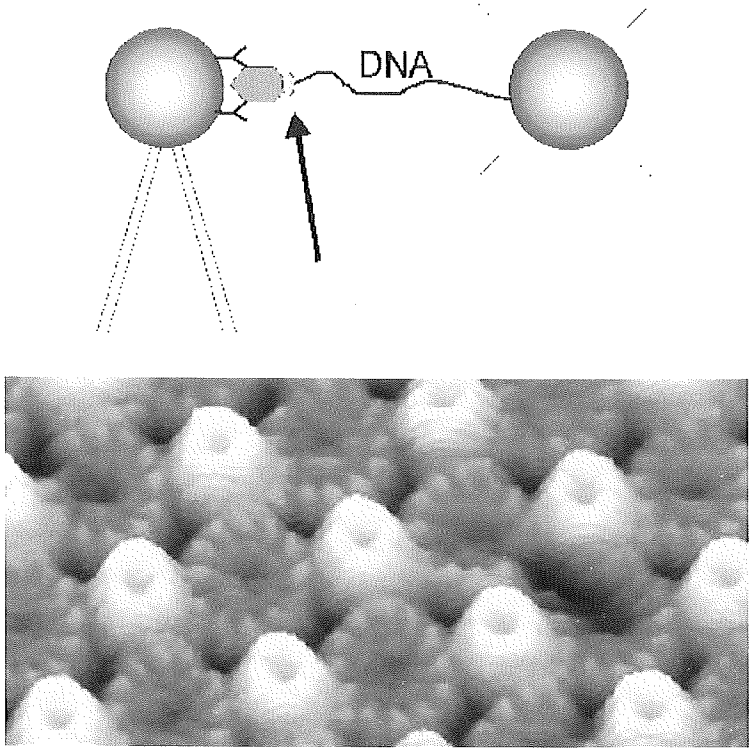


Fig. 8: ***OTS experiments.*** *A bead coated with a bacteriophage head (grey) is held by suction on top of a micropipette (top). This head is pulling on a DNA molecule attached on a second bead trapped by an optical tweezers. The DNA is recognized by molecular motor on the phage head responsible for DNA packaging in to the head. The motor spools the single DNA molecule into the head and is exerting a rather high force on the bead sitting in the trap). The bottom panel shows an AFM image of a two-dimensional crystal of ϕ29 connectors recorded in buffer solution [50]. Because connectors are packed in opposite orientations, both sides are visible: one exhibiting a cylindrical protrusion and the other showing a wheel-like structure with 12 radial petals. The cylinder connects to the phage tail, whereas the wheel connects to the icosahedral phage head via a set of molecular motors that hydrolize ATP. The cylindrical protrusions are 16.5 nm apart. Full gray level range is 4 nm, respectively.*

Unbinding events are caused by thermal fluctuations rather than by mechanical instability. Therefore unbinding forces show a distribution whose width σ is mainly determined by the force scale F^0, *i.e.* $\sigma=F^0(\Delta x)$.
When approaching the tip to the surface, many non-specific attachments may occur, even in the presence of passivated surfaces or pure polymer samples. Therefore, it is imperative to test the specificity of the interaction (see inset Fig.7). Unspecific interactions can be minimized using linkers (e.g. poly(ethylene)glycol (PEG) linkers) that shift the region where unbinding takes place away from the surface. Finally, to quantify the most probable value for the

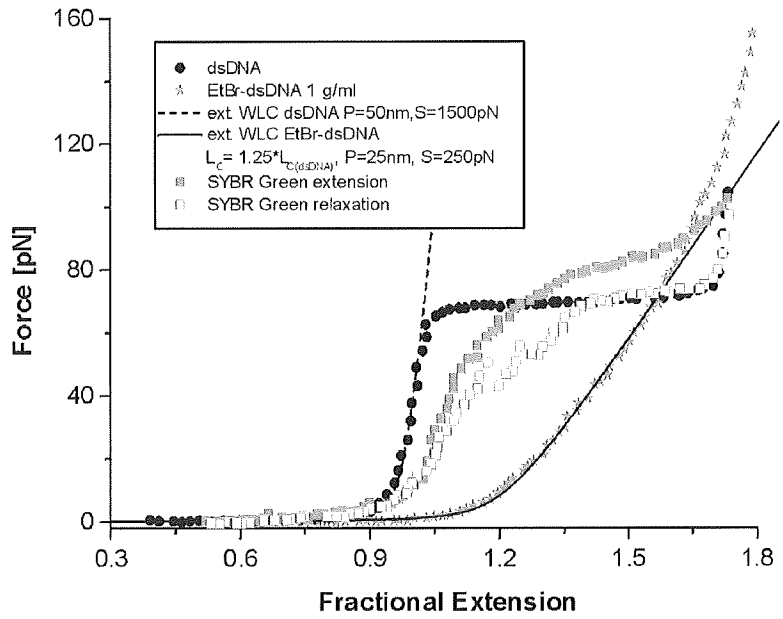


Fig. 9: ***DNA mechanics****: Typical force versus fractional extension x/L curves obtained for bare dsDNA (circles); and EtBr (stars). SYBR® Green I dsDNA (squares, full squares extension, hollow squares relaxation), Also shown is the fit to the data using an extensible worm like chain model (dashed lines). From this fit relevant enthlapic parameters (stretch modulus) are obtained.*

unbinding force of a single complex, conditions must be chosen in which the probability that two or more duplexes are attached to the tip is small.These conditions are fulfilled for a low concentration and when the linkers have a length that is comparable to the diameter of the AFM-tip (about 50 nm). In this case, it is very unlikely that two or more linkers are extended to the same length when stretched. However, subsequent rupture events may be found. But still, the last rupture event will occur for an applied force equal to F^*.

DFS measurements have been performed on complementary DNA strands of different length {10, 20, and 30 base pairs (bp)}, which were pulled apart at their opposite 5'-ends. The base sequences of the oligonucleotides were designed to favor the binding to its complementary oligonucleotides in the ground state with respect to intermediate duplexes in which the strand is shifted relative to its complement. Oligomer **a** (5'-G-G-C-T-C-C-C-T-T-C-T-A-C-C-A-C-T-G-A-C-A-T-C-G-C-A-A-C-G-G-3'), which contains 30 bases and in which every three base motive occurs only once in the sequence was chosen as a bait. For this sequence, self-complementarities are avoided because the complement of each three-base motive is not contained in the sequence. **a** was tested against its complement **b** (30 bp) and against truncated components **c** (20 bp) and **d** (10 bp), respectively. As expected, a F^* *versus* ln(v) plot shows a linear behavior for each duplex (Eq. 2, see Fig.7). For each duplex, the distance Δx from the ground state to the energy barrier and the thermal off rate k_{off} were determined according to Eq. 2.

$$F^* = F^0 \ln\left(r / F^0 k_{off}\right) \qquad (2)$$

The Δx distance was found to follow the linear relation: $\Delta x = [(0.7 \pm 0.3) + (0.07 \pm 0.03) \times n]$ nm, where n is the number of base pairs. This increase of Δx with n clearly indicates cooperativity in the unbinding process. Measurements of k_{off} can be described by: $k_{off} \approx 10^{\alpha - \beta n} s^{-1}$, where $\alpha = 3 \pm 1$ and $\beta = 0.5 \pm 0.1$. The obtained k_{off} values are in good agreement with thermodynamical data [36]. It should be noted that an exponential decrease of the thermal off-rate with the number of base pairs is expected due to the increase of the activation energy for dissociation.

Temperature dependent DFS measurements can also be performed. Sequence **e** (5'-T-A-T-T-A-A-T-A-T-C-A-A-G-T-T-G-3')[37] was attached to the tip and its complement **f** immobilized on the surface. As previously, PEG linkers were used and DNA strands were pulled apart at their opposite 5'-ends. The specificity of the interaction was comparable to the one obtained in base-pair dependent measurements (Fig.7). One striking difference of these energy landscape measurements is visible when the temperature is varied. The slope of the F^* versus ln(r) plots changes as a function of temperature, which documents a strong temperature dependence of Δx. This result emphasizes the fact that for the DNA-duplex, the energy landscape is much more complicated than that of ligand-receptor bonds.

Our measurements confirm that the most probable force for unbinding scales as the logarithm of the loading rate. From this dependence, both the natural thermal off-rate for dissociation K_{off} and the bond length x along the reaction coordinate were determined. The K_{off} values measured by DFS are in agreement with bulk temperature measurements, underlining the validity of the results. The base pair dependent measurements indicate that unbinding of DNA strands is a cooperative process. The decrease of Δx upon increasing the temperature, which is not expected in the case of one dimensional energy landscape with a sharp energy barrier, indicates entropic contributions to the unbinding of DNA and the unfolding of RNA or proteins[30].

5 Experiments using optical tweezers

Force extension-curves may also be acquired with an optical tweezers setup [38,39]. In addition, single molecule force measurements can be combined with distance measurements using fluorescence resonance energy transfer as shown in Fig.2 (FRET) [11]. One of the most impressive applications of OTs is the study of molecular motors on a single molecule level. These molecular motors can be linear motors (Kinesin, Myosin) [38,39], DNA/RNA polymerase enzymes [40,41] or DNA packaging viruses (bacteriophage ϕ29) [42].

Kinesin and myosin are two ATPase motor proteins. Kinesin, which is used for organelle transport or chromosome segregation, moves along microtubules. In contrast, myosin interacts with actin filaments and is used not only for muscle contraction but also is involved in many forms of cell movement. For these studies, OTs are used to interact single kinesin or myosin *in vitro* with either a microtubule or an actin filament. These experiments allowed the ATP hydrolyzed per cycle and the force generated to be measured at each step, giving insight into the possible mechanisms involved in the movement.

Other experiments assessed the function of motor enzymes used in DNA transcription or DNA polymerization. In this case a single DNA molecule is tethered between two beads, and the rate of transcription or polymerization can be followed in real time by applying a constant tension (force feedback) and allowing the distance between the beads to change accordingly [40,41]. Such studies have direct implications for the mechanism of gene regulation or force-induced exonuclase activity. In a recent study Smith and colleagues measured the forces involved in the packaging of DNA into bacteriophage ϕ29 heads [42]. A schematic picture of these experiments is given in Fig.8 (top). The key elements of such an experiment are the

native immobilization of the bacteriophage head onto the surface of one bead and the attachment of the DNA molecule [43] to the second microsphere. A high resolution AFM image of the bacteriophage ϕ29 packging motor protein complex is also shown in Fig.8.
Due to the high force sensitivity, OTs have been used to study (i) mechanical properties of DNA [44] (ii) protein or RNA unfolding [45] (iii) the polymerization of individual RecA-DNA filaments [46]. Again, these experiments provided new insights in biochemical processes on a single molecule and are of great relevance to biology. A recent publication demonstrates that the kinetics of binding of small ligands to dsDNA can directly be measured using optical tweezers experiments [47]. By recording force extension curves the mechanical parameters of the modified dsDNA molecule could be elucidated (Fig.8). The parameters obtained indicated the mode of binding: intercalation affected the contour length. Additionally the occupancy of the ligand on the DNA could be determined from such measurements. If compounds bind to dsDNA, which aren't intercalating then the binding is directly revealed by the way the modified dsDNA is going through it's overstretch transition. Such experiments can give direct insight into the binding of small ligands to DNA and can be of general importance for screening of DNA binding compounds.

6 Perspectives

Progress in instrumentation and sample preparation methods have opened avenues to image, manipulate and assess the nanomechanics of single molecules. Such novel experimental possibilities find a wide range of applications in the study of structure and function of biomolecules, the building blocks of life. While the amazing S/N-ratio and resolution of the AFM allows the conformational states of single protein surfaces to be probed at a lateral resolution of 0.4-1 nm lateral and 0.1-0.2 nm vertical resolution, the force sensitivity (>20 pN) of this instrument also opens the door to measure the forces required to unfold a protein. Combining single molecule imaging and force spectroscopy gives an unprecedented insight into the nature of intra- and intermolecular interactions in and between biomolecules.
Optical tweezers enhance the force sensitivity compared to the AFM by an order of magnitude. Therefore, OTs are suited to monitor molecular motors that generate forces in a range of 0.3 to 50 pN while at work. Using light, particles of some nm up to some μm in diameter can be trapped and manipulated. Simultaneously, fluorescence techniques allow small distance changes to be measured, allowing a full analysis of the nanomechanical features of a single molecule.
Using DFS measurements, the energy landscape of molecular bonds can be mapped. Moreover, relevant parameters such as the location and height of the barriers and the thermal off-rates can be determined. Since the limited range of loading rates available in an AFM experiment does not allow one to map the whole energy landscape, such experiments should be combined in the future with other DFS setup such as bio-membrane force probe or optical tweezers setups. An additional solution is to apply small cantilevers, which allow faster pulling and exhibit less thermal noise; so smaller unbinding forces can be detected. These small cantilevers are still experimental [48] and great efforts are being made to commercialize them in the future. These developments will also ask for instrument development so it is expected to last a few years before they are widely used. One could envision that the dynamic force spectroscopy will be applied in the future to assess the binding affinity of biomolecules on bio-arrays but the experimentalists have to catch up to step in this direction.

References

[1] Binnig G, Quate CF, Gerber C. *Phys. Rev. Lett.* **1986**, *56*:930-933.
[2] Mou JX, Yang J, Shao ZF. *J. Mol. Biol.* **1995**, *248*:507-512.
[3] Müller DJ, Amrein M, Engel A. *J. Struct. Biol.* **1997**, *119*:172-188.
[4] Czajkowsky DM, Sheng S, Shao Z. *J. Mol. Biol.* **1998**, *276*:325-330.
[5] Scheuring S, Reiss-Husson F, Engel A, Rigaud JL, Ranck JL. *Embo J* **2001**, *20*:3029-3035.
[6] Müller DJ, Fotiadis D, Scheuring S, Müller SA, Engel A. *Biophys. J.* **1999**, *76*:1101-1111.
[7] Möller C, Allen M, Elings V, Engel A, Müller DJ. *Biophys. J.* **1999**, *77*:1050-1058.
[8] Hansma PK, Cleveland JP, Radmacher M, Walters DA, Hillner PE, Bezanilla M, Fritz M, Vie D, Hansma HG, Prater CB, et al. *Appl. Phys. Lett.* **1994**, *64*:1738-1740.
[9] Han W, Lindsay SM, Dlakic M, Harrington RE. *Nature* **1997**, *386*:563.
[10] Viani MB, Schaffer TE, Paloczi GT, Pietrasanta LI, Smith BL, Thompson JB, Richter M, Rief M, Gaub HE, Plaxco KW, et al. *Rev. Sci. Instr.* **1999**, *70*:4300-4303.
[11] Grange W, Husale S, Güntherodt H-J, Hegner M. *Rev. Sci. Instr.* **2002**, *73*:2308-2316.
[12] Ashkin A. *Phys. Rev. Lett.* **1970**, *24*:156.
[13] Ashkin A, Dziedzig JM, Bjorkholm JE, Chu S. *Opt. Lett.* **1986**, *11*:288-290.
[14] Ashkin A, Dziedzig JM, Yamane T. *Nature* **1987**, *330*:769-771.
[15] Ashkin A, Dziedzig JM. *Science* **1987**, *235*:1517-1520.
[16] Svoboda K, Block SM. *Ann. Rev. Biophys. Biomol. Struct.* **1994**, *23*:247-285.
[17] Leitz G, Fällmann E, Tuck S, Axner O. *Biophys. J.* **2002**, *82*:2224-2231.
[18] Neuman KC, Chadd EH, Liou GF, Bergman K, Block SM. *Biophys. J.* **1999**, *77*:2856-2863.
[19] Afzal RS, Treacy EB. *Rev. Sci. Instr.* **1992**, *63*:2157-2163.
[20] Engel A, Muller DJ. *Nat Struct Biol* **2000**, *7*:715-718.
[21] Müller DJ, Engel A. *J. Mol. Biol.* **1999**, *285*:1347-1351.
[22] Schabert FA, Henn C, Engel A. *Science* **1995**, *268*:92-94.
[23] Scheuring S, Müller D, Stahlberg H, Engel H, Engel A. *European Biophys J* **2002**,:in press.
[24] Fotiadis D, Scheuring S, Müller SA, Engel A, Müller DJ. *Micron* **2002**, *33*:385-397.
[25] Scheuring S, Stahlberg H, Chami M, Houssin C, Rigaud JL, Engel A. *Mol Microbiol* **2002**,:in press.
[26] Fotiadis D, Müller DJ, Tsiotis G, Hasler L, Tittmann P, Mini T, Jenö P, Gross H, Engel A. *J. Mol. Biol.* **1998**, *283*:83-94.
[27] Allen S, Chen XY, Davies J, Davies MC, Dawkes AC, Edwards JC, Roberts CJ, Sefton J, Tendler SJB, Williams PM. *Biochemistry* **1997**, *36*:7457-7463.
[28] Evans E, Leung A, Hammer D, Simon S. *Proc. Natl. Acad. Sci. U S A* **2001**, *98*:3784-3789.
[29] Moy VT, Florin EL, Gaub HE. *Science* **1994**, *266*:257-259.
[30] Schumakovitch I, Grange W, Strunz T, Bertoncini P, Güntherodt H-J, Hegner M. *Biophys. J.* **2002**, *82*:517-521.
[31] Evans E. *Ann. Rev. Biophys. Biomol. Struct.* **2001**, *30*:105-128.
[32] Galligan E, Roberts CJ, Davies MC, Tendler SJB, Williams PM. *J. Chem. Phys.* **2001**, *114*:3208-3214.
[33] Heymann B, Grubmüller H. *Phys. Rev. Lett.* **2000**, *84*:6126-6129.
[34] Hutter JL, Bechhöfer J. *Rev. Sci. Instr.* **1993**, *64*:3342-3342.
[35] Strunz T, Oroszlan K, Schäfer R, Güntherodt H-J. *Proc. Natl. Acad. Sci. U S A* **1999**, *96*:11277-11282.

[36] Pörscke D, Eigen M. *J. Mol. Biol.* **1971**, *62*:361-.
[37] Tibanyenda N, Debruin SH, Haasnoot CAG, Vandermarel GA, Vanboo JH, Hilbers CW. *Eur. J. Biochem.* **1984**, *139*:19-27.
[38] Bustamante C, Macosko JC, Wuite GJL. *Nature Rev. Mol. Cell Biol.* **2000**, *1*:130-136.
[39] Mehta AD, Rief M, Spudich JA, Smith DA, Simmons RM. *Science* **1999**, *283*:1689-1695.
[40] Davenport RJ, Wuite GJL, Landick R, Bustamante C. *Science* **2000**, *287*:2497-2500.
[41] Wuite GJL, Smith SB, Young M, Keller D, Bustamante C. *Nature* **2000**, *404*:103-106.
[42] Smith DE, Tans SJ, Smith SB, Grimes S, Anderson DL, Bustamante C. *Nature* **2001**, *413*:748-752.
[43] Hegner M. *Single Mol.* **2000**, *1*:139-144.
[44] Williams MC, Rouzina I. *Curr. Op. Struct. Biol.* **2002**, *12*:330-336.
[45] Liphardt J, Dumont S, Smith SB, Tinoco I, Bustamante C. *Science* **2002**, *296*:1832-1835.
[46] Hegner M, Smith SB, Bustamante C. *Proc. Natl. Acad. Sci. U S A* **1999**, *96*:10109-10114.
[47] Husale S, Grange W, Hegner M. *Single Mol.* **2002**, *3*:91-96.
[48] Viani MB, Schäfer TE, Chand A, Rief M, Gaub H, Hansma PK. *J. Appl. Phys.* **1999**, *86*:2258-2262.
[49] Müller DJ, Heymann JB, Oesterhelt F, Moller C, Gaub H, Buldt G, Engel A. *Biochim. Biophys. Acta* **2000**, *1460*:27-38.
[50] Müller DJ, Engel A, Carrascosa JL, Velez M. *Embo J* **1997**, *16*:2547-2553.

B 5 Introduction to computer simulations

G. A. Vliegenthart
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Methodology **4**
2.1 General considerations 4
2.2 Error estimation & observables 7
2.3 Molecular Dynamics 9
2.4 Monte Carlo 11

3 Example: Molecular Dynamics of a mixture of spheres **14**

4 Conclusions **16**

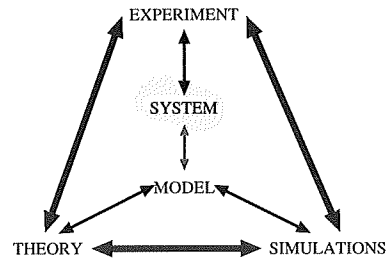


Fig. 1: *Relation between experiments, theory and computer simulations for studying a system.*

1 Introduction

The physics of complex systems deals with structures and processes in multi component mixtures. Complex systems are found in industry (paint, lubricants, food-products, detergents, polymers), academia (colloidal model systems) and nature (cells, biomacromolecules). Blood for instance, is a suspension of blood cells, biomacromolecules, small molecules and ionic species in water. Complex fluids have many interesting aspects and a challenging task is to reveal and understand the microscopic mechanisms leading to the macroscopic structure, dynamics and functionality observed.

There are different approaches to this problem. The most direct, is to study the system of interest by doing experiments. However, the design and interpretation of an experiment always involves some theoretical input. One can also try to construct a theory from first principles (using statistical mechanics) which describes the system under consideration. This requires modeling of the experimental system which in general involves a certain number of approximations and abstractions of the experimental system. Except for some special cases, it is hard or impossible to find exact solutions for the theoretical models. As computers have become increasingly more available since the 1950's, new computer based techniques have been developed which are complementary to theory and experiments. These methods are powerful in the sense that many of the approximations necessary in 'conventional' theoretical approaches do not have to be made in computer simulations. In fact, the level of complexity and detail on which a problem is studied is a matter of choice. The combination of the three methods, which is schematically drawn in Fig. 1, forms an effective approach for studying complex fluids as well as biological systems.

The two computer simulation techniques discussed in this chapter are based on *classical statistical thermodynamics* and are from a historical perspective the most important ones. The first method, molecular dynamics (MD), is based on the dynamical evolution of a many particle system. Given initial particle positions, velocities, interactions and boundary conditions, Newton's equations of motion are integrated in discrete time steps. This gives for each of the particles, the position and velocity as a function of time. The second technique discussed is the Monte Carlo (MC) technique. In this method, a sequence of particle configurations is generated which satisfies the equilibrium distribution of particle configurations. The transition from one to another configuration is established by performing a random trial-displacement of a randomly chosen particle and the acceptance or rejection of this trial move.



Fig. 2: *DNA-histon complexes in: A experiment, B atomistic simulation, C theory.*

Motivation Computer simulations provide a useful tool for many reasons. The fact that one can tune the complexity of the problem gives a direct method to check approximations made in a particular theory. In this sense, simulations can be seen as an experimental method for testing theories. Moreover, simulations can 'easily' be performed for conditions that are experimentally difficult or impossible to realise (high/low temperature, pressure *et cetera*). In combination with experimental data (NMR, X-ray), simulations are being utilised to reveal structure and function of many complex structures such as proteins.

In Fig. 2 an example is shown of a system which has been investigated by experiment, computer simulation and theory. This system is a mixture of negatively charged DNA and positively charged histon proteins. The questions addressed here concern the structural properties of the DNA/histon complex. Experimentally, the aggregates have been imaged by fluorescence microscopy where the histons were stained with fluorescein isothiocyanate. A microscopy image of this is given in Fig. 2A [1]. The experimental image shows that the histons are aggregated in a chain-like structure. However, the picture is blurred and microscopic information is absent. In order to get more insight in the microscopic details of the structure, simulations have been performed at an atomic level as shown in Fig. 2B [2]. All the atoms in the protein and the DNA were individually modeled as well as the surrounding water molecules and ions. The theoretical picture shown in Fig. 2C, was obtained using a free energy minimisation procedure and predicts how the strength of the DNA/histon complex changes with salt concentration [3]. At physiological salt concentrations ($\approx$0.1 M), the DNA is tightly bound to the histon, but it is released at low salt concentrations.

Comparing the three methods, only the atomistic simulation can give information about which of the amino acid groups are involved in the interaction with the histon protein and how the DNA is, on a microscopic level, wrapped around the protein.

In this chapter we will give an introduction to computer simulations. It is a topic with many aspects and subtleties which by no means can be addressed here comprehensively. However, there is a vast literature in books and papers available dealing with the different techniques and applications.

2 Methodology

In this section the two main simulation methods, molecular dynamics and Monte Carlo will be introduced. Before going into more detail, we discuss some general concepts common to both simulation techniques.

2.1 General considerations

Independent of the method, all computer simulations require the specification of the interactions between the particles. Firstly, the choice of these interactions is crucial for identifying the simulated system. For example, the interactions will be different when studying a gas, a charged system, a polymer or a membrane. Secondly, the interactions determine the level of detail at which the simulation is performed. This can be either atomistic, taking into account all the atoms in the system (all atoms in a DNA molecule) or more coarse-grained (the DNA chain is a string of charged beads). In addition, one can think about how to model the solvent in a convenient way or whether to disregard it. For instance, from Fig. 2B it is obvious that the majority of the particles are devoted to represent water molecules. This means that most of the computational time is being spend on the least interesting part of the system.

Interactions In this paragraph we will introduce some of the basic interactions commonly used in simulations. The simplest possible interaction is given by the hard sphere potential, which is zero everywhere except for distances r smaller than the diameter σ of a particle,

$$V(r) = \begin{cases} \infty & \text{for } r < \sigma \\ 0 & \text{for } r \quad \sigma. \end{cases} \tag{1}$$

The hard sphere interaction, which is sketched in Fig. 3A, introduces in a simple manner excluded volume interactions between particles. One could think of particles that behave as a group of billiard balls in two or three dimensions. This is a common picture when simulating a liquid at densities and temperatures where further interactions can be neglected. Apart from a model system for the liquid state [4], hard spheres have found an important application in colloid chemistry. It has been demonstrated that the hard sphere interaction gives rise to a fluid-crystal phase transition around a volume fraction of 50% of hard spheres. This phase transition was the subject of a lot of discussion in the early 1950's and was for the first time discovered in computer simulations [5]. In more recent times, this transition was observed experimentally in dense colloidal suspensions [6].
In order to reproduce the properties of atomic fluids a slightly more realistic potential is required. Such a potential should be repulsive for short distances, and attractive for larger ones. An example is the Lennard-Jones potential [7] which is given by,

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2}$$

where ϵ is the depth of the attractive well, σ the 'diameter' of the particles and r the center to center distance between the particles. In Fig. 3B the Lennard-Jones potential is shown. The attractive tail ($\sim r^{-6}$) originates from dispersion interactions (van der Waals interactions) between the atoms. The algebraic dependence of the repulsive core ($\sim r^{-12}$) has a historical rather than

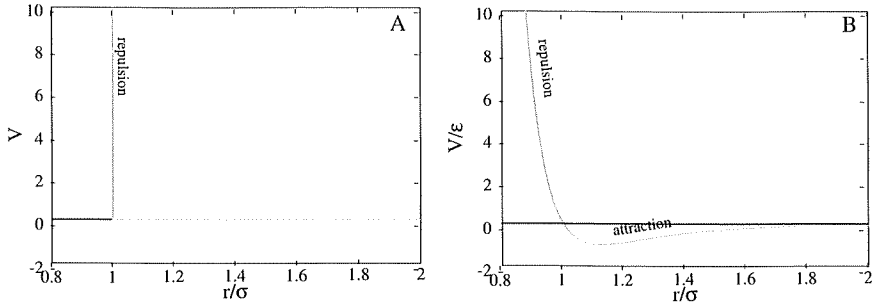


Fig. 3: *The hard sphere interaction (A) contains a hard (infinitely repulsive) core while attractions between the particles are absent. The Lennard-Jones potential (B) models long-ranged van der Waals attractions between atoms, on top of a 'soft' repulsive core.*

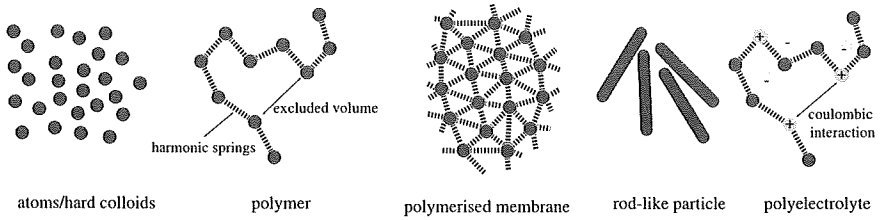


Fig. 4: *Sketch of some possible structures combining different interactions and particle shapes.*

a physical origin. It is a computationally convenient simplification for the exponential repulsion two atoms experience upon approaching.

A complete overview of all possible interactions is beyond the scope of this chapter. Just a glimpse of what can be done is summarised in Fig. 4. Ideal polymers can be modeled by attaching monomers in a string using for instance harmonic springs. Excluded volume interactions and/or charges on the chain are then easily incorporated to model more realistic polymers (see chapter C.2) or polyelectrolytes (C.3). By making multiple connections between monomers, membranes can be modeled (section D) in a polymerised as well as in a fluid state. On a more coarse-grained level one can modify the particle shape to obtain asymmetric particles like rods or discs which can represent liquid crystal molecules, rod-like viruses (Tobacco Mosaic Virus, fd virus (C.11)), rod-like inorganic colloids (boehmite) or plate-like colloids.

Interaction cut-off Long-ranged interactions like the Lennard-Jones potential Eq. (2) are computationally expensive since the state of one particle involves interactions with all the other particles in the system. Therefore, the interactions are in practice cut off at some distance r_c where the potential has decayed to a small value. In other words, the potential is set to zero for $r > r_c$. In addition, it is convenient to shift the potential such that $V(r_c) = 0$. It is important to realise that thermodynamic properties of a system with a cut-off and shifted potential are in general different. For instance, the critical temperature of a Lennard-Jones fluid goes down

from $k_BT/\epsilon \approx 1.32$ for the full potential to $k_BT/\epsilon \approx 1.09$ for a system with $r_c = 2.5$ and which is shifted [8].
Special care has to be taken with intrinsically long-ranged interactions like electrostatic or dipolar interactions. These interactions, which decay slower than r^{-d} with d the dimensionality of the system, have to be handled by special methods such as the Ewald summation technique [9, 10, 11, 12]

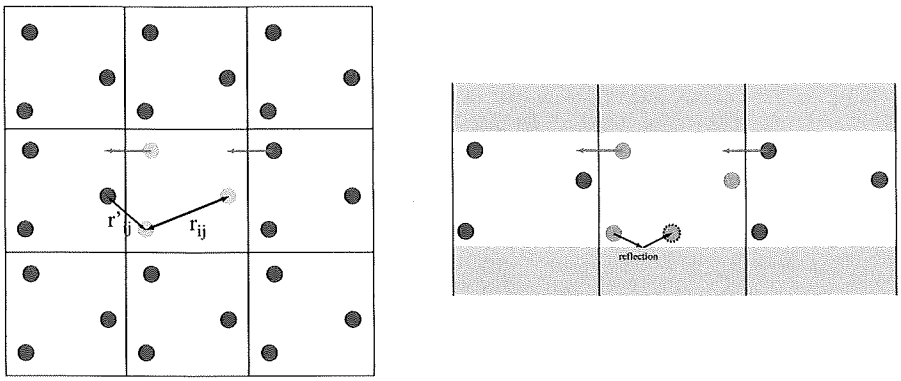
Units Most of the parameters used in simulations have in principle arbitrary values. However, these numbers should be related with values in real systems such that comparisons can be made. The first step is to choose three independent parameters as a consistent set of units. In this chapter, we have chosen one commonly used option which is to express length in units of σ, mass in units of m and energy in units of ϵ. In this way, time is expressed as $t/\sigma\sqrt{m/\epsilon}$, temperature as k_BT/ϵ, pressure as $P\sigma^3/\epsilon$, force as $F\sigma/\epsilon$ and so on.

Finite size effects The systems used in computer simulations always have a finite size and typically the number of particles is between 10^3 and 10^5. This means that the thermodynamic limit is never reached and thus we may expect the properties of our system to depend on system size. Finite size effects can be estimated using finite size scaling, i.e. plotting the observables as a function of $1/L$, with L the linear dimension of our system, and extrapolating to $1/L \rightarrow 0$. In particular systems with a large correlation length, like close to a critical point, strongly suffer from finite size effects.

Boundary conditions An important aspect of simulating a system is the implementation of the boundaries. The way this is done depends on the problem under consideration. In case one is interested in the bulk properties of a system, it is convenient to introduce periodic boundary conditions. The idea is that particles leaving the box at one side, enter the box at the opposite side. This is shown in Fig. 5A. The number of particles in a simulation box is thus conserved. The central box is repeated infinitely many times in all directions. Without potential cut-off, the particles in the central box interact with all the particles in their own box as well as with all the particles in the periodic images.
In case one is interested in confinement effects, walls with prescribed geometry can be introduced as is illustrated in Fig. 5B for flat walls.

Equilibration In general, the initial values of the different variables (positions and velocities) are not in equilibrium. Therefore, it is essential that the initial configuration has relaxed to thermodynamic equilibrium before one starts to sample observables to study the system. For instance, one necessary condition for equilibrium is that the velocities are Maxwell-Boltzmann distributed. To establish a preferred temperature in a molecular dynamics simulation, periodic velocity rescaling can be applied during equilibration.

Solvent Many systems of interest contain for a substantial part solvent molecules. These solvent molecules can act as a dielectric background or a Brownian heat-bath among others. The solvent is also the source of hydrodynamic interactions between the particles in solution. These hydrodynamic interactions are important for many dynamical processes. However, the incorporation of hydrodynamic effects via explicit solvent simulations is in most cases not feasible (computationally too demanding). New mesoscopic methods that overcome this problem are



A: periodic B: reflective & periodic

Fig. 5: *Different implementations of boundary conditions in a computer simulation.*

currently being developed. Different solutions have been proposed where the common idea is that the simulated particles are not atoms or molecules but coarse-grained portions of solvent that interact in a particular way. The most representative mesoscopic techniques are Lattice Boltzmann [13], Dissipative Particle Dynamics [14] and Multi-Particle-Collision Dynamics [15]. We wil not further discuss any of these methods here.

2.2 Error estimation & observables

Before we introduce the simulation methods, we discuss some of the properties one can study using computer simulations. We demonstrate how these properties can be calculated from the particle positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$.

Averages and errors Suppose we have measured our observable A a large number of times, say M, in the simulation. The average value of A is then straightforwardly calculated using,

$$\langle A \rangle = \frac{1}{M} \sum_{k=1}^{M} A_k \tag{3}$$

where the A_k are *independent* measurements of the quantity of interest. It is important for the accuracy of the average to ensure that these measurements are indeed independent. One way to check this is by measuring the time correlation function of the observable. This correlation function, in general, decays exponentially with a characteristic correlation time. When sampling data, the sampling interval has to be larger than the correlation time. More information can be found in [9, 10, 11, 12].
In order to estimate the accuracy of the average one calculates the standard deviation of the average value,

$$\delta^2(\langle A \rangle) = \frac{1}{M^2} \sum_{k=1}^{M} (A_k - \langle A \rangle)^2 . \tag{4}$$

Thermodynamic properties The most elementary properties of a system are its energy and momentum. The average kinetic energy which defines the temperature in the system is given by,

$$E_k = \left\langle \frac{1}{2} \sum_{i=1}^{N} m_i v_i^2 \right\rangle = \frac{3}{2} k_B T. \tag{5}$$

The average total potential energy,

$$E_p = \left\langle \sum_{i=1}^{N-1} \sum_{j>i}^{N} V(r_{ij}) \right\rangle \tag{6}$$

is obtained by summing over (in this case) pairwise additive interactions. The sum of kinetic and potential energy should be constant. The average linear and angular momentum can similarly be calculated as $\mathbf{p} = \langle \sum_i m_i \mathbf{v}_i \rangle$ and $\mathbf{J} = \langle \sum_i \mathbf{r}_i \wedge \mathbf{v}_i \rangle$.
Another standard thermodynamic quantity, the pressure, can be calculated from the virial equation,

$$P = \rho k_B T + \frac{1}{3V} \left\langle \sum_{i<j} \mathbf{F}(\mathbf{r}_{ij}) \cdot \mathbf{r}_{ij} \right\rangle. \tag{7}$$

There are also properties like the free energy, chemical potential and entropy that cannot directly be calculated from the particle coordinates or velocities. These properties need special techniques which we will not discuss here. The interested reader is referred to for instance the books of Allen and Tildesley [9] or Frenkel and Smit [10] and references therein for more information.

Structural properties Equally interesting are the structural properties of a system. Static information in real space can be obtained in the form of the pair correlation function $g_2(r)$ which measures the probability to find a particle at distance r from a given particle. The pair correlation function is zero for $r < \sigma$ (particles can not overlap). For low densities, $g_2(r)$ is constant and equals unity for all $r > \sigma$. In contrast, for a crystal $g_2(r)$ will show periodically distributed peaks which reflect the long-ranged order. In fluids, some structure of particles is found in layers around the central particle. In this case, $g_2(r)$ exhibits a first peak around the contact value $r \simeq \sigma$ followed by damped oscillatory behaviour. An example of the pair correlation function will be seen in Fig. 9.
The static structure factor is the equivalent of $g_2(r)$ in Fourier space and is defined as,

$$S(\mathbf{k}) = \frac{1}{N} \langle \rho(\mathbf{k}) \rho(-\mathbf{k}) \rangle \tag{8}$$

where k is the scattering vector and $\rho(\mathbf{k}) = \sum_j \exp(i\mathbf{k} \cdot \mathbf{r}_j)$ is the Fourier transform of the density. The structure factor is relevant since it is the quantity that one measures in a static scattering experiment, for instance with static light-scattering or neutron scattering. In case of an isotropic system, the structure factor is a function of the magnitude of the wavevector $k = |\mathbf{k}|$ only. The range of accessible k-vectors depend on the system size via $k_m = 2\pi/L$ where k_m is the smallest k-vector that can be probed and L is the linear dimension of the box. This demonstrates the limitation of simulations in the sense that large wave length fluctuations (small k-values) are cut-off.

Dynamical properties In systems where the time evolution can be defined, the dynamical dependence of some quantities present valuable information. One of these quantities is the mean squared displacement $\langle \sum_i |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \rangle$. In absence of external fields, the dynamics of molecules of a liquid is ballistic ($\sim t^2$) at short time scales (before collisions between molecules have taken place) and it becomes Brownian ($\sim t$) at longer time scales. It is for these longer times that the self-diffusion coefficient in three dimensions can be defined as,

$$\lim_{t\to\infty} D_s = \frac{\langle \sum_i |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \rangle}{6t}. \tag{9}$$

A detailed discussion of Brownian motion is given in chapter A.2 by J.K.G. Dhont.
Effects of collective dynamic processes can be captured by calculation of the dynamic structure factor

$$F(\mathbf{k}, t) = \frac{1}{N} \langle \rho(\mathbf{k}, t)\rho(-\mathbf{k}, 0) \rangle \tag{10}$$

with $\rho(\mathbf{k}, t) = \sum_j \exp(i\mathbf{k} \cdot \mathbf{r}_j(t))$. The dynamic structure factor describes how the different modes of density fluctuations decay with time. Calculation of the dynamic structure factor allows for a direct comparison of simulation data with dynamic scattering experiments.

2.3 Molecular Dynamics

In a molecular dynamics simulation Newton's equations of motion,

$$\mathbf{F}_i = m_i \mathbf{a}_i \tag{11}$$

are integrated for a system of N interacting particles with mass m_i. Here $\mathbf{F}_i = -\sum_j \nabla_{\mathbf{r}_i} V(|\mathbf{r}_i - \mathbf{r}_j|)$ is the force on particle i and $\mathbf{a}_i$ is the acceleration of particle i. The MD procedure generates the N positions $\mathbf{r}_i(t)$ and velocities $\mathbf{v}_i(t)$ as a function of time t. The force $\mathbf{F}_i$ on each particle can be calculated once the potential $V(r_{ij})$ is specified and the positions of all particles are known.
Schematically a molecular dynamics simulation is done as follows:

- `initialise`: The initial configuration is created by putting N particles in a box of volume V and choosing initial velocities (from a Maxwell-Boltzmann distribution).
- `calculate force`: In the next step the force is calculated. Without tricks, this involves a N^2 operation (one has to calculate the distance between each particle and all particles) for long-ranged interactions such as in a Lennard-Jones system.
- `update positions and velocities`: Once the position, velocity and force for each particle is known the new positions and velocities can be calculated using for instance the Euler scheme,

$$\begin{aligned} \mathbf{r}_i(t + \Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 \\ \mathbf{v}_i(t + \Delta t) &= \mathbf{v}_i(t) + \mathbf{F}_i(t)\Delta t. \end{aligned} \tag{12}$$

- `sample observables`: Quantities of interest such as the internal energy are collected with some sampling frequency to calculate the average value.

These steps are then repeated to generate a time series. It is important to keep in mind that one can start to sample observables *only* if the system has equilibrated.

Molecular Dynamics algorithms In the example in the previous paragraph we have used the so-called Euler algorithm to update the positions and velocities. The idea is straightforward, the positions and velocities are Taylor expanded around some initial value and the expansion is truncated after a finite number of terms. The new coordinates are then estimated using the information available from current time and previous time steps.
The choice of the integration algorithm is important for obtaining sensible results. Although many integration algorithms are available, only a few give good (correct) results. What are the important factors deciding which algorithm to choose? Note that the MD procedure as described here is at constant total energy (microcanonical ensemble), and so the first requirement is that the total energy and momentum (linear and angular) are conserved. Advantageous is the possibility to use a large time step without losing too much numerical stability. Finally, and this might be counter intuitive, the accuracy of the particle trajectories is not so relevant because all systems we consider here show chaotic behaviour. This means that the two phase space trajectories of two configurations which are initially infinitesimally close, diverge exponentially.
Taking into account the previous considerations, it turns out that the Euler algorithm used in Eq. (12) is particularly bad. A much better choice is the so-called Verlet algorithm [16] which can be easily derived from the Taylor expansion of the coordinates in $+\Delta t$ and $-\Delta t$,

$$\begin{aligned}\mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \dot{\mathbf{r}}_i(t)\Delta t + \frac{\ddot{\mathbf{r}}_i(t)}{2!}\Delta t^2 + ... \\ \mathbf{r}_i(t-\Delta t) &= \mathbf{r}_i(t) - \dot{\mathbf{r}}_i(t)\Delta t + \frac{\ddot{\mathbf{r}}_i(t)}{2!}\Delta t^2 - ....\end{aligned}$$

Using $\ddot{\mathbf{r}}_i(t) = \mathbf{F}_i(t)/m_i$ and $\dot{\mathbf{r}}_i(t) = \mathbf{v}_i(t)$ and summing these equations we get the expression for the new positions while subtracting the two equations gives the expression for the new velocities,

$$\begin{aligned}\mathbf{r}_i(t+\Delta t) &= 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 \\ \mathbf{v}_i(t) &= \frac{\mathbf{r}_i(t+\Delta t) - \mathbf{r}_i(t-\Delta t)}{2\Delta t}.\end{aligned} \tag{13}$$

The disadvantages of this algorithm are that positions and velocities are not known at the same time and that the velocities do not enter in the update of the positions. The latter is inconvenient because periodic velocity rescaling in the equilibration stage is often used to establish a preferred initial temperature. The major advantage of the Verlet scheme over the Euler algorithm is that it conserves energy very well. At the same time, the algorithm allows for using a large integration time step Δt. Alternatively, one can use the velocity Verlet algorithm which yields positions and velocities at the same time,

$$\begin{aligned}\mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 \\ \mathbf{v}_i(t+\Delta t) &= \mathbf{v}_i(t) + \frac{\mathbf{F}_i(t+\Delta t) + \mathbf{F}_i(t)}{2m_i}\Delta t.\end{aligned} \tag{14}$$

Note that in this case one has to calculate the new forces before the velocities can be updated.
In Fig. 6 we compare the time dependence of the temperature of a Lennard-Jones system integrated with the Euler algorithm and with the velocity Verlet algorithm. The system consists of 1000 particles at dimensionless density $\rho\sigma^3 = 0.844$ (with ρ the number density $\rho = N/V$) and with cut-off $r_c = 3.0$. The initial temperature was set at $k_BT/\epsilon = 1.0$ and the time step in both simulations was $\Delta t/\sigma\sqrt{m/\epsilon} = 0.001$. Clearly, the Euler algorithm does not conserve energy.

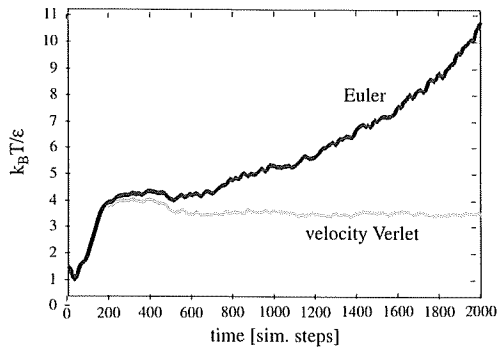


Fig. 6: *The time evolution of the temperature for the same initial configuration integrated with the velocity Verlet algorithm and the Euler algorithm.*

Note that the initial behaviour of the two algorithms is identical but that the Euler scheme starts to deviate from the Verlet scheme after a relatively short time. Finally, the results of the two algorithms diverge. Figure 6 also illustrates that the temperature of the system integrated with the velocity Verlet algorithm relaxes to a constant (equilibrated) value. To obtain a different temperature one could just re-scale the velocities with a constant factor.

2.4 Monte Carlo

The Monte Carlo method was historically the first method for simulating the properties of a fluid using a computer [17]. The main idea of the method consists in obtaining the (thermodynamic) properties of a system by sampling particle configurations proportional to their Boltzmann weight. Let us consider a system in the canonical ensemble, so we have N particles at certain positions $\mathbf{r}^N = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ at temperature T in a volume V. It is known that averages in the canonical ensemble can be calculated via,

$$\langle A \rangle_B = \frac{\int d\mathbf{r}^N A(\mathbf{r}^N) e^{-E(\mathbf{r}^N)/k_BT}}{\int d\mathbf{r}^N e^{-E(\mathbf{r}^N)/k_BT}} \tag{15}$$

where A is some observable, $E(\mathbf{r}^N)$ is the potential energy of a configuration $\mathbf{r}^N$ and the subscript B denotes a Boltzmann weighted average. The Boltzmann factor is given by,

$$p(\mathbf{r}^N) = \frac{e^{-E(\mathbf{r}^N)/k_BT}}{\int d\mathbf{r}^N e^{-E(\mathbf{r}^N)/k_BT}} = \frac{e^{-E(\mathbf{r}^N)/k_BT}}{Q(N,V,T)} \tag{16}$$

where $Q(N,V,T)$ is the partition function of the system. Note that the average in Eq. (15) is obtained by weighting the particular values $A(\mathbf{r}^N)$ with their Boltzmann weight in contrast to the average in Eq. (3).

Simple sampling & importance sampling To introduce the Monte Carlo method we first discuss the simplest Monte Carlo approach in the calculation of integrals. This method is called random or simple Monte Carlo sampling. Suppose one has an integral of the form

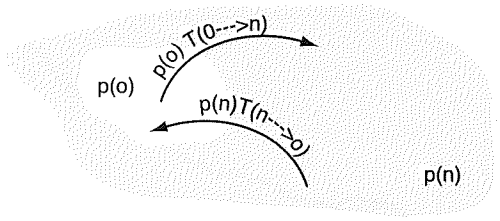


Fig. 7: *In equilibrium, the flow from an old state (o) is balanced by the flow from all new states (n).*

$I = \int_a^b f(x)dx$. Using random sampling, the value of this integral is determined by evaluating $f(x)$ at M randomly chosen values of x in the interval $x \in (a, b)$ and averaging the corresponding $f(x)$. The integral is obtained by $I = (b-a)\langle f(x)\rangle$ with the average defined in Eq. (3). This method works for low dimensional integrals under condition that the integrand behaves properly, i.e. is a smooth, slowly varying function of the variables.
However, in our application Eq. (15), the integrand is a rapidly varying function of the configurations $\mathbf{r}^N$. Moreover, the integrand is zero for most configurations and sharply peaked around the average value of the energy. This behaviour of the integrand can be understood because if we were to generate configurations randomly, then most of these configurations would contain overlapping pairs of particles for which the energy is large (infinite). These configurations would have a zero Boltzmann factor. Using the simple (random sampling) Monte Carlo scheme we would thus spend a lot of time in calculating zero contributions to the integral.
For these kind of situations it is necessary to employ what is called importance sampling. In this method, the integrand is sampled often when the Boltzmann factor in Eq. (16) is large and less frequent when the Boltzmann factor is small. In other words, the Monte Carlo method is an efficient scheme to generate a sequence of configurations proportional to their Boltzmann factor. Using these configurations, an average is calculated in which the weighting is contained in the selection of configurations $\mathbf{r}^N$. A new configuration is created from a preceding one according to a specified transition probability.

Metropolis scheme Let us now derive the Metropolis scheme [17] for determining the transition probabilities from a state $\mathbf{r}^N = o$ (o = old) to another state $\mathbf{r}'^N = n$ (n = new) and thus for generating a sequence of states obeying the Boltzmann distribution. State o has a Boltzmann factor given by $\exp\left[-E(o)/k_BT\right]/Q(N,V,T)$, where $E(o)$ is the potential energy of this configuration and k_BT the prescribed thermal energy. In equilibrium there is no net flow between states o and states n. This means that, *on average*, the number of accepted trial moves resulting in leaving states o must be exactly balanced by the number of accepted trial moves arriving in states o from *any other state* n (see Fig. 7). In practice, a stronger condition is imposed, namely that in equilibrium the number of accepted trial moves from a state o to a particular state n should be exactly balanced by the number of accepted trial moves from a particular state n to o. This condition is called detailed balance and can be written as,

$$p(o)T(o \Longrightarrow n) = p(n)T(n \Longrightarrow o) \tag{17}$$

where $T(\alpha \Longrightarrow \beta)$ denotes the transition probability to go from a state α to a state β.

The subtle difference is thus that in equilibrium the net flow from each state is zero (which means that there can be still flow from this state to a specific other state under the condition that it is balanced by flow backwards from other states) while detailed balance implies that the net flow from each state to each other state is exactly zero.
The transition probability is in itself the product of two processes, the creation of a trial move C and the acceptance of this trial move A,

$$T(\alpha \Longrightarrow \beta) = C(\alpha \Longrightarrow \beta)A(\alpha \Longrightarrow \beta) \tag{18}$$

In many Monte Carlo applications the creation of a trial move is a symmetric process, i.e. the creation of the forward and backward moves have the same probability and thus $C(\alpha \Longrightarrow \beta) = C(\beta \Longrightarrow \alpha)$. Using this form, inserting Eq. (18) in Eq. (17) and using the Boltzmann weight $p(\alpha) = \exp\left[-E(\alpha)/k_BT\right]$ gives,

$$\frac{p(n)}{p(o)} = \frac{A(o \Longrightarrow n)}{A(n \Longrightarrow o)} = \exp\left[-\left(E(n) - E(o)\right)/k_BT\right] \tag{19}$$

The algorithm determines the final expression for the acceptance probability as,

$$\begin{array}{rcll} A(o \Longrightarrow n) & = & p(o)/p(n) & \text{if} \quad E(n) > E(o) \\ & = & 1 & \text{if} \quad E(n) \leq E(o) \end{array} \tag{20}$$

In practice, a Monte Carlo translational move is performed as follows. Calculate the energy of the present state $E(o)$, then select a particle at random, move this particle over a random distance $\Delta R = \text{RAN} \times R_m$ with RAN a random number in the interval $\{-1, \ldots, 1\}$ and R_m the maximum displacement (which is a tunable parameter of the program). Calculate the energy of the new configuration $E(n)$. If the new energy is lower than the old energy accept the move immediately, in case the new energy is higher than the old energy then accept the move with probability $p(o)/p(n)$. This is implemented by generating a random number RAN in the interval $\{0, ..., 1\}$, if RAN $< p(o)/p(n)$ the move is accepted. If RAN $\geq p(o)/p(n)$ the move is rejected. Schematically, a Monte Carlo simulation is done as follows,

- `initialise`: the initial configuration is created by putting N particles in a box of volume V.
- `update positions`: generate new positions following Metropolis scheme
- `sample observables`: Quantities of interest are collected with some sampling frequency.

The magnitude of the parameter R_m determines the efficiency of the MC procedure. If R_m is large, many of the trial moves are rejected. On the other hand, if R_m is too small we will only sample phase space very slowly. A good choice for R_m is a value such that half of the trial moves are accepted.

Random Numbers Clearly, random numbers play a crucial role in Monte Carlo simulations (hence the name Monte Carlo) and therefore one should use a good random number generator. Most random number generators are algorithms that generate a sequence of numbers given an initial value of the 'seed'. The randomness of this sequence, characterised by the distribution of the numbers in the sequence, and the correlations between the numbers, determines the quality of the random number generator. A useful discussion of random number generators can be found in [11], some useful algorithms are available in [18].

Other thermodynamic ensembles In the discussion so far we have only considered the calculation of properties in the microcanonical ensemble (constant number of particles N, volume V and total energy E) and the canonical ensemble (constant number of particles N, volume V and temperature T). However, computer simulations are by no means restricted to these ensembles. Many important extensions of MD and MC exist in for instance the constant pressure, constant chemical potential, constant stress and grand canonical ensemble [9, 10, 11, 12].

3 Example: Molecular Dynamics of a mixture of spheres

In this section we present an example of a molecular dynamics simulation and show some of the properties that can be calculated.
The understanding of the basic properties of mixtures of particles differing in size and/or shape is important for biological systems. In chapter C.4, a specific biological application namely the phase behaviour of mixures of polymers and proteins is discussed. Depletion interactions are an important concept in these mixtures. These are the averaged attractive interactions which two protein particles experience in the presence of polymer molecules. The depletion interaction can, if strong enough, give rise to phase separation in a protein rich (polymer poor) and a protein poor (polymer rich) phase. The structure of these phases (gas, liquid, crystalline, gel) depends on the protein and polymer concentrations as well as on the size ratio between polymer and protein. In this section we will discuss some aspects of the structure and dynamics of a highly simplified model of such a protein-polymer mixture namely a mixture of freely overlapping spheres and 'hard' spheres.

System Consider a cubic box containing N_1 particles with diameter σ_{11} (representing the protein) and N_2 particles with diameter σ_{22} (representing the polymer). The interactions in the mixture are given by,

$$\begin{aligned} V_{11} &= \begin{cases} 4\epsilon_{11}\left[\left(\frac{\sigma_{11}}{r}\right)^{12} - \left(\frac{\sigma_{11}}{r}\right)^{6} + \frac{1}{4}\right] & \text{for } r < 2^{\frac{1}{6}}\sigma_{11} \\ 0 & \text{for } r \quad 2^{\frac{1}{6}}\sigma_{11} \end{cases} \\ V_{22} &= 0 \\ V_{12} &= \begin{cases} 4\epsilon_{12}\left[\left(\frac{\sigma_{12}}{r}\right)^{12} - \left(\frac{\sigma_{12}}{r}\right)^{6} + \frac{1}{4}\right] & \text{for } r < 2^{\frac{1}{6}}\sigma_{12} \\ 0 & \text{for } r \quad 2^{\frac{1}{6}}\sigma_{12} \end{cases} \end{aligned} \tag{21}$$

Where a subscript $\alpha\beta$ denotes the interaction between species α and species β. The cross-interaction strength is defined as $\epsilon_{12} = \sqrt{\epsilon_{11}\epsilon_{22}}$ and the cross-interaction diameter as $\sigma_{12} = (\sigma_{11} + \sigma_{22})/2$. The interaction between particles of species 1, and the interactions between particles of species 1 and 2 are repulsive for small separations and zero elsewhere. The particles of species 2 are non-interacting i.e. they behave like an ideal gas.
The interaction strength of species 2 is expressed as $\epsilon_{22} = q\epsilon_{11}$ with $q = \sigma_{22}/\sigma_{11}$ the diameter ratio of species 2 and 1. The mass of the particles of species 2 is related to the mass of particles of species 1 via $m_2 = q^3 m_1$. Here we choose $\sigma_{11} = 1$, $\epsilon_{11} = 1$ and $m_1 = 1$ so that the magnitude of the other parameters only depends on the size ratio q.
Molecular dynamics simulations have been carried out employing the velocity Verlet algorithm Eq. (14) discussed in section 2.3 with an integration time step of $\Delta t/\sigma_{11}\sqrt{m_1/\epsilon_{11}} = 0.001$. The

q	N_1	η_1	N_2	η_2
0.25	500	0.26	0-50000	0-0.41
0.60	500	0.26	0-3000	0-0.34

Table 1: *Particle numbers and corresponding volume fractions of species 1 and 2 used in the simulations described here.*

systems we have studied in more detail are mixtures with diameter ratio $q = \sigma_{22}/\sigma_{11} = 0.25$ and with $q = 0.60$. The remaining simulation parameters are listed in Tab. 1. The volume fraction of particles of species i is defined as $\eta_i = \pi\sigma_i^3 N_i/6V$.
We define a η_2 dependent self diffusion coefficient $D(\eta_2)$ and a normalised self-diffusion coefficient $D(\eta_2)/D(0)$ for species 1. These are calculated from the mean squared displacement (Eq. 9) of species 1 averaged over many realisations. The diffusion coefficient is expected to depend on η_1, η_2 and q.
In our simulations, we make a scan through the phase diagram at fixed η_1 while varying η_2. In the phase diagrams of chapter C.4, this corresponds to a vertical scan at $\eta_1 = 0.26$, where $\eta_1 = \phi_c$ in C.4. All the systems were prepared in a homogeneous state by randomly inserting the particles of species 1 and 2 in the box such that there were no overlaps between the particles of species 1 and between particles of species 1 and 2. Subsequently, the equations of motion were integrated for a large number of time steps, until a stable final state was reached. By a stable state we mean that the peak positions and amplitudes of the pair correlation function had reached a constant value. All results presented are related to this final state.

Results For illustration we show in Fig. 8A the mean squared displacement of species 1 divided by $6t$ for $q = 0.6$ and for some volume fractions η_2. These volume fractions were chosen such that the systems are stable. For small times, the curves have a slope $\sim t$ indicating ballistic behaviour. For longer times, this crosses over to Brownian motion (the plateau). The plateau value ($D(\eta)$) and the cross-over time both decrease with increasing η_2 i.e. the motion becomes slower. Figure 8B shows the normalised diffusion coefficient as a function of η_2 for $q = 0.25$ and for $q = 0.6$. Clearly, the diffusion of species 1 is decreased in case of a smaller size ratio at fixed η_2. This can be readily understood as for small polymers at the same volume fraction, the number density is much higher which increases the effective friction for the particles.
In Fig. 9 we show the pair correlation function for $q = 0.25$ at $\eta_2 = 0.41$ and $q = 0.6$ at $\eta_2 = 0.34$. Both these systems are quenched into the two phase region of the phase diagram. For $q = 0.25$ a significant first peak is found indicating strong aggregation of the particles. The second peak shows signs of splitting into two peaks which is characteristic for crystallisation. The system with larger size ratio $q = 0.6$ shows the characteristic features of a dense liquid, in the form of broad, well-pronounced peaks in the pair correlation function.
It is difficult to imagine the real space structure of our system from the pair correlation function alone. Therefore we illustrate our results by showing two snapshots in Fig. 10. The yellow spheres represent the protein particles (species 1) while the blue spheres (not at scale) represent the polymer particles (species 2). The two snapshots show that, as expected, both systems have become inhomogeneous. In the case of $q = 0.25$, we see that the system has undergone a microscopic phase separation into 'protein' rich and 'protein' poor domains with a small characteristic length scale of about 4 to 5 particle diameters. For $q = 0.6$, the phase separation has proceeded further and has led to macroscopic phase separation into two major domains, one

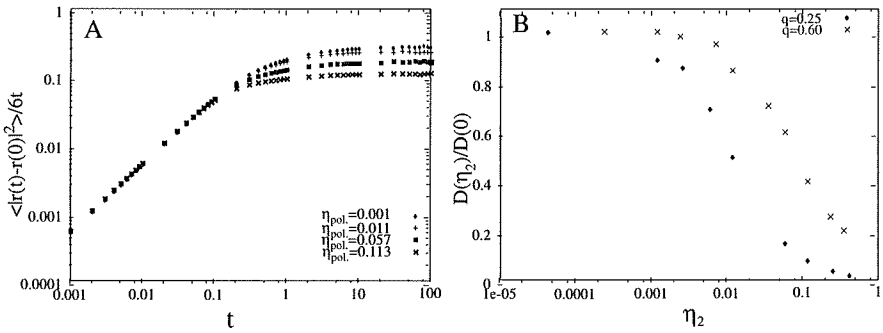


Fig. 8: *Diffusion coefficient for species 1 as a function of η_2. A: results for $q = 0.25$, $0 < \eta_2 < 0.12$, B: results for $q = 0.25$ and $q = 0.6$ plotted as a function of η_2.*

rich in 'protein', and one rich in 'polymer'. From the snap shots, no evidence for ordering is found.

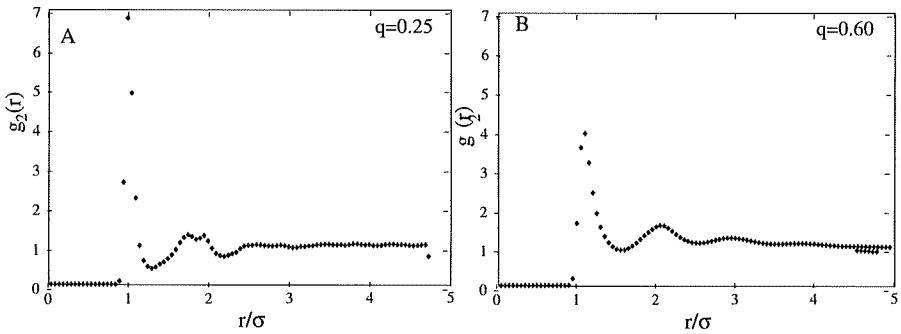


Fig. 9: *A: pair correlation function for a mixture with $q = 0.25$ at $\eta_2 = 0.41$. B: pair correlation function for a mixture with $q = 0.6$ at $\eta_2 = 0.34$.*

4 Conclusions

We have shown that computer simulations form a useful tool to study complex systems. In the example, we have illustrated that using a simple model consisting of freely overlapping spheres and repulsive spheres, we were able to capture some of the main features of a phase separating protein-polymer mixture. Computer simulations allow systems to be studied at different levels of detail. Starting with the simplest models, the complexity can be slowly increased to systematically investigate different contributions of the model to the behaviour of the system. In combination with experiments and theoretical methods, computer simulations give a powerful tool to study complex systems. An appealing advantage of computer simulations over most experimental techniques and theoretical models is that the system can be visualised dynamically.

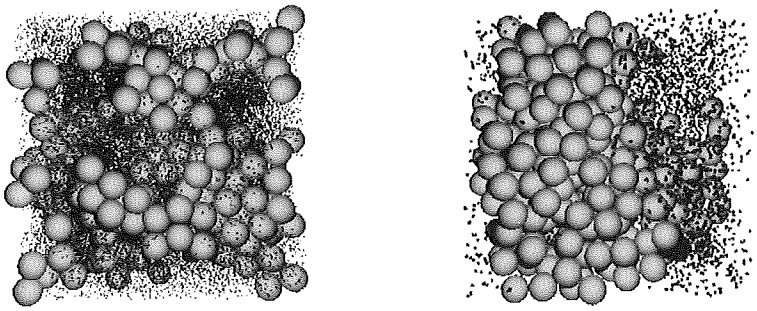


Fig. 10: *A: simulation snap shot of a mixture with* $q = 0.25$ *at* $\eta_2 = 0.41$. *B: simulation snap shot of a mixture of a mixture with* $q = 0.6$ *at* $\eta_2 = 0.34$.

Acknowledgement I would like to thank Marisol Ripoll for a critical reading of the manuscript and many helpful suggestions.

References

[1] Y. Yoshikawa, Y.S. Velichko, Y. Ichiba and K. Yoshikawa, *Eur. J. Biochem.* **268**, 2593 (2001)

[2] A.E. Garcia, *Theoretical and biophysics group, Los Alamos National Laboratory* www.t10.lanl.gov/profiles/garcia.html

[3] K.-K. Kunze and R.R. Netz, *Phys. Rev. Lett.* **85**, 4389 (2000)

[4] J.-P. Hansen and I.R. McDonald, *Theory of simple liquids* (2nd Edition, Academic Press, London, 1986)

[5] B.J. Alder and T.E. Wainwright, *J. Chem. Phys.* **27**, 1208 (1957)

[6] P.N. Pusey and W. van Megen, *Nature* **320**, 990 (1984)

[7] J.E. Lennard-Jones, *Physica* **4**, 941 (1937)

[8] B. Smit, *J. Chem. Phys.* **96**, 8639 (1992)

[9] M.P. Allen and D.J. Tildesley, *Computer simulations of liquids* (Clarendon, Oxford, 1987)

[10] D. Frenkel and B. Smit, *Understanding molecular simulation* (Academic Press, San Diego, 1996)

[11] D.P. Landau and K. Binder, *A guide to Monte Carlo simulations in statistical physics* (Cambridge University Press, Cambridge, 2000)

[12] D.C. Rapaport, *The art of molecular dynamics simulation* (Cambridge University Press, Cambridge, 1995)

[13] G.R. Mcnamara and G. Zanetti, *Phys. Rev. Lett.* **61** 2332 (1988)

[14] R.D. Groot and P. Warren, *J. Chem. Phys.* **107** 4423 (1997)

[15] A. Lamura, G. Gompper, T. Ihle and D.M. Kroll, *Europhys. Lett.* **56** 319 (2001)

[16] L. Verlet, *Phys. Rev.* **165**, 201 (1967)

[17] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.M. Teller and E. Teller, *J. Chem. Phys* **21**, 1087 (1953)

[18] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, 1986)

B 6 Single Molecule Fluorescence

J. Enderlein
Institut für Biologische Informationsverarbeitung 1
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Photophysics, Probes, and Markers 2

3 Physical Techniques 5

3.1 Modified Flow Cytometry, Microchannels, and Microdroplets 5
3.2 Confocal Detection 7
3.3 Wide-Field Imaging 9

4 Data Acquisition and Evaluation 11

4.1 Time-Tagged and Time-Correlated Photon Counting 12
4.2 Fluorescence Correlation Spectroscopy 13
4.3 Fluorescence Intensity Distribution Analysis and Related Techniques 15
4.4 Molecule-by-Molecule Analysis 15

1 Introduction

The first, indirect detection of a single molecule in solution was performed by Rotman when detecting the multiple reaction products of a single enzymatic molecule [1,2]. The next step was the direct optical detection of a single although multiply labelled molecule by Hirschfeld [3,4]. The importance of this work was to demonstrate the principal possibility of direct optical single molecule detection in a liquid, and it exercised a great influence on the further development of the field. Finally, the group of Richard A. Keller at Los Alamos National Laboratory was first able to detect individual rhodamine molecules in a fluid flow by 1996 [5,6]. Since then, the research on single molecule detection and spectroscopy (in short SMS) in solution and on surfaces under ambient conditions has seen an explosive development. Recently, there have been several, partially exhaustive reviews concerning the quickly evolving field of research namely Refs.[7-13], and several books have been published [14,15]. The present paper intends to give an overview of the actual state of the art of single molecule fluorescence spectroscopy and to present a few selected applications from our research group.

The scope of the review will be restricted to fluorescence SMS in liquids and on surfaces under ambient conditions (molecules embedded within solvents or bound to surfaces, measurements performed at room temperature), thus neglecting the broad research done on low-temperature spectroscopy (see e.g. Ref.[16]), and on atomic force and scanning tunnelling microscopy of single molecules. Moreover, near-field microscopy of single molecules [17-19] is also excluded from the considerations because it is too broad a field of research, which would need the attention of a special overview by itself (but see the recent work of van Hulst's group, Ref. [20-25]).

2 Photophysics, Probes, and Markers

Quoting the words of Richard A. Keller, it is not difficult to detect the fluorescence of a single molecule but *not* to detect anything else. 'Anything else' means, first at all, elastic and inelastic scattering of light by the surrounding medium, and fluorescence from unwanted fluorescence sources (e.g. autofluorescence from biological samples). For successfully observing a single molecule by its fluorescence, it has to have suitable photophysical properties. These properties are: absorption and emission spectrum, optical absorption cross section, quantum yield of fluorescence, fluorescence lifetime, triplet-state dynamics, 'blinking' dynamics, and photostability. In a very simplified form, a single fluorescing molecule can be described by the Yablonski scheme as depicted in Fig.1: The non-excited molecule resides in its electronic ground state S_0, from where it can be excited by single or multi-photon absorption into its first excited electronic state S_1, from where it can either return to its ground state S_0, via emission of a fluorescence photon or by a non-radiative transition induced by the interaction with the surrounding medium, either switch into a triplet state (T_1) via intersystem crossing, or may be irreversibly destructed by a photochemical reaction (photobleaching). In this simplified picture, vibrational modes within the electronic states were completely left out – a more detailed description of molecular photophysics can be found in many textbooks. Here, it is only important that, due to the vibrational mode substructure of the electronic states, fluorescence emission is red-shifted with respect to absorption (so called Stokes shift). This is crucially important for a successful implementation of any single molecule detection scheme: The Stokes shift allows for a spectral separation of the fluorescence emission from the excitation light which is elastically (Rayleigh) and inelastically (Raman) scattered back by the embedding medium.

However, for most dyes solved within a solvent like water, Raman contributions will always overlap with the fluorescence emission spectrum.

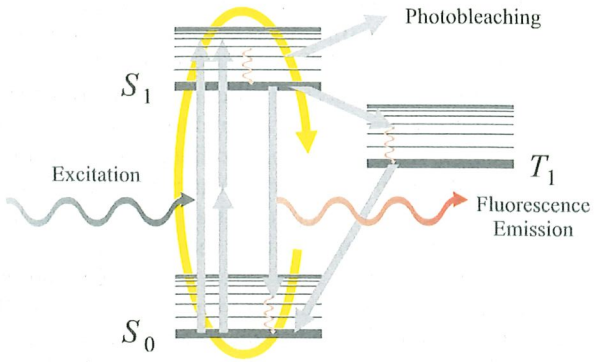


Fig. 1: *Simplified three-state Yablonski diagram of a molecule's energetic states that are involved in its photophysics. The molecule can be excited, by a one- ore multi-photon process, from its electronic ground state S_0 to the first excited state S_1. Its return to the ground state can be accompanied by the emission of a fluorescence photon. Repeated cycling of the molecule through its excited state generates a characteristic burst of detectable fluorescence photons. Competing with the S_1-S_0 transition is intersystem crossing into the first triplet state T_1, and irreversible photobleaching, which eventually leads to the complete disappearance of the molecule from detection.*

The next important property is the absorption cross section: for a one-photon fluorescence excitation process, both the fluorescence signal as well as the intensity of the scattered light are linearly proportional (below optical saturation of the fluorescence) to the excitation intensity. Increasing the absorption cross section of the fluorescing molecules thus increases the ratio of detectable fluorescence to backscattered light, improving the signal-to-noise ratio of the detection. The absorption cross sections of many organics dyes are typically between 10^{-16} and 10^{-15} cm^2, corresponding to optical extinctions between 26.000·and 260.000 l mol^{-1} cm^{-1}.

As just mentioned, the linear relation between excitation intensity and fluorescence signal is only correct for excitation intensities much below optical saturation: optical saturation occurs for excitation intensities where the molecule spends a significant art of the time not in the ground state but in the excited or the triplet states. When optical saturation starts, increase in excitation power will not produce correspondingly more fluorescence photons per time unit, and thus signal-to-noise ratio gets lower with increasing excitation intensity.

The next quantity to consider is the quantum yield of fluorescence. It refers to the probability that the return from the excited to the ground state is accompanied by the emission of a fluorescence photon (radiative transition). For many fluorescing molecules, this radiative transition is competed by non-radiative transitions, where the energy of the excited state is transferred to the solvent collisions or e.g. electron transfer. A rule of thumb is that fluorescence quantum yield decreases with increasing wavelength of the excitation/emission maximum of a dye. Thus, the standard laser dye rhodamine 6G, excitable around 520 nm, has a fluorescence

quantum yield of nearly unity, whereas the popular dye Cy5 (excitation near 640 nm) shows a fluorescence quantum yield of 0.28 (in water).

Besides cycling between the ground and excited state, a molecule can undergo an intersystem crossing to its lowest triplet state T_1. The term intersystem crossing stems from the fact that the transition is accompanied by a spin flip of the excited electron and thus symmetrically disfavoured. Correspondingly, intersystem crossing rates are usually low (ca. one crossing for every $10^5 - 10^6$ excitations). However, the average lifetime of the triplet state is usually much larger than the fluorescence lifetime (in the microsecond to millisecond range), leading to the repeated 'disappearing' of a single molecule from detection. Thus, triplet state dynamics often determines the saturation intensity. Even more problematic is the so called 'blinking' of single molecules [26-28], by which one denotes the optical disappearing of single molecules on even much larger time scales (for seconds and more), thus seriously impeding their detection. The nature of this blinking behaviour is still an active area of research, and no final and universally accepted explanation was yet put forward. Unfortunately, especially fluorescent proteins (cyan, green, yellow etc. fluorescent proteins) which are widely used and very popular in biological studies show dramatic blinking [29-31], making it partially difficult to use them as single molecule labels.

One of the most important photophysical characteristics is the photostability of a fluorescent molecule. It refers to the probability that a molecule undergoes an irreversible chemical reaction while it resides in its excited state. This photodestruction leads to the final disappearance of the molecule from detection and thus ultimately limits any SMS experiment. Typical fluorescent dye molecules survive ca. 10^5 to 10^6 excitation cycles until photodestruction, although this number can vary widely and strongly depends on the nature of solvent. Especially in biological SMS applications, the photostability is the most limiting factor, defining the maximum number of photons that can be detected from a single molecule. Again, the popular fluorescent proteins show rather high photodestruction rates, the green fluorescent protein family being the most suitable one for SMS applications [30-33]. Recently, new kinds of fluorescing labels with high photostability and low triplet state/blinking dynamics were developed, such as fluorescing nano-crystals [34-37], solid-state colour centres [38], or metallic nano-particles [39-43].

An efficient strategy of background suppression is to use fluorescent dyes with long excitation and emission wavelengths in the near-infrared and infrared wavelength region [44]. As is well known, the intensity of non-resonant light scattering decreases roughly with the fourth power of the wavelength, thus doubling the excitation wavelength decreases the scattering background by sixteen times. Additionally, using long-wavelength dyes is also advantageous when performing SMS in biological samples containing autofluorescent biomolecules. Fig.2 shows the optical absorption of several classes of synthetic dyes and those of fluorescent biomolecules. In recent years, new classes of fluorescent dyes with photophysical parameters sufficient for single molecule detection in the near-infrared region (>600 nm) have become widely available, such as the popular cyanine dyes by Amersham or the Alexa dyes by Molecular Probes. Meanwhile, even dyes beyond 700 nm are available (e.g. the IRDye family by LI-COR), although not yet characterized on the single-molecule level.

A topic of considerable interest is the development of new near-infrared and infrared dyes with properties suitable for SMS (high fluorescence quantum yield, high photo-stability), see e.g. [45]. There are two principal reasons for being interested in such dyes: firstly, scattering intensity of the excitation light, a major source of unwanted background in SMS, drops ap-

proximately off with the fourth power of the wavelength. Secondly, using NIR and IR dyes prevents the unwanted excitation of autofluorescence in biological samples. This is of significance for many SMS applications in single-cell studies or SM sensitive screening of biological fluids.

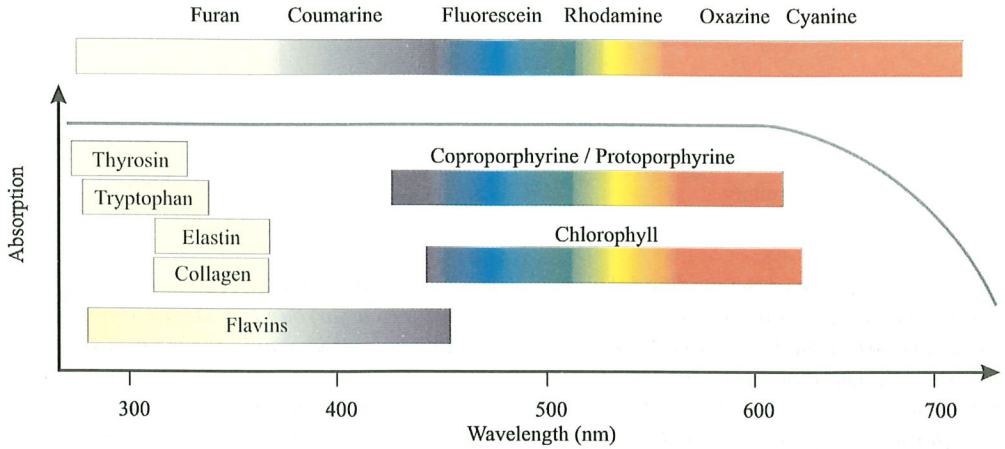


Fig. 2: *Optical absorption regions of the fluorescence of biological molecules and several synthetic dye families.*

A completely new and purely physical way of improving the photophysical properties of conventional dyes is to use the electrodynamic interaction of fluorescing dyes with metal for designing new composite fluorescing materials with strongly enhanced fluorescence properties. Lakowicz [46] coined the term 'radiative decay engineering' describing that modification of emission properties of fluorophores or chromophores in the presence of metals, leading to enhanced photostability and optical excitation efficiency. There are detailed calculations of these effects for fluorescent metallic nano-cavities available [47,48], predicting remarkable enhancement values for photostability and fluorescence brightness, although the metal is completely enclosing the fluorescing entity. The calculations clearly show that the synthesis of such metal/fluorescent composites may lead to new fluorescing labels with extraordinary optical properties.

3 Physical Techniques

3.1 Modified Flow Cytometry, Microchannels, and Microdroplets

Historically, the first successful detection of a single molecule in solution was realized with a modified flow cytometer using hydrodynamic focusing [6]. The principle of this detection scheme is depicted in Fig.3. A sample stream containing the dissolved fluorescing molecules is injected into a surrounding sheath flow providing transport of the injected molecules and simultaneous hydrodynamic focusing of the sample stream [49].

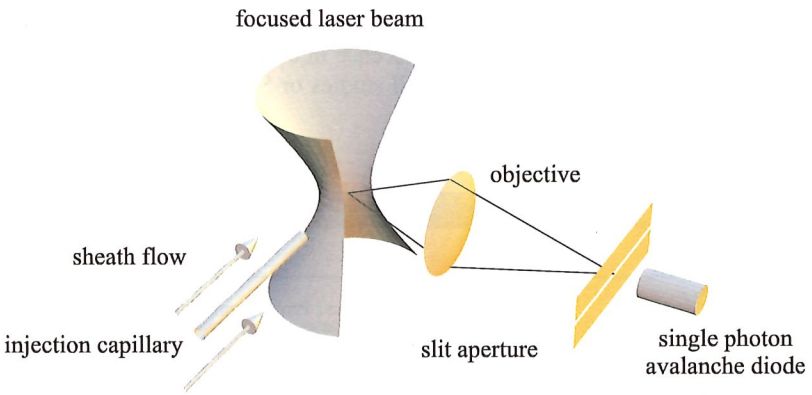


Fig. 3: *Schematic of the modified flow cytometer for SMS application. A sample stream containing the dissolved fluorescing molecules is injected into a surrounding sheath flow providing transport of the injected molecules and simultaneous hydrodynamic focusing of the sample stream. The molecules are transported through a focused laser beam, which inter-sects the fluid flow at a right angle. Fluorescence is detected through an objective with its optical axis perpendicular to the flow direction and the exciting laser beam. The fluorescence light is imaged onto a single-photon sensitive light detector, such as a single-photon avalanche diode.*

The molecules are transported through a focused laser beam (beam waist ca. 10 μm), which intersects the fluid flow at a right angle. Fluorescence is detected through an objective with its optical axis perpendicular to the flow direction *and* the exciting laser beam. Transit of a single molecule is detected as a burst in time of recorded photons. Due to the perpendicular arrangement of detection and excitation, the numerical aperture of the used objectives is much smaller than that used in confocal epi-fluorescence set-ups (high NA objectives have a very small working distance of ca. 250 μm, leaving little space for arranging two objectives at 90 degrees), so that the detection volume of the flow-cytometry set-up is ca. 10^3 larger and overall detection efficiency lower than that of confocal systems. Thus, background suppression is a much more serious issue here. The main advantage of the flow-cytometry set-up is that all molecules pass the detection region in nearly the same way, so that the burst size directly reflects the fluorescence brightness of the molecules. By evaluating the number of detected photons per molecule it is possible to distinguish between different molecular species solely by their brightness (single molecule fluorescence intensity), not using any other fluorescence property [50]. A promising application of this technique is rapid and ultra-sensitive DNA fragment sizing, employing an intercalating dye for staining DNA fragments and relying on the assumption that dye load per fragment (and thus fluorescence brightness) is proportional to fragment size [51,52]. The same working principle is also behind a new method of high-throughput detection of single nucleotide polymorphisms (SNPs) as described by Cai et al. [53]. The idea of 'guiding' molecules through the detection region can also be achieved by employing microcapillaries [54-59] or microchannels [60-63]. The advantage is its simplicity (no hydrodynamic focusing is necessary to ensure transition of the molecules through the detection region), the disadvantage may be higher background signal.
Closely related to SMD in a fluid flow is SMD in gel capillary electrophoresis [64-66]. This is an important field of research because it combines the versatility of electrophoretic separation

with the sensitivity of SMS. Most papers concerned with SMS in electrophoresis were dealing with DNA analysis (fragment sizing and/or specific sequence detection) [67-71].

One of the challenges of the flow cytometry approach is to prevent diffusion of the analyte molecules away from the flow axis. When using microcapillaries or microchannels, the molecules' motion is confined laterally, for the price of a heavily increased scattering background. A different solution to this problem is to include the analyte molecules into electrically charged micrometer sized droplets, which can then be handled with electric fields. A solution containing the analyte molecules is sprayed into a stream of microdroplets, which are subsequently directed through the detection region by means of gravity and applied electric fields. The advantage of such an approach is the facility of perfectly guiding the microdroplets through the detection region while avoiding high background signal stemming from any confining microstructures. Over the years, an impressive amount of literature was published by Mike Ramsey's group at Oak Ridge, describing the principles and numerous applications of that method [72-78].

An noteworthy variant for achieving extremely small detection volumes is to excite fluorescence by evanescent illumination together with high numerical aperture collection, achieving detection volumes as small as 0.1 μm^3 [79-82]. In a similar vein, Ref.[83-86] uses light collection above the critical angle of total internal reflection (supercritical angle fluorescence detection) for achieving comparably small detection volumes. In both cases, the extreme localization of the detection region at an interface (glass/water) is of considerable interest when trying to monitor binding kinetics on a single molecule level on the background of a rather high fluorophore concentration in solution.

3.2 Confocal Detection

As already mentioned, having a fluorescent dye with good photophysical properties is only one prerequisite for a successful detection of single molecules. Another important issue is the maximum suppression of Rayleigh and Raman scattering of the excitation light. One way, as discussed in connection with the Stokes shift of the fluorescence, is the spectral discrimination of fluorescence against scattered light by exploiting high-efficient optical filters. In this way, it is possible to completely suppress all elastic scattering contributions (Rayleigh scattering at the excitation wavelength). However, Raman scattering often shows a significant spectral overlap with the fluorescence emission. Thus, a second and important way to minimize unwanted signal contribution due to Raman scattering is to minimize the detection volume, the volume of efficient fluorescence excitation *and* detection. Whereas the fluorescence signal of a single molecule does not change with decreasing detection volume (as long as it is contained within it), the intensity of the scattered light is directly proportional to its extent (number of scattering molecules). This is the basic motivation for all confocal detection schemes. The general scheme of such a system is shown in Fig.4. A laser beam is coupled into a microscope objective with high numerical aperture, which focuses the beam into a diffraction limited spot within the sample. The emerging fluorescence is collected by the same objective (epi-fluorescence detection) and separated from the excitation beam by the dichroic mirror. After passing additional spectral filters for background suppression, the collected fluorescence light is focused onto a confocal aperture, behind which it is detected with a single-photon sensitive photoelectric detector. For technical details of a single-molecule sensitive confocal epi-fluorescence microscope the reader is referred to Ref. [87, 88].

Two effects are employed for minimizing the detection volume: the diffraction limited transversal extent of the laser beam at the focus (depending on wavelength ca. 300-500 nm), and

the restriction of the axial extent of the detection volume by detecting through the confocal aperture. The principle of confocal detection is based on the fact that only light emerging near the focus region can completely pass through the aperture. Any light coming from points farther away from that region is largely blocked by the aperture because it does not converge within the aperture plane. The axial restriction of the detection volume that can be achieved by this technique is ca. 2-4 μm.

The simplest experiment one can perform with the confocal epi-fluorescence set-up is to watch single molecules diffuse through the detection region. This supposes that one works at a sufficiently low concentration of fluorescing molecules: For a detection volume of 2 μm^3 and a concentration below ca. 10^{-9} M, one finds, on average, only single molecules (or none) within the detection region. Every transit of a single molecule through the detection region results in a characteristic burst of detected fluorescence photons.

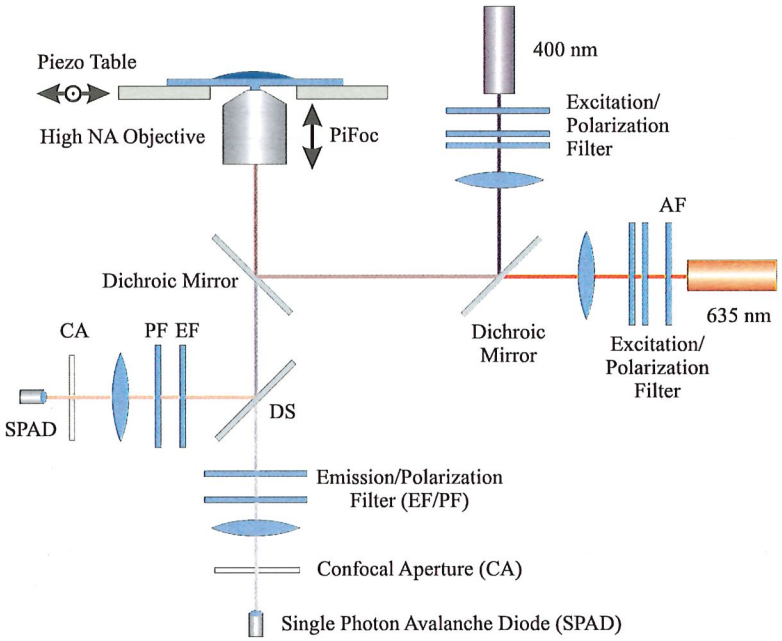


Fig. 4: *Principal scheme of a confocal epi-fluorescence microscope for single molecule spectroscopy. Shown is a set-up with two excitation laser at different wavelength for dual-colour excitation, and an additional lamp for wide-field illumination to observe the sample be-fore starting the SMS measurement. Detection is done also within tow-channels, either for two-colour or two-polarization detection, although there is no principal limitation on the total number of detection channels. The vertical position of the objective is changed by a piezo actuator (PiFoc); sample scanning is performed with a piezo table. For more details, see Ref.[87,88].*

For studying molecules that are immobilized on a surface, the focused laser is scanned over the surface, either to record the fluorescence signal along single lines or to record a complete fluorescence image by scanning in two directions. Confocal imaging and scanning constitute the basic ingredients of a confocal laser-scanning microscope (CLSM). Scanning can also be

extended into the third spatial dimension by scanning planes at different locations along the optical axis (so called z-scanning), resulting in a thee-dimensional fluorescence image of a sample (e.g. a dye-tagged cell, or a transparent substrate with embedded fluorescent molecules).

Compared with SMD by highly sensitive CCD systems, the CLSM offers several advantages. Firstly, due to the small detection volume, the CLSM has an exceptionally high signal-to-background ratio. Secondly, uniform scanning provided, the CLSM assures a homogeneous sample illumination, which may be a challenge in conventional wide field microscopy. Thirdly, the CLSM uses single point detectors for light detection, thus offering the possibility to study extremely fast processes down to the picosecond time scale. This is especially interesting for sensitive fluorescence lifetime imaging, where the florescence decay time characteristics are measured for every point on the scanned surface. Also, the splitting of the detected light into several channels for measuring other spectral characteristics such as polarization or emission wavelength is straightforward in CLSM. The main limitation of CLSM, when compared with CCD imaging, is the limited scan speed with which an image can be recorded. This scan speed is mainly determined by the maximum number of photons per time interval that can be extracted from a single molecule (optical saturation limit). Typically, if one assumes that a single molecule occupies an image area of ca. 1 μm^2, and that a realistic detectable photon count intensity of a single molecule's fluorescence is of the order of one million counts per second, an imaging speed of 0.1 μm^2/ms will yield ca. 100 photons per detected molecule. This is a reasonable number for detecting and studying individual molecules. Thus, when imaging an area of $100 \times 100 \mu m^2$, it takes a minimum of 1 s, which is by roughly three magnitudes of order more slowly than what is achievable with fast CCD systems.

3.3 Wide-Field Imaging

In the previous sections, SMD with point detectors (in most cases single-photo avalanche diodes) for detecting the fluorescence of single molecules was described. As already mentioned, this has the advantage to measure processes on practically arbitrary short time scales, however, for the prize of relatively slow imaging speed. The straightforward solution to this problem is to use a parallel detection scheme, namely a charge-coupled detector for imaging. The development of high-sensitive, low-noise CCDs has finally made imaging of single fluorescing molecules possible. A single fluorophore attached to a protein molecule in aqueous solution was first observed in 1995 by using total internal reflection fluorescence microscopy (TIRFM) and conventional epi-fluorescence microscopy [89]. That it is possible to image single dye molecules even with a conventional video camera was demonstrated by Adachi et al. [90], where individual Cy3 molecules were used to image actin against heavy meromyosin sliding.

Two types of cameras are most suitable for single-molecule imaging: intensified CCD cameras, and cooled back-illuminated CCD cameras. Intensified cameras use a solid-state multichannel plate (MCP) for amplifying the photo-electrons generated by photons on a photoelectric layer. However, they suffer from low spatial resolution (due to the coarse structure of MCPs compared with pixel size and packing on CCD chips) and high cost. Taking into account the recent emergence of affordable, high-sensitive and Peltier-cooled CCD-cameras on the market, the latter are much to be preferred for most single-molecule imaging applications. An exception may be applications where one desires to use the gating-capability of MCPs on nanosecond time-scales, allowing e.g. for time-resolved fluorescence imaging [91].

The basic schemes of single-molecule wide-field microscopy (smWFM) are shown in Fig.5. Two principally distinct excitation methods are widely used: excitation by total-internal reflection (so called total-internal reflection fluorescence microscopy – TIRF), and direct wide-field illumination. For TIRF illumination, the collimated laser beam is coupled into the microscope objective off-axis so that it is converted into plane waves hitting the sample surface above the angle of total internal reflection. For achieving that, one has to use objectives with a numerical aperture larger than the sample's index of refraction [92]. Paige et al. [81] compared both TIRFM and direct epi-fluorescence illumination for SMS applications, and Ref.[93] used alternate switching between both excitation modes for obtaining information about three-dimensional SM orientation by using the fact the that TIRF illumination produces a strong electric-field component perpendicular to the sample and, for a given in-plane angle o the molecular dipole, provides unambiguous orientation of the 3D molecular dipole moment. An important improvement in SM imaging was the ability to simultaneously record images at different wavelengths and polarizations, for dual-color and or polarization-resolved studies [94,95]. Ma et al. [96] imaged complete spectra of individual molecules by placing a grating in front of the CCD.

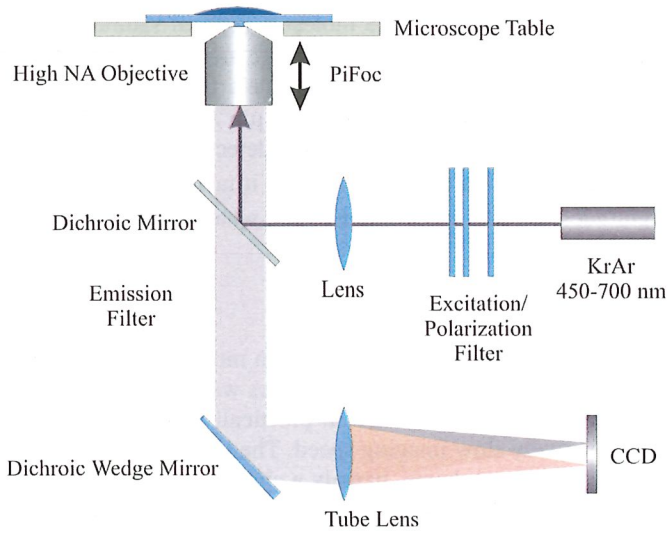


Fig. 5: *Schematic of a single-molecule sensitive wide-field microscope. The polarization and/or spectrally dividing optical element (a Wollaston prism and/or a wedged dichroic mirror as shown in the figure) splits the image into sub-images with different polarization and/or with different emission spectra, which are subsequently imaged at different positions on the same CCD chip. An essential element of the set-up is the high-sensitive CCD detector, in many cases a back-illuminated and cooled chip. In this way, multi-channel single-molecule imaging is realized; for details see e.g. Ref.[95].*

The advantage of smWFM is the high imaging speed: imaging rates of ca. one kHz can be achieved for small regions of interest. Thus, smWFM is the ideal tool where molecular motion is on the order of several μm/s. A further important feature of single molecule imaging is the possibility to localize a single molecule with an accuracy of several dozen nanometers, much below the resolution limit of the imaging optics. The principle behind is the knowledge

that a single molecule is acting as point emitter of light, generating a well defined diffraction pattern of light on the CCD. By fitting the maximum position of that pattern, the lateral position of the imaged molecules can be determined with ca. 40 nm resolution [97]. The position along the optical axis can be determined with similar accuracy: When the molecule moves away from the focal plane of the objective, its image broadens in proportion to its distance from the focal plane [98, 99]. Pioneering work on imaging the diffusion of single molecules within synthetic lipid membranes as well as within cell membranes was done by the group of Hansgeorg Schindler [100-102]. The method allowed the observation of anomalous diffusion and membrane reorganisation on a single molecule level. The basic limitation of this method is the finite number of photons that is emitted before irreversible photobleaching of the dye molecule, thus limiting the maximum length of the observable trajectory of a single molecule. A different application of the method was realized by the group of Toshio Yanagida: Single molecule imaging was used for following the motion and working of individual motor proteins and ATPase. The observation of multiply labelled sliding actin filaments stood at the beginning of this exciting development [103]. One of the most fascinating applications of the single-molecule imaging technique was the direct visualization of turnovers of single ATP molecules by ATPase [89,92,104]. Another success story was the direct imaging of the sliding motion of single kinesins along microtubules [105,106]. A powerful extension of the method is its combination with optical tweezers for performing simultaneous imaging and force measurements on a single molecule level [107].

The ultimate goal for biological applications of smWFM is the observation of single molecules within living cells. The main obstacle on the route to achieving this goal is the autofluorescence of cellular compartments. The fluorescence brightness of native biomolecules (mostly NADH and FAD) is usually very low, but it becomes a server problem at high concentrations of these molecules. There is no universal remedy to this problem, although using long-wavelength dyes together with optimized emission filters can be very helpful. There are several unique advantages that smWFM offers for single cell studies. Many cellular processes involve only small number of molecules – single cells can be as sensitive as to react to single ligand-receptor binding events. Furthermore, as already mentioned, single molecules can be localized within the cell with much higher precision than the optical resolution limit of the used optics. Last but not least, following cyclic processes on a single molecule level does not necessitate and external synchronization, as is necessary when observing on a large ensemble of molecules. A number of exciting studies has been publishing showing both the feasibility and the great potential of smWFM for *in vivo* of cells. Excellent reviews of this field can be found in Ref.[102] and [108].

Imaging techniques can be used to directly read out the three-dimensional orientation of the *emission* dipole of single molecules. This can be achieved by introducing optical aberrations when imaging, or by simply defocused imaging. Then, every molecule is imaged onto the CCD detector not as a sharp diffraction-limited image, but as a complicated interference pattern, containing information about the angular distribution of the imaged molecule. Aberrational imaging for determining molecular orientation was first demonstrated by Dickson et al. [109] and Bartko and Dickson [110,111]. A similar approach is to purposely defocus the images of the molecules [112].

4 Data Acquisition and Evaluation

There are two principally different approaches for evaluating SMS data. The first set of methods can be called *fluctuation analysis* and processes the data with no explicit recourse to the

single molecule nature of the experiment. These methods exploit the fluctuations of the measured light intensity that are caused by the fluctuating number of molecules within the detection volume. The most prominent representative of these methods is the so called *fluorescence correlation spectroscopy* (FCS), but recently new and promising techniques such as *fluorescence intensity distribution analysis* (FIDA) or *photon-counting histogramming* (PCH) have been developed. These methods will be discussed in the next section.

The second set of methods explicitly uses the fact that the measured data are mostly generated by single-molecule transits through the detection volume. Here, every single-molecule transit is identified and isolated within the continuous data stream, and every transit is then individually analyzed. These methods can be called *burst-by-burst* or *molecule-by-molecule analysis*.

4.1 Time-Tagged and Time-Correlated Photon Counting

Before discussing the above mentioned different techniques of data analysis, a few principal remarks are necessary concerning the acquisition of SMS raw data. Besides recording the temporal behaviour of the *fluorescence intensity*, a very important and useful measurement parameter is the *fluorescence lifetime*. In SMS experiments, fluorescence lifetime is exclusively measured by time-correlated single-photon counting (TCSPC)". As fluorescence correlation spectroscopy (see below), TCSPC is also a correlation technique, but on a completely different time scale. In TCSPC, fluorescence excitation is performed by a pulsed laser with sufficiently short pulse width (usually in the femtosecond through picosecond time range) and high pulse repetition rate (several MHz). For every detected photon, the time between the last exciting laser pulse and the photon arrival is measured. The time delays between laser pulses and photon detection events give direct information about the lifetime of the excited state of the molecule. For example, for a molecule following the simple photophysics depicted in Fig.1, histogramming all delay times results in an exponentially decaying curve with a decay constant equal to the lifetime of the excited state or fluorescence lifetime (for a detailed introduction to TCSPC see the classical book by O'Connor and Phillips [113]). While TCSPC measurements where still a technical challenge some 20 years ago, the last decade has seen a revolutionary development of affordable and easy-to-use pulsed diode lasers and the arrival of compact, self-contained high-speed electronics necessary for the timing of photon arrival times on a picosecond time scale.

The first successful TCSPC measurements on single molecules in solution were reported a decade ago [114,115]. In SMS applications of TCSPC, data acquisition usually proceeds in an asynchronous way: data are collected and stored only when a photon is detected, tagging every photon with its arrival time on a picosecond time scale (TCSPC time) and on a 'macroscopic' time scale with temporal resolution of ca. 100 ns [87]. This assures that data are generated only if photons are detected, preventing the generation of large data flows as in synchronous data acquisition (synchronous counting of detected photons within evenly spaced subsequent time intervals). Subsequently, any desirable analysis of the data can be performed, e.g. FCS, TCSPC histogramming etc.

As already mentioned, TCSPC addresses the fluorescence lifetime of the detected molecules, which may be of interest as a photophysical parameter by itself. Besides that, TCSPC measurements offer the additional possibility of gating off photons that arrive directly with the laser pulse (detected within a few 100 picoseconds due to the finite width of the instrumental response function), which may be an effective way of additionally discriminating Raman scattering (occurring directly after the exciting laser pulse) against the desired single-molecule fluorescence (arriving at longer time scales after the laser pulse) [116].

One of the most interesting applications of TCSPC is to use the fluorescence lifetime information for distinguishing between different species of molecules. This can be done in combination with a statistical analysis such as FCS [117], or on a molecule by molecule basis [118--120]. In this sense, TCSPC is an elegant alternative or addition to multi-colour detection measurements [121].

4.2 Fluorescence Correlation Spectroscopy

Fluorescence correlation spectroscopy is a relatively old technique, originally introduced by Elson, Magde and Webb in the early seventies [122-124]. However, it took nearly two decades until the development of new lasers with high beam quality and temporal stability, low-noise single-photon detectors, and high-quality microscope objectives with nearly perfect imaging quality at high numerical aperture, that the technique has seen a new renaissance in SMS. Achieving values of the detection volume within the range of a few μm^3 made the technique applicable for samples at reasonably high concentrations and short measurement times. An excellent and extensive description of this analysis method can be found in Ref.[125].

In FCS, the detected fluorescence intensity is correlated with a time-shifted replica of itself at different values of time shift. The result is the so-called autocorrelation function (second-order correlation function) which is calculated as

$$g(\tau) = \langle I(t) I(t+\tau) \rangle_t , \qquad (1)$$

where $I(t)$ is the fluorescence intensity at time t, and the triangular brackets denote averaging over all time values t. The physical meaning of the autocorrelation as defined above is that it is directly proportional to the probability to detect, on average, a photon at time τ if their was a detection event at time zero. This probability is composed of two basically different terms: The two photons detected at time zero and at time τ can originate from uncorrelated background or from different fluorescing molecules and therefore do not have any physical correlation (provided there is no interaction of the different fluorescing molecules). These events will contribute to a constant offset of $g(\tau)$ that is completely independent on τ. Alternatively, the two photons originate from one and the same molecule and are therefore physically correlated, leading to a time dependent component of $g(\tau)$. Thus, the temporal behaviour of the autocorrelation function is solely determined by the photophysics and dynamics of individual molecules. In this sense, FCS is a true SMS technique, although the analysis is not explicitly identifying single molecule detection events. However, although FCS lives on the small number fluctuations of molecules within the detection volume, any information about *individual* molecules is lost during the data analysis. The final result in FCS is the autocorrelation function $g(\tau)$, which is averaged over many single-molecule transits through the detection volume.

On different time scales, the temporal behaviour of the autocorrelation function is determined by different properties of the fluorescing molecules: On a nanosecond time-scale, photon antibunching can be observed, reflecting the fact that directly after the emission of a photon the molecule needs to get re-excited again for being able to emit the next photon, leading to a steep decrease of $g(\tau)$ towards short times. On a microsecond time scale, $g(\tau)$ is dominated by triplet state dynamics. If excitation and/or detection are performed with polarization filters, the autocorrelation will also show contributions from rotational diffusion dynamics of the molecules. On a millisecond to second level, the autocorrelation function shows a typical decay due to the lateral diffusion of the molecules out of the detection region. The diffusion co-

efficient is also the parameter most frequently addressed by FCS measurements. In summary, the autocorrelation function contains information about

- translational and rotational diffusion of the fluorescent molecules,
- their average number within the detection volume (thus concentration),
- triplet state dynamics,
- possible photo-isomerization kinetics,
- reaction kinetics between different fluorescing molecular species.

The advantage of FCS is the relative simplicity of the analysis. Its drawback is that it works only within a very limited concentration range: If the concentration of fluorescing molecules becomes too large (typically >> 10^{-8} M), then the contribution from correlated photons from individual molecules, scaling with the number N of molecules within the detection volume, becomes very small compared with the contribution by uncorrelated photons from different molecules, scaling with N^2. If the concentration is too low (typically < 10^{-13} M), then the probability to find a molecule within the detection region becomes extremely low. In both cases, the measurement time for obtaining a high-quality autocorrelation function gets prohibitively large. A remedy to that problem at very low concentrations is to rapidly scan the laser focus through the solution [126-128].

In contrast to single-channel FCS, where fluorescence is monitored at a single wavelength region only, multi-channel FCS using two or more detection channels at different wavelength regions and subsequently applies a cross-correlation analysis [129], or to use two-photon excitation at a single wavelength together with multi-colour detection allowing e.g. for monitoring the binding dynamics of different molecular species with different emission spectra [130], or for looking at the dissociation kinetics of bound molecules [131,132]. The two challenges of multicolor detection FCS are, firstly, to achieve an exact overlap of the detection volumes at different detection wavelengths, which may be difficult due to chromatic aberrations, and secondly the necessity to use either several lasers for efficiently exciting molecules with different emission wavelengths, or using a femtosecond pulsed laser for two-photon excitation of fluorescence, exploiting the broad two-photon absorption spectra of most molecules [133--135].

The technical challenge of multi-channel excitation/detection is to assure perfect overlap of the detection regions at different wavelengths, which is not a trivial task due to chromatic aberrations of most optical components. An alternative is to use the fluorescence lifetime instead of the fluorescence spectrum to distinguish between different molecules. By applying a sophisticated analysis that combines autocorrelation with fluorescence-lifetime measurements, a quasi-multichannel FCS without the need of multiple excitation and/or detection channels can be realized [117].

A critical aspect of FCS is that it heavily relies on the correctness of the assumed models used for data evaluation, particularly the made assumptions about molecular diffusion and the geometry of the detection region. The final result of an FCS experiment is the autocorrelation function that is condensed out of a large file of recorded fluorescence intensities. When applying a model to that autocorrelation function, it is often difficult to distinguish whether something like anomalous diffusion of the observed molecules has caused a deviation of the autocorrelation from the ideal model, or whether some technical artifact was responsible for that deviation. For example, Balakrishnan [136] showed the influence of spatial curvature effects on FCS of molecules diffusing within membranes, and Chirico et al. [137] modeled a similar

although small effect for dyes in solution, that was earlier observed by Osborne et al. [138]. Egner et al. [139] made a detailed study of the aberration effects introduced by even small refractive index mismatches, which may a problem for correct FCS data evaluation when working in biological samples, although Ganic et al. [140] could show that such aberrational effects may be reduced when using two-photon excitation. Fradin et al. [141] showed that FCS measurements inside cells can lead to erroneous values of the diffusion coefficient if the influence of membranes is not recognized, and Gennerich and Schild [142] demonstrated that FCS in small cytosolic compartments can lead to gross errors in diffusion coefficients, if confinement effects are not correctly taken into account.

FCS has been extensively used in numerous studies, only some of those can be cited here. FCS was used for studying diffusion of molecules in homogeneous and heterogeneous environment [143--145], intermolecular binding and reaction kinetics [146--150], single molecule photophysics [151--155], and conformational dynamics [156]. For recent reviews see Ref.[157,158] and the book [159].

4.3 Fluorescence Intensity Distribution Analysis and Related Techniques

A rather recent method of fluctuation analysis is the fluorescence intensity distribution analysis (FIDA) [160,161], or photon counting histogramming [162,163]. In FIDA, the detected photons are binned into time intervals of equal length (bin width usual in the range between 10 μs up to 1 ms), and the thus obtained stream of photons per time bin is subsequently histogrammed (frequency of occurrence versus number of photons per time bin). The shape of the histograms depend, in a complicated way, on the intensity of background, the fluorescence properties of the molecules, the molecule's concentration, and, at large values of time-bin width, their diffusion. The molecular properties that are addressed by FIDA are concentration and fluorescence brightness. The latter is defined as the product of absorption cross section and fluorescence quantum yield. In contrast to FCS, the exact analysis of the data, i.e. the connection between the shape of the experimentally obtained histograms and the fluorescence brightness is very complicated.

The striking advantage of this method, when compared with FCS, is that it addresses molecular brightness instead of molecular diffusion, a parameter that changes much more from molecular species to molecular species than the diffusion coefficient. Thus, it is the method of choice when analyzing mixtures of several different molecular species, as well as for studying e.g. dimerization processes. Moreover, molecular brightness is a more robust molecular parameter than the diffusion coefficient [32]. A potentially strong application of the FIDA or PCH method is the elucidation of the stoichiometry of macromolecular complexes or the number of binding sites on a macromolecule [164].

An extension of FIDA is fluorescence intensity multiple distribution analysis (FIMDA), where the data are histogrammed for different bin widths of the time intervals [165]. With increasing values of bin width, diffusion will have an increasing impact on the exact form of the histograms, so that diffusion coefficients can also be extracted from the analysis of the multiple histograms. As in the case of FCS, any combination of FIDA/FIMDA with multicolour detection and/or fluorescence lifetime measurements is feasible.

4.4 Molecule-by-Molecule Analysis

Until now, only correlation or statistical analysis techniques of SMD data were considered. FCS and FIDA perform a statistical analysis of photon detection intensities, whereas TCSPC

correlates photon arrival times with the pulses of a pulsed laser. A different approach to analyzing single molecule measurements is a molecule-by-molecule, or more correctly, a burst-by-burst analysis (a single photon burst may originate from more than only a single molecule present within the detection region). A first example of such a technique was already mentioned above in connection with using TCSPC information to distinguish single molecules on a molecule-by-molecule basis. This approach can be generalized to other photophysical parameters such as emission spectrum and polarization.

The starting point of all burst-by-burst methods is a suitable algorithm for identifying single bursts within the detected photon stream. One of the most powerful methods occurred to be the so called Lee filter as described in e.g. Ref.[166]. It identifies contiguous temporal regions where the photon detection intensities exceed a preset threshold. The algorithm takes into account that, due to shot noise in the photon detection and due to the stochastic character of diffusing molecule's transit through the detection region, detection intensities may intermittingly fall below the threshold during a single molecule' transit.

After having identified single molecule events, the photons belonging to each event can be analyzed according to their TCSPC time [120,121,167-169] polarization [170], and colour [121,169], in correspondence with parameters that were recorded by the measurement system. Finally, distributions of any desired parameters are obtained by histogramming the determined parameter values over all detected single molecule events. There are two fundamental and important differences of such burst-by-burst analysis when compared with a bulk analysis of the data (taking the information from all detected photons together without identifying single molecule events). First, cutting out single molecule events eliminates a large amount of background (low intensity scattering, detector and electronic noise) from the analysis, thus boosting the effective signal-to-background ratio of the analysis. The second difference becomes important when performing measurements on a mixture of different molecular species: In a bulk analysis, it is e.g. impossible to decide whether there is a single molecular species having a bi-modal value distribution of some parameter (e.g. a two-exponential fluorescence decay), or whether there are two different molecular species with distinct average values of the recorded parameter (e.g. two species with whether having a mono-exponential fluorescence decay time).

The technique of burst-by-burst analysis was perfected and pushed to its limit in a series of papers by the group of Claus Seidel [171-174] employing a specific burst analysis which they denote as multi-parameter fluorescence detection (MFD). While keeping the experimental advantage of monitoring single molecules diffusing through the microscopic open volume element of a confocal epi-illuminated set-up as in FCS, MFD uses pulsed excitation and time-correlated single-photon counting to simultaneously monitor the evolution of four fluorescence parameters: intensity, lifetime, anisotropy, and spectral signature. In an example that pushed the potential of the method to its limits, it was shown that one is able to distinguish between up to 16 different molecular species within one and the same sample, a task impossible to achieve by bulk analysis [175]. The method was applied for analyzing conformational dynamics in DNA [171] and ATPase [176].

Acknowledgements

I thank my colleagues Jörg Fitter, Thomas Gensch, Ingo Gregor, Benjamin Kaupp amd Digambara Patra as well as Markus Sauer from the University of Bielefeld and Johan Hofkens from the Katholeike Universiteit Leuven for many helpful discussions and suggestions. I am

grateful to the Deutsche Forschungsgemeinschaft and the Volkswagenstiftung for their financial support of my work.

References

[1] B. Rotman, Proc. Nat. Acad. Sci. USA 47 (1961) 1981-91.
[2] B. Rotman, in: Fluorescence Techniques in Cell Biology; Eds.: A.A. Thaer, M. Sernetz (Springer, Berlin, 1973) 333-7.
[3] T. Hirschfeld, Appl. Opt. 15 (1976) 2965-6.
[4] T. Hirschfeld, Appl. Opt. 15 (1976) 3135-8.
[5] N.J. Dovichi, J.C. Martin, J.H. Jett, R.A. Keller, Science 219 (1983) 845-7.
[6] E.B. Shera, N.K. Seitzinger, L.M. Davis, R.A. Keller, S.A. Soper, Chem. Phys. Lett. 174 (1990) 553-7.
[7] A.J. Meixner, Adv. Photochem. 24 (1998) 1-59.
[8] X.S. Xie, J.K. Trautman, Annu. Rev. Phys. Chem. 49 (1998) 441-80.
[9] J. Enderlein, W.P. Ambrose, P.M. Goodwin, R.A. Keller, in: Microsystem technology: A powerful tool for biomolecular studies; Eds.: M. Köhler, T. Mejevaia, H.P. Saluz (Birkhäuser, Basel, 1999) pp. 311-29.
[10] S. Weiss, Science 283 (1999) 1676-83.
[11] W.P. Ambrose, P.M. Goodwin, J.H. Jett, Orden A. Van, J.H. Werner, R.A. Keller, Chem. Rev. 99 (1999) 2929-56.
[12] Y. Ishii, T. Yanagida, Single Molecules 1 (2000) 5-14.
[13] R.A. Keller, S.R. Emory, W.P. Ambrose, P.M. Goodwin, A.A. Arias, J.J. Jett, H. Cai, Anal. Chem. 74 (2002) 316-24A.
[14] R. Rigler, M. Orrit, T. Basché (Eds.) Single Molecule Spectroscopy (Springer, Berlin, 2001).
[15] C. Zander, J. Enderlein, R.A. Keller (Eds.) Single molecule detection in solution (Wiley-VCH, Berlin, 2002).
[16] M. Orrit, J. Bernard, R. Brown, B. Lounis, in: Progress in Optics, vol. XXXV; Ed.: E. Wolf (Elsevier, Amsterdam, 1996) pp. 63-144.
[17] J.K. Trautman, J.J. Macklin, L.E. Brus, E. Betzig, Nature 369 (1994) 40-42.
[18] J.K. Trautman, J.J. Macklin, Chem. Phys. 205 (1996) 221-229.
[19] J.K. Trautman, W.P. Ambrose, in: Single Molecule Optical detection, Imaging and Spectroscopy; Eds.: T. Basché, W.E. Moerner, M. Orrit, U.P. Wild (VCH, Weinheim, 1997) pp. 191-222.
[20] A.G.T. Ruiter, J.A. Veerman, M.F. Garcia-Parajo, N.F. van Hulst, J. Phys. Chem. A 101 (1997) 7318-23.
[21] M.F. Garcia-Parajo, J.A. Veerman, S.J.T. van Noort, B.G. de Grooth, J. Greve, N.F. van Hulst, Bioimaging 6 (1998) 43-53.
[22] M.F. Garcia-Parajo, J.A. Veerman, G.M.J. Segers-Nolten, B.G. de Grooth, J. Greve, N.F. van Hulst, Cytometry 36 (1999) 239-46.
[23] H. Gersen, M.F. Garcia-Parajo, L. Novotny, J.A. Veerman, L. Kuipers, N.F. van Hulst, Phys. Rev. Lett. 85 (2000) 5312-5.
[24] H. Gersen, M.F. Garcia-Parajo, L. Novotny, J.A. Veerman, L. Kuipers, N.F. van Hulst, J. Microsc. 202 (2001) 374-8.
[25] N.F. van Hulst, J.A. Veerman, M.F. García-Parajó, L. Kuipers, J. Chem. Phys. 112 (2000) 7790-810.
[26] W. Göhde, U.C. Fischer, H. Fuchs, J. Tittel, T. Basché, C. Bräuchle, A. Herrmann, K. Müllen, J. Phys. Chem. A 102 (1998) 9109-16.
[27] O. Panzer, W. Göhde, U.C. Fischer, H. Fuchs, K. Müllen, Adv. Mat. 10 (1998) 1469-72.
[28] K.D. Weston, S.K. Buratto, J. Phys. Chem. A 102 (1998) 3635-8.
[29] M.F. Garcia-Parajo, G.M.J. Segers-Nolten, J.A. Veerman, J. Greve, Hulst N.F. Van,Proc. Nat. Acad. Sci. USA 97 (2000) 7237-42.

[30] R.M. Dickson, A.B. Cubitt, R.Y. Tsien, W.E. Moerner, Nature 388 (1997) 355-358.
[31] W.E. Moerner, J. Chem. Phys. 117 (2002) 10925-37.
[32] Y. Chen, J.D. Müller, Ruan Q.Q., E. Gratton, Biophys. J. 82 (2002) 133-44.
[33] U. Kubitscheck, O. Kuckmann, T. Kues, R. Peters, Biophys. J. 78 (2000) 2170-9.
[34] W.C.W. Chan, S. Nie, Science 281 (1998) 2016-9.
[35] M. Han, X. Gao, J.Z. Su, S. Nie, Nature Biotech. 19 (2001) 631-5.
[36] M. Dahan, T. Laurence, F. Pinaud, D.S. Chemla, A.P. Alivisatos, M. Sauer, S. Weiss, Opt. Lett. 26 (2001) 825-7.
[37] M. Dahan, S. Levi, C. Luccardini, P. Rostaing, B. Riveau, A. Triller, Single Molecules 2 (2001) 255-60.
[38] T.H. Lee, J.I. Gonzalez, R.M. Dickson, Proc. Nat. Acad. Sci. USA 99 (2002) 10272-5.
[39] L.A. Peyser, A.E. Vinson, A.P. Bartko, R.M. Dickson, Science 291 (2001) 103-6.
[40] D. Boyer, P. Tamarat, A. Maali, M. Orrit, B. Lounis, Physica. E. 17 (2003) 537-40.
[41] A: Maali, T. Cardinalb, M. Treguer-Delapierre, Physica E. 17 (2003) 559-60."
[42] L. Cognet, C. Tardin, D. Boyer, D. Choquet, P. Tamarat, B. Lounis, "Single metallic nanoparticle imaging for protein detection in cells" Proc. Nat. Acad. Sci. USA 100 (2003) 11350-5.
[43] S.A. Soper, M. Wabuyele, C.V. Owens, R.P. Hammer, in: Single-Molecule Detection in Solution - Methods and Applications; Eds: C. Zander, J. Enderlein, R.A. Keller (Wiley-VCH, Berlin, 2002) pp. 323-62.
[44] J. Arden-Jacob, J. Frantzeskos, N.U. Kemnitzer, A. Zilles, K.H. Drexhage, Spectrochimica Acta A 57 (2001) 2271-83.
[45] J.R. Lakowicz, Anal. Biochem. 298 (2001) 1-24.
[46] J. Enderlein, Appl. Phys. Lett. 80 (2002) 315-7.
[47] J. Enderlein, Phys. Chem. Chem. Phys. 4 (2002) 2780-6.
[48] V. Kachel, H. Fellner-Feldegg, E. Menke, in: Flow cytometry and sorting; Eds.: M.R. Melamed, T. Lindmo, M.L. Mendelsohn (Wiley-Liss, New York, 1990) pp. 27-44.
[49] A. van Orden, N.P. Machara, P.M. Goodwin, R.A. Keller, Anal. Chem. 70 (1998) 1444-51.
[50] W.P. Ambrose, H. Cai, P.M. Goodwin, J.H. Jett, R.C. Habbersett, E.J. Larson, W.K. Grace, J.H. Werner, B.L. Marrone, R.A. Keller, in: Topics in Fluorescence, vol. 7, Ed.: J.R. Lakowicz (Plenum Press, New York, 2000).
[51] E.J. Larson, J.R. Hakovirta, H. Cai, J.H. Jett, S. Burde, R.A. Keller, B.L. Marrone, Cytometry 41 (2000) 203-8.
[52] H. Cai, P.S. White, D. Torney, A. Deshpande, Z. Wang, B. Marrone, J.P. Nolan, Genomics 66 (2000) 135-43.
[53] Y.H. Lee, R.G. Maus, B.W. Smith, J.D. Winefordner, Anal. Chem. 66 (1994) 4142-9.
[54] R.D. Guenard, L.A. King, B.W. Smith, J.D. Winefordner, Anal. Chem. 69 (1997) 2426-2433.
[55] C. Zander, K.H. Drexhage, K.T. Han, J. Wolfrum, M. Sauer, Chem. Phys. Lett. 286 (1998) 457-65.
[56] M. Neumann, W. Ehrfeld, K.T. Han, P. Jacob, R. Konrad, S. Siebert, J. Wolfrum, M. Sauer, Biomed. Chromatography 13 (1999) 109-10.
[57] M. Sauer, B. Angerer, K.-T. Han, C. Zander, Phys. Chem. Chem. Phys. 1 (1999) 2471-7.
[58] W. Becker, H. Hickl, C. Zander, K.H. Drexhage, M. Sauer, S. Siebert, J. Wolfrum, Rev. Sci. Instrum. 70 (1999) 1835-41.
[59] J.C. Fister, S.C. Jacobson, L.M. Davis, J.M. Ramsey, Anal. Chem. 70 (1998) 431-437.
[60] C.S. Effenhauser, G.J.M. Bruin, A. Paulus, M. Ehrat, Anal. Chem. 69 (1997) 3451-3457.
[61] H.P. Chou, C. Spence, A. Scherer, S. Quake, Proc. Nat. Acad. Sci. USA 96 (1999) 11-13.
[62] M. Foquet, J. Korlach, W. Zipfel, W.W. Webb, H.G. Craighead, Anal. Chem. 74 (2002)

1415-22.
[63] D.Y. Chen, K. Adelhelm, X.L. Cheng, N.J. Dovichi, Analyst 119 (1994) 349-352.
[64] A. Castro, E.B. Shera, Anal. Chem. 67 (1995) 3181-6.
[65] A. Castro, E.B. Shera, Appl. Opt. 34 (1995) 3218-22.
[66] A. Castro, J.G.K. Williams, Anal. Chem. 69 (1997) 3915-3920.
[67] B.B. Haab, R.A. Mathies, Anal. Chem. 67 (1995) 3253-60.
[68] B.B. Haab, R.A. Mathies, Appl. Spetrosc. 51 (1997) 1579-84.
[69] M.R. Shortreed, H. Li, W.H. Huang, E.S. Yeung, Anal. Chem. 72 (2000) 2879-85.
[70] A. Castro, in: Single-Molecule Detection in Solution - Methods and Applications; Eds: C. Zander, J. Enderlein, R.A. Keller (Wiley-VCH, Berlin, 2002) pp. 303-21.
[71] W.B. Whitten, J.M., Arnold S. Ramsey, B.V. Bronk, Anal. Chem. 63 (1991) 1027-31.
[72] M.D. Barnes, K.C. Ng, W.B. Whitten, J.M. Ramsey, Anal. Chem. 65 (1993) 2360-5.
[73] M.D. Barnes, N. Lermer, C.Y. Kung, W.B. Whitten, J.M. Ramsey, S.C. Hill, Opt. Lett. 22 (1997) 1265-7.
[74] N. Lermer, M.D. Barnes, C.Y. Kung, W.B. Whitten, J.M. Ramsey, Anal. Chem. 69 (1997) 2115-2121.
[75] C.Y. Kung, M.D. Barnes, N. Lermer, W.B. Whitten, J.M. Ramsey, Anal. Chem. 70 (1998) 658-661.
[76] M.D. Barnes, C.-Y. Kung, W.B. Whitten, J.M. Ramsey, S. Arnold, S. Holler, Phys. Rev. Lett. 76 (1996) 3931-4.
[77] C.Y. Kung, M.D. Barnes, N. Lermer, W.B. Whitten, J.M. Ramsey, Appl. Opt. 38 (1999) 1481-7.
[78] R.L. Hansen, J.M. Harris, Anal. Chem. 70 (1998) 4247-56.
[79] R.L. Hansen, J.M. Harris, Anal. Chem. 70 (1998) 2565-75.
[80] M.F. Paige, E.J. Bierneld, W.E. Moerner, Single Molecules 2 (2001) 191-201.
[81] W.P. Ambrose, P.M. Goodwin, J.P. Nolan, Cytometry 36 (1999) 224-31.
[82] T. Ruckstuhl, J. Enderlein, S. Jung, S. Seeger, Anal. Chem. 72 (2000) 2117-23.
[83] T. Ruckstuhl, M. Rankl, S. Seeger, Biosens. Bioelectron. 18 (2003) 1193-9.
[84] A. Krieg, S. Laib, T. Ruckstuhl, S. Seeger, Chembiochem. 4 (2003) 589-92.
[85] T. Ruckstuhl, S. Seeger, Rev. Sci. Instrum. 72 (2001) 4145-52.
[86] M. Böhmer, J. Enderlein, in: Single-Molecule Detection in Solution - Methods and Applications; Eds: C. Zander, J. Enderlein, R.A. Keller (Wiley-VCH, Berlin, 2002) pp. 145-83.
[87] T. Funatsu, Y. Harada, M. Tokunaga, K. Saito, T. Yanagida, Nature 374 (1995) 555-559.
[88] K. Adachi, K. Kinosita, T. Ando, J. Microsc. 195 (1999) 125-32.
[89] M. Straub, S.W. Hell, Appl. Phys. Lett. 73 (1998) 1769-71.
[90] M. Tokunaga, K. Kitamura, K. Saito, A.H. Iwane, T. Yanagida, Biochem. Biophys. Res. Commun. 235 (1997) 47-53.
[91] M. Vacha, M. Kotani, J. Chem. Phys. 118 (2003) 5279-82.
[92] M. Mörtelmaier, E.J. Kögler, J. Hesse, M. Sonnleitner, L.A. Huber, G.J. Schütz, Single Molecules 3 (2002) 225-31.
[93] L. Cognet, G.S. Harms, G.A. Blab, P.H.M. Lommerse, T. Schmidt, Appl. Phys. Lett. 77 (2000) 4052-4.
[94] Y. Ma, M.R. Shortreed, E.S. Yeung, Anal. Chem. 72 (2000) 4640-5.
[95] G.J. Schütz, W. Trabesinger, T. Schmidt, Biophys. J. 74 (1998) 2223-6.
[96] G.J. Schütz, J. Hesse, G. Freudenthaler, V.P. Pastushenkoo, H.G. Knaus, B. Pragl, H. Schindler, Single Molecules 1 (2000) 153-7.
[97] G.J. Schütz, M. Axmann, H. Schindler, Single Molecules 2 (2001) 69-74.
[98] G.J. Schütz, H. Schindler, T. Schmidt, Biophys. J. 73 (1997) 1073-80.
[99] G.J. Schütz, G. Kada, V.P. Pastushenko, H.G. Schindler, EMBO J. 19 (2000) 892-901.

[100] G.J. Schütz, H. Schindler, in: Single-Molecule Detection in Solution - Methods and Applications; Eds: C. Zander, J. Enderlein, R.A. Keller (Wiley-VCH, Berlin, 2002) pp. 231-45.
[101] T. Yanagida, M. Nakase, K. Nishiyama, F. Oosawa, Nature 307 (1984) 58-60.
[102] K. Oiwa, J.F. Eccleston, M. Anson, M. Kikumoto, C.T. Davis, G.P. Reid, M.A. Ferenczi, J.E.T. Corrie, A. Yamada, H. Nakayama, D.R. Trentham, Biophys. J. 78 (2000) 3048-71.
[103] Y. Okada, N. Hirokawa, Science 283 (1999) 1152-7.
[104] Y. Inoue, A.H. Iwane, T. Miyai, E. Muto, T. Yanagida, Biophys. J. 81 (2001) 2838-50.
[105] Y. Ishii, A.H. Iwane, H. Yokota, Y. Inoue, T. Wazawa, M. Nishiyama, H. Tanaka, K. Kitamura, T. Yanagida, in: Single-Molecule Detection in Solution - Methods and Applications; Eds: C. Zander, J. Enderlein, R.A. Keller (Wiley-VCH, Berlin, 2002) pp. 273-92.
[106] U. Kubitschek, Single Molecules 3 (2002) 267-74.
[107] R.M. Dickson, D.J. Norris, W.E. Moerner, Phys. Rev. Lett. 81 (1998) 5322-5.
[108] A.P. Bartko, R.M. Dickson, J. Phys. Chem. B 103 (1999) 11237-41.
[109] A.P. Bartko, K. Xu, R.M. Dickson, Phys. Rev. Lett. 89 (2002) 026101.
[110] M. Böhmer, J. Enderlein, J. Opt. Soc. B 20 (2003) 554-9.
[111] D.V. O'Connor, D. Phillips, (Academic Press, London, 1984).
[112] S.A. Soper, L.M. Davis, E.B. Shera, J. Opt. Soc. Am. B 9 (1992) 1761-9.
[113] C.W. Wilkerson, P.M. Goodwin, W.P. Ambrose, J.C. Martin, R.A. Keller, Appl. Phys. Lett. 62 (1993) 2030-2.
[114] L.-Q. Li, L.M. Davis, Rev. Sci. Instr. 64 (1993) 1524-9.
[115] M. Böhmer, M. Wahl, H.J. Rahn, R. Erdmann, J. Enderlein, Chem. Phys. Lett. 353 (2002) 439-45.
[116] M. Köllner, P. Fischer, J. Arden-Jacob, K.-H. Drexhage, R. Müller, S. Seeger, J. Wolfrum, Chem. Phys. Lett. 250 (1996) 355-60.
[117] J. Enderlein, P.M. Goodwin, A. van Orden, W.P. Ambrose, R. Erdmann, R.A. Keller, Chem. Phys. Lett. 270 (1997) 464-70.
[118] J. Enderlein, M. Sauer, J. Phys. Chem. A 105 (2001) 48-53.
[119] D.P. Herten, P. Tinnefeld, M. Sauer, Appl. Phys. B 71 (2000) 765-71.
[120] D. Magde, E. Elson, W.W. Webb, Phys. Rev. Lett. 29 (1972) 705-8.
[121] E.L. Elson, D. Magde, Bioploymers 13 (1974) 1-27.
[122] D. Magde, E. Elson, W.W. Webb, Biopolymers 13 (1974) 29-61.
[123] J. Widengren, Ü. Mets, in: Single-Molecule Detection in Solution - Methods and Applications; Eds: C. Zander, J. Enderlein, R.A. Keller (Wiley-VCH, Berlin, 2002) pp. 69-95.
[124] N.O. Petersen, Biophys. J. 49 (1986) 809-15.
[125] N.O. Petersen, D.C. Johnson, M.J. Schlesinger, Biophys. J. 49 (1986) 817-20.
[126] A. Amediek, E. Haustein, D. Scherfeld, P. Schwille, Single Molecules 3 (2002) 201-10.
[127] P. Schwille, F.J. Meyer-Almes, R. Rigler, Biophys. J. 72 (1997) 1878-86.
[128] K. Bacia, I.V. Majoul, P. Schwille, Biophys. J. 83 (2002) 1184-93.
[129] U. Kettling, A. Koltermann, P. Schwille, M. Eigen, Proc. Nat. Acad. Sci. USA 95 (1998) 1416-20.
[130] A. Koltermann, U. Kettling, J. Bieschke, T. Winkler, M. Eigen, Proc. Nat. Acad. Sci. USA 95 (1998) 1421-6.
[131] K. Heinze, A. Koltermann, P. Schwille, Proc. Nat. Acad. Sci. USA 97 (2000) 10377-82.
[132] M. Rarbach, U. Kettling, A. Koltermann, M. Eigen, Methods 24 (2001) 104-16.
[133] K.G. Heinze, M. Rarbach, M. Jahnz, P. Schwille, Biophys. J. 83 (2002) 1671-81.
[134] J. Balakrishnan, Phys. Rev. E 61 (2000) 4648-51.
[135] G. Chirico, C. Fumagalli, G. Baldini, J. Phys. Chem. B 106 (2002) 2508-19.
[136] M.A. Osborne, S. Balasubramanian, W.S. Furey, D. Klenerman, J. Phys. B 102 (1998)

3160-67.
[137] A. Egner, M. Schrader, S.W. Hell, Opt. Commun. 153 (1998) 211-7.
[138] D. Ganic, X. Gan, M. Gu, Appl. Opt. 39 (2000) 3945-7.
[139] C. Fradin, A. Abu-Arish, R. Granek, M. Elbaum, Biophys. J. 84 (2003) 2005-20.
[140] A. Gennerich, D. Schild, Biophys. J. 79 (2000) 3294-306.
[141] J. Korlach, P. Schwille, W.W. Webb, G.W. Feigenson, Proc. Nat. Acad. Sci. USA 69 (1999) 8461-6.
[142] C. Gell, D.J. Brockwell, G.S. Beddard, S.E. Radford, A.P. Kalverda, D.A. Smith, Single Molecules 2 (2001) 177-81.
[143] N. Yoshida, M. Tamura, M. Kinjo, Single Molecules 1 (2000) 279-83.
[144] J. Widengren, R. Rigler, Cellular Mol. Biol. 44 (1998) 857-79.
[145] S. Björling, M. Kinjo, Z. Földes-Papp, E. Hagman, P. Thyberg, R. Rigler, Biochem. 37 (1998) 12971-8.
[146] K. Häsler, O. Pänke, W. Junge, Biochem. 38 (1999) 13759-65.
[147] P. Schwille, J. Bieschke, F. Oehlenschlager, Biophys. Chem. 66 (1997) 211-228.
[148] T. Wohland, K. Friedrich, R. Hovius, H. Vogel, Biochem. 38 (1999) 8671-81.
[149] J. Widengren, Ü. Mets, R. Rigler, Chem. Phys. 250 (1999) 171-86.
[150] U. Haupts, S. Maiti, P. Schwille, W.W. Webb, Proc. Nat. Acad. Sci. USA 95 (1998) 13573-8.
[151] A.A. Heikal, S.T. Hess, G.S. Baird, R.Y. Tsien, W.W. Webb, Proc. Nat. Acad. Sci. USA 97 (2000) 11996-12001.
[152] R. Brock, G. Vàmosi, G. Vereb, T.M. Jovin, Proc. Nat. Acad. Sci. USA 96 (1999) 10123-8.
[153] A.A. Heikal, S.T. Hess, W.W. Webb, Chem. Phys. 274 (2001) 37-55.
[154] P. Schwille, S. Kummer, A.A Heikal, W.E. Moerner, W.W. Webb, Proc. Nat. Acad. Sci. USA 97 (2000) 151-6.
[155] J. Widengren, C.A.M Seidel, Phys. Chem. Chem. Phys. 2 (2000) 3435-41.
[156] S. Huang, A.A. Heikal, W.W. Webb, Biophys. J. 82 (2002) 2811-25.
[157] F. Malvezzi-Campeggi, M. Jahnz, K.G. Heinze, P. Dittrich, P. Schwille, Biophys. J. 81 (2001) 1776-85
[158] J. Widengren, P. Schwille, J. Phys. Chem. A 104 (2000) 6416-28.
[159] P. Schwille, Cell. Biochem. Biophys. 34 (2001) 383-408.
[160] S.T. Hess, S. Huang, A.A. Heikal, W.W. Webb, Biochem. 41 (2002) 697-705.
[161] R. Rigler, E. Elson (Eds.) Fluorescence Correlation Spectroscopy (Springer, Berlin, 2001).
[162] P. Kask, K. Palo, D. Ullmann, K. Gall, Proc. Nat. Acad. Sci. USA 96 (1999) 13756-61.
[163] P. Kask, K. Palo, N. Fay, L. Brand, U. Mets, D. Ullmann, J. Jungmann, J. Pschorr, K. Gall, Biophys. J. 78 (2000) 1703-13.
[164] Y. Chen, J.D. Müller, P.T.C. So, E. Gratton, Biophys. J. 77 (1999) 553-67.
[165] Y. Chen, J.D. Müller, K.M. Berland, E. Gratton, Methods: A Companion to Methods in Enzymology 19 (1999) 234-52.
[166] Y. Chen, J.D. Müller, S.Y. Tetin, J.D. Tyner, E. Gratton, Biophys. J. 79 (2000) 1074-84.
[167] K. Palo, U. Mets, S. Jager, P. Kask, K. Gall, Biophys. J. 79 (2000) 2858-66.
[168] J. Enderlein, D.L. Robbins, W.P. Ambrose, R.A. Keller, J. Phys. Chem. A 102 (1998) 6089-94.
[169] J.R. Fries, L. Brand, C. Eggeling, M. Köllner, C.A.M. Seidel, J. Phys. Chem. A 102 (1998) 6601-13.
[170] M. Maus, M. Cotlet, J. Hofkens, T. Gensch, Schryver F.C. De, J. Schaffer, C.A.M. Seidel, Anal. Chem. 73 (2001) 2078-86.
[171] J.-P. Knemeyer, D.-P. Herten, M. Sauer, Anal. Chem. 75 (2003) 2147-53.

[172] M. Dahan, A.A. Deniz, T. Ha, D.S. Chemla, Chem. Phys. 247 (1999) 85-106.
[173] C. Eggeling, J.R. Fries, L. Brand, R. Günther, C.A.M. Seidel, Proc. Nat. Acad. Sci. USA 95 (1998) 1556-61.
[174] C. Eggeling, J. Schaffer, C.A.M. Seidel, J. Korte, G. Brehm, S. Schneider, W. Schrof, J. Phys. Chem. A 105 (2001) 3673-9.
[175] J. Schaffer, A. Volkmer, C. Eggeling, V. Subramaniam, G. Striker, C.A.M. Seidel, Phys. Chem. A 103 (1999) 331-6.
[176] R. Kühnemuth, C.A.M. Seidel, Single Molecules 2 (2001) 251-4.
[177] C. Seidel, personal communication.
[178] M. Börsch, P: Turina, C. Eggeling, J.R. Fries, C.A.M. Seidel, A. Labahn, P. Gräber, FEBS Lett. 437 (1998) 251-4.

B7 Modern Applications of Fluorescence Methods in Biology

T. Gensch
Institut für Biologische Informationsverarbeitung 1
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Structure and Properties of Biologically Relevant Fluorescent Molecules **3**
2.1 Intrinsic Fluorophores 3
2.2 Fluorescent Dye Molecules 4
2.3 Visible Fluorescent Proteins 7

3 Examples **10**
3.1 Tryptophan Fluorescence of PYP 10
3.2 Intracellular Calcium Measurements 12
3.3 Intracellular Chloride Measurements 13
3.4 GFP Fusion Proteins: Protein Localisation 16
3.5 GFP Fusion Proteins: Protein-Protein Interaction 17

1 Introduction

The interaction of electromagnetic radiation with matter (atoms, organic molecules, proteins, solid materials, etc.) is the physical basis of many methods that characterise the matter properties (e.g. spectroscopy), determine its structure (e.g. scattering, diffraction) and image the objects on the nm to µm scale (e.g. microscopy). One of the most important processes is the absorption of photons by molecules impinging on the sample. For the absorption the energy of the photon has to be in resonance with an allowed electronic transition. The electron returns from the excited state orbital via non-radiative relaxations (internal conversions) accompanied by a release of heat or by emitting a photon. The latter process is called luminescence. In most cases, the excited state will have singlet state character, i.e. the electron in the excited state orbital is paired (with opposite spin) with the second electron in the ground-state orbital and the luminescence is called fluorescence. Relaxation to the ground state is then spin-allowed and occurs with rates of 10^9 – 10^8 s^{-1}. The reciprocal, the fluorescence lifetime, is therefore in the range of a few ns (see also chapter B6 by J. Enderlein, in particular Fig.1).

The return of the electron from the excited singlet state to the ground state can take place also via deactivation pathways different from internal conversion and fluorescence, e.g. formation of metastable radicals and charge-transfer states, permanent photoreactions, and intersystem crossing. Intersystem crossing transfers the electron from an excited singlet state orbital into an excited triplet state orbital accompanied by a spin conversion. In the triplet state the molecule has two electrons with unpaired (or parallel) spins. Intersystem crossing is a spin-forbidden process which becomes partially allowed in molecules consisting of many atoms. The return of the electron from the excited triplet state orbital to the ground state orbital is also (partially) spin-forbidden. Therefore the rates are much smaller (10^5 – 10^0 s^{-1}) and the lifetimes are much longer compared to the return of the electron from the excited singlet state. Deactivation often occurs via heat release but can also take place by luminescence. This then is called phosphorescence in contrast to fluorescence. The phosphorescence intensity is typically weak. The phosphorescence quantum yields, i.e. the probability for emitting a photon after an absorption process, are in the range of 10^{-5} – 10^{-3}. However, phosphorescence has found some application in biology and in cellular imaging [1]. The long lifetime of the phosphorescence allows using a time-gated detection system, thereby eliminating the autofluorescence (see below) of proteins, cells or biological tissue. Phosphorescence measurements are used to detect singlet molecular oxygen (an important species in phototoxic processes) [2] and the (internal) dynamics of membrane proteins that may report about the metabolic status of the cell [3]. The triplet state of a molecule is also important for the fluorescence process. The intersystem crossing competes with the emission of fluorescence photons reducing the fluorescence quantum yield. Even more important, as long as the molecule persists in the triplet state, it can not be excited again, i.e. it can not fluoresce. Due to the long persistent times (µs to s), triplet-state formation reduces the cycling time and, therefore, the relative brightness of a fluorescent molecule in imaging using high excitation light intensities.

Fluorescence is a photophysical process which occurs in almost all molecules after absorption of a photon, but with very different probability. For biologically relevant molecules, however, it is not a common photophysical property. Fluorescence quantum yields (Φ_{fl}) are typically in the range of 0.001 – 0.05. For fluorophores (molecules with high absorption and emission probability) specially developed for fluorescence techniques the quantum yields are in the range of 0.1 to 0.5 and can reach even the theoretical maximum of 1, i.e. one emitted photon per 1 absorbed photon.

2 Structure and Properties of Biologically Relevant Fluorescent Molecules

In this chapter the structure and photophysical properties of fluorescent molecules will be introduced that are important for biological research. I will describe several methods used to determine protein conformational changes and protein-protein interactions by means of fluorescent molecules in whole tissue, inside cells, and *in vitro* studies.

2.1 Intrinsic Fluorophores

The first class of fluorophores used in biological fluorescence applications were the intrinsic fluorophores. From the naturally occurring amino acids, only the aromatic amino acids, tryptophan (Trp), tyrosine (Tyr) and phenylalanine (Phe) show a significant fluorescence which can be excited in the near UV spectral region (Fig.'s 1A and 1B). The fluorescence of phenylalanine (λ_{exc} = 260 nm, λ_{em} = 282 nm, Φ_{fl} = 0.02) and tyrosine (λ_{exc} = 275 nm, λ_{em} = 304 nm, Φ_{fl} = 0.14) is scarcely exploited due to the low absorption and fluorescence probability (Phe) and the low tyrosine fluorescence from proteins. This is due to the rather efficient fluorescence resonance energy transfer (FRET) from tyrosine to tryptophan residues and the fact that only a few proteins contain tyrosine but no tryptophan residues. As a consequence excitation of a protein solution in the near-UV almost always causes fluorescence from the tryptophan residues (λ_{exc} = 270-282 nm, λ_{em} = 353 nm, Φ_{fl} = 0.13). There currently exists no advanced fluorescence imaging technique based on protein fluorescence due to the large number of different proteins in a cell or a biological tissue and the need of near-UV excitation light. Protein fluorescence is used when changes in protein conformation, folding and unfolding events and protein-protein interactions are examined. There is a long history of this type of *in vitro* studies and the measurement of protein fluorescence has become a standard laboratory tool [4-6].

Why is tryptophan fluorescence such a powerful tool? An explanation may be drawn from the complex structure of tryptophan excited states. After absorption of a near-UV photon, two different excited singlet states can be populated by the electron. The two excited states have very different absorption and emission characteristics (Fig.1C). The electron switches between these two states while it is in the excited state. The ^1L_a state depends on the solvation of the tryptophan. Polar solvents form hydrogen bonds with the nitrogen atom of the tryptophan side chain. The stronger the hydrogen bonding, the lower is the energy of the ^1L_a state and the more red-shifted is the emission wavelength (from 320-350 nm). The ^1L_b state with a nearly orthogonal dipole moment has an emission spectrum shifted to the blue (300-320 nm). ^1L_b is less dependent on the polarity of the environment. Therefore, the emission wavelength of tryptophan residues in a protein will depend on their microenvironment and the degree of exposure to the (aqueous) solution. The more buried in a protein a tryptophan residue is, the more blue-shifted will be its emission spectrum. Tryptophans with fluorescence maxima below 330 nm, between 330-342 nm, and above 342 nm are classified as buried, partially buried and solvent exposed, respectively (for a more detailed discussion see [7]). The quenching by the side chains of adjacent amino acids is another factor which changes the tryptophan fluorescence (here the intensity rather then the spectral position) [8]. Both the spectral position and intensity are used to monitor conformational changes, folding and unfolding, and protein-protein interactions. In chapter 3.1 an example will be given on the conformational changes of a bacterial photoreceptor sensed by its tryptophan emission wavelength and intensity.

The fluorescence lifetime of tryptophan is also often used as a read-out parameter to characterise protein conformational changes. The interpretation of the fluorescence lifetime data, however, is more difficult since even free tryptophan shows a multiexponential decay of

its fluorescence. Therefore the interpretation of time-resolved protein fluorescence is less straight forward.

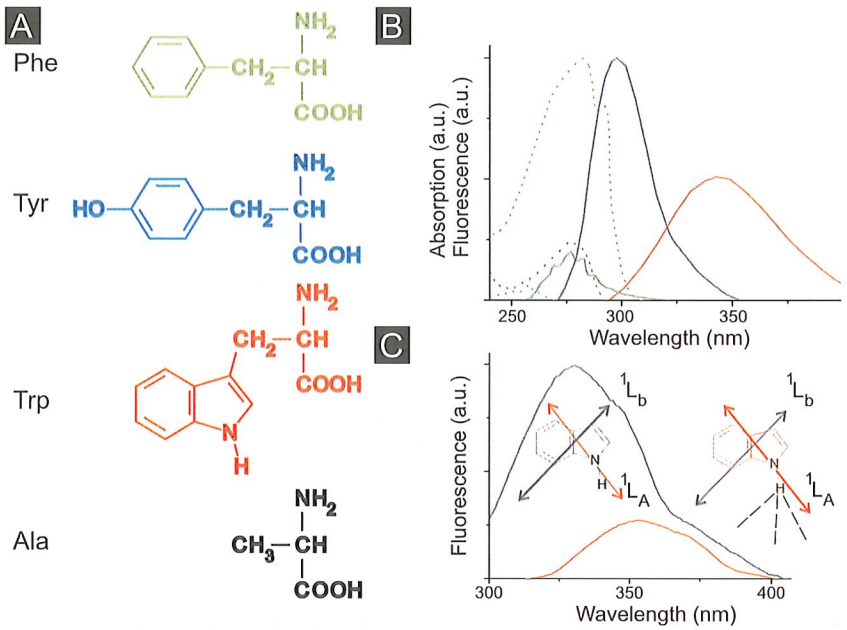


Fig. 1: A) Molecular structure of the three fluorescent aromatic amino acids phenylalanine (Phe, green), tyrosine (Tyr, blue), and tryptophan (Trp, red). Alanine (Ala) is shown as example for a non-fluorescent amino acid. B) Absorption (dotted) and fluorescence spectra (solid) of Phe (green), Tyr (blue), and Trp (red). The spectra are normalised with respect to their absorption probability and fluorescence quantum yield, respectively. C) Fluorescence spectra of proteins with buried (blue) and solvent-exposed Trp residues (red). The dipole moments of the two excited states of Trp are shown together with the structure of the Trp side chain. The bold arrow marks the dominating transition in non-polar (blue, ^1L_b) and polar environment of Trp (red, ^1L_a), respectively. The dashed lines represent hydrogen-bonds.

2.2 Fluorescent Dye Molecules

Many fluorescence applications in biology rely on the use of large organic molecules containing (several) aromatic heterocycles or polyenes (molecules containing a cyclic or linear chain of alternating single and double carbon-carbon bonds). The common feature of all these molecules is an extended system of π-bonds (see molecular structures in Fig.2A and 2B). The π-electrons are considered to be delocalised over all the atoms of the conjugated chain, i.e. the electrons are thought to move freely along the length of the carbon skeleton. This delocalisation leads to a smaller energy difference between the ground state and the excited states of the molecule. Photoexcitation becomes feasible in the visible and near-infrared spectral region with large absorption probabilities (10 – 500 times larger than that of tryptophan).

Great numbers of such molecules have been synthesised in the past 100 years. The vast majority of these molecules were not intended to be used as fluorescent dyes rather than for colouring of textiles and clothes or medical treatment. Among the thousands of molecules available, only a few were useful as fluorescent dyes. As a general rule a good fluorescent dye should combine a high fluorescence and a low triplet quantum yield. Furthermore specific absorption properties (spectral position, high absorption probability) are important. Chemical modifications of this stock of dyes lead to compounds with better absorption and fluorescence properties (ε, spectral position, Φ_{fl}, τ_{fl}), solubility, photostability, protein labelling or cell membrane permeability (see also: http://www.probes.com). Many of the known fluorescent dyes fall into families, the most important ones are: carbocyanines, coumarins, flavins, fluoresceins, rhodamines, oxazines, and porphyrins (Fig.'s 2A and 2B). The effect of delocalisation in extended π-electron systems is examplified in Figure 2B showing the fluorescence spectra of three carbocyanines. The extension of the conjugated system by one or two double bonds of carbon atoms leads to significant red-shifts of the fluorescence maximum by 100 and 200 nm, respectively.

Modern fluorescence microscopy uses excitation from the far-blue to the near-infrared spectral region (350 - 800 nm). When irradiated with light of these wavelengths cells become autofluorescent (see e.g. [9]) due to the excitation of molecules inside cells. The brightness of the autofluorescence depends on the excitation wavelength and the biological specimen. The molecules causing autofluorescence have very different chemical structures and vary in their absorption and fluorescence properties. Several molecules such as polyenes (carotenes) and aromatic heterocycles (flavins, porphyrins; Fig.2A), which serve as endogenous cofactors for enzymes (e.g. flavin- and hemeproteins) or in photosynthetic pigment-protein complexes, absorb photons with high absorption probabilities from the blue to the near-infrared spectral region and emit fluorescent light with reasonable efficiencies (Φ_{fl} = 0.01 - 0.25). Other sources of autofluorescence include molecules with low fluorescence efficiencies and (often) low absorption probability, but which are abundantly in the cell. When the emission intensities are low, care must be taken that the solutions in which the cells are examined, do not contain fluorescent substances or impurities. In general, the fluorescent dye should give a considerably brighter signal than the autofluorescence. Therefore, dyes with favourable absorption and fluorescence properties should be chosen carefully.

The fluorescence properties of some dyes depend on the environment provided by the solute. The spectral properties, fluorescence quantum yield or fluorescence lifetime, may change with temperature, pH, concentration of ions, membrane voltage, membrane composition, etc. It is for this reason that fluorescent dyes have been developed that represent an important tool for the non-destructive characterisation of cellular systems. The chemical and photophysical nature underlying the fluorescence change can be quite different. In chapters 3.2 and 3.3 the use of fluorescent indicators to monitor cellular Ca^{2+}- and Cl^-concentrations is presented.

For many studies the fluorescent dyes have to be transported into the cell. Different techniques have been developed to incorporate the dyes. Most commonly used methods include microinjection, reversible permeabilisation of the cell membrane, endocytosis, and chemical modification of the fluorescent dye to render it membrane permeable. For example, carboxylate or phenolic hydroxyl groups of the fluorescent dye are derivatised to yield acetate or acetoxymethyl esters (AM esters, see Fig.5A). Once inside the cell, the dye esters are hydrolysed by unspecific esterases, releasing the parent molecule. The dyes may be toxic to the cellular system. It is advisable to use the lowest dye concentration possible with respect to the signal-to-noise ratio of the measurement.

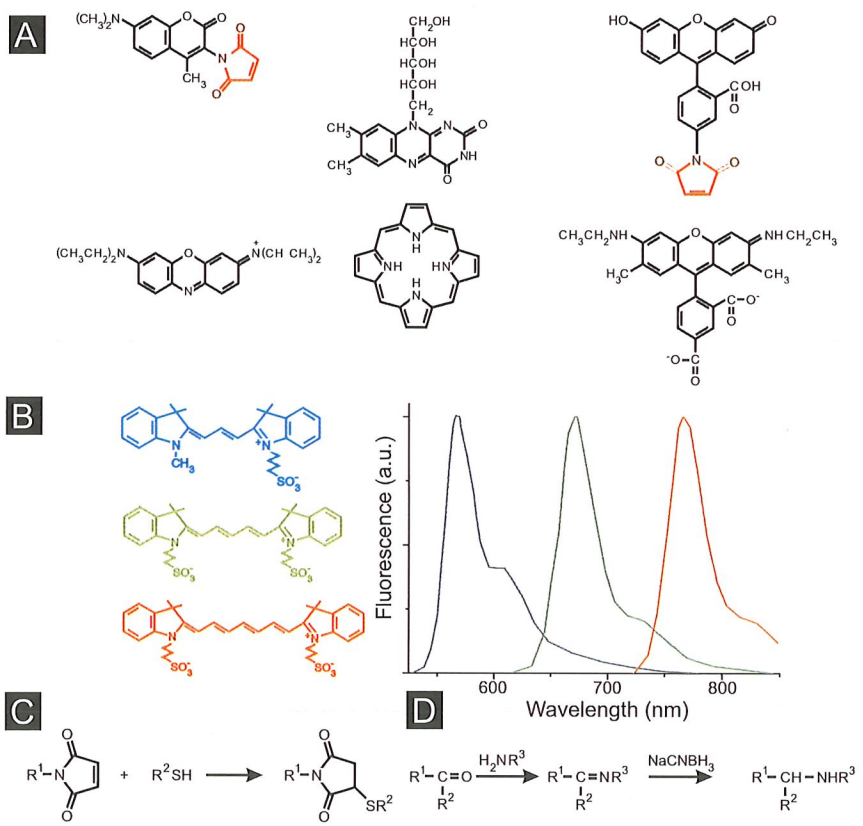


Fig. 2: A) Molecular structures of important fluorophore families: coumarin, flavin, fluorescein (first line), oxazine, porphyrin, rhodamine (second line). The side group marked in red is the maleimide group attached to the fluorophore for labelling reasons. B) Molecular structure of three carboxycyanines and their fluorescence spectra (Cy3 (blue), Cy5 (green), Cy7 (red)). The fluorescence spectrum is red-shifted (to lower energies) for the more extended system of π-bonds. C) Coupling reaction for labelling of the thiol group with a maleimide group. D) Coupling reaction for labelling of amines with a succinimidyl group.

The covalent attachment of a fluorescent dye to a protein offers the possibility to study the protein both *in vitro* (i.e. conformational changes, protein-protein interactions) and *in vivo* (i.e. localisation, observation of concentration changes, protein-protein interactions). Two types of chemical groups in the side chains of some amino acids serve as sites for modification: Primary amino groups (present in lysine, asparagine, arginine, and glutamine), and the thiol group of cysteine. Amine-reactive fluorescent dyes are widely used to modify proteins for immunochemistry, cell tracking, or receptor labelling. Thiol-reactive fluorescent dyes frequently serve as probes of protein structure and function, because they specifically label only cysteines. The two labelling reactions are depicted in Fig.'s 2C and 2D. In the simplest case, all amino acids with reactive side groups in the protein are labelled. The reactivity of side groups can differ largely due to the accessibility and microenvironment

determined by the protein structure. For specific labelling, the number of reactive side groups must be reduced. It is advisable to label only a single site, so that the position of the fluorescent dye is known. Sometimes it is necessary to introduce an amino-acid residue that can be labelled into the polypeptide chain by genetic engineering. Since the labelling reaction of amines or thiols can be sequentially achieved, one amine and one thiol group can be labelled specifically with different chromophores. This allows decoration of proteins with a FRET pair that serves to study intramolecular rearrangements of the protein.

The ability to label any position in the amino-acid sequence is the strength of this technique, in particular for *in vitro* studies of proteins. However, one disadvantage of the method is that the specific labelling of a protein must take place outside the cell. After labelling, proteins need to be introduced again into the cell by microinjection or other methods. Unfortunately a non-invasive method similar to the one using AM-ester indicators does not yet exist. Recently, two techniques for *in vivo* protein labelling have been developed. One approach is to incorporate a small tetracysteine motif (CCXXCC; C for cysteine, X for any other amino acid) into the protein. A membrane permeable biarsenical dye is then added to the cell. After entry the dye will react covalently with the CCXXCC motif [10,11]. Since the CCXXCC sequence is rarely found in proteins, side reactions are minimal. The second approach is based on the introduction of amino acids (not occurring naturally) containing specific side-groups (keto, azide, acetylene) which can be covalently attached to fluorescent dyes by specific chemical reactions that do not derivatise natural amino acids. The fluorescent dyes for labelling are added as membrane permeable substances. The successful application of this method has been shown for bacteria [12,13] and yeast [14,15]. Both of these methods are new and most of the work with dye-labelled proteins has been carried out *in vitro*.

2.3 Visible Fluorescent Proteins

Intrinsically fluorescent proteins (VFP) are a group of proteins, which absorb and emit *visible* light without being bound to a prosthetic chromophore. The fluorescent group is formed by a chain of posttranslational chemical reactions *in situ*. Because the chromophore formation is an autocatalytic process, it can take place in any cell type. If one can introduce the VFP-DNA in the cell to be studied, the VFP will be produced and its fluorescence can be detected. By linking the VFP-DNA to the DNA of another protein, a so-called fusion protein is produced. In this way any protein of interest can be fluorescently labelled *in vivo* in contrast to the labelling using organic fluorescent dyes.

The most thoroughly studied fluorescent protein is the Green Fluorescent Protein (GFP) from the jellyfish *Aequoria victoria* (see [16]). VFPs are expressed in many marine coelenterates, i.e. jellyfish, sea pens, sea squirts, and sea anemones. In the jellyfish *A. victoria*, GFP is part of a two-component bioluminescence system. Bioluminescence is the process by which visible light is emitted by an organism as a result of a chemical reaction(s). An enzyme catalyses the oxidation of a substrate molecule. In *A. victoria*, the enzyme is called Aequorin, a Ca^{2+}-binding protein. The Aequorin-Ca^{2+}-substrate complex emits broad-band blue light *in vitro* (centred at 470 nm). However, *in vivo A. victoria* does not emit blue (but green) bioluminescence: The Aequorin complex undergoes radiation less energy transfer (FRET) to GFP which emits green fluorescence light (centred at 510 nm). Although the physiological role of bioluminescence of the marine coelenterates is not entirely clear it probably serves at least one of three functions: defence, offence, or communication.

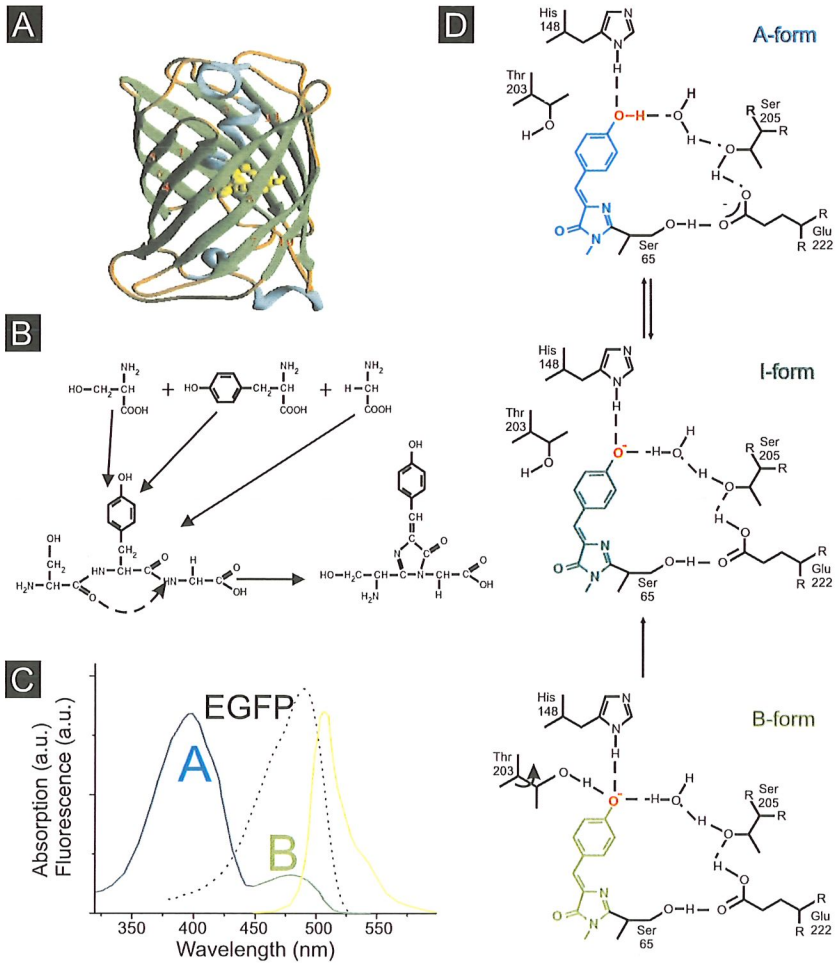


Fig. 3: A) Three-dimensional structure of GFP from A. victoria. The chromophore connencted to the central α-helix (blue) is shown in yellow in the middle of the barrel formed by β-sheets (green). B) Reaction scheme of GFP chromophore formation from Ser65, Tyr66, and Gly67. C) Absorption spectra of A (blue) and B form (green) of wild-type GFP from A. victoria and a widely used mutant (EGFP, black) with a stabilised anionic form of the chromophore. The fluorescence spectrum (yellow) is for all three forms very similar. D) Protonation state of the chromophore of wild-type GFP and hydrogen-bonding interactions for neutral A-form (blue) and the two anionic forms (I (dark green) and B (light green)). The protonation/deprotonation site is drawn in red.

The three-dimensional structure of GFP (Fig.3A) has been elucidated by X-ray crystallography. GFP has a unique barrel-like structure with a diameter of about 24 Å and a height of 42 Å. Eleven β-sheets form the walls of a can. The chromophore is in the centre of

the eleven β-sheets and is linked to the α-helix that runs through the centre of the barrel. The structures of GFP, its mutants (see below) or other VFPs are very similar and differ only in the region of the mutations. The chromophore (see below) is well protected in the centre of the barrel. Several polar residues and water molecules form a hydrogen-bonding network around the chromophore. A large cavity that occupies a volume of ~135 $Å^3$ is found on one side of the chromophore. The cavity does not open to bulk solvent, and contains at least four water molecules. The rigid structure of the "β-can" protects the chromophore efficiently against quencher molecules (including O_2 and H_3O^+). Both the protonation state of the chromophore as well as the hydrogen-bonding network plays an important role in the photophysics of GFP (see below).

The chromophore of GFP is formed by three consecutive amino-acid residues of the polypeptide chain (Serine65, Tyrosine66, and Glycine67; see Fig.3B). The reaction starts with architectural distortions after the folding of GFP has been completed. It is followed by peptide cyclisation, dehydration and a final oxidation – which is the rate limiting step - to generate an aromatic imidizalone where the two rings are conjugated [17]. This reaction (also called maturation) occurs in about 10 hours for GFP and can be considerably faster or slower for certain mutants. A substantial fraction of GFP expressed in bacteria or cells is misfolded and does not form a fluorescent protein. This is in part due to the fact that folding is optimal at temperatures considerably lower than 37 °C where bacteria and eukaryotic cell lines are propagated. *A. victoria* inhabits the cold north-western Pacific Ocean with much lower temperatures.

After maturation, GFP has a major absorption maximum at 398 nm and a minor absorption maximum at 475 nm with a shoulder on the long-wavelength flank of the 475 nm band (Fig.3C). Excitation at 398 nm results in an emission maximum at 508 nm, while irradiation at 475 nm causes an emission maximum at 503 nm. The protonation state of the phenolic oxygen atom influences the spectral properties of GFP. This is also true for a number of mutants. The 398 nm absorption band is attributed to the protonated, the 475 nm absorption band to the deprotonated chromophore. The neutral (A) and anionic (B) species of the chromophore are in ground-state equilibrium. The equilibrium can be influenced by changes in protein concentration, ionic strength, pH, and temperature. Excitation of either species A (neutral) or species B (anionic) results in similar fluorescence spectra. This is due to the fact that the phenolic oxygen of Tyr66 is more acidic in the excited state than in the ground state. Therefore an excited-state proton transfer occurs resulting in a common anionic excited state (I^*) that is responsible for the observed emission spectrum. In fact, a third state (I) has been identified that shows absorption at 490 nm [18]. This state has a deprotonated chromophore but a hydrogen-bonding network as the A form. Even more states (different chromophore isomers or ground-state A forms) have been proposed to explain the results of the many spectroscopic studies (including time-resolved fluorescence and single-molecule spectroscopy) on wild-type GFP and mutants.

Many mutants of GFP have been produced to improve several properties of GFP useful in fluorescence microscopy and spectroscopy: 1. Rate of folding and maturation and the maturation yield in different host organisms (bacteria, yeast, mammalian cells); 2. Suppression of dimerization; 3. Position of absorption and fluorescence maxima. Significant improvements concerning the first two points have been achieved. Absorption and fluorescence maxima of GFP mutants range from 360 nm to 516 nm and from 442 nm to 529 nm, respectively. Countless numbers of fusion proteins between GFP and other proteins have been generated to exploit the genetically encoded fluorescent "tag". GFP fluorescence can be used as a reporter for gene expression. The GFP fusion proteins have been used for the intracellular localisation of proteins, for the study of protein dynamics and folding in living cells, for the sensing of pH, ion concentrations, phospholipids and other molecules important

in cellular processes. The available spectral range allows the construction of FRET (see chapter 3.5) pairs (either BFP (λ_{abs} /λ_{fl} = 400 nm/450 nm) and EGFP (480 nm/515 nm) or CFP (435 nm/480 nm) and YFP (515 nm/528 nm)). By fusing each of the two GFPs with a potential interaction partner, FRET reports on protein conformational changes and protein-protein interactions. The GFP-fusion proteins can monitor the properties of sub cellular compartments by attaching short targeting signals, which direct the sensor proteins to certain cell organelles, the nucleus, or the cell membrane.
More recently, several VFPs with absorption and emission spectra that are red-shifted ($\lambda_{abs}/\lambda_{fl}$ = 550-610 nm/580-640 nm) have been identified and cloned from anemones and corals. Their overall protein structure is very similar to GFP while the chromophore is different. It has an additional conjugated carbon-carbon double bond. Some of these VFPs (including the best known DsRed) also display high fluorescence quantum yields so that, at first glance, they would form a superior FRET pair with EGFP or YFP compared to the CFP/YFP pair. The latter has two major drawbacks for microscopy applications: excitation of CFP is suboptimal with the most common Argon-ion laser source (488 or 454 nm). In addition, the CFP emission spectrum is very broad causing a significant spectral overlap with the acceptor YFP. In contrast, EGFP and YFP can be perfectly excited at 488 nm while DsRed is having a very low absorption probability. But the red VFPs have some severe disadvantages. There are only a few reports that use functional DsRed fusion proteins. The reasons are probably a more complicated and time consuming (often incomplete) maturation of the DsRed chromophore compared to GFP and a strong propensity for oligomerization (tetramers, octamers and higher oligomers even at nM concentrations). These drawbacks have to be overcome to exploit the red-absorbing VFPs with their superior spectral properties in cell biology.

3. Examples

In chapter 1 the properties of fluorophores used in biological research have been introduced. In the following chapter examples for the use of fluorescence will be given: 1. protein fluorescence to monitor protein-conformational changes (3.1); 2. Visualisation of intracellular ion concentration changes (3.2) and determination of intracellular ion concentrations (3.3); 3. Localisation (3.4) and interaction of proteins (3.5) in cells by means of VFP-fusion proteins.

3.1 Tryptophan Fluorescence of PYP

Photoreceptor proteins are light-sensing proteins which are expressed in almost all organisms (from bacteria to higher animals and plants). The most important families are rhodopsins, phytochromes, cryptochromes, phototropins as well as photoreceptors involved in setting the circadian rhythm. The light-absorbing part, a prosthetic chromophore, is covalently attached to the protein. The chemical structure of the chromophore differs significantly as does the structure of the proteins. Despite these differences, there is a general principle underlying photoreceptor proteins. A structural rearrangement of the chromophore (e.g. isomerisation, charge transfer) occurs after absorption of a photon. The local protein environment in which the chromophore is immersed reacts with a conformational change followed by large scale conformational changes of the parts that are far from the chromophore-binding site. These structural changes activate highly amplifying signalling pathways by means of protein-protein interactions.
Photoactive Yellow Protein (PYP) is a blue-light photoreceptor protein of photosynthetic bacteria. It belongs to a family of sensor proteins (the PAS domain proteins [19]) that are found in all organisms, often as one domain in multi-domain proteins involved in signal transduction. PAS domains serve diverse functions and can bind different chromophores, but many PAS domains function as light-independent elements without binding a chromophore.

The PYP from *Halorhodospira halophila* is a small, water-soluble, that mediates the blue-light repellent response of this organism. PYP is composed of 125 amino acids and a thiol-linked (to Cys-69) *p*-coumaryl chromophore. Its structure is that of a typical α/β protein. After absorption of a photon, the coumaryl chromophore isomerises in a few ps and conformational changes occur on a time scale of 50 ns to 1 ms that yield the signalling state pB. pB returns to the ground state in about 1 s thereby completing the photocycle. It has been shown by a number of spectroscopic techniques that pB has conformational differences compared to the dark state pG [20].

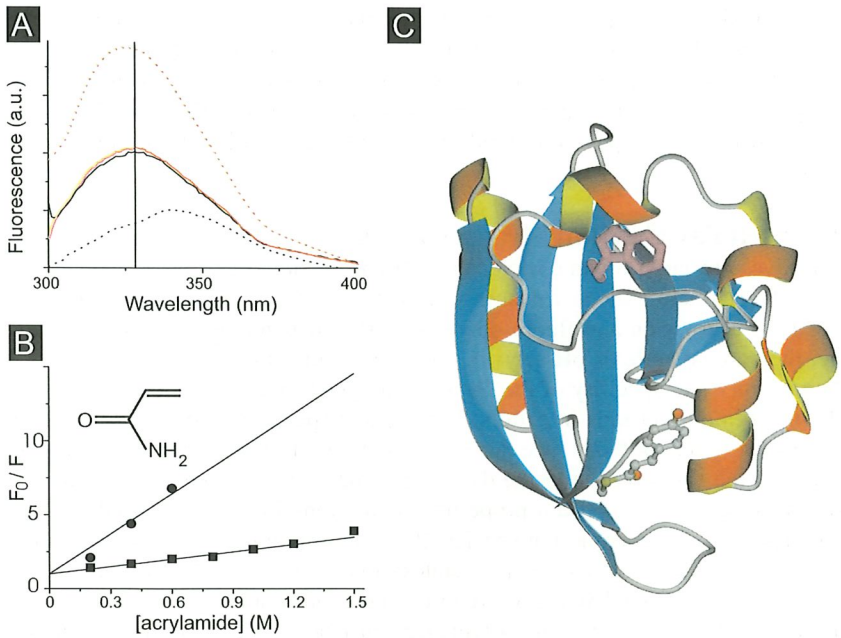


Fig. 4: A) Trp fluorescence spectra of PYP (λ_{exc} = 295 nm) in the ground state pG (solid) and in the signalling state pB (dotted) at pH 8 (black) and pH 6 (red). The vertical line marks the pH-independent fluorescence maximum of pG (328 nm). B) Quenching of Trp fluorescence of PYP by acryl amide (λ_{exc} = 295 nm). Stern-Volmer plot of the fluorescence intensity ratio (F_0 = fluorescence intensity in the absence of acrylamide, F = fluorescence intensity at different acry amide concentrations) as a function of acrylamide concentration for pG (squares) and pB (circles) at pH 8. The larger increase for pB reflects a larger accessibility of Trp119 for the quencher in the pB state. C) Three-dimensional structure of PYP showing TRP119 (pink) with the N-terminal cap to the right. The p-coumaryl chromophore is depicted in grey in the lower middle of PYP.

PYP contains a single tryptophan residue (Trp119), that is located at the interface between the central β-sheet and its N-terminal cap. The fluorescence properties of Trp119 in the dark state pG (λ_{max} = 328 nm; Φ_{fl} = 0.01; pH-independent) are characteristic of a buried tryptophan residue in a hydrophobic environment with significant quenching by adjacent amino-acid residues. Signalling-state formation leads to pH-dependent fluorescence changes: At pH values < 6.5 the fluorescence increases twofold, with a small blue shift of the emission

maximum of Trp119. At pH > 6.5, the emission maximum of TR119 shifts to the red, whereas the fluorescence intensity decreases by about 50 % [21] (Fig.4A).
These results support the idea that signalling state formation in PYP leads to significant changes in the protein structure, even at sites distant from the chromophore. The nature of the changes in pB depends on the pH: At slightly alkaline pH, the conformational change leads to the exposure of the central β-sheet harbouring Trp119. At slightly acidic pH, the polarity of the environment of Trp119 is hardly affected by the formation of the signalling state, but the quenching of the fluorescence by adjacent amino acids is reduced (Fig.4B). The accessibility of small molecule quenchers (acryl amide) in the solution is enhanced at acidic and alkaline pH in the signalling state (pB) compared to the dark state (pG). This argues for a more flexible and accessible structure of the N-terminal cap and a looser interaction with the central β-sheet in pB. These changes may be recognised by an interacting protein involved in the signalling process. The results are of importance for many members of the PAS domain family since Trp119 is a highly conserved residue and may be directly involved in the signal transduction.

3.2 Intracellular Calcium Measurements

Fluorescent dyes are large organic molecules. Often functional groups are added to improve solubility and cell membrane permeability. Some dyes (e.g. fluorescein) show significant changes of their absorption and fluorescence properties depending on their protonation state. Therefore, they are suitable as pH sensors. In general, this is not the case for other ions, i.e. the binding affinity and specifity of the fluorescent dyes is not high. Na^+, K^+, Ca^{2+}, Mg^{2+} and Cl^- are the most important ions in cellular biology and special fluorescent sensors have been developed for every one of them. The design of most dyes follows a common principle: Upon binding of ions, the fluorescence of the dye is quenched. For a fluorescent dye to function as an ion sensor it has to meet certain properties: 1. the ion-binding part should posses a high affinity and specifity for a single ion species; 2. the ion-binding moiety chain should quench the fluorescent dye only in one of two states: ion free or ion bound. In this way, the fluorescence intensity of the dye directly reports on the ion concentration.
Some dyes use the quenching of polynuclear aromatic hydrocarbons by amines. The quenching "reaction" involves the lone pair of electrons of the nitrogen atom. If this electron pair participates in chelating a cation, the electron transfer is inhibited. The fluorophore is *not* quenched and the fluorescence intensity is increased. In contrast, in the absence of cations, the amino group can interact with the fluorophore directly. This leads to a smaller ("quenched") fluorescence signal. The interaction of the ion-binding moiety with the fluorescent dye moiety may also affect properties other than the fluorescence intensity (e.g. absorption and fluorescence spectra, fluorescence lifetime, etc.) By chemical engineering of the ion-chelating group, the affinity for ion binding can be tailored to the concentration range of the ion occurring in the cell.
The determination of cellular Ca^{2+} concentrations ($[Ca^{2+}]$) has received great attention due to the special role Ca^{2+} ions play as second messenger in many cellular processes. Depending on the cell type, free $[Ca^{2+}]$ range from 50 nM to more then 10 μM. A chelator/buffer widely used to control $[Ca^{2+}]$ is called BAPTA (see Fig.5A). It is suited for $[Ca^{2+}]$ between 50 nM and 1 μM. Chemical modifications of BAPTA can tune the Ca^{2+} affinity by more then two orders of magnitude (up to 500 μM). BAPTA has been chemically coupled to fluorescent dyes that absorb from the near-UV (350 nm) up to the orange (580 nm) spectral region. After Ca^{2+} binding, some of the fluorescent Ca^{2+} indicators change their spectra (absorption and/or fluorescence) or the fluorescence lifetime, but in most experiments the fluorescence intensity is used for measuring the intracellular Ca^{2+} concentration, $[Ca^{2+}]_i$. The fluorescence intensity difference between Ca^{2+} free and Ca^{2+} bound states range from 5- to 100fold *in vitro*. Both,

the fluorescence-intensity change and the affinity to Ca^{2+} are usually different *in vitro* and *in vivo*. The different properties may be explained by the cellular environment (proteins with charged groups, ions, lipids, etc.). It is also advisable to use low intracellular concentrations of the fluorescent Ca^{2+} indicator, not to disturb the cellular Ca^{2+} homeostasis system.

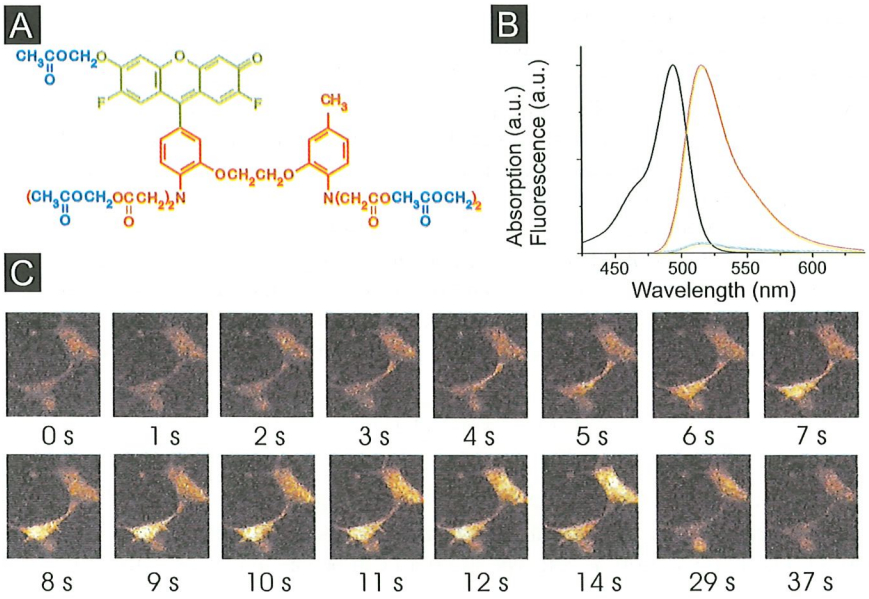


Fig.5: A) Molecular structure of the AM-ester form of the Ca^{2+} indicator dye Fluo-4. Ca^{2+} chelates to the BAPTA moiety (red), membrane permeability is achieved by the AM-esters (blue). The green moiety is the fluorescein dye (see also Fig.2A) sharing one benzoic ring with the BAPTA moiety B) Absorption (black) and fluorescence spectra at low $[Ca^{2+}]$ (green) and high $[Ca^{2+}]$ (red). C) Time sequence of two-photon fluorescence microscopy images showing $[Ca^{2+}]_i$ changes in HEK293 cells. The cells are stably transfected with an octopamine receptor and stimulated with 1 µM octopamine

Ca^{2+} imaging of cells stably transfected with a neurotransmitter receptor from Drosophila is shown in Fig.5C. Octopamine, a monophenolic amine specific to invertebrates, binds to the receptor. Upon binding, changes in $[Ca^{2+}]_i$ occur. Figure 5C shows two cells loaded with the fluorescent Ca^{2+} indicator Fluo-4 (Fig.5A). Fluo-4 belongs to the class of Ca^{2+} indicators which change their fluorescence intensity upon Ca^{2+} binding (Fig.5B). The $[Ca^{2+}]$ rises in the narrow foothill of the cell on the left. Within 5 s, the Ca^{2+} increase propagates to other parts of the cell. With a short delay, also the upper right cell shows an increase of $[Ca^{2+}]_i$. In other experiments, periodic changes $[Ca^{2+}]_i$ have been observed in several cells (see for a related study [22]). The periodicity depends on the octopamine concentration and most likely on the local concentration of the receptor protein in the cell membrane.

3.3 Intracellular Chloride Measurements

The read-out parameter most often used in fluorescence microscopy is the fluorescence intensity. It depends on many factors and obviously on the concentration of the fluorophore. When used in individual cells or whole tissue, the fluorophore distribution is all but

homogeneous. This makes comparisons of fluorescence-intensity signals from different parts of a cell or between different cells rather difficult or impossible. In contrast, fluorescence lifetime has the advantage that it does not depend on the fluorophore concentration. By using the fluorescence lifetime of a molecule, measurements from different parts of the cell or from different cells can be directly compared.

Determination of cellular Cl^- concentrations ($[Cl^-]$) asks for a different fluorescent indicator type than for measuring $[Ca^{2+}]$. The intracellular concentration of Cl^- ($[Cl^-]_i$) is in the mM range (5-100 mM). Cl^- ions cause a change in the excited state properties of the fluorescent moiety via a process called collisional quenching. If an excited molecule is colliding with another molecule or ion – named the quencher (Q), the two form a (short-lived) complex. This complex may have a much larger rate for non-radiative deactivation. Hence, the fluorescence probability decreases. The fluorescent dye can be used to determine the concentration of the quencher. Both the fluorescence intensity and the fluorescence lifetime decreases if the concentration of quencher molecules is high enough (usually mM) to ensure diffusional collision between fluorophore and quencher while the fluorophore is in its excited state. Figure 6B shows the decrease of fluorescence lifetime for increasing quencher (here Cl^-) concentrations. The lifetimes at different $[Cl^-]$ are described by the Stern-Volmer equation,

$$\tau_0 / \tau = 1 + K_{SV}\left[Cl^-\right] \qquad (1)$$

where τ_0 and τ are the decay time constants in the absence and presence of Cl^-, respectively. The Stern-Volmer-constant, $K_{SV} = k_q\tau_0$, is a measure of the Cl^- sensitivity of the probe. If K_{SV} is known for a certain solution or cell type the $[Cl^-]$ can be calculated from the lifetimes τ and τ_0.

In the following, I will show how $[Cl^-]_i$ in olfactory sensory neurons (OSNs) from rat can be determined. The OSN has asymmetric topology with the cell body incorporating the nucleus, a 10 – 20 µm long dendrite ending in the dendritic knob. Olfactory signal transduction is initiated when odorants bind to receptor proteins and trigger the synthesis of the second messenger cAMP [23]. Signal transduction in vertebrate OSNs involves an unusual way of current amplification which is based on the concerted action of two types of transduction channels: cAMP-gated cation channels and Ca^{2+}-gated Cl^- channels. Both channels are expressed in the plasma membrane of the chemosensory cilia which processes of the OSNs dendritic knob. An outstanding feature of this transduction process is that OSNs, unlike other primary neurons, need an unusually high intracellular Cl^- concentration, $[Cl^-]_i$, to support the Ca^{2+}-dependent Cl^- efflux. Cl^- transport and Cl^- homeostasis in OSNs must be organised in such a way that it promotes Cl^- accumulation.

Although $[Cl^-]_i$ plays such a central role in olfactory transduction, only little information pertaining to Cl^- homeostasis in OSNs is available. Two previous studies indicate a range for $[Cl^-]_i$ of 20 - 80 mM in isolated rat OSNs [24] and in cryosections of rat olfactory epithelium [25]. $[Cl^-]_i$ measurements from living OSNs in intact epithelium have not been possible so far. Fluorescence lifetime imaging of a Cl^- sensitive dye is applied that allows the determination of $[Cl^-]_i$ in dendritic knobs of intact OSNs. The optical sectioning is possible due to the two-photon excitation in the scanning microscope which takes place only in the focus of the laser. Fluorescence lifetime was measured using time correlated single photon counting (TCSPC) (for details see [26]). The total fluorescence intensity and the fluorescence decay time curve were recorded for every pixel in the image.

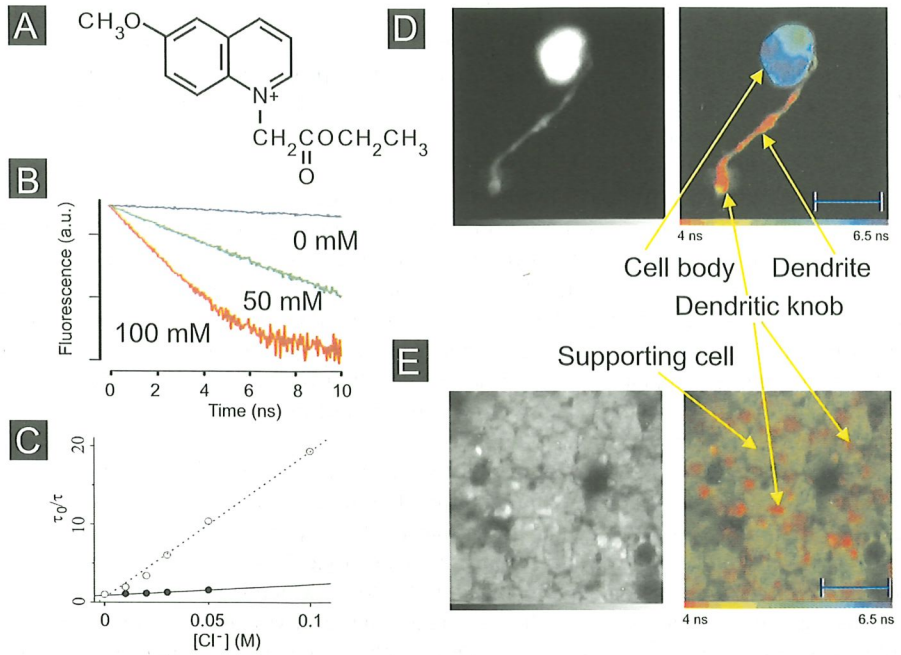


Fig. 6: A) Molecular structure of the Cl^- indicator dye MQAE. B) Time course of fluorescence decays at 0 M (blue), 50 M (green) and 100 mM (red) Cl^- concentration in vitro. C) Stern-Volmer analysis of MQAE Cl^- sensitivity in vitro (open circles, dotted line) and in the olfactory epithelium (full circles, solid line). D) Fluorescence intensity (left) and fluorescence lifetime image (right) of an isolated olfactory sensory neuron (OSN). The scale bar is 10 µm. E) Fluorescence intensity (left) and fluorescence lifetime image (right) of the olfactory epithelium from rat in the layer of the dendritic knobs of the OSN. The red and green structures in the fluorescence lifetime image are dendritic knobs (red) and supporting cells (green), respectively. The scale bar is 10 µm; the z-resolution is about 1 µm.

The fluorescent Cl^- indicator used was MQAE ([N-(6-methoxyquinolyl) acetoethyl ester], see Fig.6A). MQAE shows a large affinity for Cl^- *in vitro* as can be seen from the change of the fluorescence decays as a function of [Cl^-] (Fig.6B). MQAE is also membrane permeable and accumulates in the cells. Measurements are possible for several hours after loading. A large difference of the Stern-Volmer constant was found when calibration of MQAE was done *in vivo* (15 mM^{-1}) compared to *in vitro* experiments (185 mM^{-1}) [26,27] (see Fig.6C).
Fig.6D shows the fluorescence intensity and fluorescence lifetime image of an isolated OSN. In both one can see the overall topology: cell body with nucleus, dendrite, and dendritic knob. The cilia are not visible because they are (1) narrow and (2) mobile on the time scale of the measurement (minutes). The fluorescence lifetime decreases from the cell body to the dendritic knob. This could be interpreted with a Cl^- gradient in the cell (with high Cl^- in the dendritic knob) assuming the same Cl^- sensitivity of MQAE along the cell. Fig.6E shows the fluorescence intensity and fluorescence lifetime images of the OSNs dendritic knobs and neighbouring supporting cells in the olfactory epithelium. The difference between the two cell

types is most evident in the fluorescence lifetime image. Many dendritic knobs are only visible in the fluorescence lifetime image. It can be concluded from all these experiments that $[Cl^-]_i$ in OSNs is high (50-70 mM). Detailed investigations of this and other properties are currently under way [27]. The use of the fluorescence lifetime as a read-out parameter turned out to be a reliable method to compare different cellular properties. A purely intensity based analysis would have needed a calibration for every single cell investigated, which none of the cells would have survived.

3.4 GFP Fusion Proteins: Protein Localisation

GFPs have been fused to many different proteins (e.g. motor proteins, kinases, receptors, ion channels) in the past ten years. Often the GFPs are used as a biologically inert participant merely used as a a visible marker. This passive use as a fluorescent tag is widespread in fluorescence imaging. The barrel-like structure of GFPs that shields the chromophore from the external environment makes GFPs superior to organic dyes. Typical passive use of GFPs include monitoring the appearance, degradation, location or translocation of the proteins to which they are fused. The light-dependent translocation of phytochrome will be presented as an example in the following section.

Plants have developed sensory photoreceptors, including the phytochromes as the most important proteins. Having captured an incoming photon, the activated phytochrome molecule must relay this information to nuclear genes that mediate appropriate adjustments in growth and development. While many of the physiological phytochrome-dependent responses are known, the molecular determinants are poorly understood, which is mainly due to the missing three-dimensional structure of phytochrome. The generic phytochrome (phy) molecule is a soluble, *dimeric* chromoprotein (125 kDa) with a covalently attached tetrapyrrole chromophore. Phytochromes photosensory activity is based on light-induced, reversible switching between two conformers: the biologically inactive form P_r (red absorbing; 670 nm) and the biologically active form P_{fr} (far-red absorbing, 730 nm) (see Fig.7A). P_{fr} formation initiates an intracellular transduction process culminating in altered expression of specific genes that direct morphogenesis. P_r reconversion from P_{fr} (either light-induced or in the dark) can abrogate this process. In *Arabidopsis*, the most thoroughly investigated plant system, the phytochrome family consists of five members, designated phyA to phyE. Mutational analysis has shown that individual members of the family have different, albeit overlapping photosensory and/or physiological functions to control plant responses (see [28]).

Besides efforts to crystallise phytochrome another strategy was followed by several research groups to understand phytochrome induced signal transduction at the molecular level [29]. Fusion proteins of different phytochromes and GFPs were generated to monitor the distribution of the phytochromes under different light-illumination conditions. For both, phyA and phyB, a light-dependent translocation of the phytochrome from the cytoplasm to the nucleus was observed in Arabidopsis albeit different photosensory properties. phyB nuclear import is induced only by red light, and can be reversed by far-red light. In the nucleus, phytochromes localise to discrete subnuclear foci ("nuclear bodies", NB). Localisation to NBs directly relates to the functions of phytochromes.

In Fig.7B the nuclear localisation patterns of a phyB-GFP fusion protein under different illumination intensities of red light is shown (adapted from [30]). In the dark phyB accumulates in the cytoplasm. *Arabidopsis* seedlings were grown under different intensities of red light for four days. Due to the overlapping absorption of P_r and P_{fr} and the reversion of P_{fr} to P_r in the dark, a light-dependent equilibrium was established, with higher P_{fr} concentrations with increasing red light intensity. Four different phyB subnuclear localization patterns were observed (see Fig.7B): diffuse nuclear localization ($[P_{fr}] \leq 10$ %), small and numerous NBs only (10 % < $[P_{fr}]$ < 20 %), both small and large NBs (20 % <$[P_{fr}]$ < 45 %), and large NBs

only ([P_{fr}] > 45 %). phyB nuclear import is not sufficient for phyB accumulation in NBs. It rather is determined by the percentage of phyB protein that is in the active conformer (P_{fr}), and occurs when P_{fr}-P_{fr} homodimers are present (see Fig.7C).

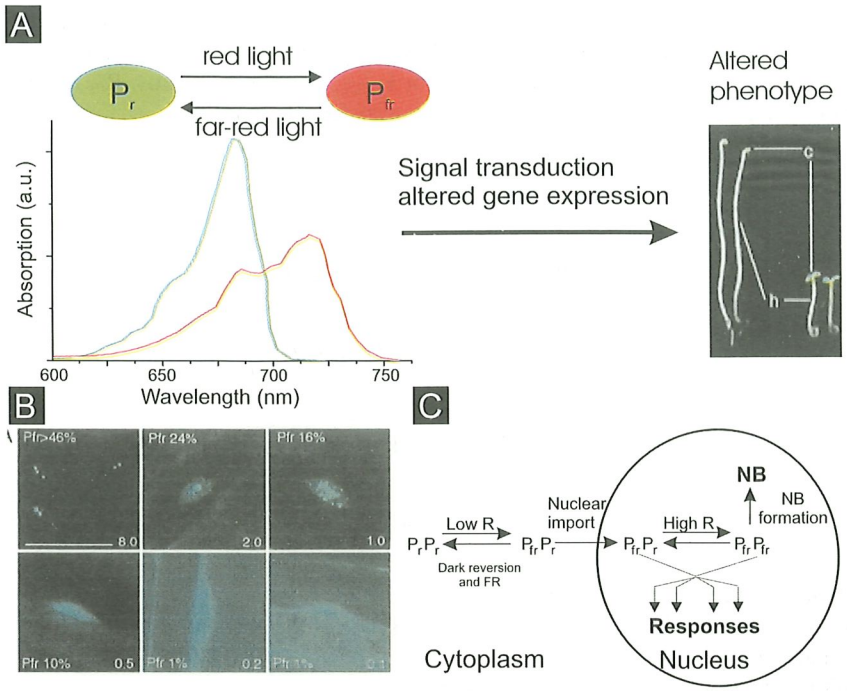


Fig. 7: A) Photoreaction scheme and absorption spectra of the two phytochrome forms, P_r (green) and P_{fr} (red). The picture on the right shows different phenotypes of Pea seedlings grown without (left) and with red light (right) due to phytochrome (P_{fr}) action. B) Translocation of phyB in Arabidopsis seedlings grown under different intensities of red light. phyB accumulates in the nucleus at low (lower right panels) and in nuclear bodies (NB, upper left panels) at high red-light intensities (adapted from [30]). C) Mechanism of phytochrome action. Nuclear import at low red-light intensities (P_r-P_{fr} heterodimers) and NB formation at high red-light intensities (P_{fr}-P_{fr} homodimers; adapted from [30]).

This example illustrates the strength of GFP fusion proteins to monitor signal transduction and regulatory processes in living cells and organisms. If a specific stimulus generates a certain pattern of (subcellular) localisation (here of phytochrome) potential interaction partners may be identified by genetic screening methods.

3.5 GFP Fusion Proteins: Protein-Protein Interaction

Förster Resonance Energy Transfer (FRET, see also [31] and references therein) is a process in which energy is transferred non-radiatively (that is, via long-range dipole-dipole coupling) from a fluorophore in an electronic excited state (donor) to another chromophore (acceptor). The latter may but need not be fluorescent. The transfer rate k_{FRET} varies inversely with the

sixth power of the donor-acceptor distance, r, over the range of 1–10 nm. Such distances are relevant for most biomolecules or their domains engaged in complex formation and conformational transition.
The FRET transfer rate, k_{FRET}, depends on three additional parameters: 1. the overlap of the donor-emission and acceptor-absorption spectra, J; 2. the relative orientation of the donors emission and the acceptors absorption transition moments, κ; 3. the refractive index of the solution, n. The quantitative treatment of FRET originated with Theodor Förster [32] and is embodied in widely disseminated formulas for k_{FRET}, the Förster radius, R_0, and the transfer quantum yield, generally denoted as the energy transfer efficiency, E:

$$k_{FRET} = 1/\tau_D \left(R_0/r\right)^6 \tag{2}$$

$$R_0^6 = a_0 \kappa^2 J n^{-4} \Phi_{flD} = a_0 \kappa^2 J n^{-4} k_{fl} \tau_D \tag{3}$$

$$E = k_{FRET}\tau_{DA} = \left(R_0/r\right)^6 / \left(1+\left(R_0/r\right)^6\right) \tag{4}$$

$$1/\tau_D = k_{fl} + k_{nr} = k_{isc} + k_{pb} \tag{5}$$

$$1/\tau_{DA} = 1/\tau_D + k_{FRET} \tag{6}$$

r - distance between donor and acceptor
τ_D – fluorescence lifetime of the donor in the absence of the acceptor
R_0 - Förster radius
a_0 - unit less constant = $8.8 * 10^{-28}$ (for R_0 in nm)
$J = 10^{17} \int q_D \varepsilon_a \lambda^4 d\lambda$ (in $nm^6 mol^{-1}$)
q_D - normalised donor emission spectrum
ε_a – spectral distribution of the acceptor extinction coefficient (in $cm^{-1} M^{-1}$)
τ_{DA} – fluorescence lifetime of the donor in the presence of the acceptor
k_{FRET} - rate constant of FRET from donor to acceptor
k_{fl} - rate constant of radiative deactivation of the donor excited state
k_{nr} - rate constant of non-radiative deactivation of the donor excited state
k_{isc} - rate constant of intersystem crossing of the donor
k_{pb} - rate constant of photobleaching of the donor

FRET can be determined in many different ways by means of the fluorescence intensity (I), fluorescence lifetime, and the anisotropy. In practice, often, only the change of fluorescence intensity or lifetime of the donor are used to quantify the FRET efficiency.

$$E = I_{DA}/I_D = \tau_{DA}/\tau_D \tag{7}$$

I_D – fluorescence lifetime of the donor in the absence of the acceptor
I_{DA} – fluorescence lifetime of the donor in the presence of the acceptor

This can lead to erroneous results because other rate constants of the excited state relaxation of the donor may change in the presence of the acceptor (for instance due to a quenching

interaction between donor and acceptor). It is advisable to use both, donor and acceptor signals for the quantification of FRET-measurement.
A major problem with cell biological applications of FRET, particularly those involving imaging techniques, is that the reference values, I_D and τ_D are generally unknown and may vary throughout the sample. In part, this is a result of changes in the environment-sensitive rate constant k_{nr}. When using environment-insensitive GFP mutants as donor, this problem can be circumvented. In addition, although the donor-separation distance, r, largely determines whether FRET is observed, one or more of the other parameters defining R_0 may also be of significance. Molecular translocations of the donor-acceptor pair between the cytoplasm and the plasma membrane (changing n) and conformational rearrangements (changing the orientational factor κ) are possible examples. Almost universally, a value 2/3 is used for κ^2.

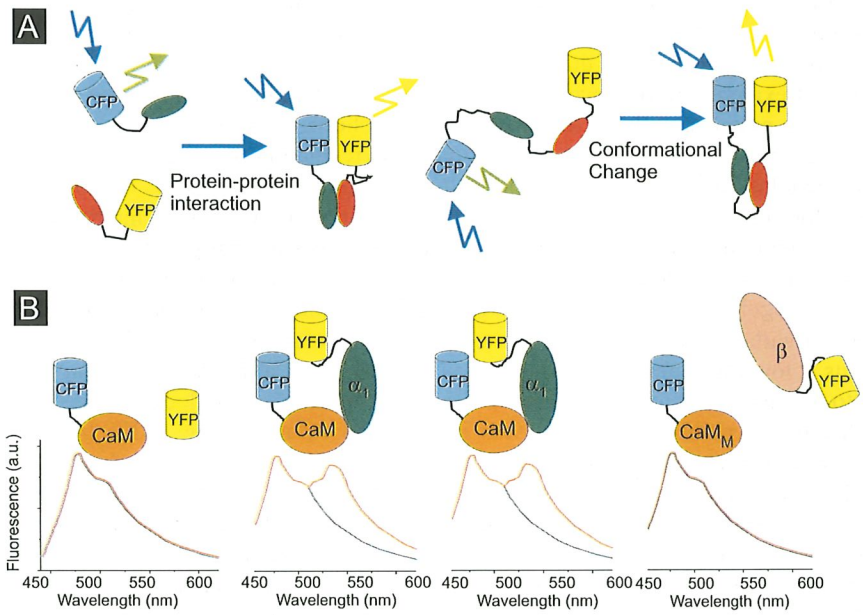


Fig. 8: A) Principle for intermolecular (left) and intramolecular FRET (right) with CFP/YFP as donor/acceptor pair. B) Intracellular FRET measurements of cells expressing Cam-CFP (donor) and several potential acceptors (from left to right): YFP (no FRET); YFP-$\alpha_I$subunit (FRET); YFP-α_I subunit (FRET; with CaM$_M$-CFP, a mutant with no Ca^{2+} binding ability); and YFP-β (no FRET) (adapted from [33]).

Unfortunately, this is valid only if the donor and acceptor molecules are oriented randomly and rotate rapidly and isotropically during the donor excited-state lifetime. Such a condition may often not exist. For example in VFPs, the rotational correlation times are fivefold longer than their fluorescence lifetimes, thereby greatly limiting rotational relaxation. κ^2 is far from 2/3. For a random yet static molecular distribution, the ensemble value of κ is not constant

rather than a function of the donor-acceptor distance, r. It follows that absolute distances determined from VFP-FRET pairs (like CFP/YFP) must be taken with caution.
The majority of the *intracellular* FRET reports use VFPs as donor and acceptor, because of the possibility of genetic labelling of proteins with VFPs. Protein-protein interactions are investigated after the synthesis of two fusion proteins (donor VFP fused to potential interaction partner 1, acceptor VFP fused to potential interaction partner 2) in the same cell (Fig.8A). A FRET signal indicates colocalization of the interaction partners on the 1 – 10 nm length scale. This determination is one to two orders of magnitude more precise than that obtained by conventional optical microscopy. Many different interaction pairs have been successfully studied including multimeric membrane receptors and ion channels, cytosolic proteins with regulatory function, and proteins involved in nodulation of gene expression.
Protein conformational changes can be monitored if donor and acceptor VFP are fused to one and the same protein (Fig.8A). If that protein changes its conformation, the FRET efficiency alters. Often a peptide sequence serving as an interaction (binding) domain is incorporated into the fusion protein. If the binding event of the protein with that domain is dependent on some environmental parameter (as the concentration of an ion or a nucleotide) the fusion protein may then report about that parameter.
An elegant example for the investigation of protein-protein interactions using FRET between VFP-fusion proteins will be discussed in the following section [33]. Among the most intriguing forms of Ca^{2+} channel modulation is the regulation of voltage-gated Ca^{2+} channels by intracellular Ca^{2+}, acting via channel-calmodulin (CaM) interactions. The channels consist of pore-forming α subunits and modulating β subunits arranged in a multimeric complex. CaM is a ubiquitous Ca^{2+}-binding protein which interacts with many proteins (e.g. enzymes or ion channels) and modulates their functional properties. It binds up to four Ca^{2+} ions. Ca^{2+}-dependent regulation of gating in voltage-gated Ca^{2+} channels is mediated by interaction of Ca^{2+}/CaM with the carboxyl tail of the α subunits. However, CaM may already bind to the α subunits at low Ca^{2+} concentrations with no Ca^{2+} bound to CaM. "Preassociation" of Ca^{2+}-free CaM to the channels would then serve to enhance response times to local changes of $[Ca^{2+}]_i$.
Whether this hypothesis is correct was investigated by measuring FRET between several VFP-fusion proteins (see Fig.8B). The fusion protein CaM-CFP was used as a donor. YFP served as the acceptor and was fused to the α and β subunit of the Ca^{2+} channel, respectively. FRET efficiencies were determined from a three cube FRET measurement scheme in which the fluorescence intensity at the acceptor fluorescence maximum after excitation at the donor absorption maximum is used together with two calibration measurements. Figure 8B shows the fluorescence spectra (recalculated from the FRET efficiencies) for the different FRET pairs after excitation of the donor, CaM-CFP. No FRET is observed for CaM-CFP and YFP, as expected. FRET occurs between CaM-CFP and α_1-YFP in resting cells (with low Ca^{2+} concentrations). It indicates a Ca^{2+}-independent association of CaM to the channel. This finding is supported by the observation of FRET between α-YFP and a mutant CaM_M-CFP which can not bind Ca^{2+}. The failure of detecting FRET for CaM-CFP and β-YFP served as a control showing that: 1. the observed FRET between CaM-CFP and α-YFP is not unspecifically caused by a high density of CaM-CFP near the cell membrane; 2. CaM-CFP is not within the vicinity (i.e. 5 nm) of β-YFP. This gives information on the architecture in the channel-CaM complex. In a number of similar FRET studies the stoichiometry and the subunit interactions of other ion channels [34,35] have been investigated. These studies offer information about the organization of multi-protein complexes for which a direct three-dimensional structural information is far or out of reach.

Acknowledgement

I want to thank all my coworkers from the IBI1 in the Research Centre Jülich, in particular Arndt Baumann, Jörg Enderlein, Benjamin Kaupp, and Timo Strünker. I am also grateful to the fruitful collaboration with colleagues from other Universities involved in the work presented here (Hiroshi Kaneko, Stefan Frings (University of Heidelberg), Johnny Hendriks (Free University Amsterdam), Klaas Hellingwerf (University of Amsterdam)).

References

[1] Marriott, G., Clegg, R.M., Arndt-Jovin, D.J., Jovin, T.M. (1991) Time-resolved imaging microscopy. Phosphorescence and delayed fluorescence imaging. Biophys. J. 60, 1374-1387.

[2] Bilski, P., Daub, M.E., Chignell, C.F. (2002) Direct detection of singlet oxygen via its phosphorescence from cellular and fungal cultures. Method Enzym.352, 41-52 Part A.

[3] Mazhul, V.M., Zaitseva, E.M., Shcharbin, D.G. (2000) Intramolecular dynamics and functional activity of proteins. Biofizika 45, 965-989.

[4] Eftink, M.R. (1991) Fluorescence Techniques for Studying Protein Structure. Methods Biochem. Anal. 35, 127-205.

[5] Eftink, M.R. (1991) The use of fluorescence methods to monitor unfolding transitions in proteins. Biochemistry-Moscow 63, 276-284.

[6] Wallace, L.A., Matthews, C.R. (2002) Sequential vs. parallel protein-folding mechanisms: experimental tests for complex folding reactions. Biophys. Chem. 101, 113-131.

[7] Reshetnyak, Y.K., Koshevnik, Y., Burstein, E.A. (2001) Decomposition of Protein Tryptophan Fluorescence Spectra into Log-Normal Components. III. Correlation between Fluorescence and Microenvironment Parameters of Individual Tryptophan Residues. Biophys. J. 81, 1735-1758.

[8] Chen, Y., Barkley, M.D. (1998) Toward Understanding Tryptophan Fluorescence in Proteins. Biochemistry 37, 9976-9982.

[9] Billinton, N., Knight, A.W. (2001) Seeing the Wood through Trees: A Review of Techniques for Distinguishing Green Fluorescent Protein from Endogenous Autofluorescence. Anal. Biochem. 291, 175-197.

[10] Zhang, J., Campbell, R.E., Ting, A.Y., Tsien, R.Y. (2002) Creating new fluorescent probes for cell biology. Nat. Rev. Mol. Cell Biol. 3, 906-918.

[11] Griffin, B.A., Adams, S.R., Jones, J., Tsien, R.Y. (2000) Fluorescent labeling of recombinant proteins in living cells with FlAsH. Method Enzym. 327, 565-578.

[12] Wang, L., Brock, A., Herberich, B., Schultz, P.G. (2001) Expanding the genetic code of Escherichia coli. Science 292, 498-500.

[13] Wang, L., Zhang, Z.W., Brock, A., Schultz, P.G. (2003) Addition of the keto functional group to the genetic code of Escherichia coli. PNAS 100, 56-61.

[14] Chin, J.W., Cropp, T.A., Anderson, J.C., Mukherji, M., Zhang, Z.W., Schultz, P.G. (2003) An expanded eukaryotic genetic code. Science 301, 964-967.

[15] Deiters, A., Cropp, T.A., , Mukherji, M., Chin, J.W., Anderson, J.C., Schultz, P.G. (2003) Adding Amino Acids with Novel Reactivity to the Genetic Code of *Saccharomyces Cerevisiae*. JACS 125, 11782-11783.

[16] Zimmer, M. (2002) Green Fluorescent Protein (GFP): Applications, Structure, and Related Photophysical Behavior. Chem. Rev. 102, 759-781.

[17] Barondeau, D.P., Putnam, C.D., Kassmann, C.J., Tainer, J.A, Getzoff, E.D. (2003) Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. PNAS 100, 12111-12116.

[18] Creemers, T.M.H., Lock, A.J., Subramanian, V., Jovin, T.M., Völker, S. (1999) Three photoconvertible forms of green fluorescent protein identified by spectral hole-burning. Nat. Struct. Biol. 6 557-560

[19] Pellequer, J.L., Wager-Smith, K.A., Kay, S.A., Getzoff, E.D. (1998) Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. PNAS 95, 5884-5890.

[20] Hellingwerf, K.J., Hendriks, J., Gensch, T. (2003) Photoactive Yellow Protein, A New

Type of Photoreceptor Protein: Will This "Yellow Lab" bring us where we want to go? J. Phys. Chem. A *107*, 1082-1094.

[21] Gensch, T., Hellingwerf, K.J., Hendriks, J. (2004) Tryptophan Fluorescence monitors Structural Changes accompanying Signaling State Formation in the Photocycle of Photoactive Yellow Protein. submitted to Photochem. Photobiol. Sci.

[22] Grohmann, L., Blenau, W., Erber, J., Ebert, P.R., Strünker, T., Baumann, A. (2003) Molecular and functional characterization of an octopamine receptor from honeybee (Apis mellifera) brain. J. Neurochem. 86, 725-735.

[23] Frings, S. (2001) Chemoelectrical signal transduction in olfactory sensory neuron of air-breathing vertebrates. Cell Mol. Life Sci. 58, 510-519.

[24] Kaneko, H., Nakamura, T., Lindemann, B. (2001) Noninvasive measurement of chloride cocentration in rat olfactory receptor cells with use of a fluorescent dye. Am. J. Physiol. Cell. Physiol. 280, C1387-1393.

[25] Reuter, D., Zierold, K., Schröder, W.H., Frings, S. (1998) A depolarizing chloride current contributes to chemoelectrical transduction in olfactory sensory neurons in situ. J. Neurosci. 18, 6623-6630.

[26] Kaneko, H., Putzier, V., Frings, S., Gensch, T. (2002) Determination of intracellular chloride concentration in dorsal root ganglion neurons by fluorescence lifetime imaging. Curr. Top. Membr. 53, 167-189.

[27] Kaneko, H., Putzier, I., Frings, S., Kaupp, B., Gensch, T. (2004) Chloride accumulation in mammalian olfactory sensory neurons. in preparation

[28] Quail, P.H. (2002) Phytochrome photosensory signalling networks. Nat. Rev. Mol. Cell Biol. 3, 85-93.

[29] Nagy, F., Schäfer E. (2002) Phytochromes control photomorphogenesis by differentially regulated, interacting signaling pathways in higher plants. Annu. Rev. Plant Biol. 53, 329-355

[30] Chen, M., Schwabb, R., Chory, J. (2003) Characterization of the requirements for localization of phytochrome B to nuclear bodies. PNAS 100, 14493-14498

[31] Jares-Erijman, E.A., Jovin, T.M. (2003) FRET Imaging. Nat. Biotechnol. 21, 1387 – 1395.

[32] Förster, T. (1965) Delocalized excitation and excitation transfer. in Modern Quantum Chemistry Part III: Action of Light and Organic Crystals (ed. Sinanoglu, O.) 93–137 (Academic Press, New York, 1965) oder

[33] Erickson, M.G., Alseikhan, B.A., Peterson, B.Z., Yue, D.T. (2001) Preassociation of Calmodulin with Voltage-Gated Ca^{2+}Channels Revealed by FRET in Single Living Cells. Neuron 31, 973–985.

[34] Zheng, J., Varnum, M.D., Zagotta, W.N. (2003) Disruption of an intersubunit interaction underlies Ca2+-calmodulin modulation of cyclic nucleotide-gated channels. J. Neurosci. 23, 8167-8175.

[35] Biskup, C., Zimmer, T., Benndorf, K. (2004) FRET between cardiac Na^+ channel subunits measured with a confocal microscope and a streak camera. Nat. Biotechnol. *22*, 220-224.

B 8 Exploring the inner space of cells by cryoelectron-tomography

W. Baumeister, M. Beck, A. Briegel, M. Cyrklaff,
Th. Keil, J. Kürner, V. Lucic, O. Medalia,
St. Nickell, and J. Plitzko
Max-Planck-Institute of Biochemistry
Martinsried

Summary

Electron Tomography (ET) is uniquely suited to obtain three-dimensional (3D) images of large pleiomorphic structures, such as supramolecular assemblies, organelles, or even whole cells. While the principles of ET have been known for decades, its use has gathered momentum only in recent years. Technological advances (namely computer controlled transmission electron microscopes and large area CCD cameras) have made it possible to develop automated data acquisition procedures. This, in turn, allowed to reduce the total electron dose to levels low enough for studying radiation sensitive biological materials embedded in vitreous ice. As a result, we are now poised to combine the power of high-resolution 3-D imaging with the best possible preservation of the specimen [1,2]. In the past, ET has mainly been used to examine thin sections of plastic-embedded materials. This approach has provided valuable insights into cellular architecture, but it falls short of revealing the macromolecular organization inside cells. Chemical fixation, staining with contrast enhancing heavy atom compounds, dehydration and plastic embedding affect the specimen significantly and make the interpretation of such tomograms at the molecular level very problematic, if not impossible. The use of high-pressure freezing instead of chemical fixation improves specimen preservation significantly but it does not eliminate the problems arising from the intricate interactions between heavy atom stains and molecular structures. Obviously, cryo-sectioning of vitrified material is the method of choice for large (> 1 μm) objects (cells, tissues), but it remains a challenging task despite recent progress [3]. It is often possible to isolate macromolecular complexes, virus particles or organelles using procedures that maintain their structural integrity. Such nanoscale specimens are suitable for direct analysis by ET [4,5]. The resolution obtained allows the docking of high resolution component structures obtained by X-ray crystallography or NMR. Hybrid approaches of this kind can be used to generate pseudoatomic maps of assemblies that are too large or variable for direct high-resolution structural studies. ET of frozen-hydrated whole prokaryotic cells or thin eukaryotic cells grown directly on EM girds provides 3-D images of macromolecular structures unperturbed and in their functional environment [6,7]. Currently resolution is limited to 4-6 nm but with instrumental advances, such as liquid He cooling and CCD cameras optimized for intermediate voltage TEMs, we are now entering the realm of molecular resolution (2-4 nm) [8]. High resolution tomograms of organelles or cells contain vast amount of information; essentially they are 3-D images of the cell's entire proteome and they should ultimately enable us to map the spatial relationships of macromolecules in a cellular context, the 'interactome'. However, it is no trivial task to retrieve this information because of the poor signal-to-noise ratio of such tomograms and the crowded nature of the cytoplasm and many organelles. Denoising procedures can help to combat noise and to facilitate visualization, but advanced pattern recognition methods are needed for detecting and identifying with high fidelity specific macromolecules based on their structural signature (size and shape, for example). Provided that high- or medium-resolution structures of the molecules of interest are available, they can be used as templates for a systematic interrogation of the tomograms [9,10]. Experiments with phantom cells, i.e. lipid vesicles encapsulating a known set of proteins have shown that such template-matching approach is feasible. Once the challenges of obtaining sufficiently good resolution and of creating efficient data-mining algorithms are met, and comprehensive libraries of template structures become available, we will be able to map the supramolecular landscape of cells systematically and thereby provide a new perspective for the function of cellular systems [11].

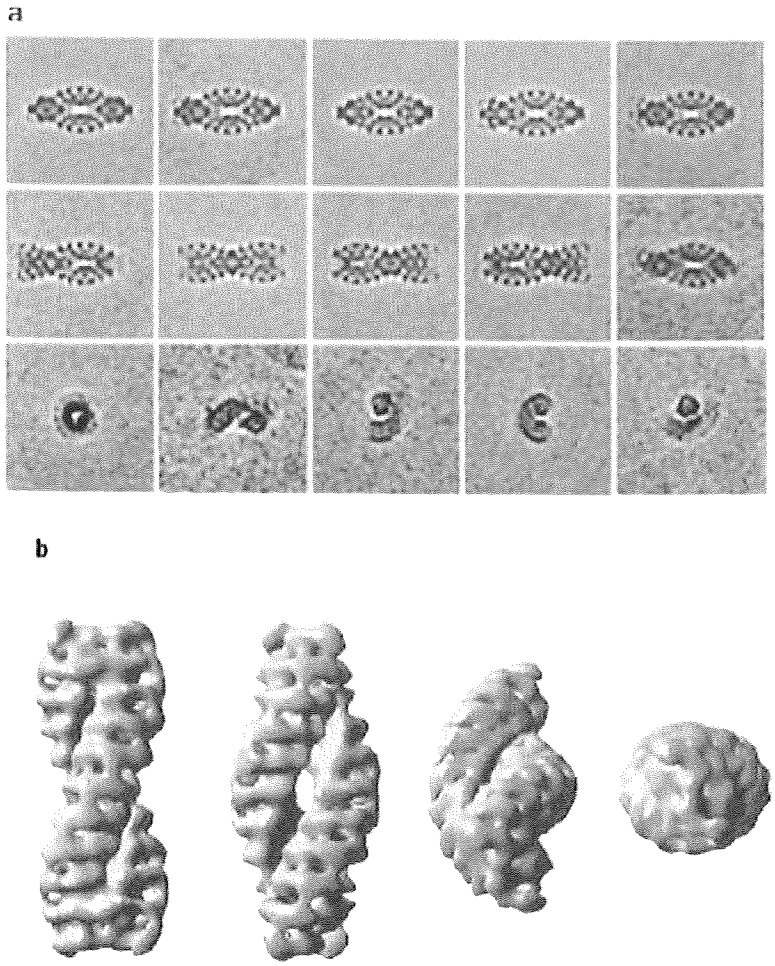


Fig. 1: *Representative example that illustrates the type of 3D reconstructions that can be obtained with large macromolecular complexes by single-particle cryo-EM. In this example, the specimen is a giant assembly of Drosophila melanogaster tripeptidyl peptidase II (TPP II). The protein monomers have a relative molecular mass of* $150,000$ *($M_r = 150K$) and the intact assembly has a particle mass of* $\sim 6,600K$. *(a) Some of the distinct views that are obtained by averaging many equivalent projections of individual particles randomly oriented within a thin film of vitreous ice. A wide variety of side views can be distinguished, corresponding to rotation of the particle around its long axis. In addition, other projections, shown in the bottom row, correspond to particles that are viewed directly on axis or at a small tilt relative to the axial view. (b) 3D surface representation of the TPP II complex at 3.3-nm resolution, in which the particle is first rotated about its long axis and then it is tilted to bring the long axis perpendicular to the page [11].*

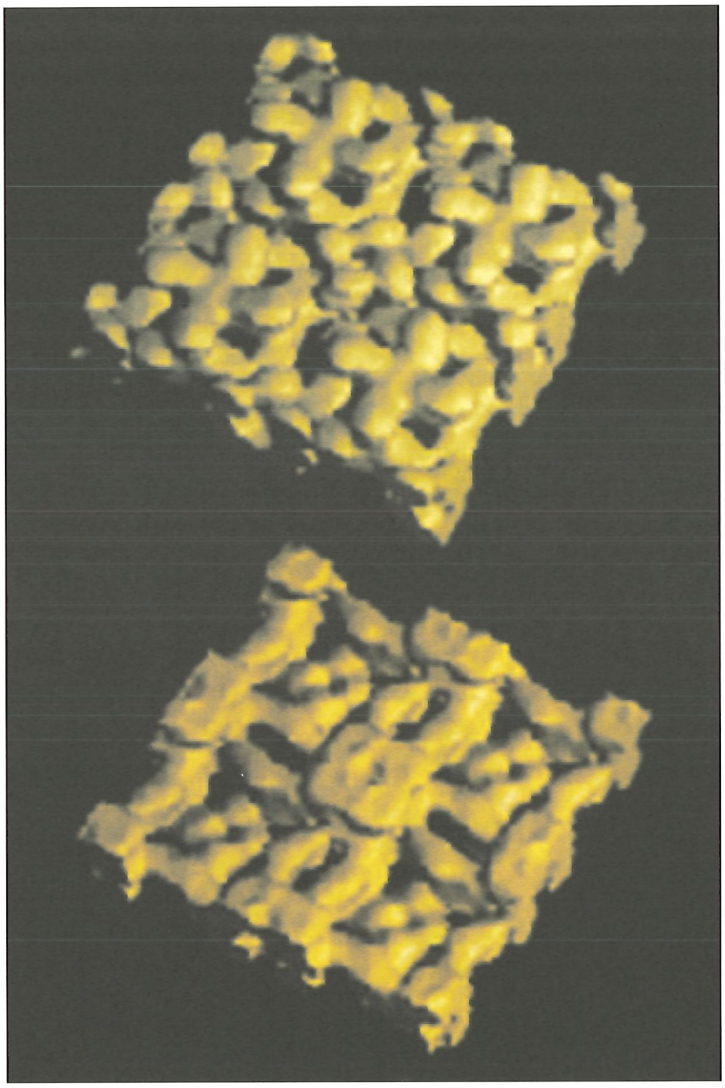


Fig. 2: *S-layers are the outermost component of the cell wall of many bacteria and most of archaea. S-layer proteins form natural two-dimensional protein crystals covering the cell completely and conferring stability in addition to other structures of the cell envelope.*
S-layers occur in almost all phylogenetic branches and represent an important cell envelope structure which protects the cell from larger protein complexes (lytic enzymes) or even from predator attack by Bdellovibrio bacteriovorans. Many S-layers are molecular sieves with pores of about 2 to 4 nm in size and they define a quasi-periplasmic space in Archaea.
The figure shows a 3D reconstruction of the S-layer from the Gram-positive bacterium Sporosarcina ureae, viewed from the outer surface (top) and the inner surface (bottom) which is in contact with the peptidoglycan. The distance between periodic tetrameric structures is 12.9 nm.

References

[1] A. Koster, R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz, and W. Baumeister, J. Struct. Biol. **120** (1997), p. 276.

[2] W. Baumeister, R. Grimm and J. Walz, Trends Cell Biol. **9** (1999), p. 81.

[3] W. Baumeister, Curr. Opin. Struct. Biol. **12** (2002), p. 679.

[4] M. Nitsch, J. Walz, D. Typke, M. Klumpp, L.-O. Essen and W. Baumeister, Nat. Struct. Biol. **5** (1998), p. 855.

[5] K. Grünewald, P. Desai, D.C. Winkler, J.B. Heymann, D.M. Belnap, W. Baumeister and A.C. Steven, Science **302** (2003), p. 1396.

[6] R. Grimm, H. Singh, R. Rachel, D. Typke, W. Zillig and W. Baumeister, Biophys. J. **74** (1998), p. 1031.

[7] O. Medalia, I. Weber, A.S. Frangakis, D. Nicastro, G. Gerisch and W. Baumeister, Science **298** (2002), p. 1209.

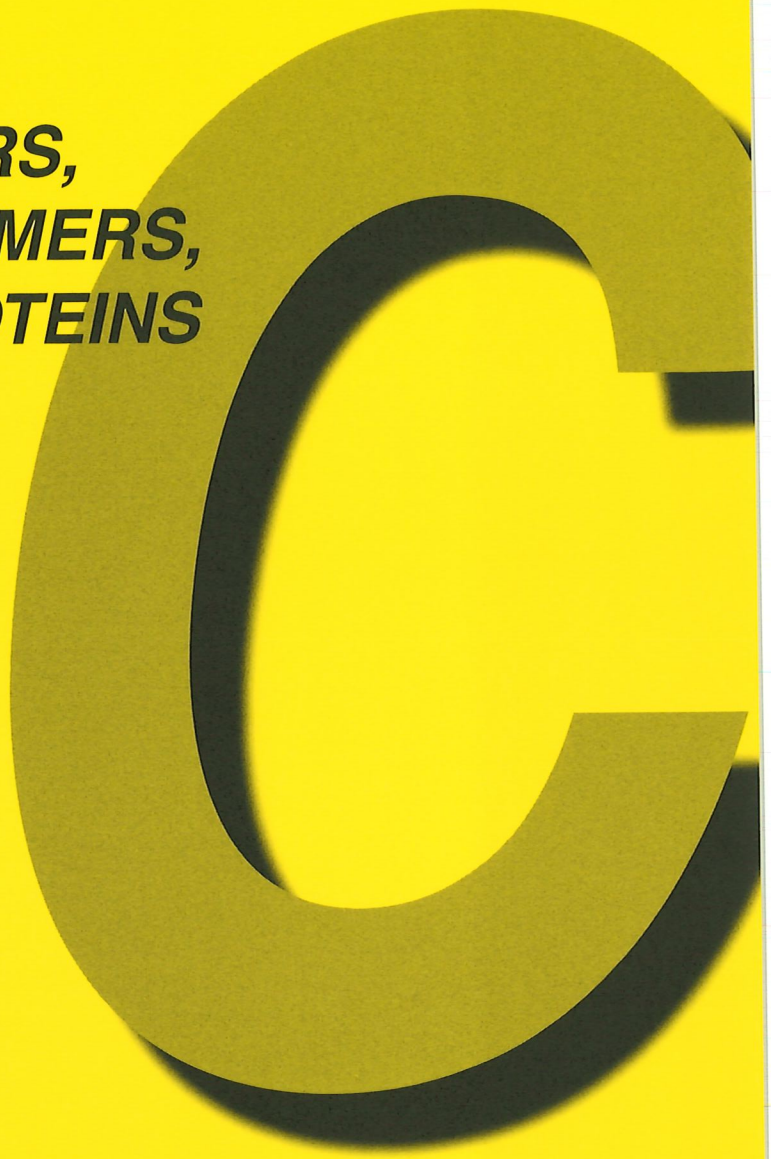
[8] J. Plitzko, A.S. Frangakis, S. Nickell, F. Förster, A. Gross and W. Baumeister, Trends Biotechnol. **20** (2002), p. S40.

[9] J. Böhm, A.S. Frangakis, R. Hegerl, S. Nickell, D. Typke and W. Baumeister, P. Natl. Acad. Sci. USA **97** (26) (2000), p. 14245.

[10] A. S. Frangakis, J. Böhm, F. Förster, S. Nickell, D. Nicastro, D. Typke, R. Hegerl and W. Baumeister, P. Natl. Acad. Sci. USA **99** (2002), p. 14153.

[11] A. Sali, R. Glaeser, T. Earnest and W. Baumeister, Nature **422** (2003), p. 216.

POLYMERS, BIOPOLYMERS, AND PROTEINS



C 1 Flexible Polymers

H. Frielinghaus
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Chain Models 2
2.1 The Freely Jointed Chain . . . 3
2.2 The Freely Rotating Chain . . . 6
2.3 The Kuhn Chain . . . 8
2.4 Semiflexible Chains . . . 9
2.5 Summary of Chain Models . . . 11
2.6 Distribution of the End-to-End Distance . . . 12
2.7 The Gaussian Chain . . . 14
2.8 The Continuous Description . . . 15

3 Phenomena of Polymers 17
3.1 Polymer near a hard wall . . . 17
3.2 Rubber elasticity . . . 19
3.3 Scattering of chains . . . 22
3.4 Deviations – The Flory Exponent . . . 24
3.5 Semidilute Solutions – The Blob Concept . . . 27

4 Conclusion 29

A Appendix: Theorem of Lagrange 31

1 Introduction

Polymers occur everywhere – in nature and industrial products. Examples are the DNA, wood, silk, synthetic polymers, often called plastics, fibres and gels. In biology the conformation of many molecules plays an important role, say of proteins or the DNA, which assigns to the molecule a specific function. Arrangements of proteins may appear as a polymeric structure on a larger scale, again for specific tasks. For technological applications the easy tailoring of polymeric materials is important. Finally, technical/synthetic plastics are irreplaceable in everyday life. Examples are housing of electronics, wire isolation, glue, rubber or medical materials. Some of them are evidence of high-tech products.

Many synthetic polymers are rather simple. They consist of a single unit which is chemically lined up (or chained) to form a linear polymer. One example of such a unit is the monomer ethylene (Fig. 1). The double bond gives rise to chemical reactions with other molecules of the same species. If only two monomers have reacted a dimer is formed. If the reaction can be held up, a polymerization process takes place, and finally a long chain molecule is formed. Whereas the molecules look straight on the paper, they are not in reality, however. Many conformations are possible, which will be the subject of this lecture.

Additionally, polymers have an unconventional phase diagram. There is no gas phase. The "liquid" phase appears to be highly viscous. But besides the viscous contributions there are also elastic properties. These become especially obvious for rubber. Polymers also never crystallize completely. For many polymers the glassy state seems to be preferred. Besides the neat polymers, solutions of polymers also show a rich variety of properties. The large number of degrees of freedom per molecule is often responsible for the unfailing nature of polymers.

Polymers in Biology differ from synthetic plastics. Often there is a larger variety of monomers building up the chain molecule (DNA for example). The chain structure usually becomes visible on larger scales – for DNA beyond the helix structure. Proteins (and DNA) are complicated due to the folding, which leads to different hierarchial stages. Some of these stages might appear as simpler polymeric structures (chromatin loops). Finally, special proteins make up the cytoskeleton by lining up to long filaments. The diameters of such filaments lie in the range of 7-25nm. These filaments are responsible for cell stability, motion, and transport of material within the cell (see contribution of A. Bausch C.9). In the living cell the polymerization and depolymerization can be a continuous process. Through stabilization mechanisms the structures are highly resistent. In many cases the properties of filaments on larger length scales can be described by two parameters – the elasticity and the contour length. The coarse graining averages over the elementary properties of the microscopic pattern.

In chapter 2 different chain models are developed to describe conformations of polymers. The first model is rather simple and refinements allow for more realistic modelling. Chapter 3 describes more phenomena of polymers where the developed concepts of microscopic conformations are used. The underlying theories and concepts are a collection of different textbooks [1–9].

2 Chain Models

The simplest synthetic polymer is polyethylene. While we already described the polymerization process, we now want to consider the conformation of such a chain. The main task will be to describe the position of the carbon atoms – or more generalized the arrangement of the

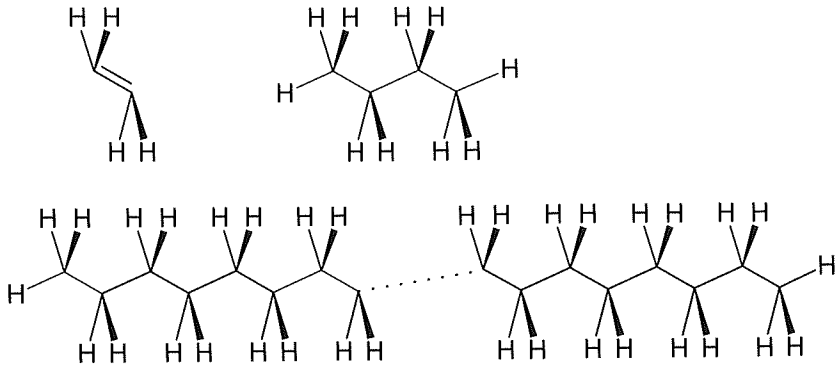


Fig. 1: *The steps from a monomer to a polymer. (top left) the monomer ethylene – actually flat, (top right) the saturated dimer, (bottom) the polymer polyethylene.*

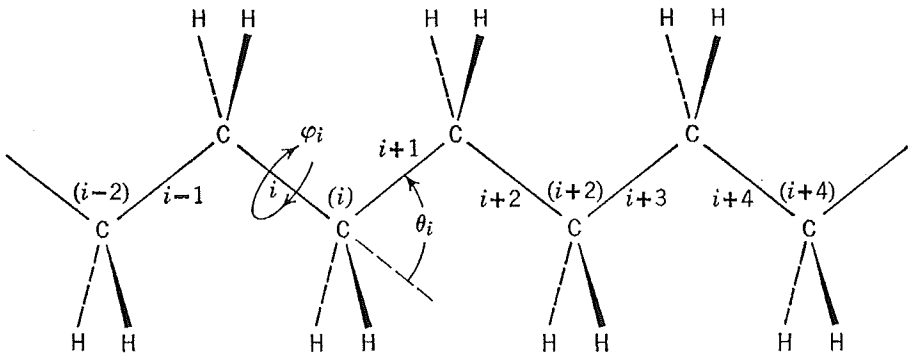


Fig. 2: *The polyethylene chain with definitions of angles: θ_i bond angle, φ_i torsional angle.*

polymer backbone. Each backbone atom is numbered by the index i. There already exist some restrictions due to the bond angles θ_i and the torsional angles φ_i. During the refinement of the models the properties become more and more realistic. Finally, the positions of the hydrogen atoms – or of larger side groups – give rise to steric hindrance interactions and are considered in the section of semiflexible polymers.

2.1 The Freely Jointed Chain

This particular chain is described by N consecutive backbone bonds, which do not have any restrictions [1, 10]. This means that the bond length is fixed (ℓ), but the angle between consecutive bonds is not restricted at all – restrictions due to the bond angles are neglected. The path of this backbone resembles that of a random walk. A drunken man on his way home would follow a similar path since he does not remember which way he took a minute ago. Therefore, there is an obvious link to diffusion of particles already now, but we will discuss this later (see section 2.8).

One observable of a chain is the end-to-end distance. It simply can be described by the sum of all bond vectors (see Fig. 3) connecting two consecutive monomers ($\vec{r}_i \equiv \vec{r}_{i-1,i}$; the generalized

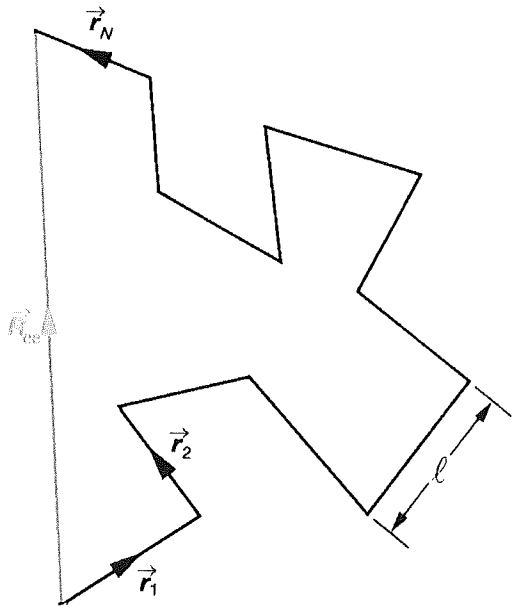


Fig. 3: *A sketch of a freely jointed chain. The bond vector has a constant length ℓ, but the angle is totally free. The end-to-end distance $\vec{R}_{ee}$ is described by the sum of all bond vectors.*

vector $\vec{r}_{i,j}$ connects two monomers i and j, which is shown in Fig. 4)

$$\vec{R}_{ee} = \sum_{i=1}^{N} \vec{r}_i \tag{1}$$

Statistically, the average of the end-to-end distance is simply zero. This is true since the free chain does not have a preferential orientation, and this means that there is no orientation of any vector. This applies for every single bond vector and the end-to-end vector.

$$\left\langle \vec{R}_{ee} \right\rangle = \vec{0} \tag{2}$$

The equation above describes the first moment of the end-to-end distance distribution (see also section 2.6). The next moment is the second moment $\langle \vec{R}_{ee}^2 \rangle$. But before averaging the end-to-end distance, some rearrangements of the terms should be made. The square of the end-to-end distance is a scalar, and a square of a sum contains all possible pairs of products.

$$\vec{R}_{ee}^2 = \vec{R}_{ee} \cdot \vec{R}_{ee} = \sum_{i,j=1}^{N} \vec{r}_i \cdot \vec{r}_j \tag{3}$$

Those pairs can be sorted for diagonal terms (r_i^2) and off-diagonal terms, where the indexing even can be sorted.

$$\vec{R}_{ee}^2 = \sum_{i=1}^{N} \vec{r}_i^{\,2} \; + \; 2 \cdot \sum_{1 \le i < j \le N} \vec{r}_i \cdot \vec{r}_j \tag{4}$$

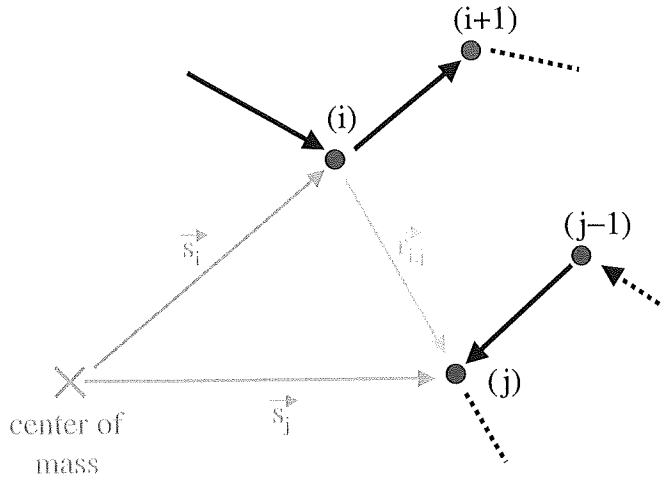


Fig. 4: *The definition of the vectors needed in the radius of gyration calculation. The vectors $\vec{s}_i$ describe the position of the monomer i relatively to the center of mass. The vectors $\vec{r}_{i,j}$ connect the monomers i and j, and therefore describe a sub-chain.*

The averaging of these two different sums yields a simple result: The diagonal terms consider the squared length of a single bond (ℓ^2), the summation of which simply adds up N times the same number. The other contribution considers the correlation between different bonds. Our simple model neglects any correlations for now and the term vanishes.

$$\left\langle \vec{R}_{ee}^2 \right\rangle = \sum_{i=1}^{N} \left\langle \vec{r}_i^{\,2} \right\rangle \; + \; 2 \cdot 0 \; = N\ell^2 \tag{5}$$

The characteristic size of the polymer chain is given by the square root of the second moment. There is no preferential orientation – but the polymer needs some space.

$$\sqrt{\left\langle \vec{R}_{ee}^2 \right\rangle} = \sqrt{N}\,\ell \tag{6}$$

The for polymers characteristic dependence of the average end-to-end distance $\langle \vec{R}_{ee}^2 \rangle^{1/2}$ on the degree of polymerization N occurs already for this over-simplified model. The average chain size is proportional to $\sqrt{N}$ only and therefore is much smaller than the full extension. This is due to the large number of degrees of freedom of such a molecule.
Another frequently used observable is the radius of gyration. It measures the size of a molecule in the manner of a moment of inertia. The vectors of all monomers relatively to the center of mass are squared (assuming the same weight of all monomers, see Fig. 4). This sum is normalized by the number of monomers ($N+1$). The radius of gyration therefore measures the average extension of a chain relatively to the center of mass.

$$\vec{R}_g^2 = \frac{1}{N+1} \sum_{i=0}^{N} \vec{s}_i^{\,2} \tag{7}$$

The Theorem of Lagrange allows for the rearrangement of this sum (Appendix A). Now all vectors connecting monomers i and j occur. Note that the normalization is $(N+1)^2$ due to the

larger amount of addends now.

$$\vec{R}_g^2 = \frac{1}{(N+1)^2} \sum_{0 \le i < j \le N} \vec{r}_{i,j}^2 \tag{8}$$

This would not help much, unless the connecting vectors can be interpreted as end-to-end vectors of chain parts. The single connecting vector is a sum of bond vectors

$$\vec{r}_{i,j} = \sum_{k=i+1}^{j} \vec{r}_k \tag{9}$$

and in the averaging (of the radius of gyration) the square of the connecting vector yields simply $(j-i)\ell^2$.

$$\left\langle \vec{R}_g^2 \right\rangle = \frac{1}{(N+1)^2} \sum_{0 \le i < j \le N} (j-i)\, \ell^2 \tag{10}$$

$$= \frac{\ell^2}{(N+1)^2} \sum_{j=1}^{N} \sum_{k=1}^{j} k \tag{11}$$

$$= \frac{\ell^2}{(N+1)^2} \sum_{j=1}^{N} \frac{j \cdot (j+1)}{2} \tag{12}$$

$$= \frac{N \cdot \ell^2}{6} \cdot \frac{N+2}{N+1} \quad \xrightarrow[N \gg 1]{} \quad \frac{N \cdot \ell^2}{6} \tag{13}$$

The term $(j-i)$ describes a length of a sub-chain, and all possible combinations have to be summed up. This summation can be sorted by summing only terms within a chain of length j (instead of N). In this chain all possible lengths k of sub-chains are summed up. The inner sum is an arithmetic sum. The outer sum can be determined by mathematical tables or algebraic programs (Maple, Mathematica, ...). In the limit of long chains the second factor of eq. 13 reduces to unity. Therefore there is a fixed ratio between the squares of the radius of gyration and the end-to-end distance.

$$\left\langle \vec{R}_g^2 \right\rangle = \frac{\left\langle \vec{R}_{ee}^2 \right\rangle}{6} \tag{14}$$

The radius of gyration is smaller since it measures relatively to the center of mass, whereas the end-to-end distance is a measure for the whole chain.
In the calculation of the radius of gyration we used the argument that sub-chains behave similar to the whole chain (see eq. 9), except that the chain length is just $(j-i)$. This means that the chain has some self-similarity when looking inside the chain. Scattering experiments (sections 3.3 and 3.4) will display this behavior. A chain is a fractal object therefore.

2.2 The Freely Rotating Chain

The freely rotating chain model accounts for the fact that consecutive carbon-carbon-bonds enclose a bond-angle θ typical for a tetrahedron (see Fig. 2). This is due to the sp^3-hybridization of the carbon atom. The torsional angle φ is still assumed to be free. The effect of steric

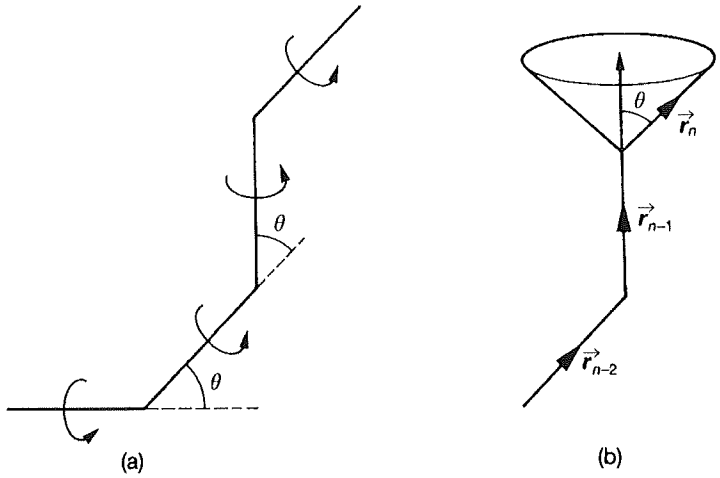


Fig. 5: *A sketch of a freely rotating chain. The bond angle θ is fixed, whereas the torsional angle is free.*

hindrance of hydrogens or even side groups will be discussed later. The resulting chain model is depicted in Fig. 5. It is directly obvious that the projection of two successive bonds is:

$$\langle \vec{r}_i \cdot \vec{r}_{i+1} \rangle = \ell^2 \cos\theta \tag{15}$$

In detailed calculations [10] it can be shown that the correlation between bond vectors of larger distances decays in the following way:

$$\langle \vec{r}_i \cdot \vec{r}_j \rangle = \ell^2 (\cos\theta)^{|j-i|} \tag{16}$$

For the already derived formula of the end-to-end distance (compare eq. 4) the second addend has to be taken into account now and is finite.

$$\left\langle \vec{R}_{ee}^2 \right\rangle = \sum_{i=1}^{N} \langle \vec{r}_i^{\,2} \rangle + 2 \cdot \sum_{1 \le i < j \le N} \langle \vec{r}_i \cdot \vec{r}_j \rangle \tag{17}$$

$$= N\ell^2 + 2\ell^2 \sum_{1 \le i < j \le N} (\cos\theta)^{j-i} \tag{18}$$

$$= N\ell^2 + 2\ell^2 \sum_{k=1}^{N-1} (\cos\theta)^k (N-k) \tag{19}$$

The decay of correlations can be directly inserted and the power series can directly be executed. The number k tells us how long a considered sub-chain is, and the factor $(N-k)$ accounts for the number of possibilities where a sub-chain has the first segment. The remaining sum is a modified power series and again can be determined by mathematical tables or algebraic programs.

$$\left\langle \vec{R}_{ee}^2 \right\rangle = N\ell^2 \left[\frac{1+\cos\theta}{1-\cos\theta} - \frac{2\cos\theta}{N} \cdot \frac{1-(\cos\theta)^N}{(1-\cos\theta)^2} \right] \tag{20}$$

$$\xrightarrow[N\gg 1]{} N\ell^2 \left[\frac{1+\cos\theta}{1-\cos\theta} \right] \tag{21}$$

While the full result looks overwhelming, we first consider the case of long chains. Then the second term of eq. 20 can be neglected. Later we will also consider shorter chains (section 2.4) where the lower flexibility plays an important role. Comparing eq. 21 with 5 we see, that the principal dependence of the end-to-end distance on the bond number N is conserved. The characteristic ratio of the freely rotating chain is defined for large bond numbers relatively to the freely jointed chain:

$$C_\infty = \lim_{N\to\infty} \left\langle \vec{R}_{ee}^2 \right\rangle / (N\ell^2) \tag{22}$$

And for the case of a freely rotating chain we simply obtain:

$$C_\infty = \frac{1+\cos\theta}{1-\cos\theta} \tag{23}$$

For tetrahedral bond angles ($\theta = 70.53°$) the characteristic ratio is $C_\infty = 2$. This is still an idealized result. Polymeric sulphur might come close to the idealization. For real polymers the characteristic ratio can be defined by experiments, and one typically obtains values for C_∞ in the range of 4 to 10 [11, 12]. Larger values correspond to a higher stiffness. The reason for this result is a steric interaction of hydrogens in the case of polyethylene or even larger side groups. Models exist which take this steric hindrance into account and obtain quite good results for the experimentally observed C_∞ values [1, 10]. For our purpose we can argue that the steric hindrance leads to a higher stiffness, which leads to a slower decay of the correlations than the freely jointed chain (eq. 16). The argument $\cos\theta$ has to be substituted then. In section 2.4 we will introduce a rigidity as a coarse grained parameter to describe the slower decay of correlations. Note that the characteristic ratio C_∞ bases on the microscopic bond with a typical length of 1.54Å and N counts the number of chemical bonds in the molecule.

2.3 The Kuhn Chain

We have argued that the correlations between bonds of longer distances $|j-i|$ decay in the manner of equation 16. This means that sub-chains can define an effective segment (the Kuhn segment, see Fig. 6) with negligible correlations.

$$\left\langle \vec{R}_{ee}^2 \right\rangle_K = \sum_{i=1}^{N_K} \langle \vec{r}_i^{\,2} \rangle_K + 2 \cdot \sum_{1\le i<j\le N_K} \langle \vec{r}_i \cdot \vec{r}_j \rangle_K \tag{24}$$

$$= N_K \ell_K^2 \tag{25}$$

The segment length of the Kuhn chain is not a constant anymore, but the average length can be given by ℓ_K. The Kuhn bond number is called N_K. Now, to relate the new effective parameters ℓ_K and N_K with microscopic models we need to find the model-independent observables. One of them is the contour length (given for either model):

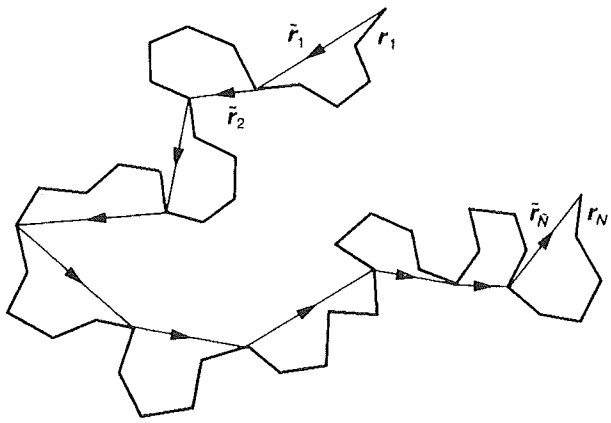


Fig. 6: *The Kuhn chain embraces several segments to form a single effective segment. This effective segment has lost all correlations with respect to the others.*

$$L = N_K \ell_K = N\ell \tag{26}$$

The other one is the end-to-end distance:

$$\left\langle \vec{R}_{ee}^2 \right\rangle_K = L\ell_K = C_\infty L\ell \tag{27}$$

From the model-independent observables one directly can solve for the Kuhn chain defining parameters, and one obtains:

$$\ell_K = C_\infty \ell \qquad \text{and} \qquad N_K = N/C_\infty \tag{28}$$

In literature [2] there exist also other expressions.[1] Here, the basis model is the freely rotating chain, and the full extension of this chain model is taken as the model-independent observable. The slightly corrected Kuhn chain parameters read then $\ell_K = \sqrt{1.5}\, C_\infty \ell$ and $N_K = N/(1.5\, C_\infty)$. This example demonstrates that the Kuhn model depends on the microscopic definition of a contour length. Whenever possible, theoreticians like to take the end-to-end distance as the only model-independent observable. On the other hand, the Kuhn segment length provides a good estimate over which distance correlations between bond vectors are lost. We will therefore use the Kuhn length in the following chapters, and some results actually will be independent on the definition of a contour length (see section 3.1).

2.4 Semiflexible Chains

Real chains have finite length and the result of eq. 20 has to be discussed in detail. The mentioned result bases on a microscopic model (freely rotating chain). Another model [13] considers the chain in a coarse grained manner. The decay of correlation of tangential vectors over a contour length segment ℓ is then described by:

[1] There exist even more effective segment lengths, which leave the number of segments N constant instead of the contour length. Those models just compare a freely rotating chain with a freely jointed chain with identical N. Scattering experiments have access to this number as we will see in section 3.3 (especially eq. 83).

$$\langle \cos\theta \rangle = \exp(-\alpha) \tag{29}$$

The parameter α is then given by the coarse grained rigidity κ:

$$\alpha = \ell \frac{k_B T}{\kappa} \tag{30}$$

Even though the coarse grained model allows to calculate the end-to-end distance in a consistent way, we just substitute the $\cos\theta$ terms and obtain for small α (the product αN is not necessarily small):

$$\left\langle \vec{R}_{ee}^2 \right\rangle = N\ell^2 \left[\frac{1+\exp(-\alpha)}{1-\exp(-\alpha)} - \frac{2\exp(-\alpha)}{N} \cdot \frac{1-\exp(-\alpha N)}{(1-\exp(-\alpha))^2} \right] \tag{31}$$

$$\xrightarrow[\alpha \ll 1]{} N\ell^2 \left[\frac{2}{\alpha} - \frac{2}{N} \cdot \frac{1-\exp(-\alpha N)}{\alpha^2} \right] \tag{32}$$

$$= 2\left[N_p \ell_p^2 - \ell_p^2 (1-\exp(-N_p)) \right] \tag{33}$$

The last line involves the persistence length ℓ_p and the corresponding degree of polymerization N_p. The definitions are:

$$\ell_p = \ell/\alpha = \frac{\kappa}{k_B T} \qquad \text{and} \qquad N_p = \alpha N \tag{34}$$

Depending on the rigidity κ, the parameter α can be large or small, and correspondingly N_p is large or small. The limits correspond to a flexible chain or a rod, and the end-to-end distance becomes for these limits:

$$\left\langle \vec{R}_{ee}^2 \right\rangle \xrightarrow[N_p \gg 1]{} 2N_p\ell_p^2 \qquad \text{and} \qquad \left\langle \vec{R}_{ee}^2 \right\rangle \xrightarrow[N_p \ll 1]{} 2N_p^2\ell_p^2 \tag{35}$$

The flexible chain limit has the typical proportionality to N_p and ℓ_p^2. The rod-like limit is proportional to the squared contour length $\propto (N_p\ell_p)^2$. Examples in biology are found for different filaments of the cytoskeleton or the neuronal network (see Fig. 7 and reference [14]). Microtubule are rather stiff, and therefore N_p is rather small. The considered actin filaments are of intermediate stiffness, whereas neurofilaments appear to be highly flexible, which implies a large N_p. Here, the coarse graining means that the properties of the microscopic filament units appear only in the manner of an averaged flexibility.
Another option to express the behaviour of a semiflexible chain is using the Kuhn length and degree of polymerization. Since we declared the Kuhn model to be the most useful we will continue in this notation. The end-to-end distance then is:

$$\left\langle \vec{R}_{ee}^2 \right\rangle = N_K\ell_K^2 - \tfrac{1}{2}\ell_K^2(1-\exp(-2N_K)) \tag{36}$$

The basic units are connected by:

$$\ell_K = 2\ell_p \qquad \text{and} \qquad N_K = \tfrac{1}{2}N_p \tag{37}$$

The limits for flexible chains and rigid rods read then:

$$\left\langle \vec{R}_{ee}^2 \right\rangle \xrightarrow[N_K \gg 1]{} N_K\ell_K^2 \qquad \text{and} \qquad \left\langle \vec{R}_{ee}^2 \right\rangle \xrightarrow[N_K \ll 1]{} N_K^2\ell_K^2 \tag{38}$$

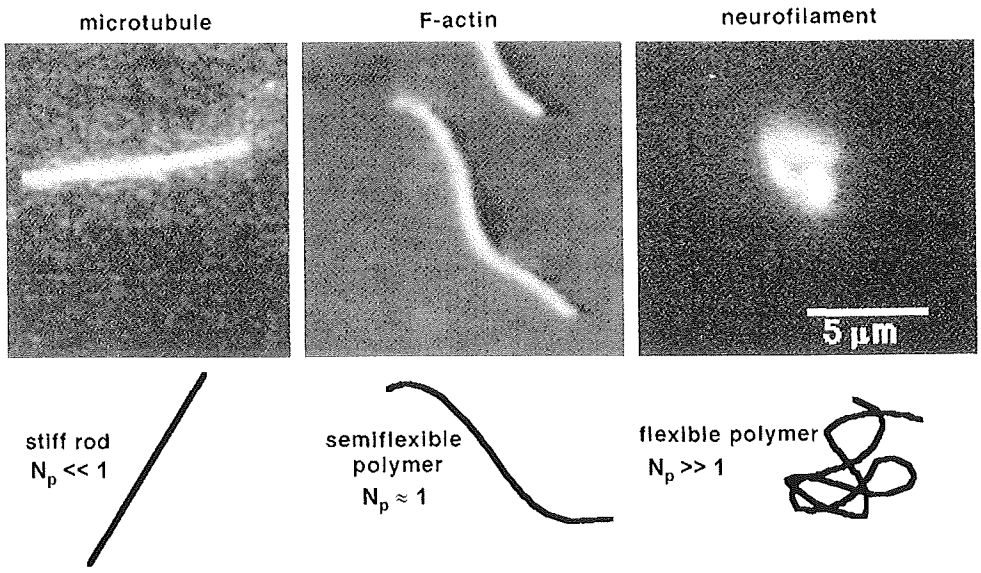


Fig. 7: *Biological examples of semiflexible chains [14]: Microtubule are rather stiff; the F-actin is semiflexible; the neurofilament is highly flexible. The number of persistence lengths N_p represents a measure of the flexibility.*

In the same manner as eqs. 31-33 one can derive an expression for the radius of gyration. But also for a continuous contour length [15] (and the Wiener integral method [16]) one obtains the identical result. We will just report the result which reads:

$$\left\langle \vec{R}_g^2 \right\rangle = \frac{N_K \ell_K^2}{6} - \frac{\ell_K^2}{4} + \frac{\ell_K^2}{4N_K} - \frac{\ell_K^2}{8N_K^2}\left[1 - \exp(-2N_K)\right] \tag{39}$$

The limits of flexible chains and rigid rods for the radius of gyration read then:

$$\left\langle \vec{R}_g^2 \right\rangle \underset{N_K \gg 1}{\longrightarrow} \tfrac{1}{6} N_K \ell_K^2 \qquad \text{and} \qquad \left\langle \vec{R}_g^2 \right\rangle \underset{N_K \ll 1}{\longrightarrow} \tfrac{1}{12} N_K^2 \ell_K^2 \tag{40}$$

A full comparison of the end-to-end distance and the radius of gyration is depicted in Figure 8. In the limit of large degrees of polymerization one finds the flexible chain limit with a ratio of 6 between the two radii R_g and R_{ee}. In the limit of small degrees of polymerization one finds the rigid rod limit with a ratio of 12 between the two radii.

2.5 Summary of Chain Models

For all chain models we can summarize the characteristic behavior of flexible chains where the characteristic size increases with the square root of the degree of polymerization N.

$$\sqrt{\left\langle \vec{R}_{ee}^2 \right\rangle} \propto \sqrt{\left\langle \vec{R}_g^2 \right\rangle} \propto \sqrt{N} \cdot \ell \tag{41}$$

This indicates that the characteristic size is much smaller than the full extension. The explicit prefactors depend on the flexibility of the chain and can be expressed in terms of a microscopic

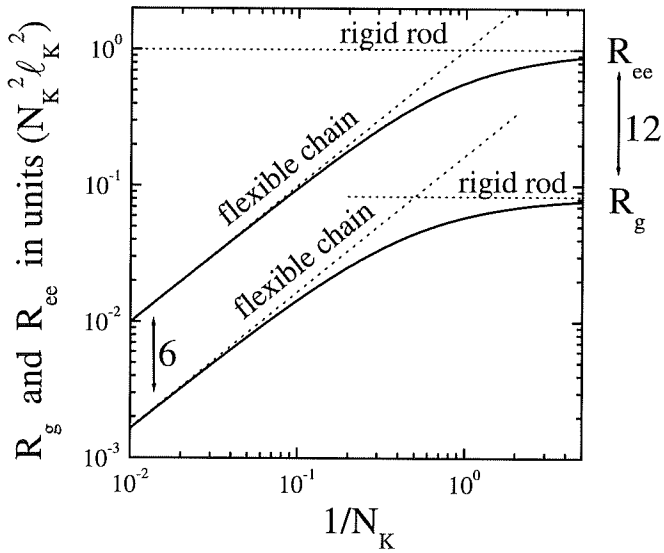


Fig. 8: *The comparison of the two typical radii R_g and R_{ee} as a function of the number of Kuhn segments. On the left (large N_K) the flexible chain limit is reached. On the right (small N_K) the rigid rod limit is reached.*

model C_∞ (eq. 21), a coarse grained model (eq. 35) with a rigidity κ, or more sophisticated microscopic models [1, 10]. In the rigid rod limit, the characteristic sizes are proportional to the full extension of the chain:

$$\sqrt{\left\langle \vec{R}_{ee}^2 \right\rangle} \propto \sqrt{\left\langle \vec{R}_g^2 \right\rangle} \propto N \cdot \ell \tag{42}$$

For this limit the flexibility does not play any role, and R_{ee} measures exactly the full extension (see eq. 26 and discussion below).

2.6 Distribution of the End-to-End Distance

Previously we have calculated averages of the characteristic radii. In this section we want to derive an expression for the distribution of the end-to-end distance [3]. Without many restrictions, we want to start with (a) either the freely rotating chain or (b) the Kuhn chain. This means we want to use the statistical independence of the chain segments by (a) definition or (b) for large enough segments, which have lost correlations. For the individual distribution of a single segment the normalization holds:

$$\int d^3r_i \, p_i(\vec{r}_i) = 1 \tag{43}$$

Later we will see that the final result does not depend on the shape of the single segment distribution. The first moment expresses that there is no preferred orientation of the segments:

$$\langle \vec{r}_i \rangle = \int d^3r_i \, \vec{r}_i p_i(\vec{r}_i) = \vec{0} \tag{44}$$

The second moment yields the average size of the single segment, as we have seen before.

$$\langle \vec{r}_i^{\,2} \rangle = \int d^3 r_i \, (\vec{r}_i^{\,2}) p_i(\vec{r}_i) = \ell^2 \tag{45}$$

These considerations focused on moments of a single segment distribution. The statistical independence of the segments is expressed by:

$$\langle \vec{r}_i \vec{r}_j \rangle = \delta_{ij} \ell^2 \tag{46}$$

Now the observable of the end-to-end distance is introduced by $\vec{R} = \sum_i \vec{r}_i$. The distribution of this observable is simply expressed by the average under the restriction that the condition between $\vec{R}$ and $\sum_i \vec{r}_i$ holds:

$$P(\vec{R}) = \left\langle \delta(\vec{R} - \sum_i \vec{r}_i) \right\rangle \tag{47}$$

Now one uses the trick of the Fourier transformation of this distribution. The advantage is, that the Fourier variable $\vec{Q}$ directly couples with each individual segment variable independently.

$$\tilde{P}(\vec{Q}) = \left\langle \int d^3 R \, \exp(\mathbf{i}\vec{Q}\vec{R}) \cdot \delta(\vec{R} - \sum_i \vec{r}_i) \right\rangle \tag{48}$$

$$= \left\langle \exp(\mathbf{i}\vec{Q} \sum_i \vec{r}_i) \right\rangle \tag{49}$$

i is the imaginary unit. The averaging of this expression is now written explicitly, and transformations of the exponential are done consequently. The higher order terms must be taken into account as well and will be discussed below.

$$\tilde{P}(\vec{Q}) = \int d^3 r_1 \, p_1(\vec{r}_1) \cdots \int d^3 r_N \, p_N(\vec{r}_N) \, \exp(\mathbf{i}\vec{Q} \sum_i \vec{r}_i) \tag{50}$$

$$= \int d^3 r_1 \, p_1(\vec{r}_1) \cdots \int d^3 r_N \, p_N(\vec{r}_N) \left[1 - \tfrac{1}{2} (\vec{Q} \sum_i \vec{r}_i)^2 + \cdots \right] \tag{51}$$

$$= \int d^3 r_1 \, p_1(\vec{r}_1) \cdots \int d^3 r_N \, p_N(\vec{r}_N) \left[1 - \tfrac{1}{2} \sum_i (\vec{Q}\vec{r}_i)^2 + \cdots \right] \tag{52}$$

The first transformation just involves the Taylor series of the exponential. The first moments of $\vec{r}_i$ vanish since there is no preferential orientation. The next line uses the statistical independence of the chain segments. In the following we do not use the explicit formulation of the averaging, but use the brackets again:

$$\tilde{P}(\vec{Q}) = 1 - \tfrac{1}{6} \sum_i \left\langle \vec{Q}^2 \vec{r}_i^{\,2} \right\rangle + \cdots \tag{53}$$

$$= 1 - \tfrac{1}{6} N \vec{Q}^2 \ell^2 + \cdots \tag{54}$$

$$\approx \exp(-\tfrac{1}{6} N \vec{Q}^2 \ell^2) \tag{55}$$

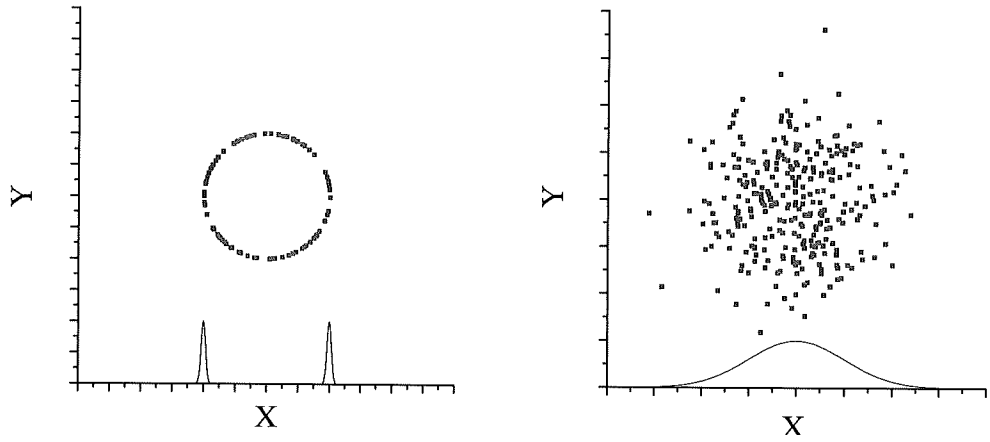


Fig. 9: *The distribution of a single segment (freely jointed chain) and the whole chain. While for a single segment the distribution is rather pronounced, for a large amount of segments the central limit theorem yields a Gaussian distribution.*

The first line is valid for squared averages of individually projected segments. Please note the additional factor $\frac{1}{3}$. Now, the individual second moments occur, the segment length appears as the explicit average. Finally, the first two terms of the Taylor series appear equally well for an exponential.
Actually one had to take all higher order terms of eq. 51 into account to be sure that the final expression holds. In the higher order terms there occur some expressions with higher order moments of single chain segments and *many* expressions involve only second order moments of different segments. This *"many"* becomes the more important the longer the chain is. These *many* terms finally make the exponential exact in the limit of long chains.
Finally using the definition of the mean end-to-end distance $R_{ee}^2 = N\ell^2$, and Fourier transforming the last expression back into real space, one obtains:

$$P(\vec{R}) = (\tfrac{2}{3}\pi R_{ee}^2)^{-3/2} \exp(-\frac{3}{2}\frac{\vec{R}^2}{R_{ee}^2}) \tag{56}$$

This distribution is a Gaussian with the respective normalization, which is independent of the initial distribution. The justification of eq. 55 might seem to be hand-waving; the final result is exact in the context of the central limit theorem [17].
While other explicit derivations for the freely rotating chain exist (see "saddle point approximation" [10]), we explicitly want to demonstrate the meaning of the central limit theorem. For the freely rotating chain the distribution of a single segment covers a sphere (see Fig. 9), while for the whole chain the end-to-end distance distribution is a Gaussian. So, even a pronounced distribution of a single segment ends up in a Gaussian distribution for a whole chain, when a large number of segments is considered.

2.7 The Gaussian Chain

The step to the Gaussian chain is small taking all previous sections into account [4]. The idea is, that the number of segments in a Kuhn segment is already large enough, so that the distribution

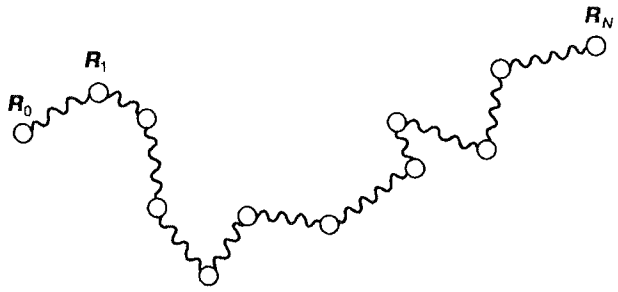


Fig. 10: *The equivalent picture for a Gaussian chain is a number of effective monomers connected by elastic springs.*

function of a single Kuhn segment can be approximated by a Gaussian distribution. Then the distribution function of N_K segments can be written:

$$W(\vec{r}_1, \cdots, \vec{r}_{N_K}) = (\tfrac{2}{3}\pi\ell_K^2)^{-3N_K/2} \cdot \exp(-\frac{3}{2\ell_K^2}\sum_i \vec{r}_i^{\,2}) \tag{57}$$

The limit of a Gaussian distribution might be achieved for slightly larger segments. So the numbers ℓ_K and N_K should be taken as approximate values, which they are anyway (see section 2.3). Finally we write for the effective bond vectors the difference of the monomer vectors:

$$W(\vec{R}_0, \cdots, \vec{R}_{N_K}) = (\tfrac{2}{3}\pi\ell_K^2)^{-3N_K/2} \cdot \exp\left(-\frac{3}{2\ell_K^2}\sum_i (\vec{R}_i - \vec{R}_{i-1})^2\right) \tag{58}$$

We will need this expression in the next section in the manner of an equivalence. Please note that the normalization does not cover the initial vector $\vec{R}_0$, which will be discussed in one example below (section 3.1). The now obtained distribution function can be compared with a thermodynamic partition function by rewriting:

$$W \propto \exp\Big(-\frac{1}{k_BT}\cdot\underbrace{\frac{3k_BT}{2\ell_K^2}\sum_i (\vec{R}_i - \vec{R}_{i-1})^2}_{\text{Hamiltonian}}\Big) \tag{59}$$

k_B is the Boltzmann constant. The identified Hamiltonian describes N_K Hookean springs with a single temperature dependent spring constant. The Gaussian chain therefore can be drawn as beads connected by N_K springs (see Fig. 10). The temperature dependence of the spring constant is important for the rubber elasticity as we will see in the example below (section 3.2). Furthermore, the temperature dependent Hamiltonian can be called entropic since the temperature dependence actually cancels out.

2.8 The Continuous Description

While we had a certain number of beads in the previous chapter, we now want to consider the limit of a continuous chain path in space. The sum in the exponential of eq. 58 is simply replaced by an integral and the difference of the monomer vectors is substituted by a differential, and so we obtain for the distribution function:

Table 1: *Equivalence of the discrete and continuous description of a Gaussian chain. Note that the continuous variables k or s are still notated in units of indices.*

Discrete	Continuous
$\sum_{i=k}^{k'}$	$\int_{k}^{k'} ds$
$\vec{R}_i - \vec{R}_{i-1}$	$\partial\vec{R}/\partial s$
$\vec{R}_{i+1} - 2\vec{R}_i + \vec{R}_{i-1}$	$\partial^2\vec{R}/\partial s^2$
Kronecker delta δ_{ij}	Dirac delta function $\delta(s-s')$
integral $\int d^3R_1 \cdots \int d^3R_{N_K}$	path integral $\int \delta^3 R$

$$W = \exp\left(-\frac{3}{2\ell_K^2}\int ds\left(\frac{\partial\vec{R}(s)}{\partial s}\right)^2\right) \tag{60}$$

Again, we can identify a Hamiltonian with a temperature dependent spring constant. Generally, one introduces a potential for the beads U_e.

$$H = \frac{3k_BT}{2\ell_K^2}\int_0^N ds\left(\frac{\partial\vec{R}}{\partial s}\right)^2 + \int_0^N ds\cdot U_e(\vec{R}) \tag{61}$$

The partition function or the distribution function is then obtained in the manner of classical thermodynamics:

$$G \propto \exp\left(-\frac{1}{k_BT}H\right) \tag{62}$$

The change in the nomenclature going from the discrete to the continuous description of chains is summarized in Table 1: The summation over monomers is substituted by an integration. The difference between neighbored monomer vectors is substituted by the differential. The given difference between three neighbored monomer vectors is substituted by the second derivative. The discrete Kronecker delta is replaced by the Dirac delta function. In a partition function the integration over the monomer vectors is replaced by a path integral. Taking all these rules into account and using the correct normalization the simple equation 62 is more precisely given by:

$$G(\vec{R},\vec{R}',N) = \frac{\int_{\vec{R}}^{\vec{R}'}\delta^3R\cdot\exp\left(-\frac{3}{2\ell_K^2}\int_0^N ds\left(\frac{\partial\vec{R}}{\partial s}\right)^2 - \frac{1}{k_BT}\int_0^N ds\cdot U_e(\vec{R})\right)}{\int d^3R'\int_{\vec{R}}^{\vec{R}'}\delta^3R\cdot\exp\left(-\frac{3}{2\ell_K^2}\int_0^N ds\left(\frac{\partial\vec{R}}{\partial s}\right)^2\right)} \tag{63}$$

This Green function has the constraints of the initial vector (the first monomer vector) and the final vector (last monomer) and the degree of polymerization. This formalism is used to describe the thermodynamic behavior of polymers [4] partly in blends, partly in solutions as we will see in the next section. One important property of this distribution function is, that it is a Green function, which solves the following differential equation:

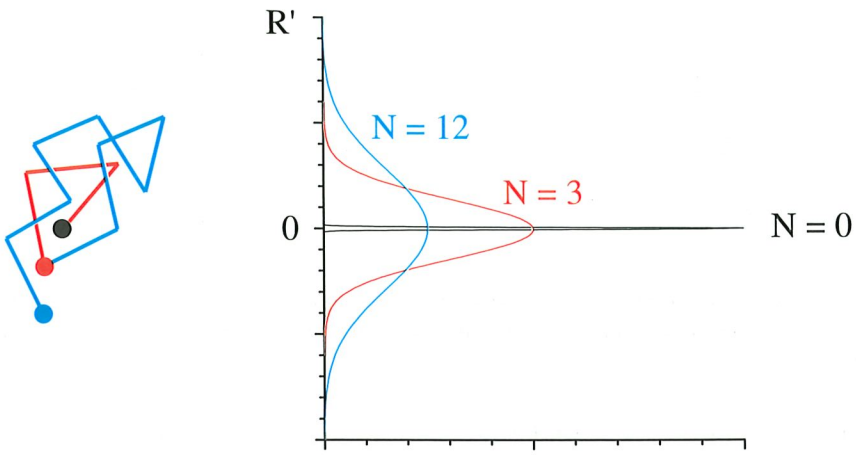


Fig. 11: *Comparison of the Green function of polymers with the diffusion process: (left) discrete picture, (right) continuous picture.*

$$\left(\frac{\partial}{\partial N} - \frac{\ell_K^2}{6}\nabla_{\vec{R}}^2 + \frac{1}{k_B T}U_e(\vec{R})\right) G(\vec{R}, \vec{R}', N) = \delta(\vec{R} - \vec{R}')\delta(N) \tag{64}$$

The left hand side term is the homogenous part and the right hand side is the inhomogeneity typical for Green functions. The solutions for $N > 0$ solve the homogenous part alone, and some interpretations of the left terms can be given: First of all we are talking about *polymer thermodynamics*. The first two addends are typical for a *diffusion equation* where N corresponds to time. And finally the whole left part looks like a *Schrödinger equation*, but the time is imaginary (N). A comparison of the polymer thermodynamics and the diffusion interpretation is given in Figure 11. The first step ($N = 0$) corresponds to the initial condition ($t = 0$) – the distribution function is a delta function at the origin. The next shown stage corresponds to three steps in N or time. The chain has three monomer units or the particle has diffused by three "steps". The next shown stage corresponds to twelve steps - the chain is even longer or the particle has diffused for a longer time.

3 Phenomena of Polymers

While we have focused on the conformation of a single chain in the last chapter we now will work out phenomena of polymers. There we will describe effects of polymers in dilute solution – the chains are isolated – in semidilute solutions and in blends. Whereas the first state obviously can be treated as idealized, the last state behaves also quite ideal. The semidilute solution is the most complicated state and will be treated in the last section.

3.1 Polymer near a hard wall

The modelling of a polymer near a hard wall bases on the Green function of section 2.8 [18]. We assume the polymer to be in dilute solution, which means that the polymer coils do not intersect. Furthermore we assume an idealization which is true for θ-solvents. The detailed

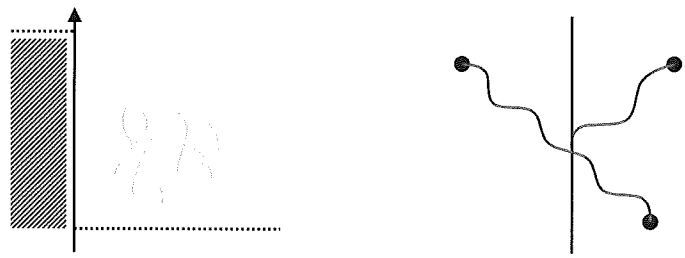


Fig. 12: *(left) The polymer in the presence of a hard wall potential. (right) Disregarded conformation corrected by an allowed conformation. This correction is done with the method of images.*

meaning is discussed in section 3.4. For the present modelling it simply means that we can use the Green function as already derived.
The Green function of a totally free polymer is already given in equation 56. Now the presence of a plain wall limits the polymer to the half space $z > 0$ only. A possible attractive or repulsive potential in the vicinity of the wall is not considered here – the boundary condition just prohibits the presence of the polymer beyond the wall. This also implies that $G = 0$ at the boundary for $z = 0$ or $z' = 0$. One could imagine to solve the Green function for a potential being large or infinity in negative half space $z < 0$ (see Fig. 12, left). But fortunately a method developed for electrodynamics yields a simple solution. By the method of images (image charges) one directly obtains a solution with the correct boundary condition:

$$G(\vec{R},\vec{R}',N_K) = (\tfrac{2}{3}\pi N_K \ell_K^2)^{-3/2} \cdot \exp\left(-\frac{3}{2N_K\ell_K^2}(x-x')^2\right)\exp\left(-\frac{3}{2N_K\ell_K^2}(y-y')^2\right) \quad (65)$$
$$\cdot \left[\exp\left(-\frac{3}{2N_K\ell_K^2}(z-z')^2\right) - \exp\left(-\frac{3}{2N_K\ell_K^2}(z+z')^2\right)\right]$$

While this solution is formally correct a plausible interpretation should be given (Fig. 12, right) furthermore. The solution with the initial vector being in the correct half space ($z > 0$) and the final vector being in the wrong half space is prohibited. The mirrored solution of the final vector adds one possible path where the polymer touches the wall and makes the originally considered path impossible. Possible paths contribute positively while impossible paths cancel out. To get an idea of the polymer concentration near the wall, one has to integrate the Green function over possible end-vectors $\vec{R}'$ for a given initial vector.

$$\begin{aligned}\rho(z) &= \int d^3R'\, G(\vec{R},\vec{R}',N_K) \qquad (66)\\ &= (\tfrac{2}{3}\pi N_K\ell_K^2)^{-1/2}\int_0^\infty dz' \left[\exp\left(-\frac{3}{2N_K\ell_K^2}(z-z')^2\right) - \exp\left(-\frac{3}{2N_K\ell_K^2}(z+z')^2\right)\right]\\ &= \operatorname{erf}\left(\sqrt{\frac{3}{2}}\cdot\frac{z}{R_{ee}}\right)\end{aligned}$$

The two Gaussian integrals in x and y-direction directly cancel out with the normalization. The integral in positive half space yields the error function. This solution is depicted in Figure 13.

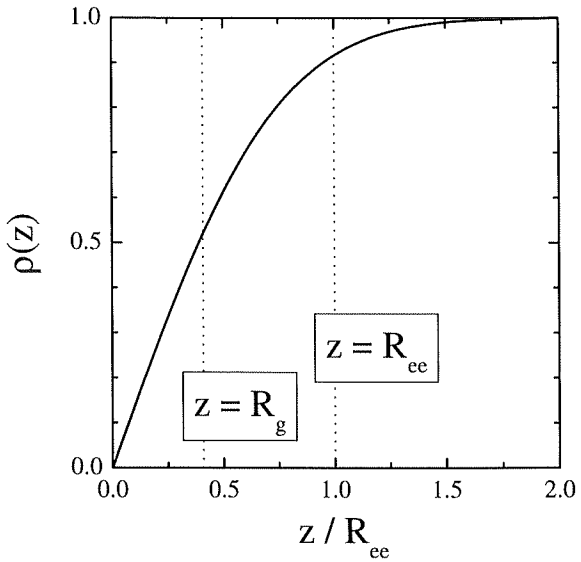


Fig. 13: *The distribution of an end-monomer relatively to a hard wall. The bulk concentration is normalized to be unity. The two distances R_{ee} and R_g correspond to $\approx$ 10% and $\approx$ 50% deviations of the bulk value.*

While for a distance of the average end-to-end vector the concentration of the polymer deviates by $\approx$10% from the bulk value only, the distance of the average radius of gyration approximately predicts where the concentration reaches the half-maximum value. The normalization of the bulk value to unity is due to the missing normalization of the initial vector (see eqs. 58 and 63).

It is more likely to find the polymer in the bulk phase than close to the wall. The polymer obviously is depleted by the wall. Near a wall, the polymer has less allowed configurations. This means a loss of entropy close to the wall, which is unfavorable. Therefore, the depletion effect is of entropic origin.
This depletion effect is important for a mixture of dissolved colloidal particles and polymers. The effective depletion zones of isolated colloidal particles is large, while for closely packed colloidal particles the depletion zones overlap and therefore are effectively smaller. Therefore the polymer depletion can lead to a phase separation of colloidal particles and polymers [19]. This effect is highly interesting for the formulation of paints for instance.

3.2 Rubber elasticity

Rubber consists of high molecular weight polymers [5, 20]. To obtain a stable rubber the molecules are chemically cross-linked in a vulcanization process (see Fig. 14) and a polymer network is obtained. Rubber can be strained by several 100%, and its original shape is reversibly obtained when the stress is released. The elastic modulus increases linearly with temperature. Only at very low temperatures the elastic modulus becomes comparable to normal solids and the rubber becomes fragile. The deformation is defined by the elongation say in z-direction (see Fig. 15).

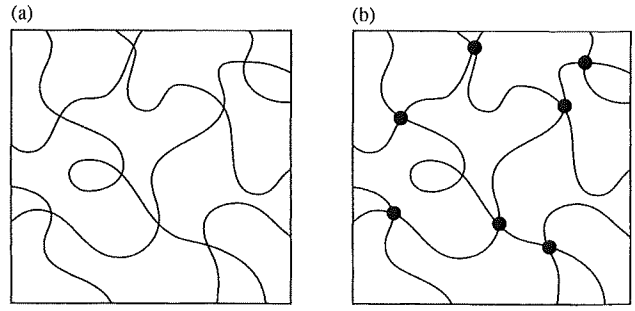


Fig. 14: *Principle of a polymer network. (a) non-vulcanized – just topological cross-links are present, (b) vulcanized – some cross-links are chemically bond.*

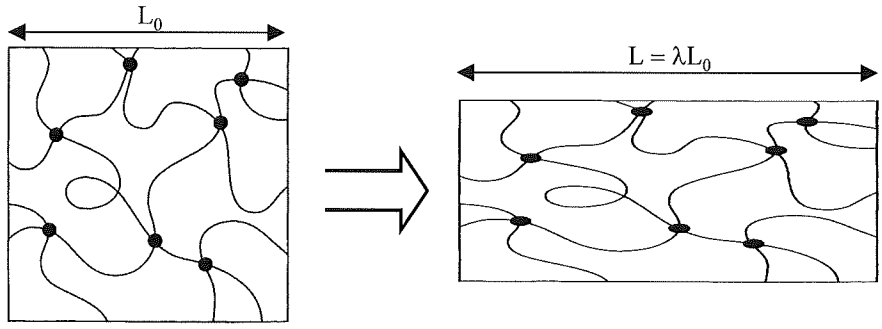


Fig. 15: *The polymer network is stretched in one dimension by a factor of λ. Due to the conservation of volume the polymer contracts in the other directions.*

$$\lambda \equiv \lambda_z = L/L_0 \tag{67}$$

It is a very good approximation, that the volume of a rubber is conserved under strain. Therefore:

$$\lambda_x \cdot \lambda_y \cdot \lambda_z = 1 \tag{68}$$

Usually the rubber is isotropically deformed in the perpendicular directions and contracted due to the conservation of volume.

$$\lambda_x = \lambda_y = \lambda^{-1/2} \tag{69}$$

Now we assume that the positions of all cross-links are uniformly transformed by the same transformation. This behavior is called affine. The transformed coordinates read then:

$$(x', y', z') = (\lambda^{-1/2}x, \lambda^{-1/2}y, \lambda z) \tag{70}$$

As the thermodynamic potential we consider the free energy difference. The difference considers the stretched relatively to the non-stretched rubber.

$$\Delta F(\lambda) = F(\lambda) - F(\lambda = 1) = \Delta U(\lambda) - T\Delta S(\lambda) \tag{71}$$

It is obvious that the free energy has energetic and entropic contributions. At low temperatures the energy becomes dominant, which is mainly determined by the monomer properties. This energy is responsible for the behavior similar to normal solids. At high temperatures the entropy is dominant, and we will focus on this contribution. The Green function (see eq. 56) of a chain segment between two cross-links is then given by:

$$P(\vec{R}') = (\tfrac{2}{3}\pi R_{ee}^2)^{-3/2} \exp\left(-\frac{3}{2R_{ee}^2}(\lambda^{-1}x^2 + \lambda^{-1}y^2 + \lambda^2 z^2)\right) \tag{72}$$

The used coordinates are the stretched coordinates (eq. 70). From this Green function (or probability) the free energy can be calculated.

$$\begin{aligned} \Delta F(\lambda) &= -T\Delta S(\lambda) = -k_B T \cdot n \left\langle \ln P(\lambda) - \ln P(\lambda = 1) \right\rangle \\ &= \tfrac{1}{2} k_B T \cdot n \cdot (2/\lambda + \lambda^2 - 3) \end{aligned} \tag{73}$$

The entropy is dominant. To account for the number of chain segments, the number n is multiplied with the single segment contributions. The averaging over the squared coordinates cancels with the prefactor of the argument in eq. 72. The force due to the strain is calculated by the derivative of the free energy:

$$\partial \Delta F / \partial L = k_B T \cdot n \cdot (\lambda - \lambda^{-2})/L_0 \tag{74}$$

The stress is obtained by the force normalized to the area, and one obtains:

$$\sigma_z(\lambda) = k_B T \cdot n \cdot (\lambda - \lambda^{-2})/V_0 = RT \cdot \rho / M_C \cdot (\lambda - \lambda^{-2}) \tag{75}$$

The final result is slightly rewritten by expressing the number of chain segments as $n = N_A \rho V_0 / M_C$. The gas constant is $R = k_B N_A$ with N_A being the Avogadro number. The factor temperature is responsible for the higher elastic modulus at elevated temperatures. The density ρ being always of the order of $1\mathrm{g/cm}^3$ does not change the elastic modulus. The molar mass of a chain segment between two cross links M_C is an important factor. Highly cross-linked rubbers, i.e. M_C small, are harder.

The idealized and real dependence of the stress σ on the strain λ is discussed in the context of Figure 16. For stresses around 1 and below the theory fits the experimental data nicely. For intermediate strains the stress is lower than theory. This is due to topological cross-links which can slide away upon external stress. Another factor is the fluctuations of the cross-link positions, which we assumed to be transformed strictly. These fluctuations also account for a decreased stress. At even higher strains the chains do not behave Gaussian anymore – the chain segments are strongly stretched – and therefore a higher stress is found than in our simple theory.

The origin of rubber elasticity is entropy, and the elastic modulus increases with temperature. In this context we can go back to the Hamiltonian of eq. 59. The elastic constant $k(T)$ also linearly increases with temperature, and can be seen as the origin for the rubber elasticity. The intermediate filaments of the cytoskeleton form also a network in the cell, which is responsible for the reversible deformability of a cell. It can be discussed if this elasticity is purely entropic, but surely the cross-link density (or M_C) would affect the rigidity of the network.

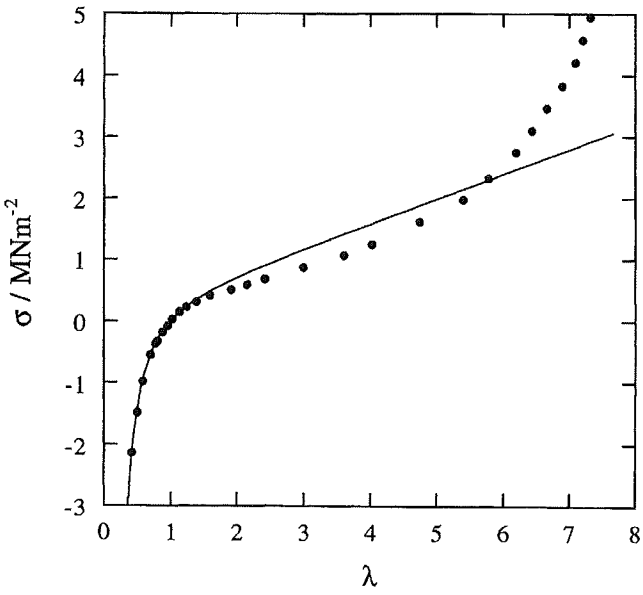


Fig. 16: *Stress strain curve of a rubber [21]. The details of this curve are discussed in the text.*

3.3 Scattering of chains

The scattering techniques make up the best way to observe molecular (and atomic) structures.[2] One uses a certain kind of radiation (light, x-rays, neutrons), which usually includes many wavelengths. A monochromator selects a certain wavelength λ. The divergence of the beam is then usually reduced by apertures (collimation). This well defined radiation hits the sample. While usually larger amounts of the radiation are transmitted without being diffracted, the interesting information about the sample structure is obtained by the scattered intensity. Large area detectors allow to collect as much information as possible. The intensity is measured as a function of the scattering angle θ, which can be expressed in terms of a momentum transfer $Q = 4\pi \sin(\theta/2)/\lambda$. The momentum transfer Q is the variable of the Fourier space, which we already introduced in section 2.6. The scattering intensity is proportional to the form factor of a single chain $S(Q)$ [6, 7, 22]. The form factor of a single chain considers interferences of all possible pairs of "monomers". For the time being the "monomer" units shall build up a Gaussian chain. Further refinements are discussed below.

$$S(Q) \propto \frac{1}{N} \sum_{i,j=0}^{N} \left\langle \exp\left(\mathrm{i}\vec{Q} \cdot (\vec{R}_i - \vec{R}_j)\right) \right\rangle \tag{76}$$

$$\propto \frac{1}{N} \sum_{i,j=0}^{N} \exp\left\langle -\tfrac{1}{2} \left(\vec{Q} \cdot (\vec{R}_i - \vec{R}_j)\right)^2 \right\rangle \tag{77}$$

[2]Further details of a scattering experiment are discussed in the contribution of W. Schweika B1. We restrict ourselves to the scattering of polymer chains.

$$\propto \frac{1}{N}\sum_{i,j=0}^{N} \exp\left\langle -\tfrac{1}{6}\vec{Q}^2\cdot(\vec{R}_i-\vec{R}_j)^2\right\rangle \tag{78}$$

Equation 77 has moved the averaging from outside to the inside of the exponential. This Gaussian transformation in principle uses similar arguments as we used between equations 50 and 55, and it becomes exact for Gaussian chains. The next rearrangement just makes use of the average projection (see also eq. 53). In the present stage we use the property of a Gaussian chain and replace the summation by an integration, which is correct for large enough length scales (see discussion below).

$$S(Q) \propto \frac{1}{N}\int_0^N di \int_0^N dj \exp\left(-\tfrac{1}{6}\vec{Q}^2\cdot|i-j|\cdot\ell_K^2\right) \tag{79}$$

$$= N\cdot f_D(Q^2R_g^2) \tag{80}$$

$$f_D(x) = \tfrac{2}{x^2}\left(\exp(-x)-1+x\right) \tag{81}$$

In principle the same arguments were used for eq. 33. The final result yields the Debye function f_D, which is often observed in scattering experiments on polymeric systems. It has two important regions, namely for small and relatively large scattering vectors Q.

$$S(Q) \propto N(1-\tfrac{1}{3}Q^2R_g^2) \qquad \text{for small } Q \tag{82}$$

$$\propto N\cdot 2/(Q^2R_g^2) \qquad \text{for large } Q \tag{83}$$

For small scattering vectors one obtains a maximum intensity at $Q \to 0$, which contains the number of correlated monomers N. Experienced researchers prefer to use the molar volume that a polymer chain covers $S(Q\to 0)\propto V$. The dependence of the intensity at small scattering vectors reflects the radius of gyration. So at small scattering angles one obtains information about the chain as a whole molecule. At larger scattering angles, one obtains a Q^{-2} dependence with a prefactor, which depends on a segment length ℓ_σ. This segment length is defined by the ratio of the radius of gyration and the degree of polymerization $6R_g^2/N=\ell_\sigma^2$, and compares the real chain with the idealized freely jointed chain. If N is defined with respect to a backbone bond as the basic unit, one obtains $\ell_\sigma=\sqrt{C_\infty}\,\ell$. For details further reading in reference [2] is recommended. Therefore, at larger scattering angles one observes intra-chain parts. The self-similarity of chain parts on different length scales is observed. Also information about the rigidity C_∞ is obtained.

The described behavior can be fully observed from a small angle neutron scattering experiment on polymers [23]. An example for polystyrene is shown in figure 17. At small scattering vectors $Q<3\cdot10^{-3}\text{Å}^{-1}$ the intensity levels off. Since the molar volume $\bar{V}_W$ is relatively large, the full off-levelling is not observed in this experiment. At larger scattering vectors one nicely observes the Q^{-2}-behavior, which tells about the self-similarity of the chain over many length scales.

One interesting detail of this experiment is that the sample consists of protonated and deuterated polymer only. When discussing the scattering of a polymer blend then interferences between different molecules occur as well. These contributions have the same Q-dependence as a single polymer [7, 24], and so the scattering of a polymer blend looks like the scattering of isolated polymers (in a θ-solvent, see section 3.5).

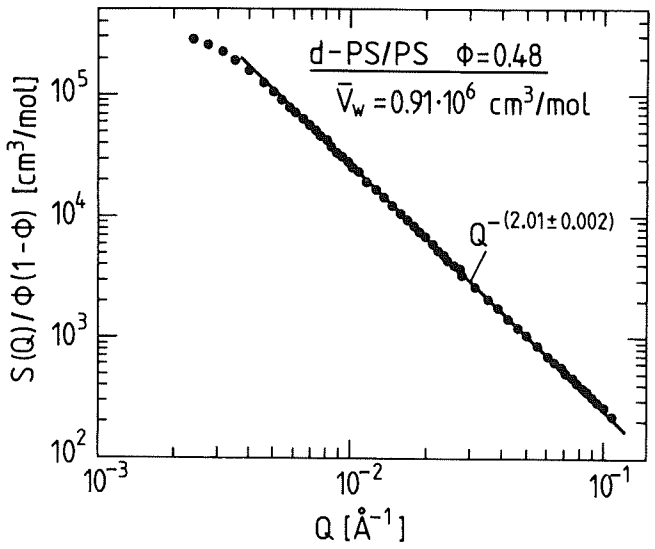


Fig. 17: *Scattering of a polystyrene / deuterated polystyrene blend [23]: The structure factor as a function of the scattering vector.*

The whole calculation assumed "small" scattering vectors, which are small compared to the inverse monomeric size $Q \ll \ell^{-1}$. Approaching this limit would allow one to observe a real rigid part of the chain. At even larger scattering vectors one would observe the chemical structure, i.e. each atom within the monomer. The whole Q-range would therefore span the following sizes: whole chain – self-similarity of chain – rigid part of the chain – atomic structure. There exist approaches, which cover the rigid part of a chain as well [25, 26] (there applied for polystyrene). More interesting are these theories in studies of worm-like micelles [27], where the apparent Q-range is shifted to lower values.

The for all scattering experiments interesting Q-ranges can be also discussed in the context of the continuous description (eq. 63) (see contribution R. Winkler C3 and [28]). The first stage of approximation, which is presented in this lecture, just considers the Gaussian chain with $\partial\vec{R}/\partial s$ terms. The only parameter of this chain model is the end-to-end distance $\langle R_{ee}^2 \rangle$. The corresponding Q-range covers the whole chain and the self-similarity of the chain. For this Q-range the expression of the end-to-end distance can base on any segment length ℓ. The next stage of approximation considers a semiflexible chain with additional $\partial^2\vec{R}/\partial s^2$ terms in the hamiltonian. The corresponding Q-range additionally covers the rigid part of the chain. The contour length L comes into play, which makes a segment length distinguishable. The natural segment length of such a chain is given by $\ell = \langle R_{ee}^2 \rangle / L$, and the natural degree of polymerization is given by $N = L^2/\langle R_{ee}^2 \rangle$. A higher stage of the continuous description would introduce the torsional energy. For the present lecture we can safely neglect the higher stages.

3.4 Deviations – The Flory Exponent

Differences from the behavior above occur when polymers are dissolved in good solvents [8]. An example of scattering curves is shown in Figure 18. For this example the behavior at large

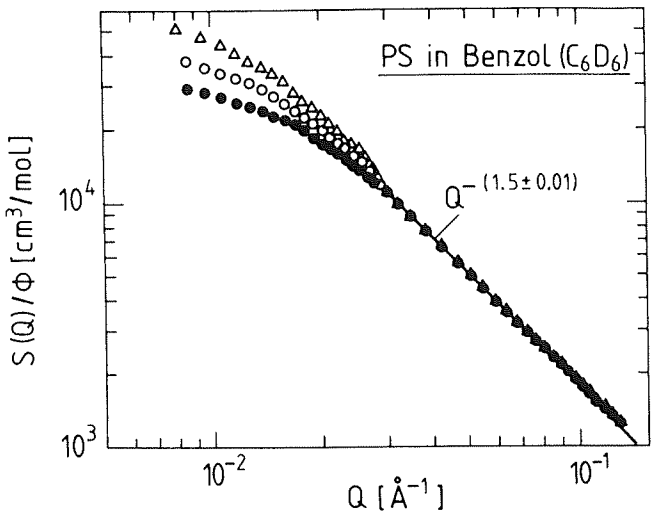


Fig. 18: *Scattering of a polystyrene solution in deuterated benzene [23]: The structure factor as a function of the scattering vector.*

scattering vectors is more like $Q^{-1.5}$. This behavior extends over a large Q-range and occurs for many samples, which speaks for the significance of the effect. In viscosity measurements one observes the end-to-end distance R_{ee} as a function of the degree of polymerization N with deviations from the ideal coil behavior (eq. 41). Generally one can write for the end-to-end distance:

$$\left\langle \vec{R}_{ee}^2 \right\rangle = N^{2\nu} \cdot \ell_e^2 \qquad \nu \in [\tfrac{1}{2}..1] \tag{84}$$

The exponent ν lies in the range between $\frac{1}{2}$ and 1, which corresponds to ideal coils and rods. The exact value of polymers in good solvents is the issue of this section. For curiosity the relation to the radius of gyration should be mentioned as well:

$$\left\langle \vec{R}_g^2 \right\rangle = \frac{\aleph}{6} \left\langle \vec{R}_{ee}^2 \right\rangle \qquad \text{with} \qquad \aleph = \frac{6}{(2\nu+1)(2\nu+2)} \tag{85}$$

The prefactor $\aleph$ is of the order of unity, and was calculated theoretically in reference [3]. It agrees with our results in the limits of ideal coils and rods (eqs. 38 and 40):
While we considered the local rigidity of a chain – and this can be a result of steric hindrance of side groups – we neglected hindrance of chain parts with very different indices n and m (see Fig. 19). This interaction is called the excluded volume interaction. The Flory argument allows to calculate the exponent ν within a very good approximation. While we already lost the precise reference of some numbers (say N) to the polymer unit (be it monomers or backbone bonds) sometimes, we now will deliberately argue on the basic dependence. This arguing is called scaling argument then.
The occupied volume of a diluted, isolated polymer chain in d dimensions is then $V \cong R^d$. The radius R is a typical size of the chain. The monomer density in this isolated volume is given by $\rho \cong N/R^d$ then. Now the average repulsive energy can be calculated by:

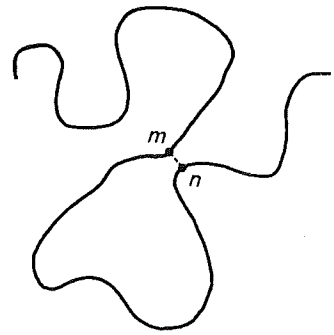


Fig. 19: *Whereas the chain flexibility results from steric hindrance of neighbored monomers, the excluded volume interaction bases on repulsive interactions of distant monomers n and m.*

Table 2: *Comparison of the Flory exponent based on the Flory argument with exponents from renormalization calculations. The agreements are surprisingly good.*

d	ν_F	ν (best value)
1	1	1 exact
2	3/4	3/4 exact
3	3/5	0.588
4	1/2	1/2 exact

$$E_{\text{rep}} \cong \varepsilon \cdot V \cdot \rho^2 \cong \varepsilon \cdot N^2/R^d \tag{86}$$

The expression corresponds to a Coulomb interaction of a (homogenous) charge density ρ within a volume V. The factor ε just takes care for the correct unit of the repulsive energy. Now the entropy of a chain can be calculated in the manner of previous calculations (see eq. 73). It basically is the logarithm of the Green function:

$$S \cong k_B \cdot \ln P(R) \cong -\tfrac{3}{2} k_B R^2/(N\ell^2) \tag{87}$$

The free energy is given by the energy and entropy then:

$$F = E_{\text{rep}} - TS \cong \varepsilon \cdot N^2/R^d - \tfrac{3}{2} k_B T R^2/(N\ell^2) \tag{88}$$

The minimization of the free energy is done via the first derivative $\partial F/\partial R = 0$. The expression can be sorted by R and N-terms then, and one finally obtains:

$$R \propto N^\nu \qquad \text{with} \qquad \nu = 3/(2+d) \tag{89}$$

This result can be compared with more sophisticated calculations with the formalism of a renormalization group (see Table 2). It is very astonishing that the simple scaling argument of Flory delivers the exact values except for 3 dimensions. The more precise value of $\nu = 0.588$ only slightly deviates from the Flory value 0.6. The deviation from the ideal coil behavior ($\nu = 0.5$) means, that the polymer swells due to the repulsive excluded volume interaction.

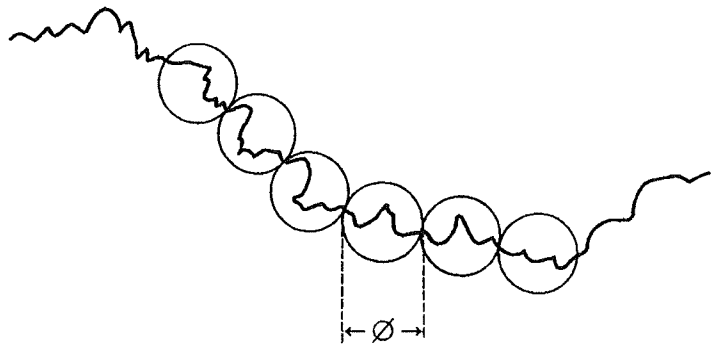


Fig. 20: *A Flory chain with sections of the diameter Ø. The scattering power law is derived in the text.*

The connection to scattering experiments is drawn in the following (see also Fig. 20). A chain can be divided into 3-dimensional sections of diameter Ø. Within such a section a sub-chain is found with the Flory-behavior (eq. 84), and so the number of monomers of the sub-chain is $N_{Ø}^{\nu} \propto Ø$. The scattering intensity is proportional to the numbers of monomers within such a sub-chain $S(Q) \propto N_{Ø}$, and the respective scattering vector is connected with the diameter via $Q \propto Ø^{-1}$. Finally, one obtains for the scattering at large enough Q-values:

$$S(Q) \propto Q^{-1/\nu} \tag{90}$$

Therefore the Flory exponent can be observed in a single scattering experiment. This is a much simpler way than measuring the end-to-end distance by viscosity for several molar mass polymers (see eq. 84).
It shall be mentioned that the Flory exponent occurs for polymers in good solvents. There, the contact of the polymer with itself is unfavorable (Fig. 19), which leads to the effect described above. But solvents exist (θ-solvents), for which the interaction solvent-polymer is in equilibrium with the interaction polymer-polymer. Energetically, there is no difference for the polymer-polymer interaction then, and the ideal coil behavior is retrieved ($\nu = \frac{1}{2}$). In a polymer blend there is no unfavorable interaction neither, and one obtains the ideal coil behavior as well, as we will also see in the next section.

3.5 Semidilute Solutions – The Blob Concept

The characteristic issue of semidilute polymer solutions is the contact between different polymers, which leads to many topological cross-links. Therefore, such a solution behaves gel-like. The cytoskeleton certainly is a semidilute polymer solution with topological cross-links. Partially, aspects of biological cross-linking play an additional role. Furthermore, semidilute solutions are interesting for glues. A poor mans glue consists only of polymer and solvent. When the solvent evaporates the polymer hardens and fixes two pieces together. More sophisticated glues consist of a ready made polymer and a solvent, which is a monomer. This monomer can be polymerised, and after full conversion the whole glue is hardened.
We will derive expressions for the blob concept of DeGennes [9, 29] in the following, again in the manner of scaling arguments. The semidilute polymer solution is a network, with many polymer-polymer contacts of a typical distance ξ (see Fig. 21). Within such a blob the polymer

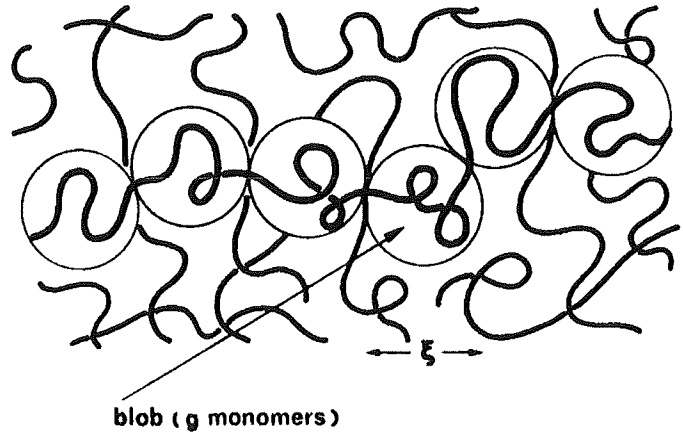


Fig. 21: *A polymer network formed by a semidilute polymer solution. A blob is defined by the intersection of a chain segment with other chains. The numbers ξ and g are related in the text.*

behaves like in a good solvent, and so one finds g monomers inside the blob according to $\xi \propto g^{\nu}$ with $\nu = 0.6$. Beyond the blob the correlation is lost (ξ is also called correlation length), and one says that the excluded volume interaction has been screened out. Beyond the blob the polymer behaves like an ideal coil:

$$R_g = \sqrt{3} \cdot \xi \cdot \left(\frac{N}{g}\right)^{1/2} \tag{91}$$

The fraction N/g is the number of blobs, which behave like an ideal coil, and ξ plays a role of a segment size. For low concentrations $\phi < \phi^*$ the blob becomes large, and finally covers a full single polymer $g \cong N$. Then, we recover the situation of a polymer in a good solvent (section 3.4). For very high concentrations $\phi \cong 1$ the blob sizes shrinks $g \cong 1$ and a blob contains a monomer approximately. Then we recover the situation of a polymer blend (compare section 3.3, Fig. 17). The overlap concentration ϕ^* is reached, when the monomer concentration of an isolated polymer reaches the overall polymer concentration:

$$\phi^* \propto N \cdot R_g^{-3} \propto N^{1-3\nu} \tag{92}$$

The respective N-dependence has been used then. Now the correlation length is expressed in the manner of a scaling argument:

$$\xi \propto R_g \left(\frac{\phi}{\phi^*}\right)^{y} \tag{93}$$

At the overlap concentration ξ has to agree with the radius of gyration, and at higher concentrations it behaves with the concentration like ϕ^y with an unknown exponent y. Another scaling argument says, that the blob-size does not depend on the degree of polymerization if the concentration remains constant. The polymer is assumed to be much larger than a blob and then there is no influence on the blobs anymore. From the N-dependence one obtains:

$$\xi \propto N^{\nu} \left(\frac{\phi}{N^{1-3\nu}} \right)^{y} \propto N^{0} \tag{94}$$

And finally one can solve for the unknown exponent y:

$$\xi \propto \phi^{y} \qquad y = \nu/(1-3\nu) = -3/4 \tag{95}$$

This result can be used to derive the same dependence for the radius of gyration. One simply goes back to equation 91 and uses all known dependencies:

$$R_g^2 \propto \xi^2 \frac{N}{g} \propto \xi^{2-1/\nu} \tag{96}$$

$$\propto \phi^{(1-2\nu)/(3\nu-1)} = \phi^{-1/4} \tag{97}$$

The two dependencies of ξ and R_g as a function of the concentration ϕ were observed experimentally by neutron scattering [29] (see Figs. 22 and 23). For the measurement of the correlation length one uses protonated polymer in deuterated solvent. Since correlations are screened out beyond the blob, one just observes the blob-scattering, the scattering from the inside of the blob. The measurement of the radius of gyration is more tricky. One uses small amounts of deuterated polymer amongst protonated polymer dissolved in a normal solvent (equivalent to protonated polymer). Then the neutrons are sensitive to the small amounts of deuterated polymer, whose chains appear to be isolated. At small scattering vectors one observes whole chain properties such as the radius of gyration. The presented experiments agree well with the blob model of DeGennes.

4 Conclusion

In chapter 2 we have derived the chain conformation in terms of observables like the end-to-end distance and the radius of gyration. We developed several theories, some of which included rigidity. All models agree in the general conclusion that the typical size R scales with $\sqrt{N}\,\ell$ for ideal coils. N is the degree of polymerization and ℓ is a typical size for the monomer (or backbone length). On the other hand the typical size of rigid rod molecules scales like $N\ell$, which is the contour length. The model of semiflexible chains allows to interpolate between the two limits.

The distribution of the end-to-end distance is a Gaussian, independent of the distribution of a single segment. This is due to the central limit theorem. The distribution of the end-to-end distance is a Gaussian function as shown for the discrete and continuous description. The Green function of polymers has similarities with the diffusion process, and quantum mechanics in imaginary time.

In chapter 3 we described phenomena of chain molecules. The underlying theories always base on the developed concepts of chapter 2. The depletion effect in colloidal systems with dissolved polymer plays an important role for formulations of paints. The elasticity of polymer networks is important for intermediate filaments of the cytoskeleton, and for rubber. Scattering is of special interest for the observation of polymers, since neutrons and x-rays are sensitive to structural sizes of polymers. Deviations of polymer conformations are observed in good solvents – the excluded volume interaction leads to polymer swelling. Polymers in semidilute

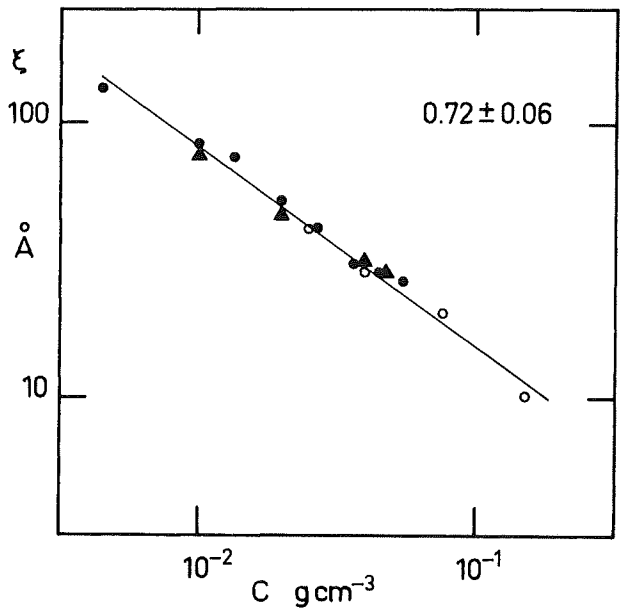


Fig. 22: *Neutron scattering experiment on semidilute polymer solutions [29]. The parameter ξ (single blob) is accessible by labelling the polymer network against the solvent.*

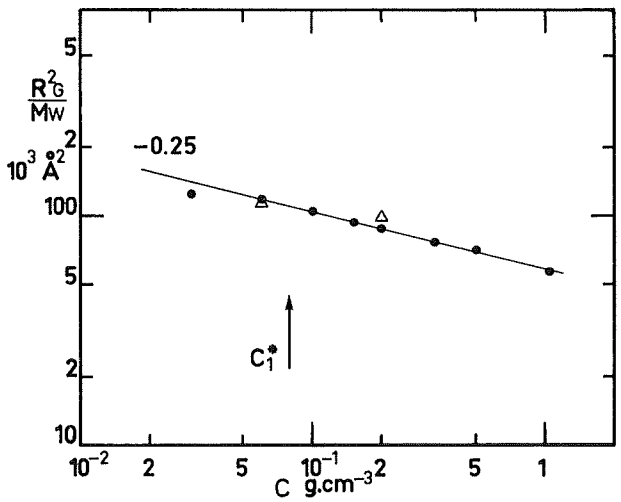


Fig. 23: *Neutron scattering experiment on semidilute polymer solutions [29]. The parameter R_g of the whole chain is accessible by labelling a few polymers against the polymer network and the solvent.*

solution form networks which are interesting for the cytoskeleton and glues. The here developed pictures and concepts are widely applied for many other phenomena of polymers, and therefore may give the reader a solid fundament in the field of polymer physics.

Appendices

A Appendix: Theorem of Lagrange

The theorem of Lagrange is used to eliminate the coordinates $\vec{s}_i$ relatively to the center of mass in the expression for the radius of gyration. Each such vector can be reformulated by the initial vector $\vec{s}_0$ plus a chain segment vector $\vec{r}_{0,i}$:

$$\vec{s}_i = \vec{s}_0 + \vec{r}_{0,i} \tag{98}$$

The square within the radius of gyration expression is then expanded:

$$\vec{R}_g^2 = \frac{1}{N+1}\sum_{i=0}^{N}(\vec{s}_0 + \vec{r}_{0,i}) \cdot (\vec{s}_0 + \vec{r}_{0,i}) \tag{99}$$

$$= \vec{s}_0^2 + 2 \cdot (N+1)^{-1}\vec{s}_0 \sum_{i=1}^{N} \vec{r}_{0,i} + (N+1)^{-1}\sum_{i=1}^{N} \vec{r}_{0,i}^2 \tag{100}$$

Now one should insert, that the sum of all center of mass coordinates is zero. This can be expressed in other ways using eq. 98:

$$\sum_{i=0}^{N} \vec{s}_i = 0 \quad \Rightarrow \quad \vec{s}_0 = -(N+1)^{-1}\sum_{i=1}^{N} \vec{r}_{0,i} \tag{101}$$

$$\vec{s}_0^2 = (N+1)^{-2}\sum_{i,j=1}^{N} \vec{r}_{0,i} \cdot \vec{r}_{0,j} \tag{102}$$

For the expanded expression of the radius of gyration one obtains then:

$$\vec{R}_g^2 = -(N+1)^{-2}\sum_{i,j=1}^{N} \vec{r}_{0,i} \cdot \vec{r}_{0,j} + (N+1)^{-1}\sum_{i=1}^{N} \vec{r}_{0,i}^2 \tag{103}$$

Another insertion is the law of cosines:

$$\vec{r}_{0,i} \cdot \vec{r}_{0,j} = \tfrac{1}{2}(\vec{r}_{0,i}^2 + \vec{r}_{0,j}^2 - \vec{r}_{i,j}^2) \tag{104}$$

This finally allows to reformulate the radius of gyration further. After completing the sums to run from 0 instead of 1 again, one finally obtains the desired result:

$$\vec{R}_g^2 = \frac{1}{2}(N+1)^{-2}\left(\sum_{i,j=1}^{N}\vec{r}_{i,j}^{\,2} + 2\sum_{i=1}^{N}\vec{r}_{0,i}^{\,2}\right) \tag{105}$$

$$= \frac{1}{2}(N+1)^{-2}\sum_{i,j=0}^{N}\vec{r}_{i,j}^{\,2} \tag{106}$$

$$= (N+1)^{-2}\sum_{0\le i<j\le N}\vec{r}_{i,j}^{\,2} \tag{107}$$

This theorem is not to be confused with the theorem of Steiner/Huygens, which only considers the expansion (eq. 100) with the center of mass coordinates (eq. 101).

References

[1] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969)

[2] R. G. Larson, *The Structure and Rheology of Complex Fluids* (Oxford University Press, Oxford, 1999)

[3] J. des Cloizeaux, G. Jannink, *Polymers in Solution: Their Modelling and Structure* (Clarendon Press, Oxford, 1990)

[4] M. Doi, S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, Oxford, 1986)

[5] I. W. Hamley, *Introduction to Soft Matter* (Wiley, New York, 2000)

[6] R. J. Roe, *Methods of X-ray and Neutron Scattering in Polymer Science* (Oxford University Press, Oxford, 2000)

[7] J. S. Higgins, H. C. Benoît, *Polymers and Neutron Scattering* (Clarendon Press, Oxford, 1994)

[8] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953)

[9] P. G. DeGennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979)

[10] K. Sturm in 22. IFF-Ferienkurs, *Physik der Polymere* (Forschungszentrum Jülich GmbH, Jülich, 1991)

[11] J. Bandrup, E. H. Immergut, E. A. Grulke, *Polymer Handbook* (Wiley, New York, 1999), especially VII/47

[12] L. J Fetters, D. J. Lohse, R. H. Colby in J. E. Mark *Physical Properties of Polymers Handbook* (American Institute of Physics, Woodbury, 1996)

[13] L. D. Landau, *Statistische Physik* (Akademie Verlag, Berlin, 1979), especially chapter XII §127

[14] J. Käs, H. Strey, J. X. Tang, D. Finger, R. Ezzell, E. Sackmann, P. A. Janmey, Biophys. J. **70**, 609 (1996)

[15] H. Benoît, P. Doty, J. Chem. Phys. **57**, 958 (1953)

[16] M. G. Bawendi, K. F. Freed, J. Chem. Phys. **83**, 2491 (1985)

[17] F. Schwabl, *Statistical Mechanics* (Springer, Berlin, 2002)

[18] E. Eisenriegler in 33. IFF-Ferienkurs, *Soft Matter* (Forschungszentrum Jülich GmbH, Jülich, 2002)

[19] X. Ye, T. Narayanan, P. Tong, J. S. Huang, M. Y. Lin, B. L. Carvalho, L. J. Fetters, Phys. Rev. E **54**, 6500 (1996)

[20] T. Springer in 25. IFF-Ferienkurs, *Komplexe Systeme zwischen Atom und Festkörper* (Forschungszentrum Jülich GmbH, Jülich, 1994)

[21] L. R. G. Treloar, Trans. Faraday Soc. **40**, 59 (1944)

[22] S. Hoffmann, in 33. IFF-Ferienkurs, *Soft Matter* (Forschungszentrum Jülich GmbH, Jülich, 2002)

[23] D. Schwahn in 22. IFF-Ferienkurs, *Physik der Polymere* (Forschungszentrum Jülich GmbH, Jülich, 1991)

[24] D. Schwahn in T. Brückel, G. Heger, D. Richter, R. Zorn, *5th Laboratory Course on Neutron Scattering* (Forschungszentrum Jülich GmbH, Jülich, 2001), see Lecture 14

[25] J. S. Pedersen, P. Schurtenberger, Macromol. **29**, 7602 (1996)

[26] D. Pötschke, P. Hickl, M. Ballauff, P. O. Åstrand, J. S. Pedersen, Macromol. Theory Simul. **9**, 345 (2000)

[27] G. Jerke, J. S. Pedersen, S. U. Egelhaaf, P. Schurtenberger, Phys. Rev. E **56**, 5772 (1997)

[28] R. G. Winkler, P. Reineker, L. Harnau, J. Chem. Phys. **101**, 8119 (1994)

[29] M. Daoud, J. P. Cotton, B. Farnoux, G. Jannink, G. Sarma, H. Benoît, R. Duplessix, C. Picot, P. G. DeGennes, Macromol. **8**, 804 (1975)

C 2 Polymer Dynamics

D. Richter
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Entropy Driven Dynamics 3
2.1 Gaussian Chains 3
2.2 The Rouse Model 4
2.3 Neutron Spin Echo Results 9

3 Reptation 14
3.1 Mean-Squared Displacements 15
3.2 Self-Correlation Function 16
3.3 Single Chain Dynamic Structure Factor 16
Neutron Spin Echo Results on Chain Confinement 18
Contour Length Fluctuations 21
3.4 Reptation of a Biopolymer 25

4 Hydrodynamic Interactions and Polymer Solutions 28
4.1 The Zimm Model 28
4.2 Results on Dilute Solutions 30
4.3 Semidilute Solutions 31
4.4 Chain Specific Effects Limiting Universal Behaviour 35
4.5 Dynamics of Biopolymers in Solution 40

5 Summary and Conclusion 43

1 Introduction

In our every day life plastics or polymers play a very important role. Polymeric materials are used, because they are durable, cheaply to produce, easily to process and because they exhibit very versatile and favourable mechanical properties, e.g. depending on temperature or time the the same polymer may be viscous or rubber elastic, very tough with high impact strength or even brittle. In the simplest case polymers are long linear chain molecules, build from one repeating unit, the monomer; such polymers are called linear homopolymers. Other than synthetic polymers, biopolymers are highly complex molecular structures with a sharp molecular weight and an uniquely defined sequence of building blocks. They are the molecular manifestation of life with a high variety of structure and specific function. Nevertheless, at least in their denaturated state they exhibit similarities with synthetic materials, while in their folded state the specific function evolves. Also the general principles of self-assembly and structure formation are common to both species. Finally, for biopolymers aside of structural aspects, also dynamic features are playing a key role in their function.

This lecture aims to identify general principles of chain motion on a molecular scale which underlie the macroscopic properties, and presents concepts and experimental results on these motional mechanisms in space and time. We will mainly address the dynamics of homopolymers and give a few examples for studies on the molecular dynamics of biopolymers.

Neutron scattering with its space time sensitivity on a molecular and atomic scale unravels the details of the molecular motions in question. Commencing at the scale of the single bond, where movements take place at a pace as in normal liquids, quasielastic neutron scattering (QENS) provides insight into local relaxation processes. At larger length scales first the entropy driven Rouse motion and at even larger distances the effect of entanglement constraints due to the mutual interpenetration of chains comes into the observation range. The most powerful technique suitable for these investigations, the neutron spin echo spectroscopy (NSE) operates in the time domain and uncovers a time range from about *2ps* to *200ns* and accesses momentum transfers between 0.01Å^{-1} and 3Å^{-1}.

Coherent quasi- and inelastic neutron scattering reveals the dynamic structure factor $S(Q,t)$ or its Fourier transformed counterpart $S(Q,\omega)$ [1]

$$S(Q,t) = \frac{1}{N}\sum_{ij}\left\langle \exp\left(-i\underline{Q}\underline{r}_j(t)\right)\exp\left(i\underline{Q}\underline{r}_i(0)\right)\right\rangle \qquad (1)$$

where $\underline{r}_j(t)$ and $\underline{r}_i(0)$ are the position vectors of the scatterers at time t and time $t=0$ respectively, N is the number of scatterers and $\hbar\underline{Q}$ is the momentum transfer during scattering. The brackets denote the thermal average. $S(Q,t)$ reflects the pair correlation function and relates to the collective properties of a material. In the neutron cross section it is weighted by the average scattering length $\left|\bar{b}\right|^2$.

Incoherent scattering is related to scattering length disorder which may either origin from spin dependent scattering lengths like in the case of hydrogen or from isotope mixtures with different scattering properties. This disorder prevents constructive interference of partial

waves scattered at different atoms and reveals the self correlation function. Eq.[1] provides the self correlation function, if in the double sum only terms with $i = j$ are considered. In the cross section $S_{inc}(Q,t)$ is weighted by the average scattering length fluctuations $\left(\left\langle b^2\right\rangle - \left|\bar{b}^2\right|\right)$.

In Gaussian approximation which is widely used for the calculation of neutron dynamic structure factors for polymer dynamics Eq.[1] is approximated by

$$S(Q,t) = \frac{1}{N}\sum_{ij}\exp\left[-\frac{Q^2}{6}\left\langle\left(r_i(t)-r_j(0)\right)^2\right\rangle\right] \qquad (2)$$

This lecture naturally is not able to review exhaustively the contribution of high resolution neutron scattering to the field of polymer dynamics, but rather wants in an exemplary way to display important contributions by example.

We will start with the standard model of polymer dynamics, the so called Rouse model, which considers the entropy driven dynamics of chains in a heat bath. Then we will inspect the influence of topological constraints arising from the mutually interpenetrating chains in a dense ensemble. This will lead us to the concept of reptation with manifestations in melts of flexible polymers and the motion of biopolymers in semidilute solution. Thereafter we address the role of hydrodynamic interactions in polymer solutions and discuss the dynamics of polymer chains in dilute and semidilute solutions. Finally, we will briefly consider neutron scattering experiments on proteins in solution and discuss the diffusion of myoglobin in dense environment.

2 Entropy Driven Dynamics

The dynamics of a generic linear, Gaussian chain – as described in the Rouse model [2] – is the starting point and standard description for the Brownian dynamics in polymer melts. In this model the conformational entropy of a chain acts as a resource for restoring forces for chain conformations deviating from thermal equilibrium. In this chapter we deal with these entropy driven dynamics in terms of the Rouse model and present NSE results on the space-time evolution of the Rouse relaxation. Here, we like to exemplify the mathematical treatment of chain dynamics problems. Therefore, we detail the description such that it may be followed in all steps. In the discussion of further models later on for details we refer to the relevant literature.

2.1 Gaussian Chains

The conformation of a flexible linear polymer chain is - on scales somewhat larger than the main chain bond length ℓ_0 – close to a random walk. The conformations of such a chain are described by a set of segment vectors $\{\underline{r}(n)\}$ with $\underline{r}(n) = (\underline{R}(n) - \underline{R}(n\text{-}1))$, where $\underline{R}(n)$ is the position vector of segment n. Following the central limit theorem, the length distribution of a vector $\underline{r}$ connecting segments that have a topological distance of n steps is Gaussian.

$$\phi(r,n) = \left(\frac{3}{2\pi n\ell^2}\right)^{3/2}\exp\left(-\frac{3r^2}{2n\ell^2}\right) \qquad (3)$$

where ℓ is the segment length. The Rouse model bases on a further idealization of the chain statistics assuming that the bond vector $\underline{r}$ of hypothetic connecting points of the chain has a Gaussian distribution of length

$$\psi(\ell) = \left(\frac{3}{2\pi b^2}\right)^{3/2} \exp\left(-\frac{3\ell^2}{2b^2}\right) \tag{4}$$

Yielding $\langle \ell^2 \rangle = b^2$ For simplicity throughout the paper we will take $b^2 = \ell_0^2 C_\infty = \ell^2$ keeping however in mind that the building block of a Gaussian chain may well contain a larger number of main chain bonds. C_∞ is the characteristic ratio accounting for the local stiffness arising from the non-random bond angle distribution of the bonds of length ℓ_0.

The probability of a conformation $\{\underline{r}(n)\}$ follows as

$$P(\{\underline{r}(n)\}) = \prod_{n=1}^{N} \psi(r_n) = \left(\frac{3}{2\pi\ell^2}\right)^{\frac{3}{2}N} \exp\left[-\sum_{n=1}^{N} \frac{3\left(\left(\underline{R}(n) - \underline{R}(n-1)\right)^2\right)}{2\ell^2}\right] \tag{5}$$

with N counting the number of segments of the chain. The free energy of a Gaussian chain is entirely described by its conformational entropy

$$S = k_B \ell n\left[P(\{r(n)\})\right] \tag{6}$$

The Gaussian chain model yields a spring constant even for a single "bond" $k = 3k_BT/\ell^2$, where k_B is the Boltzmann constant. From Eq.[5] the chain extension between arbitrary points along the chain may be computed to $\left\langle \left(R(n) - R(m)\right)^2 \right\rangle = |n - m|\,\ell^2$.

2.2 The Rouse Model

The Rouse model starts from such a Gaussian chain representing a coarse grained polymer model, where springs represent the entropic forces between hypothetic beads [3] (Fig.1).

We are interested in the motion of segments on a length scale $\ell < r < R_E$, where $R_E^2 = N\ell^2$ is the end-to-end distance of the chain. The segments are subject to an entropic force resulting from Eq.[6] (x-components):

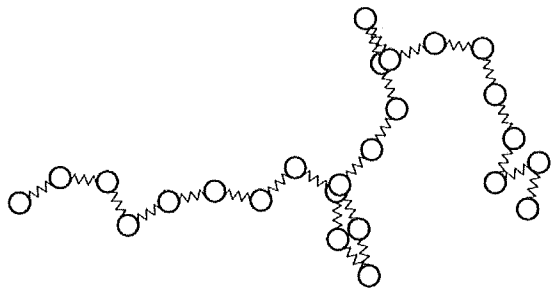


Fig. 1: *Spring-bead model of a Gaussian chain as assumed in the Rouse model. The beads are connected by entropic springs and are subject to a frictional force* $\zeta_0 v$*, where* v *is the bead velocity and* ζ_0 *the bead friction coefficient.*

$$\frac{\partial}{\partial x(n)} k_B T \ell n P\left(\{x(n)\}\right) = \frac{3k_B T}{\ell^2}\left[x(n+1) - 2x(n) + x(n-1)\right] \tag{7}$$

and a stochastic force $f_x(n,t)$ which fulfils:
$\langle f_x(n,t)\rangle = 0$ and $\langle f_\alpha(n,t)\ f_\beta(m,0)\rangle = 2k_B T \zeta_0\ \delta_{nm}\ \delta_{\alpha\beta}\ \delta(t)$; ζ_0 denotes a friction coefficient and α, β the Cartesian components.

With this the Langevin equation for segment motion assumes the form

$$\zeta_0 \frac{\partial x(n)}{\partial t} = \frac{3k_B T}{\ell^2}\left[x(n+1) - 2x(n) + x(n-1)\right] + f_x(n,t) \tag{8}$$

Regarding the index n as a continuous variable we arrive at

$$\zeta_0 \frac{\partial x}{\partial t} = \frac{3k_B T}{\ell^2} \frac{\partial^2 x(n)}{\partial n^2} + f_x(n,t) \tag{9}$$

The boundary condition of force free ends requires $\left.\frac{\partial x(n)}{\partial n}\right|_{n=0,N} = 0$. The partial differential equation is solved by cos-Fourier transformation to normal coordinates fulfilling the boundary conditions

$$\tilde{x}(p,t) = \frac{1}{N}\int_0^N dn \cos\left(\frac{p\pi n}{N}\right) x(n,t) \tag{10}$$

In normal coordinates the Langevin Eq.[9] becomes

$$\zeta_p \frac{\partial}{\partial t}\tilde{x}(p,t) = -k_p \tilde{x}(p,t) + \tilde{f}(p,t) \tag{11}$$

where $k_p = \frac{2\pi^2 p^2}{N} k$, $\zeta_p = 2N\zeta_0$. For the stochastic forces we have

$$\left\langle f_\alpha(p,t)\, f_\beta(q,0)\right\rangle = 2\zeta_p\, k_B T\, \delta_{pq}\, \delta_{\alpha\beta}\, \delta(t-t') \tag{12}$$

Eq.[11] is readily solved by a single exponential

$$\tilde{x}(p,t) = \frac{1}{\zeta_p} \int_{-\infty}^{t} dt \exp\left((t-t')/\tau_p\right) \tilde{f}(p,t') \tag{13}$$

where the mode relaxation time τ_p is given by

$$\tau_p = \frac{\zeta_p}{k_p} = \frac{\zeta_0 N^2 \ell^2}{3\pi^2 k_B T\, p^2} = \frac{\tau_R}{p^2} = \frac{N^2}{W\pi^2 p^2};\; W = \frac{3k_B T}{\ell^2 \zeta_0} \tag{14}$$

where τ_R is the Rouse time - the longest time in the relaxations spectrum – and W is the elementary Rouse rate. The correlation function $\left\langle \tilde{x}(p,t)\, \tilde{x}(p,0)\right\rangle$ of the normal coordinates is finally obtained by

$$\begin{aligned}
\left\langle \tilde{x}(p,t)\, \tilde{x}(p,0)\right\rangle &= \frac{1}{\zeta_p^2} \int_{-\infty}^{t} dt_1 \int_{-\infty}^{0} dt_2 \exp\left[-(t-t_1-t_2)/\tau_p\right] \left\langle f(p,t_1)\, f(p,t_2)\right\rangle \\
&= \frac{1}{\zeta_p^2} \int_{-\infty}^{t} dt_1 \int_{-\infty}^{0} dt_2 \exp\left[-(t-t_1-t_2)/\tau_p\right] 2\zeta_p\, k_B T\, \delta(t_1-t_2) \\
&= \frac{k_B T}{k_p} \exp\left(t/\tau_p\right) = \frac{N\ell^2}{6\pi^2 p^2} \exp\left(-t/\tau_p\right)
\end{aligned} \tag{15}$$

For the center of mass coordinate one finds

$$\left\langle \tilde{x}(0,t)\, \tilde{x}(0,0)\right\rangle = \frac{2k_B T}{N\zeta_0} t \tag{16}$$

Scattering experiments are sensitive to the mean-square segment correlation functions.

$$B(n,m,t) = 3\, \left\langle \left[x(n,t) - x(m,0)\right]^2 \right\rangle \tag{17}$$

They are obtained by back-transformation of the normal coordinates (inverse of Eq.[10])

$$\begin{aligned}
B(n,m,t) &= 6D_R t + |n-m|\ell^2 + \frac{4N\ell^2}{\pi^2} \\
&\times \sum_{p=1}^{N} \frac{1}{p^2} \cos\left(\frac{p\pi m}{N}\right) \cos\left(\frac{p\pi m}{N}\right) \left[1 - \exp\left(-\frac{tp^2}{\tau_R}\right)\right]
\end{aligned} \tag{18}$$

In order to arrive at Eq.[18], we have used $D_R = \frac{1}{6t}\left\langle \left(x(0,t) - x(0,0)\right)^2 \right\rangle$ (D_R: Rouse diffusion coefficient) and $\sum_{p=1}^{\infty} \frac{1}{p^2}\left[\cos\left(\frac{p\pi m}{N}\right) - \cos\left(\frac{p\pi n}{N}\right)\right]^2 = \frac{\pi^2}{2N}|n-m|$. For the special case of the self correlation function ($n = m$) $B(n,n,t)$ reveals the mean-square displacement of a polymer segment. For large p the $\cos^2$ in Eq.[18] is a rapidly oscillating function which may be replaced by the mean-value ½. With this approximation we convert the sum into an integral and obtain

$$\begin{aligned} B(n,n,t) = \left\langle r^2(t)\right\rangle &= \frac{4N\ell^2}{\pi^2}\int_0^{\infty} dp \frac{1}{p^2}\frac{1}{2}\left(1 - \exp\left(-tp^2/\tau_R\right)\right) + 6D_R T \\ &= 2N\ell^2\left(\frac{t}{\pi^3\tau_R}\right) + 6D_R t = 2\ell^2\left(\frac{3k_B T}{\pi\zeta_0\ell^2}\right)^{1/2} + 6D_R t \end{aligned} \tag{19}$$

In contrast to normal diffusion in the segmental regime the mean-square displacement does not grow linearly, but with the square route of time. For the translational diffusion coefficient $D_R = k_B T / N\,\zeta_0 = \frac{W\ell^4}{3N\ell^2} = \frac{W\ell^4}{3R_E^2}$ is obtained; D_R is inversely proportional to the number of friction performing segments.

The self correlation function relates directly to the mean-square displacement of the diffusing segments. Inserting Eq.[19] into Eq.[2] for $t < \tau_R$ we have

$$S_{self}(Q,t) = \exp\left[-Q^2 D_R t\right]\ \exp\left\{-\frac{2}{\sqrt{\pi}}\left(\frac{k_B T\ell^2}{12\zeta_0}Q^4 t\right)^{1/2}\right\} \tag{20}$$

In the case of coherent scattering, which observes the pair correlation function, interferences from scattering waves emanating for various segments complicate the scattering function.

$$\begin{aligned} S_{chain}(Q,t) = \frac{1}{N}\exp\ \left[-Q^2 D_R t\right]\ \sum_{n,m}\exp\ \left\{-\frac{1}{6}|n-m|\ Q^2\ell^2\right\}\ * \\ \exp\left\{-\frac{2}{3}\frac{R_E^2 Q^2}{\pi^2}\sum_p \frac{1}{p^2}\left\{\cos\left(\frac{p\pi m}{N}\right)\cos\left(\frac{p\pi m}{N}\right)\left(1 - \exp\left(-\frac{tp^2}{\tau_R}\right)\right)\right\}\right\} \end{aligned} \tag{21}$$

for small Q ($QR_E < 1$) the second and third terms are negligible and $S_{chain}(Q,t)$ describes the centre of mass diffusion of the chain.

$$S_{self}(Q,t) = \frac{1}{N}S_{chain}(Q,t) = \exp\left(-D_R Q^2 t\right) \tag{22}$$

For $QR_E > 1$ and $t < \tau_R$ the internal relaxations dominate. For $t = 0$ we have $S_{chain}(Q,t) = S_{chain}(Q)$; i.e. the structure factor corresponds to a snapshot of the chain structure.

$$S_{chain}(Q) = \frac{1}{N}\sum_{n,m}\exp\left(-\frac{1}{6}Q^2|n-m|\ell^2\right) \tag{23}$$

Replacing the summations by integrals and observing the relation $R_g^2 = \frac{1}{6}N\ell^2$ for the radius of gyration Eq.[23] immediately leads to the well known Debye function

$$\begin{aligned} S_{chain}(Q) &= N f_{Debye}\left(Q^2R_g^2\right) \\ f_{Debye}(x) &= \frac{2}{x^2}\left(e^{-x}-1+x\right) \end{aligned} \tag{24}$$

The shape of the Debye function corresponds to the uppermost curve in Fig.2.

Important scaling properties are revealed by an approximate computation of the high-Q behaviour of $S_{chain}(Q,t)$ [4]. Replacing sums by integrals and performing some simplifications.

$$\begin{aligned} S_{chain}(Q,t) &= \frac{12}{Q^2\ell^2}\int_0^\infty du\exp\left\{-u-(\Omega_R t)^{1/2}\ h\left(u(\Omega_R t)^{-1/2}\right)\right\} \\ \text{with}\qquad h(y) &= \frac{2}{\pi}\int_0^\infty dx\ \frac{\cos(xy)}{x^2}\left(1-\exp\left(-x^2\right)\right) \end{aligned} \tag{25}$$

is obtained. Note that Eq.[25] only depends on one variable, the Rouse variable

$$(\Omega_R t)^{1/2} = \frac{Q^2}{6}\sqrt{\frac{3k_BT\,\ell^2 t}{\zeta_0}} = \frac{Q^2\ell^2}{6}\sqrt{Wt} \tag{26}$$

Since the Rouse model does not contain an explicit length scale, for different momentum transfers the dynamic structure factors are predicted to collapse to one master curve, if they are represented as a function of the Rouse variable.

Further note that for $t = 0$ Eq.[25] does not resemble the Debye function but yields its high Q-limiting behaviour $\propto Q^{-2}$ -, it is only valid for $QR_g >> 1$. In that regime the form of Ω_R immediately reveals that the intra-chain relaxation increases $\propto Q^4$ in contrast to normal diffusion $\propto Q^2$. Finally, Fig.2 illustrates the time development of the structure factor.

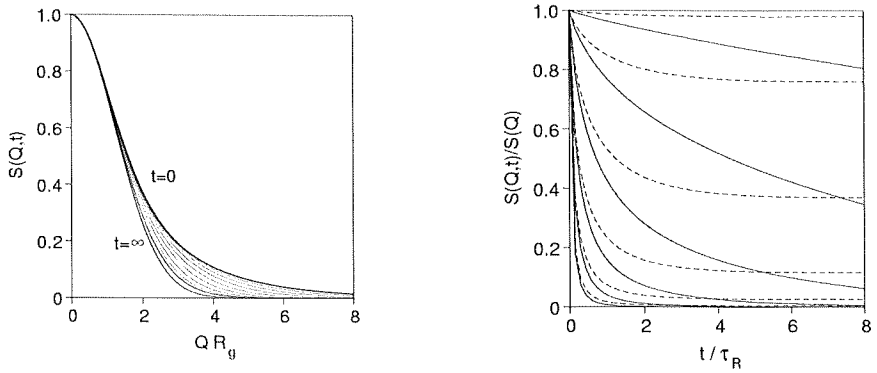


Fig. 2: *Development of S_{chain} (Q,t) for different times (left) and the normalized relaxation function $S_{chain}(Q,t)/S(Q)$ (right) for QR_g = 1, 2, ... 6. The dashed lines contain only the intra-chain relaxation, whereas the solid lines include the center-of-mass diffusion. Note that for short chains, respectively for small Q the diffusion dominates the observed dynamics.*

2.3 Neutron Spin Echo Results

The self correlation function of a Rouse chain was first observed on polydimethylsiloxane (PDMS) [5]. Since a straight forward study of the incoherent scattering by NSE is very difficult – due to spin flip scattering a severe loss of polarization occurs leading to very weak signals – the measurements of the self correlation function were performed on high molecular weight deuterated PDMS chains which contained short protonated labels at random positions. In such a sample the scattering essentially originates from the contrast between the protonated sequence and a deuterated environment and therefore is coherent. On the other hand the sequences are randomly distributed, so that there is no constructive interference of partial waves arising from different sequences. Under these conditions the scattering experiments measures the self correlation function.

In Fig.3 the corresponding NSE spectra are plotted against the scaling variable of the Rouse model. As predicted by Eq.[20] the results for the different momentum transfers follow a common straight line.

Recently it became also possible to observe directly to incoherent cross section from a protonated chain. Fig.4 displays the time dependent mean-square displacement obtained from a high molecular weight (M_w = 80000) monodisperse polyethylene-propylene (PEP) melt at 492K [6].

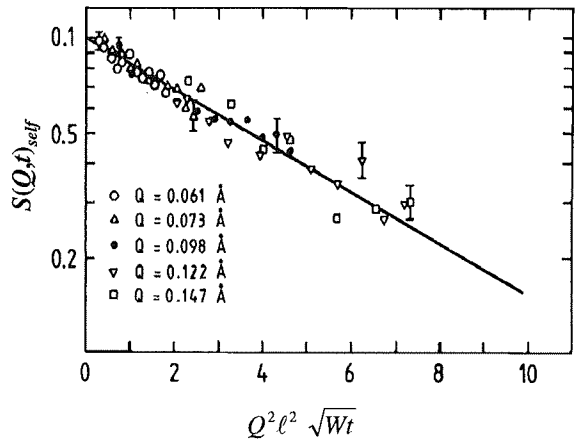


Fig. 3: *Self correlation for a PDMS melt T = 100C. The data at different momentum transfers are plotted vs. the scaling variable of the Rouse model.*

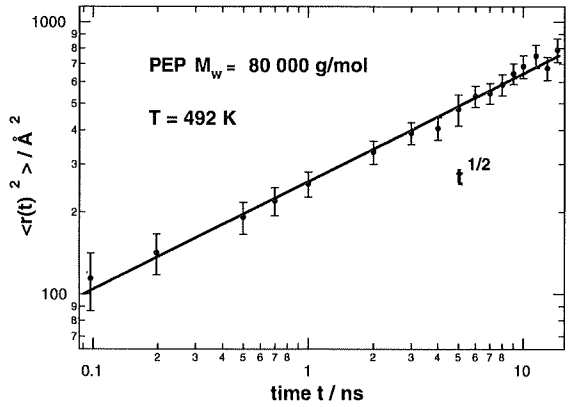


Fig. 4: *Time dependent mean-square displacement of a PEP segment. The solid line indicates the prediction of the Rouse model.*

In Gaussian approximation for negligable translational diffusion Eq.[19] in connection with Eq.[20] yields

$$\left\langle r^2(t) = -\frac{6}{Q^2}\, \ell n\left[S_{self}(Q,t)\right]\right\rangle \tag{27}$$

The mean-square displacement follows with high accuracy the predicted square root law in time. Since neutron quasielastic scattering resolves dynamic processes in space and time, these measurements give direct information about the segment displacement at a given time,

e.g. at 10ns $< r^2(10ns) > = 620Å^2$ i.e. the average proton has traveled 25Å during this time interval.

The pair correlation function arising from the segment motion within one given chain is observed, if some protonated chains are dissolved in a deuterated matrix. Fig.5 displays the observed spectra from polyethylethylene (90% dPEE, 10% hPEE) with a molecular weight of $M_w^h = 21.5kg/mol$; $M_w^d = 24.5kg/mol$ and a narrow molecular weight distribution [7]. The solid lines give the prediction of the dynamic structure factor of Eq.[21]. Obviously very good agreement is achieved.

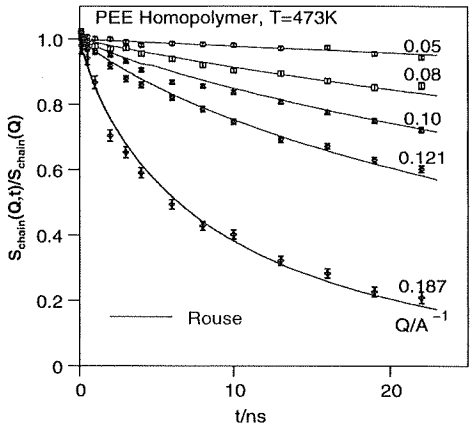


Fig. 5: *Single chain structure factor from a PEE melt at 473K. The numbers along the curves represent the experimental Q-values in [Å^{-1}]. The solid lines are a joint fit with the Rouse model (Eq.[21]).*

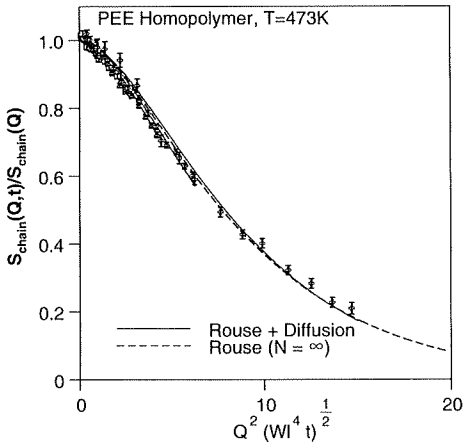


Fig. 6: *Single chain structure factor from PEE melts as a function of the Rouse scaling variable. The dashed line displays the Rouse prediction for infinite chains, the solid lines incorporate the effect of translational diffusion. The different symbols relate to the spectra displayed in Fig.5.*

We now use these data, in order to investigate the scaling prediction inherent in Eq.[25]. Fig.6 presents a plot of the data of Fig.5, now as a function of the Rouse scaling variable (Eq.[26]).

The data follow with satisfying precision the scaling prediction. The small deviations are related to the translational diffusion of the chains. This becomes evident from Fig.7, where the obtained relaxation rates $\Gamma(Q)$ are plotted versus Q in a double logarithmic fashion.

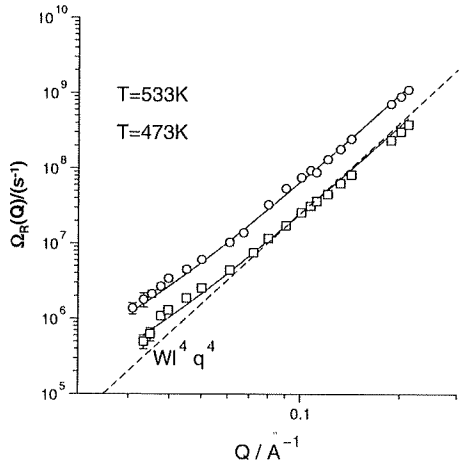


Fig. 7: *Relaxation rates from PEE melts vs. Q for two different temperatures. The dashed line represents the $\Omega_R \sim Q^4$ prediction of the Rouse model. The solid lines include the contribution from translational diffusion (Eq.[28]).*

The dashed line gives the Rouse prediction $\Omega_R \propto W\ell^4Q^4$. While at larger momentum transfers the experimental results follow very well this prediction, towards lower Q, a systematic relative increase of the relaxation rate is observed. Including translational diffusion, we have

$$\Omega_R(Q) = Q^2\left[D + Q^2\frac{W\ell^4}{6}\right] = Q^2W\ell^2\left[\frac{\ell^2}{3R_E^2} + \frac{Q^2\ell^2}{6}\right] \tag{28}$$

The solid lines in Fig.7 represents the prediction of Eq.[28]. Perfect agreement is obtained.

The above expressions provide an universal description of the dynamics of a Gaussian chain and are valid for real linear polymer chains on intermediate length scale. The specific (chemical) properties of a polymer enter only in terms of two parameters $N\ell^2 = R_E^2$ and ℓ^2/ζ_0. The friction parameter is governing the Rouse variable in terms of $W\ell^4 = \frac{3k_BT\,\ell^2}{\zeta_0}$. As eluded to in Eq.[28] also the center of mass diffusion coefficient may be expressed in these terms $D = k_BT\,\ell^2/\left(\zeta_0R_E^2\right)$. Since the Rouse model does not contain an inherent length scale, the parameter N (chain length) and ℓ^2 (segment length squared) are somewhat arbitrary as long as

the physical values of ℓ^2/ζ_0 and R_E^2 are kept constant. The NSE experiments measure directly the friction coefficient/length squared.

Setting the time scale, the monomeric friction coefficient is a basic quantity in all rheological measurements. There this quantity is inferred indirectly (i) either from measurements of the dynamic modulus $G(\omega)$ in the first relaxation regime. Here in terms of the Rouse model the relaxation spectrum H is given by [8]

$$H(\omega) = \frac{\sqrt{\langle \ell^2_{mon} \rangle}\, \rho N_A}{2\pi\, M_0} \left(\frac{\zeta k_B T}{6} \right)^{1/2} \omega^{1/2} \tag{29}$$

M_0 is the monomer mass, $\sqrt{\langle \ell^2_{mo} \rangle} = \sqrt{n C_\infty \ell_0^2}$ the size of the monomer, n the number of bonds/monomer, ρ the polymer density, N_A the Avogadro number, ζ the friction coefficient related to the monomer size and ω the relaxation frequency; or (ii) from viscosity measurements, where in the Rouse regime

$$\eta = \frac{N_A}{36} \frac{\rho}{M_0} \zeta\, \ell^2_{mon}\, N_{mon} \tag{30}$$

is predicted with $N_{mon} = N/n$. Very often also viscosity studies at high molecular weight are used. Then the effect of entanglements has to be considered in an empirical fashion. The NSE experiments in the Rouse regime provide direct microscopic access to ζ. From $W\ell^4 = \frac{3k_B T\, \ell^2}{\zeta_0}$ we obtain

$$\zeta = \left(\frac{3k_B T}{W \ell^4} \right) \ell^2_{mon} \tag{31}$$

ℓ^2_{mon} is obtained from measurements of the chain end-to-end distance by $\ell^2_{mon} = \frac{R_E^2}{M} M_0$. A detailed tabulation is e.g. given in ref. [9]. NSE experiments on polymer melts in the Rouse regime have been performed by now on seven different polymers including two studies of a systematic temperature dependence [7,10-14]. Table 1 presents the monomeric friction coefficient originating from these measurements and compares them with rheological data taken mostly from Ferry's book. While in some instances very good agreement is evident (PDMS, PVE) in other cases substantial deviations are visible. Also the method, how the rheological results were obtained, seems to determine significantly the rheological friction coefficient e.g. in the case PEP a rescaling of the high molecular weight viscosity at 373K gives $\zeta^{VIS} = 6.3 \cdot 10^{-11}\ {}^{Ns}/_{m}$ while an evaluation from the dynamic modulus yields $\zeta^{Mod} = 0.98 \cdot 10^{-11}\ {}^{Ns}/_{m}$. The microscopic value of $3.8 \cdot 10^{-9}\ {}^{Ns}/_{m}$ is just in between.

Table 1: Monomeric friction coefficients from NSE experiments compared to rheological data from the compilation of Ferry [8].

Polymer	Temperature	$\zeta^{NSE}\left[10^{11}\frac{Ns}{m}\right]$	$\zeta^{rheo}\left[10^{11}\frac{Ns}{m}\right]$
PVE [10]	408	1.59	1.26
PI [10]	408	0.569	0.19
1,4 PB [11]	353	0.60	3.1
PEP [12,13]	373	3.82	0.98 (6.3*)
PE [12,13]	415	0.18	0.255*
PDMS [14]	373	0.298	0.282
PEE [7]	473	0.695	

*)J. Luettmer-Strathmann, Int. J. Thermophys. **22**, 1507 (2001)

3 Reptation

The dynamic modulus of a polymer melt is characterized by a plateau in frequency which broadens with increasing chain length. In this plateau regime the polymer melt acts like a rubber, where the elastic properties are derived from the entropy elasticity of the chains between permanent cross links. The modulus of a rubber is inversely proportional to the mesh size and proportional to the temperature. In analogy, it is suggestive to assume that the entanglement or topological interactions between the chains in a melt lead to the formation of a temporary network which then displays rubber elastic properties. Other than in the rubber, for long times the chains may disentangle and the melt flows. This flow process is characterized by a melt viscosity $\eta \sim N^{3.2\ldots3.6}$ and a translational chain diffusion coefficient $D \sim N^{-2\ldots2.3}$ [3]. Using the analogy to the modulus of a rubber, we may estimate the distances between entanglement points from the plateau modulus G_N^0. For different polymers they come out to be between 30Å and 100Å. On the basis of such assumptions, a number of theories of viscoelasticity have been developed [3,15-17]. The most famous among them is the reptation model of DeGennes [18] and Doi and Edwards [15]. In this model, the dominating chain motion is a reptile like creep along the chain profile. The lateral restrictions are modeled by a tube with a diameter d parallel to the chain profile. d relates to the plateau modulus of the melt.

$$d^2 = \frac{4}{5}\frac{R_E^2}{M}\frac{k_B T}{G_N^0} \tag{32}$$

The restrictions of the motions by the presence of other chains are not effective on a monomer scale but rather permit lateral freedom on intermediate length scales. The experimental observations for viscosity and diffusion can be made directly comprehensible in this simple model. The viscosity relates to the longest relaxation time in a system. If we consider Rouse diffusion along the tube with a Rouse diffusion coefficient $D_R \approx 1/(N\zeta_0)$ then an initial tube configuration is completely forgotten, when the mean-square displacement along the tube $\langle r^2(t)\rangle_{tube} = (contour\ length\ L)^2$. Thus, for the longest relaxation time, we obtain

$$\tau_\eta \approx \frac{L^2}{D_R} \approx N^3 \zeta_0 \tag{33}$$

The diffusion coefficient is found by considering that during this time in real space the mean-square displacement just amounts to the end to end distance of the chain squared. Thus, we have

$$D_{rep} = \frac{R_E^2}{\tau_\eta} \approx \frac{N}{N^3 \zeta_0} \approx \frac{1}{N^2 \zeta_0} \tag{34}$$

3.1 Mean-Squared Displacements

We now consider the predictions of the reptation model for the mean-square displacement of the chain segments.

For short times, when the chain segment has not yet realized the topological constraints ($r < d$), we expect unrestricted Rouse motion $\left\langle \Delta r^2(t) \right\rangle \approx t^{1/2}$ (Eq.[19]). Experimentally this was the case for PEP (Fig.4) where for an entangled chain for times up to 20ns and displacement up to about 30Å Rouse dynamics was observed. At a time τ_e, $\tau_e = \frac{d^4}{\pi^2 W \ell^4}$, the mean-square displacement reaches the order of the tube diameter. Then motional restrictions are expected.

For times $t > \tau_e$ one dimensional curve linear Rouse motion along the tube has to be considered. Displacements along the tube are described by Eq.[18] where we have to change real space coordinates to coordinates $s(t)$ along the tube.

If a segment is displaced along the tube by $\left\langle \left(s_n(t) - s_n(0)\right)^2 \right\rangle$ then the mean-square displacement in 3-d real space is $d\left\langle \left(s_n(t) - s_n(0)\right)^2 \right\rangle^{1/2}$. With that we obtain

$$\left\langle r^2(t) \right\rangle = \begin{cases} 2d\left(\dfrac{k_B T\, \ell^2 t}{\zeta_0 \pi}\right)^{1/4} & \tau_e < t < \tau_R \quad \textit{Rouse relaxation along the tube} \\ \\ 2d\left(\dfrac{k_B T\, t}{N \zeta_0}\right)^{1/2} & \tau_R < t < \tau_d \quad \textit{Diffusion along the tube} \end{cases} \tag{35}$$

In Fig.8 the two situations correspond to the second and third processes. The second process, where the chain performs Rouse motion along the tube is called "local reptation" while the creep like diffusion along the tube which eventually leads to a complete tube renewal is also termed pure reptation. The terminal time τ_d after which the chain has left its original tube determines to a large extend the viscosity of the melt ($\tau_d \approx \tau_\eta$; see Eq.[33]). Beyond that time reptation diffusion prevails.

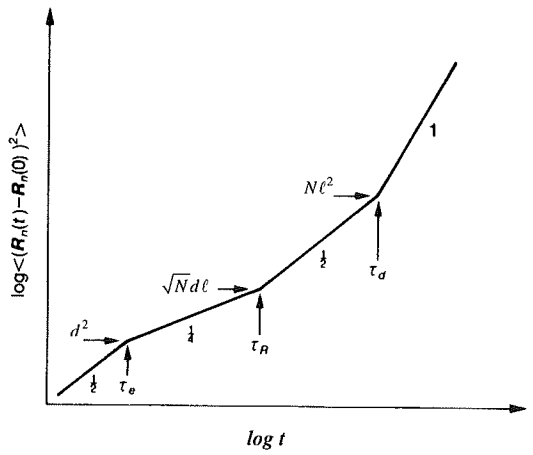


Fig. 8: *Mean-square displacement of a chain segment in the reptation model.*

3.2 Self-Correlation Function

In Gaussian approximation the self correlation function of a reptating chain directly relates to the above calculated mean-square displacements. This supposes that the Gaussian width after a diffusion *time t* of the single segment distribution along the 1-dimensional tube contour may be taken for the time dependent displacement. Projecting this on the Gaussian contorted tube then would again correspond to a Gaussian sublinear diffusion in real space.

However, the real process has to be modelled by projecting the segment probability distribution due to curvilinear Rouse motion with the curve linear coordinate - s - onto the random walk like contour path of the contorted tube. This leads to a non-Gaussian probability distribution of the segment at times $t > \tau_e$. The necessity to perform the proper averaging has first been shown by Fatkullin & Kimmich in the context of interpretation of field-gradient NMR diffusometry data [19]. Their result invalidates the Gaussian approximation for times above τ_e. We note that Eq.[36] is strictly valid only for $t >> \tau_e$ when $\langle r^2(t)\rangle >> d^2$.

$$S_{self}(Q,t > \tau_e) = \exp\left[\frac{Q^4 d^2}{72}\frac{\langle r^2(t)\rangle}{3}\right] erfc\left[\frac{Q^2 d}{6\sqrt{2}}\sqrt{\frac{\langle r^2(t)\rangle}{3}}\right] \qquad (36)$$

The effect on the scattering function is that if (wrongly) interpreted in terms of the Gaussian approximation the cross over to local reptation appears to occur at significant lower values of τ_e. The general asymptotic $t^{1/4}$ law remains untouched.

3.3 Single Chain Dynamic Structure Factor

Now we turn to the single chain dynamic structure factor, which is also strongly affected by the topological tube constraints. Qualitatively we would expect the following behavior

(i) At short times $t < \tau_e$ the chain will perform unrestricted Rouse motion and the dynamic structure factor of Equ's [21] or [25] should well describe the dynamics. This e.g. has been exemplified in early measurements of the Rouse dynamic structure factor of entangled PDMS melts - in this material d is large [20].

(ii) In the regime of local reptation the chain has already explored the tube laterally and further density fluctuations of the labeled chain will only be possible via Rouse relaxation along the tube. Under such circumstances the structure factor to a first approximation will mirror the form factor of the tube $S_{chain}(Q,\tau_R > t > \tau_e)/S_{chain}(Q) \approx \exp\left(-Q^2d^2/36\right)$. In this regime the experiment should reveal the size of the topological constraints without applying any detailed model.

(iii) In the creep regime $t > \tau_R$ the memory of the tube confinement will be gradually lost and the dynamic structure factor should reveal the fraction of the still confined polymer segments.

(iv) Finally, in the diffusive regime at very small Q ($QR_g << 1$) an equivalent to Eq.[22] will be valid, where now the reptation diffusion coefficient will be measured.

De Gennes [21] and Doi and Edwards [3] have formulated a tractable analytic expression for the dynamic structure factor. Thereby they neglected the initial Rouse regime i.e. the derived expression is valid for $t > \tau_e$ once confinement effects become important. The dynamic structure factor is composed from two contributions $S^{loc}(Q,t)$ and $S^{esc}(Q,t)$ reflecting local reptation and escape processes (creep motion) from the tube

$$\frac{S_{chain}(Q,t)}{S_{chain}(Q)} = \left[1-\exp\left(-\frac{Q^2d^2}{36}\right)\right] S^{loc} + \exp\left(-\frac{Q^2d^2}{36}\right) S^{\sec} \tag{37}$$

The local reptation part was calculated as

$$S^{loc}(Q,t) = \exp\left(\frac{t}{\tau_0}\right) erfc\left(\sqrt{t/\tau_0}\right) \tag{38}$$

where $\tau_0 = 36/\left(W\ell^4Q^4\right)$ and $erfc$ is the complementary error function $erfc(x) = 2/\sqrt{\pi} \int_x^\infty e^{-t^2} dt$. The general expression for $S^{esc}(Q,t)$ due to pure reptation was given by Doi and Edwards.

$$S^{\sec}(Q,t) = \sum_{p=1}^{\infty} \frac{\sin^2(\alpha_p)\, 2N\mu}{\alpha_p^2\left(\mu^2 + \alpha_p^2 + \mu\right)} \exp\left(-\frac{4t\alpha_p^2}{\pi^2\tau_d}\right) \tag{39}$$

where $\mu = Q^2 N \ell^2 / 12$ and α_p are the solutions of equation $\alpha_p \tan(\alpha_p) = \mu = \frac{3R_E^6}{\pi^2 W \ell^4 d^2}$; $\tau_d = \frac{3N^3 \ell^2}{\pi^2 W d^2}$ is the reptation time.

For short times $S_{chain}(Q,t)$ decays mainly due to local reptation (first term), while for longer times (and low Q) the second term resulting from the creep motion dominates. The ratio of the two relevant time scales τ_0 and τ_d is proportional to N^3. Therefore, for long chains at intermediate times $\tau_e < t < \tau_d$ a pronounced plateau in $S_{pair}(Q,t)$ is predicted. Such a plateau is a signature for confined motion and is present also in other models for confined chain motion. Besides the reptation model also other entanglement models have been brought forward.

1. In generalized Rouse models, the effect of topological hindrance is described by a memory function. In the border line case of long chains the dynamic structure factor can be explicitly calculated in the time domain of the NSE experiment [22].

2. Rubbers like models take entanglements as local stress points acting as temporary cross links. De Cloizeaux [23] has proposed such a model, where he considers infinite chains with spatially fixed entanglement points at intermediate times.

   Qualitatively the rubber like and the generalized Rouse model of Ronca arrive at similar results for $S_{pair}(Q,t)$ in the transition region. Quantitatively the Q-dispersion of the plateau heights, however, is more pronounced in the Ronca model.

3. Recently in a mode coupling approach a microscopic theory describing the polymer motion in entangled melts has been developed. While these theories describe well the different time regimes for segmental motion, unfortunately as a consequence of the necessary approximations a dynamic structure factor could not yet been derived [24].

3.4 Neutron Spin Echo Results on Chain Confinement

Fig.9 compares the dynamic structure factors from two polyethylene melts both measured at 509K for two different molecular weights (9(a); $M_w = 2$kg/mol [25,26]; 9(b): $M_w = 12.4$kg/mol) [27]. The solid lines in Fig.9(a) display a fit with the Rouse dynamic structure factor according to Eq.[21]. Very good agreement is achieved. Fig.9(b) presents equivalent results from the higher M_w melt with the solid lines again showing the prediction of the Rouse model.

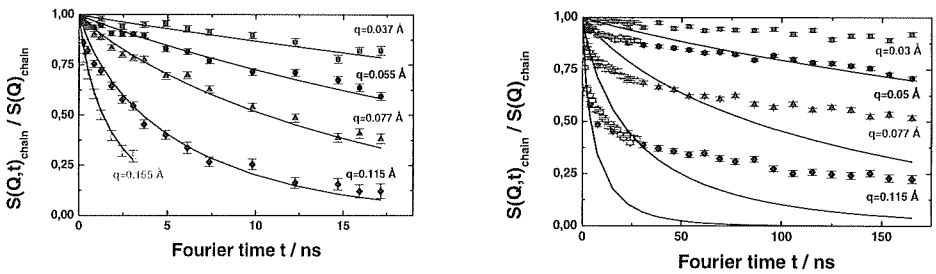


Fig. 9: *Dynamic structure factors from PE-melts at 509K (a) $M_w = 2000$; (b) $M_w = 12400$. The solid lines display the predictions of the Rouse model.*

Please note that the time scale for the 12.4kg/mol sample is extended by one order of magnitude compared to the short chain case. While for the short chain melts, the Rouse model describes well the experimental observations, for the longer chains the model fails completely. Only in the short time regime the initial decay of the dynamic structure factor is depicted while for longer times the relaxation behavior is strongly retarded signifying the confinement effects.

In Fig.10 dynamic structure factor data from a $M_w = 36$kg/mol PE-melt are displayed as a function of the Rouse variable $Q^2\ell^2\sqrt{Wt}$ (Eq.[26]) [28]. Other than in Fig.6 where the scaled data followed a common master curve, here they split into different branches which only at small values of the scaling variable are coming close together. This splitting is a consequence of the existing dynamic length scale which invalidates the Rouse scaling properties. We note, that this length is of purely dynamical character and cannot be observed in static equilibrium experiments.

In the spirit of Eq.[37] and neglecting the ongoing decay of $S_{chain}(Q,t)$ due to local reptation, from the heights of the achieved plateaus we may obtain a first estimate for the amount of confinement. Identifying the plateau levels with a Debye-Waller factor describing the confinement we get $d = 46$Å a value which is a lower estimate for the true tube diameter, since S^{loc} is not fully relaxed. The horizontal dashed lines in Fig.10 are the predictions from this Debye-Waller factor estimate.

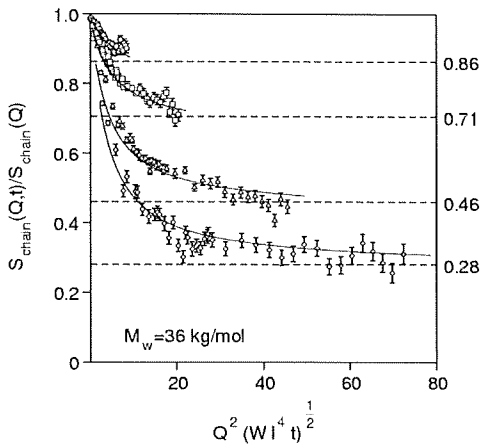


Fig. 10: *Scaling presentation of the dynamic structure factor from a $M_w = 36000$ PE-melt at 509K as a function of the Rouse scaling variable. The solid lines are a fit with the reptation model (Eq.[3.39]). The Q-values are from above $Q = 0.05,\ 0.077,\ 0.115,\ 0.145$Å^{-1}. The horizontal dashed lines display the prediction of the Debye-Waller factor estimate for the confinement size (see text)*

In Fig.11 the same data are presented linearly as a function of time. The figure compares them with the predictions of the various models. It is evident that these data clearly favour the reptation model. We note that the fits with the reptation model were done with only one free parameter, the entanglement distance d. The Rouse rate $W\ell^4$ was determined earlier through

NSE data taken for $t < \tau_e$. With this one free parameter, quantitative agreement over the whole range of Q and t using the reptation model with $d = 46.0 \pm 1.0$Å was found.

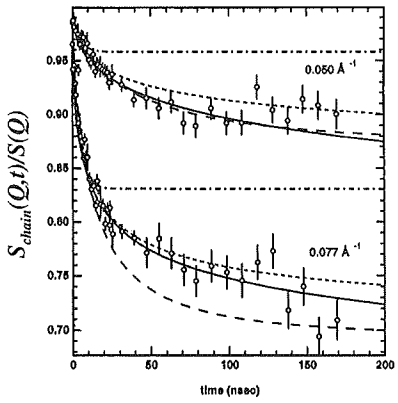


Fig. 11: *Plot of $S_{chain}(Q,t)$ vs. t at $Q = 0.050$ and 0.077Å^{-1} with a comparison between the predictions of reptation (solid lines) [21], local reptation (dotted lines) [21], the model of des Cloizeaux [23] (dashed lines) and the Ronca model [22] (dot-dashed lines) [28].*

We now turn to the mean-square displacement of a chain segment, which may be observed by incoherent quasielastic scattering. In the local reptation regime we expect to observe the predicted cross over of the time dependent mean-square displacement from a $t^{1/2}$- to a $t^{1/4}$-law (Eq.[35]). Experiments were performed on PE and PEP samples at the same temperatures, where dynamic structure factors were studied [6]. Fig.12 displays these data in terms of the mean-square displacement $\langle r^2(t)\rangle = -6\ell n\left(S_{self}(Q,t)\right)/Q^2$. Thereby the Gaussian assumption is implied.

Inserting the Rouse rate $W\ell^4(509\text{K}) = 7 \pm 0.7\text{nm}^4/\text{ns}$ (Table 1) obtained from single chain structure factor measurement into Eq.[20] the solid line $\sim t^{1/2}$ is obtained. It quantitatively corrobates the correctness of the Rouse description at short times. The data also reveal clearly a transition to a $t^{1/4}$ law, though Eq.[35] would predict the dotted line. The discrepancy explains itself in considering the non-Gaussian character of the curve linear Rouse motion (Eq.[36]). Fixing $W\ell^4$ and d to the values obtained from the single chain structure factor measurement, the dashed lines in Fig.12 reveals the prediction of the non-Gaussian treatment. For $Q = 0.1$Å^{-1} the unrestricted Rouse regime ($t < \tau_e$) as well as the local reptation regime is perfectly reproduced. For $Q = 0.15$Å^{-1} the prediction of Eq.[36] lies slightly outside the error band of the data points (lower dashed line). With increasing Q-values the spatial resolution increases and agreement with theory may only be expected for Q-values $< 2\pi/d$. For PE with a tube diameter of 47Å the limiting wave vector would be $Q = 0.13$Å^{-1} which may explain why at $Q = 0.15$Å^{-1} deviations become visible.

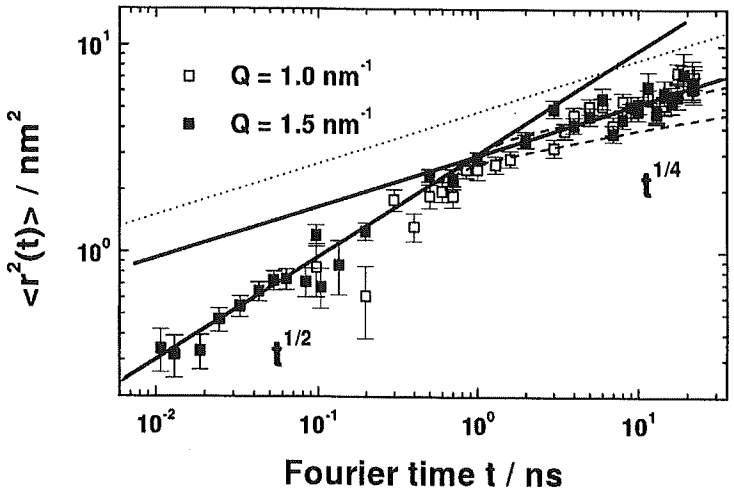


Fig. 12: *NSE data obtained from the incoherent scattering from a fully protonated PE-melt in a representation of -6 ℏ[$S_{self}(Q,t)$]/Q^2 which is the mean-square displacement* $\langle r^2(t) \rangle$ *as long as the Gaussian approximation holds. Solid lines describe the asymptotic power laws. Dotted lines: prediction form the Gaussian approximation, dashed lines: see text.*

Summarizing, from the so far reported NSE results on the dynamics of long highly entangled polymer melts a consistent picture evolves, where both pair and self-correlation functions are quantitatively described with identical parameters in terms of the structure factors for reptation. The data now cover a substantial region of the time domain where the reptation concept is in principle applicable. Compared with other phenomonological entanglement models, reptation is now the only approach that provides a consistent description of all the NSE data. This is compelling evidence that reptation is indeed the principle relaxation mechanism in entangled linear-chain systems.

3.5 Contour Length Fluctuations

Though in general the reptation predictions Eq.[33] and [34] for a number of experimental observables like the viscosity or the translation diffusions coefficient are qualitatively in good agreement with experiment, in detail deviations are evident. For the viscosity instead of the $\eta \sim M^3$ power law in general higher exponents in the order of 3.4 are found. Similarly, recently it become evident that also the translational diffusion coefficient does not follow Eq.[34] but displays a stronger molecular weight dependence. Outside the experimental error these data show that similar to the viscosity result the pure reptation prediction is violated ($D \sim M^{-2.30\pm0.05}$ instead of $D \sim M^{-2}$). On the other hand extensive viscosity measurements on PB by Colby et al. [29] show that above about 300 entanglement lengths the slope of η vs M turns from $M^{3.4}$ to M^3 indicating that the deviations from reptation are due to finite size effects.

Again NSE experiments unravelled the molecular origin of these finite size effects on chain confinement [27]. Experiments were performed on PE-melts spanning a molecular weight

range from 12.4kg/mol up to 190kg/mol. Fig.13 displays the results. We first qualitatively discuss the data and compare the results at $Q = 0.115\text{Å}^{-1}$. For the highest molecular weights at $M_w = 190$kg/mol and $M_w = 36$ kg/mol the spectra are characterized by an initial fast decay reflecting the unconstrained dynamics at early times and very pronounced plateaus of $S(Q,t)$ at later times, signifying the tube constraints. The plateau values for the two samples are practically identical. Inspecting the results for smaller M_w, we realize that (i) the dynamic structure factor decays to lower values (0.5 at $M_w = 190$kg/mol), 0.4 at $M_w = 24.7$kg/mol, and nearly 0.2 at $M_w = 12.4$kg/mol); (ii) furthermore the long time plateaus start to slope more and more the smaller M_w becomes, with $M_w = 12.4$kg/mol nearly losing the two-relaxation step character of $S(Q,t)$. Apparently the chain is disentangling from the tube and constraints are successively removed. If the data are analyzed using the pure reptation model of Eq.[39] a tube diameter d, which increases with decreasing molecular weight is obtained. This is displayed in Fig.14, where the such obtained tube diameter is plotted as a function of molecular weight. We observe and increase from about 45 to 60Å.

As it has been inferred from the broad a cross over phenomena observed in the macroscopic chain dynamics such as the molecular weight dependence of the melt viscosity very important limiting mechanism must exist which effect the confinement limit of the reptation process. These processes increase in importance as the chain length decreases. Two main mechanisms are proposed

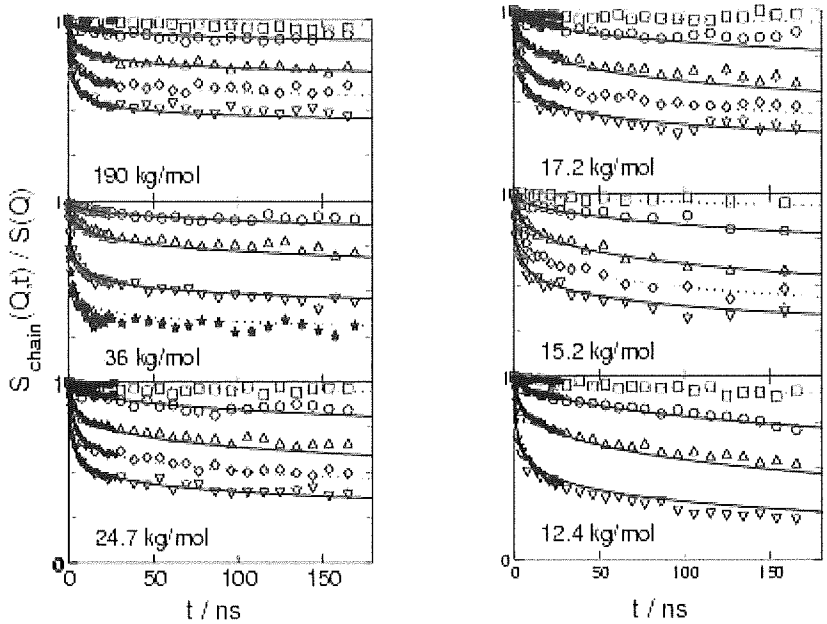


Fig. 13: *Neutron-spin echo spectra from polyethylene melts of various molecular weights. The Q values correspond to squares, $Q = 0.03\text{Å}^{-1}$; circles, $Q = 0.05\text{Å}^{-1}$; triangles (up), $Q = 0.077\text{Å}^{-1}$; diamonds, $Q = 0.096\text{Å}^{-1}$; triangles (down), $Q = 0.115\text{Å}^{-1}$; crosses, $Q = 0.15\text{Å}^{-1}$. Filled symbols refer to a wavelength of the incoming neutrons $\lambda = 8$Å and open symbols refer to $\lambda = 15$Å. For lines, see explanation in text.*

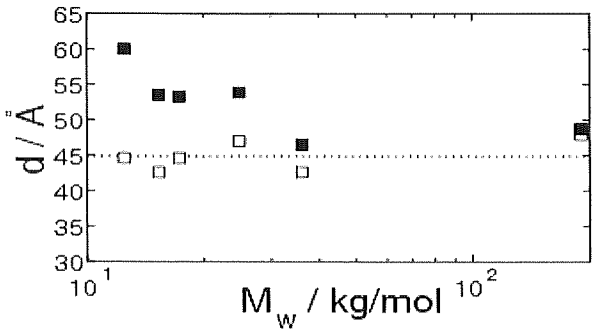


Fig. 14: *Tube diameters from model fits with pure reptation (filled squares) and reptation and contour-length fluctuations (open squares) as a function of molecular weight. The dotted line is a guide for the eye.*

(i) ***Constraint Release (CR).*** It takes place, if a confining chain moves out of the way of a given chain and thus opens some freedom for lateral motion. This phenomenon is an intrinsic many body effect and for monodisperse polymer melts becomes significant mainly in the creep regime and

(ii) ***Contour Length Fluctuations (CLF)*** of the primitive path length of a confined chain. CLF are already important for the removal of confinement in the local reptation regime i.e. for times $t < \tau_R$. If a chain contracts in its tube, then after subsequent expansion it looses memory of the original tube. In this way, the effect of tube confinement is gradually diminished with increasing time, the tube confinement will be removed from the ends. Other than constraint release CLF [30] is a single body effect and more easily treatable.

In order to calculate effects of CLF we have to ask how does the fraction of monomers grow with time which is released through CLF at the chain ends. This first passage problem for the chain end has been solved revealing that at early times the fraction of still confined tube segments behaves as

$$\mu(t) = 1 - \frac{C_\mu}{Z}\left(\frac{t}{\tau_e}\right)^{1/4} \tag{40}$$

where $Z = N/N_e$ is the number of entanglements/chain. One dimensional stochastic simulation led to $C_\mu = 1.5 \pm 0.02$. The released segment fraction is then $\psi(t) = 1 - \mu(t)$.
Following an approach of Clark and McLeish [31], this result may be incorporated into the structure factor of Eq.[37]. We assume that after time t all monomers between 0 and $s(t)$ and between 1 - $s(t)$ and 1 have escaped from the tube. Here $s(t) = \Psi(t)/2$. This statement contains two approximations, first that the fraction of the chain escaped from the tube is the same from each end of the chain, and, second, it ignores the distribution of $s(t)$ replacing it by the single average value. If we make these two assumptions the rest of the calculation is straightforward and leads to a modified escape term in Eq.[37]

$$S^{esc}(q,t) = \frac{N}{2\mu^2}\left[2\mu + e^{-2\mu} + 2 - 4\mu s(t) - 4e^{-2\mu s(t)} + e^{-4\mu s(t)}\right] \quad (41)$$

The lines in Fig.13 display the result of a fit with Eq.[41]. A group of three common Q values for all molecular weights is displayed by thick solid lines to facilitate an easy comparison of the molecular weight dependence of the curves. Additional Q values only available for some M_w's are represented by dotted lines.

The open squares in Fig.14 show the resulting tube diameters. Aside from some small fluctuations they now stay constant, independent of M_w. If we compare the experimental spectra with the model prediction (Eqs.[37] and [41]), we generally find good agreement. The gradually increasing decay of $S_{chain}(Q,t)$ with decreasing M_w is depicted very well both with respect to the magnitude of the effect as well as to the shape of $S_{chain}(Q,t)$.

Thus, the comparison between the experimental chain length- dependent dynamic-structure factor and theoretical predictions clearly shows, that in the time regime $t \leq \tau_R$, contour-length

fluctuations are the leading mechanisms that limit the chain confinement inherent to the reptation picture. Without any further assumption or fitting parameter - the tube diameter d stays constant with M_w - it is possible to describe the full M_w dependence of $S_{chain}(Q,t)$ in terms of local reptation and the contour-length fluctuation mechanism. Even for chain lengths corresponding to only 6–7 entanglements the tube diameter appears to be a well-defined quantity assuming the same value as for asymptotically long chains. The confinement is lifted from the chain ends inwards, while the chain center remains confined in the original tube.

Concluding this paragraph we emphasize that CLF also effect the macroscopic melt properties as a eluded to in the beginning in a significant way.

(i) From the relation between $G(\omega)$ and the total tube survival probability it is immediately clear that the modifications of $S^{esc}(Q,t)$ should also be mirrored by $G(\omega)$. It has been shown that CLF introduce an $\omega^{1/4}$ regime into the spectrum of $G''(\omega)$. However, the other limiting mechanism (CR) adds further significant modifications [32].

(ii) The translational chain diffusion is necessarily also effected. First the terminal time defining the diffusion "step" is reduced since reptation has only to relax the not yet released central parts of the tube and secondly the diffusive length is reduced, since only the displacement of the central part counts. Both effects do not cancel and the net effect gives $D \sim Z^{-2.4}$ up to $Z \cong 300$ in very good agreement with experiment.

(iii) Finally also the anomalous power law exponent of 3.4 for the viscosity mass relation has been attributed to CLF. The treatment of Millner and McLeish even predicts the cross over to pure reptation within the proper range of Z [33].

It is the virtue of the NSE experiments that they confirmed the CLF mechanism quantitatively on a molecular level in space and time. On the other hand constraint release process still remain to be probed on a molecular scale and it is expected that future NSE experiments on bimodal melts will clarify these processes as well on a molecular scale.

3.6 Reptation of a Biopolymer

Polymerized actin (F-actin) plays an essential role in cell mechanics and cell motility, and it is an attractive model system for studying the fundamental physical properties of semiflexible polymers [34]. Monomeric actin (G-actin, molecular mass $M_r = 42.000$) polymerizes in physiological salt solutions (pH 7.5, 2mM $MgCl_2$, 150mM KCl) to double-stranded filaments (F-actin). Such filaments exhibit a polydisperse length distribution ($4 - 70\mu m$ in the considered case). The size of the polymers make it possible to observe them by video microscope in real space and to analyze there motion. For this purpose single filaments are labelled by fluorescence markers and are placed into an environment of unlabelled filaments.

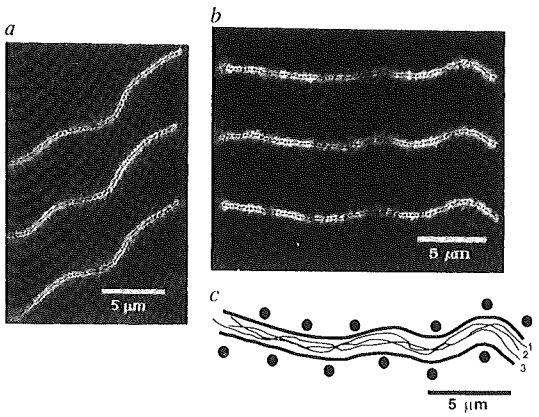


Fig. 15: *Time sequences of an undulating labelled actin filament in a free state (a) and embedded in a non-labelled solution of actin filaments (b), where the actin monomer concentration of the solution $C_A = 0.1mg\ ml^{-1}$, taken at time intervals of 0.1s. The dashed black lines in the micrographs show the contour lines determined by image processing. C, Superimposition of transient contour lines (numbered 1, 2, 3) of the filament shown in b. The thick black lines represent the maximum possible undulation amplitudes, which are restricted by the surrounding chains (shown schematically by black circles).*

Fig.15 displays the transient contours of a free actin filament (Fig.15a) and an actin filament embedded in a semidilute F-actin solution (Fig.15b) A decrease in the amplitudes of the thermally excited undulations is clearly visible for the embedded filament. The restricted chain motions can be understood in terms of the undulations of a filament in a tube formed by the surrounding entangled filaments. As indicted in Fig.15c, the undulations probe the tube diameter, which allows one to determine its local diameter by measuring the maximum flicker amplitudes.

The tube may be imaged and its width determined (Fig.16) by a superposition by a sufficient number of contours. The mean diameter $<d>$ of the tube has been determined as a function of the actin concentration c_A revealing $< d > \sim C_A^{-0.5 \pm 0.15}$. This result is consistent with what would be expected for a mesh of stiff rods.

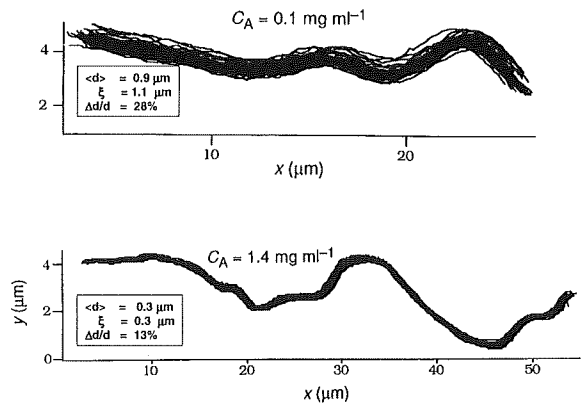


Fig. 16: *Determination of the local tube diameter formed around a labelled actin filament by surrounding filaments. The tubes were imaged by the superimposition of 60 traces of the contour taken at 0.1-s intervals. The mean tube diameter <d> is determined by averaging the width of the dark bands along the contour. Above each tube the actin monomer concentration C_A of the surrounding F-actin solution is noted. In the boxes are given values of the average tube diameter <d>, together with the mesh size ξ determined by permeation studies.*

Fig.17 displays time sequences of fluorescence micrographs demonstrating the snake-like self-diffusion of the filaments in the F-actin solutions. At first the filament slides partially out of the tube at the right side. Then it retracts back into the tube. By this fingering out of the tube the filament explores new tube segments. Over a longer interval of time (~ 10min), shown in the example in Fig.17b, substantial portions of the filament have left the original tube at both ends by repeatedly sliding backwards and forwards. The central section of the molecule, however, remains very close to its original configuration. Thus the reptation movement predicted by the de Gennes can be directly imaged for a semiflexible polymer. The central part of the tube is not entirely static, however, Fig.17b shows a local deviation in the tube contour, indicating the release of a topological constraint when one of the surrounding filaments reptates away. Such events are infrequent, as is expected when the entangling filaments have a size similar to the labelled filaments. The diffusion coefficient along the tube D, was determined by evaluation of the random fingering motion of the chain ends (compare Fig.17Aa,b,c). The chain end positions the (xi ,y_i) with respect to a local coordinate system with y-axes parallel to the tube axes at the ends, were recorded at time intervals $\Delta t = 0.84$s. D was determined as the arithmetic mean of the diffusion coefficients parallel to the tube at both ends (D_1 and D_2) according to

$$D = \frac{1}{2}(D_1 + D_2)$$
$$= \frac{1}{2}\left(\frac{1}{(N-1)} \sum_{i=2}^{N} \frac{\left(y_1(i) - y_1(i-1)\right)^2}{2 \times \Delta t} \right. \qquad (42)$$
$$\left. + \frac{1}{(N-1)} \sum_{i=2}^{N} \frac{\left(y_2(i) - y_2(i-1)\right)^2}{2 \times \Delta t} \right)$$

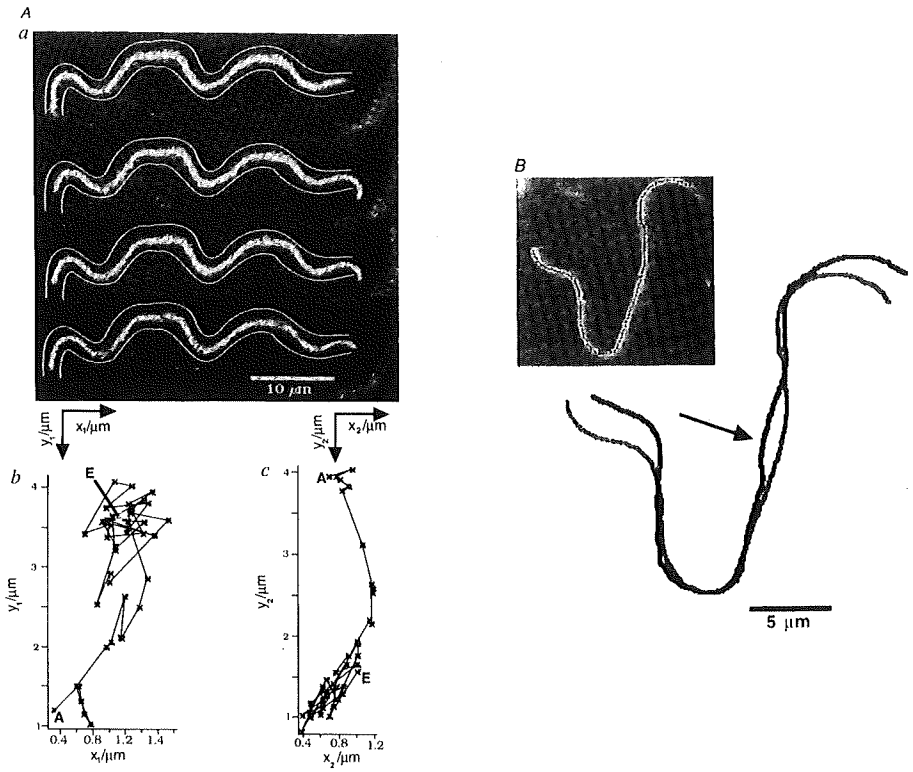


Fig. 17: *(Aa) time sequence of the reptation motion of a semixflexible actin filament. Snapshots of the filament embedded in a solutin of actin filaments of monomer concentrations $c_A = 0.1mg\ ml^{-1}$ are taken at time intervals of 20s. The schematic diameter of the tube drawn to guide the eye (thin white lines) is three times larger then the real mean diameter ($<d> = 0.9\mu m$). Ab, Ac, Random walk of chain ends (x_i, y_i) ($n = 1,2$) of the filament of contour length $L_C = 51\mu m$ shown in a. Positions (x_i, y_i) of end-points were recorded at time intervals of $\Delta t = 0.84s$. The time sequences started at points A and ended at points E. Note that single jumps of the two ends must not be correlated, because they are driven by the thermal undulations, but on long timescales the motion of the two ends is correlated and proves that the filament is really sliding out of the tube. B, Traces of the contour of a tube surrounding the filament shown in the photograph. The first trace (grey) is taken at the beginning of the experiment. The second trace (black) is taken after 7.5 min. It may be seen that the filament has left the original tube at both ends, which happens by the filament repeatedly sliding backward and forward. The central part of the tube is not entirely static. The arrow shows a local deviation in the tube contour, indicating release of constraint when one of the surrounding filaments reptates away.*

The diffusion coefficients depends on the chain length and decrease approximately linearly with increasing contour length L_c rendering values between $D_{tube} = 18 \cdot 10^{-14} m^2/s$ ($L_c = 20\mu m$) to $D_{tube} = 8 \cdot 10^{-14} m^2/s$ ($L_c = 50\mu m$). Though these study is dealing with very stiff polymers

yielding results which are not applicable to flexible polymers in the melt, the virtue of these measurements is the direct visualization of a tube like motion under confined geometry.

4 Hydrodynamic Interactions and Polymer Solutions

In the two preceding chapters, we have learned about two generic features of polymer dynamics, the entropic forces and the topological interactions. Here, we will discuss a phenomenon occurring in solutions only. As discussed broadly in the lecture on "Brownian Motion" in solutions, hydrodynamic interactions transmitted by the solvent are an important ingredient for the understanding of the dynamics of such systems. If a monomer moves, via friction it exerts a force on the solvent creating a flow field. This flow field then interacts again via friction with the other polymer segments and couples the motions of all monomers with each other. In this paragraph we will first discuss the conceptual basis in terms of the Oseen tensor and the Zimm model, then we will present experimental results on dilute solutions, discuss the effect of concentration in regarding semidilute solutions, where the screening of the hydrodynamic interactions and collective diffusion are important. We will briefly consider chain specific effects limiting the universal dynamic behaviour described in the Zimm model and finally, we will consider some dynamic experiments on biopolymers in solution.

4.1 The Zimm Model

The flow field exerted by a force $\underline{F}$ in the solvent is described by the Oseen tensor $\underline{\underline{T}}$. $\underline{\underline{T}}$ is derived from the incompressible Navier-Stokes equation (see Eq.[115] in the lecture on "Brownian Motion").

$$\underline{v}(\underline{r}) = \underline{\underline{T}} \cdot \underline{F} \tag{43}$$

$$\underline{\underline{T}}(r) = \frac{1}{8\pi\eta_s |r|}\left\{\underline{\underline{E}} - \frac{\underline{r} \otimes \underline{r}}{r^2}\right\} \tag{44}$$

$\underline{v}(\underline{r})$ is the flow field, $\underline{F}$ the force, η_s the solvent viscosity, and r the distance between force and flow field. $\underline{\underline{E}}$ is the unity tensor and the circle with the cross designates the outer product.

For a polymer in solution the velocity of the solvent $\underline{v}(\underline{r}_n)$ at the monomer position $\underline{r}_n$ is the consequence of the forces which are exerted by all other moving monomers on the solvent.

$$\underline{v}\ (\underline{r}_n) = \sum_{m \neq n} T_{nm}\ \ \underline{F}_m^{chain} \tag{45}$$

$\underline{F}_m^{chain}$ is the force which is exerted by segment m on the solvent. This force is given by the Rouse equation (Eq.[9]). Inserting Eq.[9] into Eq.[43], we obtain

$$\underline{v}(\underline{r}_n) = \sum_{m \neq n} T_{nm}\left(k\frac{\partial^2 \underline{r}_m}{\partial m^2} + \underline{f}_m\right) \tag{46}$$

This velocity field exerts a force on segment n. Including the hydrodynamic force term in the Rouse Eq.[9], we obtain the Zimm equation [35].

$$\zeta \frac{d\,\underline{r}_n}{dt} = \left(k \frac{\partial^2 \underline{r}_n}{\partial n^2} + \underline{f}_n \right) + \zeta \sum_{m \neq n} \underline{\underline{T}}_{nm} \left(k \frac{\partial^2 \underline{r}_m}{\partial m^2} + \underline{f}_m \right) \tag{47}$$

This is a coupled non-linear differential equation which cannot be solved without approximations.

The standard solution is obtained after an orientational averaging of the Oseen tensor. Furthermore, all intrachain distances are averaged with their equilibrium distribution. For a theta solvent this equilibrium distribution is Gaussian. With that, the Oseen tensor becomes

$$\langle T_{nm} \rangle = \frac{1}{\left(6\pi^3 |n-m|\right)^{1/2} \eta_s \ell} \tag{48}$$

Introducing Eq.[48] into Eq.[47] we linearlize the differential equation. Fourier transformation using the Rouse normal modes (Eq.[8]) approximately diagonalizes Eq.[47]. Neglecting the very small off-diagonal elements the Zimm equation may be easily solved. For the relaxation time spectrum, we obtain

$$\frac{1}{\tau_p} = \frac{1}{\tau_Z} p^{3/2} \text{ with } \tau_Z = \frac{\eta_s N^{3/2} \ell^3}{\sqrt{3\pi} k_B T} = 0.325 \frac{\eta_s R_e^3}{k_B T} \tag{49}$$

Compared to the Rouse model, the relaxation time spectrum is changed qualitatively ($1/\tau_p \approx p^{3/2}$ instead of $1/\tau_p \approx p^2$ in the Rouse case). This qualitative change is the consequence of the long range hydrodynamic interaction. For the mean square displacement of a single monomer a qualitatively different behaviour compared to the Rouse case is found.

$$\Delta r_{nn}^2 = \Gamma\left(\frac{1}{3}\right) \frac{2}{\pi^2} \left(\frac{\sqrt{3\pi} k_B T}{\eta_s} t \right)^{2/3} \tag{50}$$

Table 2: Comparison of predictions of the Rouse and the Zimm models

Properties	**Zimm**	**Rouse**
Diffusion coefficient	$0.196 \frac{k_B T}{\eta_s R_e}$	$\frac{k_B T}{N \zeta_0}$
Longest relaxation time	$\tau_z = 0.325 \frac{\eta_s R_e^3}{k_B T}$	$\tau_R = \frac{\zeta_0 R_e^4}{3\pi^2 k_B T \ell^2}$
Dispersion relation	$\frac{1}{\tau_p} \approx \frac{1}{p^{3/2}}$	$\frac{1}{\tau_p} \approx \frac{1}{p^2}$
Mean-square segment displacement	$\Gamma\left(\frac{1}{3}\right) \frac{2}{\pi^2} \left(\frac{\sqrt{3\pi} k_B T}{\eta_s} t \right)^{2/3}$	$2\ell^2 \left(\frac{3 k_B T}{\pi \zeta \ell^2} t \right)^{1/2}$

For the diffusion coefficient of the whole chain

$$D_z = \frac{8k_BT}{3\left(6\pi^3\right)^{1/2}\eta_s \ell\sqrt{N}} = 0.196\,(0.203)\,\frac{k_BT}{\eta_s R_e} \tag{51}$$

is obtained, where the first prefactor stands for θ- and the second for good solvents. Polymer coils exhibit a diffusion coefficient following Einstein's law for hard spheres! Recapitulating table 2 compares the predictions of the Rouse and the Zimm models.

4.2 Results on Dilute Solutions

For polydimethylsiloxane (PDMS) bromobenzene is a theta solvent at $T = 84°C$. In such a solvent the polymer-polymer and the polymer solvent interactions compensate such that like in the melt the chain conformations follow an ideal Gaussian random walk. Fig.18 presents NSE results under such conditions. The solid lines display the prediction of the Zimm model. For the data description the only parameter of the model, the solvent viscosity, was fitted. The resulting value agrees within 10% with that for the solvent viscosity [36].

We now inspect the detailed lineshapes of the dynamic structure factor. For that purpose the negative logarithm of the normalized scattering function is plotted logarithmically vs. the logarithm of time. In such a plot power laws in the exponent become straight lines. The solid lines display the asymptotic $t^{2/3}$ behaviour of the self-correlation function which for the single chain dynamic structure factor should be reached at long times.
The dashed line present the dynamic structure factor of the Zimm model [36]. For long times and large momentum transfer, where the structure factor has already decayed significantly, the experimental points approach to the asymptotic law. The insert displays the Q dependence of the characteristic relaxation rate. The data display with good precision the expected power law behaviour $\Omega_Z \propto 1/\tau(Q) \propto Q^3$.

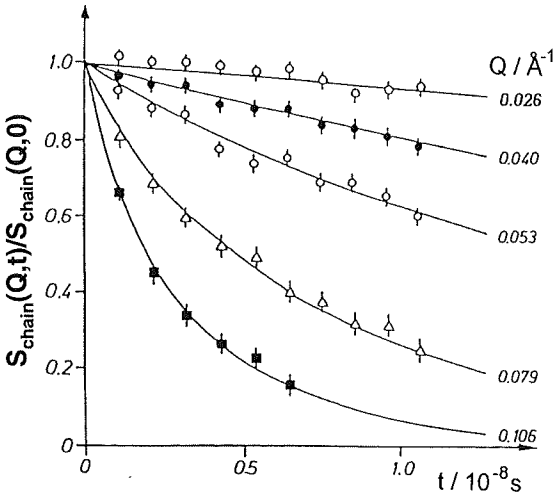


Fig. 18: *Dynamic structure factor obtained from a dilute solution of PDMS in bronobenzene at $T = 84$ °C under θ conditions.*

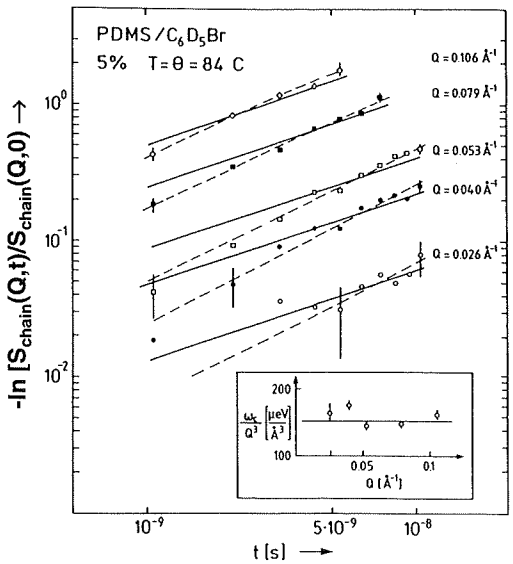


Fig. 19: *The spectral shapes of the data from Fig.18. Solid lines: asymptotic $t^{2/3}$ behaviour; dashed lines: dynamic structure factor of the Zimm model according to Dubois-Violette [37]. Insert characteristic relaxation rate ω_c/Q^3 vs. Q.*

4.3 Semidilute Solutions

In semidilute solutions the polymer chains start to interpenetrate each other and to overlap. On short scales corresponding to distances smaller than the average distance ξ_c between points of overlap - we expect that the polymer will behave as in dilute solution. For larger distances $r > \xi_c$, we need to consider two different effects.

1. The transient network formed by the overlapping polymers will relax with a collective diffusion coefficient D_{coop}.

2. Since the flow field scatters at the intervening chains, such distances the hydrodynamic interactions will be screened.

We will discuss first the collective diffusion. The transition from the spatial heterogeneity at small distances towards homogenous behaviour at $r = \xi_c$ changes the dynamic properties. At short distances we will observe single chain dynamics with the characteristic relaxation rate, $1/\tau(Q) \approx T/\eta_s \cdot Q^3$ while at larger distances collective relaxation with $1/\tau(Q) \approx D_{coop} \cdot Q^2$ should prevail. The scaling theory proposed by de Gennes [17] predicts that the collective diffusion in good solvents should be proportional to the polymer volume fraction $\phi^{3/4}$. Analogously, the correlation length should scale with $\xi_c \approx \ell\phi^{3/4}$.

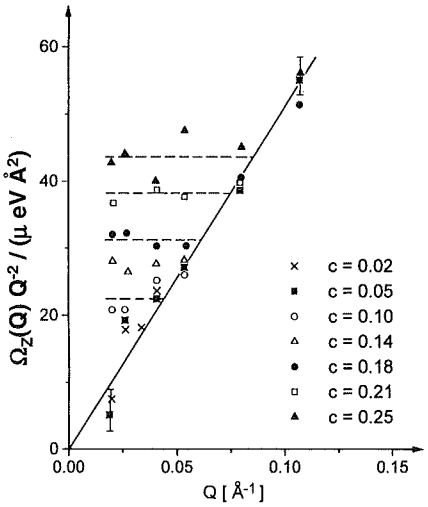


Fig. 20: *Collective reduced relaxation rate $\Omega_Z(Q) \equiv 1/\tau(Q)$ for a PDMS ensemble dissolved in benzene at various volume fractions $\phi \equiv c$. The solid line represents the dilute solution scaling $\Omega_Z(Q) \sim Q^3$, while the dashed lines denote the Q-dispersion of the collective diffusion.*

Fig.20 displays the reduced relaxation rates $1/(\tau(Q) \cdot Q^2)$ for varying volume fractions of PDMS in benzene as a function of Q. In this presentation the Zimm characteristic $1/\tau(Q) \approx Q^3$ becomes a straight line, while the collective diffusion changes into Q independent plateaus. We observe that for the two lowest polymer volume fractions 2% and 5% for the whole momentum transfer regime, single chain behaviour is observed. With increasing polymer concentration, at increasing Q we observe a transition to a plateau characterized by the collective diffusion coefficient. Let us consider the points at $\phi = 0.18$. Here the first four data points ($Q < 0.05\text{Å}^{-1}$) define the plateau of the collective diffusion coefficient, while the data points at higher momentum transfer agree with those of the dilute polymer solution. In this Q regime single chain dynamics is observed. Fig.21 displays the concentration dependence of the such determined collective diffusion coefficient. It exhibits a power law with an exponent 0.67 ± 0.05 slightly smaller than the value of 0.75 predicted by the scaling theory [36].

While the collective behaviour of the polymer ensemble is characterized by the collective diffusion, the relaxation of a single chain is dominated by the screening of the hydrodynamic interaction. While at short distances $r < \xi_c$ the unperturbed hydrodynamic flow field leads to Zimm relaxation, for distances $r > \xi_c$ the hydrodynamic interaction is screened and Rouse relaxation is expected.

An experimental verification of this prediction requires the observation of the dynamics of a single polymer chain within the ensemble. This may be achieved either by labelling a few chains within an ensemble [38] which is matched by the solvent or by experiments at zero average contrast [39]. In this case a mixture of the same amount of identical protonated and deuterated polymers in a solvent with the exact average contrast is investigated. Then, all interference terms between different chains in the structure factor compensate and only the

single chain dynamic structure factor is visible (the interference terms between congeneric polymers appear with positive, those between different polymers with negative sign).

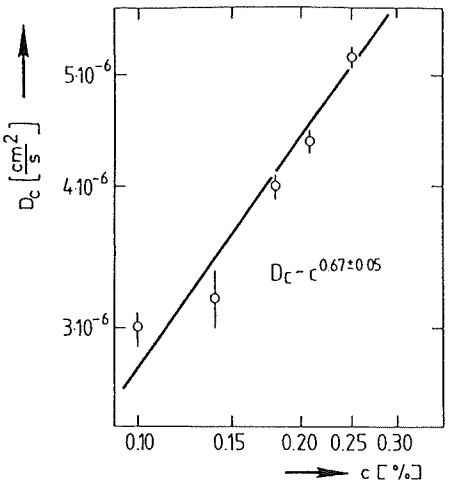


Fig. 21: *Collective diffusion coefficients D_{coop} from PDMS in benzene for various polymer volume fraction [36].*

Since the polymers are identical and the solvent scattering power is exactly in between that of the polymers all interference terms compensate.
Fig.22 presents the static structure factor of a semidilute solution (i) under average contrast zero (Iz: 50% h/50% d-polymer) and (ii) of 100% protonated polymers in a deuterated solvent.

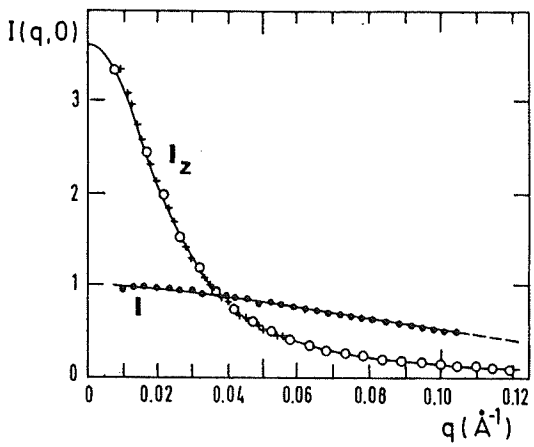


Fig. 22: *Static structure factor of PDMS solutions under full (I) and zero average contrast (I_z) (PDMS concentration 0.2g/cm^3).*

While in the case (i) the single chain structure factor is observed in the case (ii) that of the transient polymer network is seen. There only the polymer parts in between overlap points, the so called blobs, scatter leading to a very much reduced intensity, at low Q.

Fig.23 compares experiments under full contrast with others at zero average contrast above and below the crossover condition $Q\xi_c = 1$. For $Q\xi_c < 1$ in the region of the screened hydrodynamic interaction the experiment under full contrast (100% hPDMS) exhibit a single exponential decay (solid line). In this regime, collective diffusion is observed. The single chain structure factor (50% hPDMS, 50% dPDMS) under zero average contrast displays the square root behaviour of the Rouse dynamics.

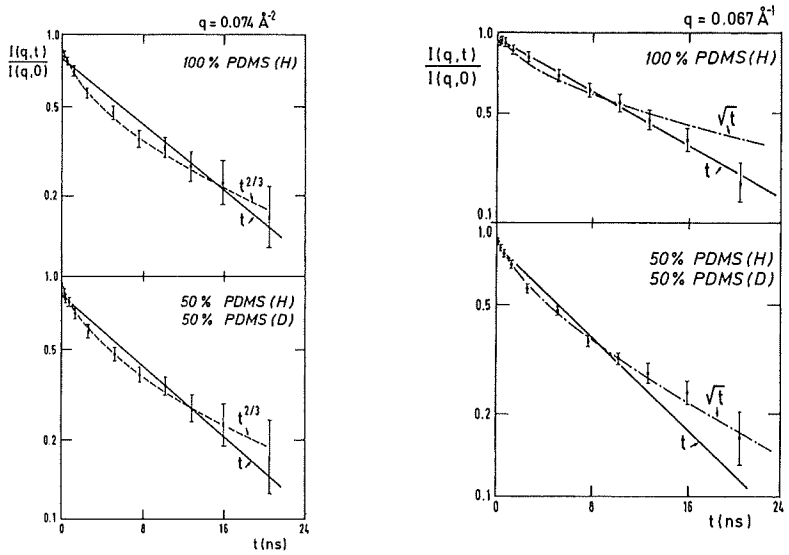


Fig. 23: *Comparison of the dynamic structure factors of semidilute PDMS-solutions under full (100% PDMS) or zero average contrast (50% PDMS(H), 50% PDMS (D)) above and below the cross over $Q\xi_c = 1$.*

In this regime, the hydrodynamic interaction is screened and only the Rouse dynamic remains. Going to Q values above the cross over ($Q\xi_c > 1$) under both contrast the $t^{2/3}$ time dependence of the Zimm model is approximately observed. In both cases, the single chain dynamic dominates.

Finally, Fig.24 presents the reduced relaxation rates $\Gamma(Q)/Q^2$. Within regime *B* Zimm dynamic prevails. This is valid for both labelings. On the other hand in the regime of small momentum transfers dependent on the contrast, either the plateau of the collective diffusion (A) or the strong reduction of the relaxation rate as a consequence of the Rouse dynamics is seen (C). These experiments are a particularly beautiful example for the opportunities offered by the h-d contrast variation in neutron scattering.

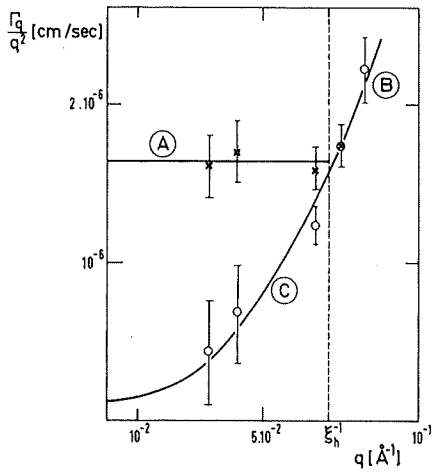


Fig. 24: *Dispersion relation of the characteristic relaxation processes (A) collective diffusion (B) single chain Zimm dynamics (C) Rouse relaxation with screened hydrodynamic interaction.*

4.4 Chain Specific Effects Limiting Universal Behaviour

From a conceptual point of view the Zimm-Rouse model is limited towards large scales by the onset of topological interactions – the associated confinement and reptation mechanism has been discussed in chapter 3 – and towards smaller scales by chain specific effects. There, the simplifying assumptions of the Zimm-Rouse model cease to be valid and the local chain structure comes into play. Locally the chain is stiff; this rigidity leads to deviations from the universal behaviour. Furthermore as a consequence of the rotational potentials local relaxation mechanisms across the rotational barriers come into play leading to an internal viscosity. Even more locally, we approach the regime of the elemental relaxation processes, the α-process and possibly higher order relaxations, which are not discussed here.

Chain rigidity is described in introducing a bending force into the Rouse Eq.[9] which is described by a fourth order derivative with respect to the contour coordinate. This fourth order term changes the relaxation time spectrum of the Rouse model (Eq.[14]) approximately to

$$\frac{1}{\tau_p^b} = \frac{W\pi^2}{N^2}\left[p^2 + \alpha_B p^4\right] \tag{52}$$

where $\alpha_B = \dfrac{\pi^2 C_\infty^2}{4N^2 \sin(\theta/2)}$ and θ is the bond angle [40].

Eq.[52] in its first part agrees with the Rouse result of Eq.[14]. The term proportional to p^4 is the correction due to rigidity effects. For higher mode numbers p the relaxation rates $\left(\tau_p^b\right)^{-1}$ are increasingly accelerated.

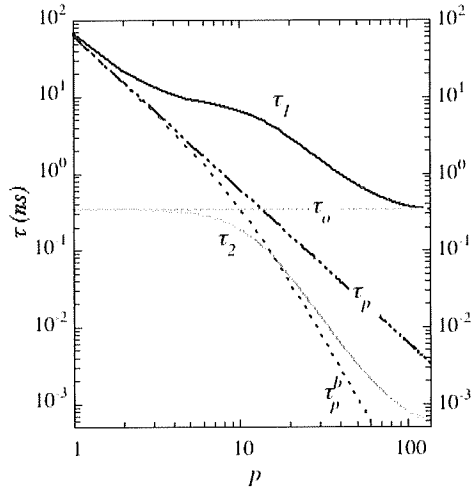


Fig. 25: *Mode number dependence of the relaxation times τ_1 and τ_2(solid lines). The dashed-dotted line shows the relaxation time τ_p in the Rouse model (Eq.[14]). The horizontal dashed line displays the value of τ_a. The dashed and the dotted lines represent the relaxation time when the influence of the chain stiffness is considered: bending force model τ_p^b (dotted).*

Furthermore, because of the equipartition theorem [see e.g. Eq.[13]) the amplitude of the corresponding correlator will diminish. We notice that under these conditions also the eigenfunctions become very similar to those of the Rouse problem. In a first order approximation therefore, the mean square displacements (MSD) (Eq.[19]) may be calculated in replacing the characteristic relaxation times $\tau_p = \tau_R/p^2$ by τ_p^b. With these new expressions for the MSD we then may evaluate the dynamic structure factor following the prescription of Eq.[21]. Using the same approach as for the Rouse model, we also may calculate the full solutions.

The description of the chain dynamics in terms of the Rouse model is not only limited by the local stiffness effects but also by local dissipative relaxation processes like jumps over the barrier in the rotational potential. Allegra et al. [41] introduced an internal viscosity as a frictional force which arises due to a departure from configurational equilibrium.
Thereby, the rotational relaxation processes are described by one single relaxation rate τ_a. From an expression for the difference in free energy due to small excursions from equilibrium an explicit expression for the internal viscosity force in terms of a memory function is derived. The approach leads to two branches of the relaxation time spectrum τ in dependence of the mode number p. They are displayed in Fig.25.

While at low p the solution for the longer relaxation time τ_1 coincides with that of the Rouse equation (multidot line), at higher p such a coincidence is observed for the shorter relaxation time τ_2. In between both modes repel each other and for small and large p reach the value of τ_a. We also include the relaxation times of the bending model τ_p^b (dotted line). For high p these times decrease much more strongly than those of the Rouse model. With similar results

the internal viscosity approach may also be extended towards of hydrodynamic interactions again leading to two each other repelling relaxation modes.

The two models may be approached experimentally in comparing the dynamics of two polymers with very similar stiffness (C_∞) and significantly different rotational barriers. PDMS and Polyisobutylene (PIB) constitute such a pair. Their characteristic ratio are very similar $C_\infty = 6.73$ (6.19) while the rotational barriers are vastly different: 3-3.5 kcal/mol for PIB and ~ 0.1kcal/mol for PDMS. Experiments in the melt showed that for PDMS the Rouse model describes the dynamics of this polymer perfectly over the full accessible Q-regime ($Q \leq 0.4\text{Å}^{-1}$), while for PIB a slowing down of the relaxations (compared to the Rouse prediction) is already observed for $Q \geq 0.15\text{Å}^{-1}$ [42,43]. Given the nearly identical chain rigidity these experiments qualitatively demonstrated that for flexible polymers dissipation through jumps over rotational barriers limits the universal dynamics. In the crowded environment of the melt the obtained internal relaxation times τ_a however could not easily be related to single rotational jumps.

In order to more deeply understand this process, experiments in dilute solutions, where the interchain friction is weak were performed [43]. Again PIB and PDMS were used. In order to obtain directly comparable data, chains with nearly identical chain dimensions were employed (PDMS: $R_g = 21.3\text{Å}$, PIB: $R_g = 19.2\text{Å}$).

For small chains in solution the translational diffusion significantly contributes to the overall decay of $S(Q,t)$. Therefore precise knowledge on the diffusional properties are essential. Combing dynamic light scattering (DLS) and NSE revealed effective collective diffusion coefficients. Measurements at different concentrations showed that up to a polymer volume fraction of $\phi = 10\%$ no concentration dependence could be detected. All data are well below the overlap volume fraction of $\phi^* = 0.23$. Since no ϕ-dependence was seen, the data may be directly compared with the Zimm prediction for dilute solutions (Eq.[51]).

Fig.26 displays the dependence of $D(\phi) \equiv \Gamma / Q^2$ on T/η_s. In accordance with Eq.[51] a very good linear relationship is observed. A comparison of the absolute values with Eq.[51] is done best by looking on the slope of $D(\phi)$ vs T/η_s. Using $R_e = 5.22\text{nm}$. Eq.[39] predicts a slope of $5.37 \times 10^{-13} N/K$ which agrees very well with the measured value of $5.4 \times 10^{-13} N/K$.

Fig.27 presents typical spectra taken on both polymer solutions at 300K (a) and 378K (b). The PDMS data are represented by open symbols, while the PIB data are shown by full symbols. Let us first look at the data at 378K. At $Q = 0.04\text{Å}^{-1}$ ($QR_g \equiv 0.8$) we are in the regime of translational diffusion, where the contributions of the intrachain modes amount to $\approx 1\%$ only. There the spectra from both polymers are identical. In going now to larger Q-values, gradually the spectra from the PIB-solutions commence to decay at later times. This effect increases with increasing Q and is maximal at $Q = 0.4\text{Å}^{-1}$ (Fig.27a).

The Allegra-approach bases on the physical idea of an intrachain viscous resistance to configurational equilibration after an excursion from equilibrium. In solution the corresponding relaxation time should relate to jumps across the rotational barriers within the chain.

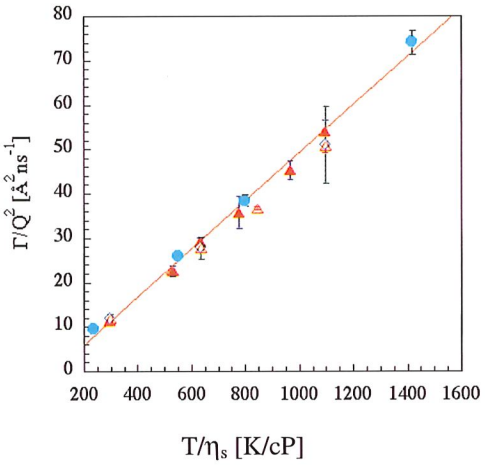


Fig. 26: *Relaxation rate Γ divided by Q^2 for PDMS in toluene solution vs. the reduced variable T/η_s. Results correspond to NSE measurements, $\phi = 0.1$ (●) and DLS for $\phi = 0.1$ (▲), $\phi = 0.07$ (△), and $\phi = 0.02$ (◇). Solid line shows a linear regression fit of the $\phi = 0.1$ NSE and DLS data.*

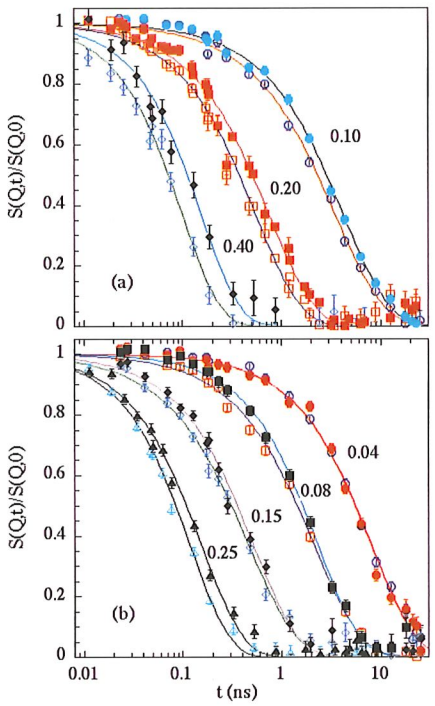


Fig. 27: *Chain dynamic structure factor of PDMS (empty symbols) and PIB (full symbols) in toluene solution at 300 K (a) and 378 K (b). The corresponding Q-values [$Å^{-1}$]are indicated. Lines through the points represent the single exponential fits which describe the data well.*

For the PIB analysis the results of the Rouse-Zimm evaluation of PDMS were taken as an input. The analysis was performed using with the full solution of the Allegra-model. Fig.28 displays the results of such a fit at 327K with τ_a being the only fit parameter – the diffusion coefficient again was extracted from the low Q-results and kept fixed. The fits are excellent for all temperatures. The resulting values for τ_a are displayed in Fig.29.

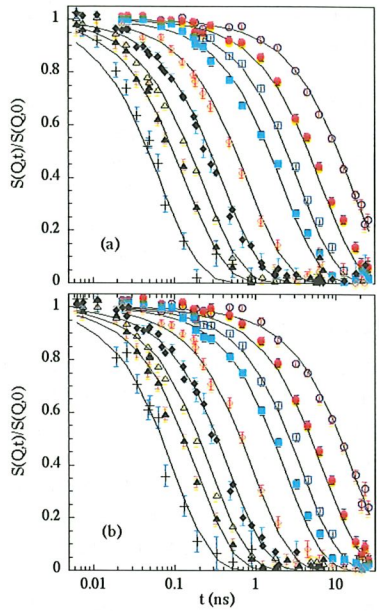


Fig. 28: *Chain dynamic structure factor of PDMS (a) and PIB (b) in toluene solution at 327K at the Q-values: 0.04 Å^{-1} (●); 0.06 Å^{-1} (●); 0.08 Å^{-1} (□); 0.10 Å^{-1} (■); 0.15 Å^{-1} (◇); 0.20 Å^{-1} (◆); 0.25 Å^{-1} (△); 0.30 Å^{-1} (▲); 0.40 Å^{-1} (+). Solid lines correspond to fitting curves: Rouse-Zimm model for PDMS and Rouse-Zimm with intrachain viscosity for PIB (see text).*

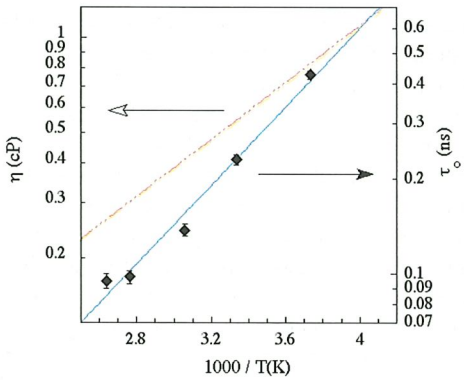


Fig. 29: *T-dependence of the solvent viscosity (dashed dotted line) and the characteristic time τ_a deduced for the conformational transitions in PIB (◆). The solid line through the points corresponds to the fit to an Arrhenius law.*

$\tau_a(T)$ follows an activated behavior with an activation energy $E = 3.1 \pm 0.3$ Kcal/mol,

$$\tau_a = 1.27 \times 10^{-12} \exp\left[\frac{3.1}{RT}\right] [\text{sec}] \tag{53}$$

The pre-exponential factor of about 1 ps lies well in the microscopic range. For comparison we also display the temperature dependence of the solvent viscosity η_s which is characterized by an activation energy of 2kcal/mol.

The activation energy of 3.1kcal/mol lies well in the range of activation energies for torsional jumps for PIB found by other techniques [45] and simulation [46] and allows to identify the mechanism behind the intrachain friction as rotational jumps which change the chain conformation. The finding corroborates the physical picture behind the Allegra mechanism and shows that the intrachain viscosity appears to be the leading mechanism causing deviations from universal relaxation in going to local scales.

4.5 Dynamics of Biopolymers in Solution

Proteins are of particular importance in biology. They serve as catalysts of special chemical reactions, as transport molecules and in many other functions. After the detection of their three dimensional structure with a high density of atoms inside, the picture of a solid structure arose. Later investigations showed that the dynamics of proteins are essential for their functions, e.g. a O_2 molecule would not be able to bind to (or leave) a myoglobin (or haemoglobin) molecule without fluctuations of the protein. Also practically all catalytic reaction are accompanied by drastic conformational changes.

Below typically 200K a protein behaves very much like a solid; the mean-square displacement raises proportional to temperature with values comparable to solids. But above this temperature the protein gets much more flexible. This was found in many investigations using different measurements techniques. The theory is that proteins undergo conformational changes with rising frequency and/or amplitude. These conformational changes seem to be the functional important fluctuations of the protein.

This so called dynamic transitions was detected by Doster et al. [47] in 1989 on myoglobin and has induced a large amount of follow up experiments. The main finding is that the mean-square proton displacement (MSD) increases linearly with temperature up to about 200K. Above this temperature a strong increase of the MSD sets in which is associated with increasing flexibility and protein function. Fig.30 displays such a measurement on the purple membrane (PM) containing the membrane protein bacteriorhodopsin (BR). MSD's for a completely dry and a hydrated PM are compared. Only for the hydrated system a dynamical transition at approximately 230K between harmonic and non-harmonic motion is observed [47]. Below this temperature the hydrated sample, following a harmonic behaviour, is not active, whereas above the transition temperature non-harmonic diffusive motions appear and the BR photocycle can be completed. In contrast to this, the dry PM sample exhibits harmonic motions up to room temperature and BR is not functional in the entire temperature range. The fact that non-harmonic motions are not necessarily liberated at high temperatures indicates that temperature alone is not sufficient to obtain high amplitudes of motion. These results significantly support the hypothesis that a global soft environment obtained through temperature and hydration, which allows larger diffusive thermal motions, is required to ensure the protein activity.

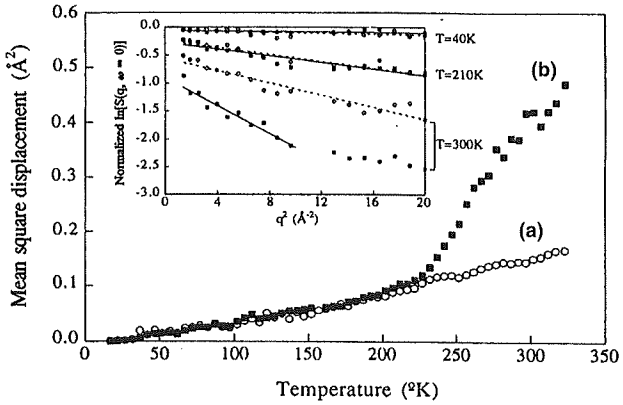


Fig. 30: *Temperature dependence of the isotropically averaged mean-square displacements of hydrogens in PM at very low hydration (a) and fully hydrated (b). The inset shows linear fits of* $S(Q)\ \omega = 0 = exp(-1/3\ Q^2\langle u^2\rangle)$ *as a function of* Q^2 *for different temperatures.*

As may be realized from Fig.30 the amplitudes associated with the dynamical transition are in the Å-regime. The large scale dynamics of proteins relating to conformational fluctuations of large proteins parts are basically unexplored. The nanomechanics of protein substructures that may be involved in biological function will have the dimension of nanometers. At present the window of opportunity for such experiments is not yet open and will have to wait for the next generation of ultra high resolution NSE-instruments to appear at the new MW-spallation sources.

Today e.g. NSE-experiments are undertaken to study the translational diffusion of proteins also in crowded environments. Such information is important e.g. for the mechanistic understanding of the oxygen exchange in red blood cells which are packed with haemoglobin up to a volume fraction of 0.35. An effective oxygen exchange requires proximity to the cell surface, implying fast diffusion of haemoglobin. Similarly the supply of muscle fibres is supported by diffusion of oxygen which is bound to the storage protein myoglobin. Protein-facilitated oxygen diffusion is usually measured using tracer methods on a macroscopic scale. Dynamic light scattering experiments on coloured proteins at high concentrations are complicated by absorption and multiple scattering. The light scattered by blood cells is dominated by the cell membrane and not by intra-cell diffusion. Neutron spin echo (NSE) spectroscopy allows to overcome these difficulties and in addition provides information on diffusion on a molecular length scale.
Here we discuss a concentration dependent diffusion study on myoglobin solutions in D_2O [49]. With increasing protein volume fraction protein interactions are important and knowledge on the structure factor $S(Q)$ is mandatory. Fig.31 displays the concentration dependent static structure factors obtained from SANS experiments. Fig.32 presents the concentration and Q-dependent effective myoglobin diffusion coefficients as obtained from NSE-experiments.

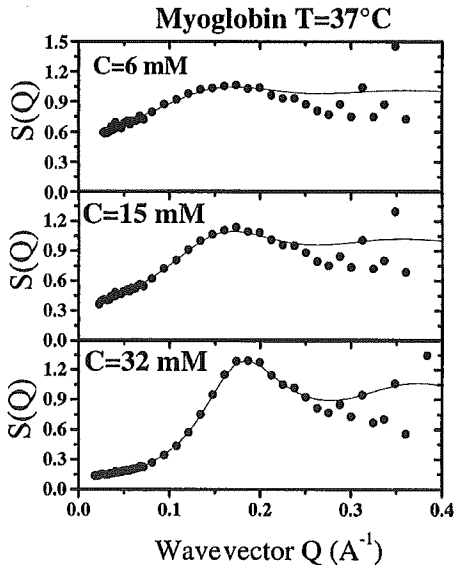


Fig. 31: *Structure factor S(Q) of myoglobin solutions ranging from 6 to 30mM (corresponding to volume fractions $0.1 \leq \phi \leq 0.4$. The points are deduced from experiments and continuous lines are the results from a refinement using the mean spherical approximation (MSA).*

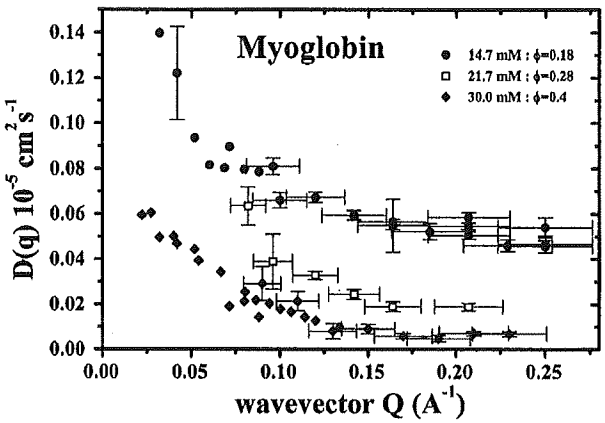


Fig. 32: *Wave vector dependence of the apparent diffusion coefficient D(Q) for three different myoglobin concentrations. One can observe a plateau at high Q, which is related to the self-diffusion coefficient $D(Q) \cong D_s^{\phi}$. The increase at low wave vector is due to the direct electrostatic interaction (see text).*

For each protein volume fraction Φ the wave vector dependence of the apparent diffusion coefficient $D(Q)$ is qualitatively similar. $D(Q)$ tends to a plateau value D^{ϕ} at high wave

vectors and increases at low wave vectors. Furthermore D^{ϕ} is strongly volume fraction dependent.

$D(Q)$ measures the collective diffusion of the protein in solution over distances $d \cong 2\pi/Q$. Structural (direct interactions due to inter-particle potential) and hydrodynamic interactions (effect of the motions of a particle on others via solvent) are generally separated using

$$D(Q) = D_{\infty} \frac{H(Q)}{S(Q)} \tag{54}$$

where D_{∞} is the infinite dilution diffusion coefficient given by Stokes-Einstein relation: $D_{\infty} = k_B T/6\pi\eta R$ for a spherical particle. $H(Q)$ denotes the hydrodynamic factor. $H(Q) = 1$ if there is no hydrodynamic effects. Using the analytical description of $S(Q)$ displayed in Fig.31 the product $D(Q) \cdot S(Q)$ may be computed. It is presented for $\phi \cong 0.2$ in Fig.33. The top of Fig.33. shows that the hydrodynamic factor is wave vector dependent. It seems to oscillate in phase with the structure factor. For a given wave vector $H(Q)$ is also very strongly protein volume fraction dependent. Finally, as evidenced in Fig.33 the self-diffusion coefficient, of an almost spherical protein with radius R, can be measured for $QR >> 1$ within the high plateau regime of D^{ϕ} by coherent neutron scattering.

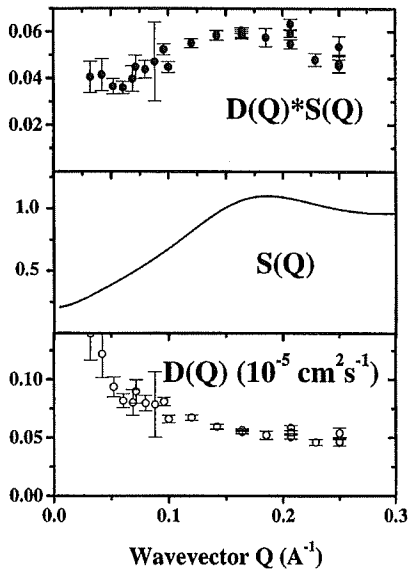


Fig. 33: *Apparent diffusion coefficient D(Q) deduced from NSE measurements (bottom), structure factor deduced from MSA analysis (middle) and S(Q) · D(Q) for myoglobin solutions of 14.7mM (Φ ~ 0.2) (top).*

5 Summary and Conclusion

In this lecture we have discussed the dynamics of linear polymers covering the universal dynamics from dilute solutions to the melt. We have identified the mechanism limiting the

universal dynamics of flexible polymers towards small scales and in some instances we have addressed also dynamical features of biopolymers, though their investigation by neutron scattering is still in its infancy. The lecture attempted to transmit a flavor of what can be achieved with high resolution neutron spectroscopy permitting the access to molecular motion simultaneously in space and time.

The lecture commenced with a description of the standard model of polymer motion, the entropy driven dynamics, covered by the so called Rouse model. In the spatial range, where the Rouse approximations are valid, the NSE measurements have confirmed the predictions of the Rouse model both for the self and the pair correlation function.

Towards larger scales topological interactions resulting from the mutually interpenetrating chains gain dominant influence and confine the chain motion to a tube along the chain profile. We have presented measurements of the dynamic structure factor of a reptating chain which unequivocally confirm the picture of local reptation i.e. Rouse relaxation along a contorted tube. A measurement of the self correlation function corroborates the picture – both experiments may be interpreted by the same set of parameters. Finally, with decreasing chain length, the tube confinement looses its grip and tube length fluctuations free the motions of the chain ends. This picture inherent in the reptation approach was also corroborated by NSE studies. The reptative motion of semiflexible F-actin filaments in semidilute solution as observed qualitatively in real space by video microscopy exemplified the use of biopolymers as model system for polymer physics problems.

Considering dilute and semidilute solutions we discussed the effect of hydrodynamic interactions in terms of the Zimm model. Single chain relaxation including hydrodynamic screening and cooperative motion were addressed and related to the relevant theoretical concepts.

Towards shorter length scales in the regime, where the picture of universal chain dynamics governed by entropic and hydrodynamic forces breaks down, the first order limiting mechanism was identified by NSE studies. It turned out that intrachain dissipation related to rotational jumps between different isomeric states is the leading process limiting universal chain dynamics.

The large scale intramolecular dynamics of proteins has not yet been touched by neutron scattering and will possibly have to wait for more powerful and more sensitive instruments at the MW-spallation sources to come. Here we have discussed briefly (i) the so called dynamic transition relating dynamical softness and function and (ii) some experiments on the diffusion of myoglobin on a molecular scale in crowed environment revealing the self diffusion coefficients relevant for oxygen exchange in red blood cells.

Though this lecture restricted itself to the dynamics of linear polymers, we hope to have shown that the investigation of polymer dynamics is a broad and rich field with a great potential. Neutrons play a key role in such investigations, because they allow to perform observations on the proper length and time scales as well as to exercise hydrogen deuterium contrast variation. Further developments will go in the direction of time dependent phenomena and real time experiments. We will look on nonequilibrium and transient phenomena, on the complexity posed by multicomponent systems, on molecular rheology, in order to understand the functioning of materials, and we will develop a bridge to biology.

References

[1]G.L. Squires, *Introduction to the theory of thermal neutron scattering*, Cambridge University Press, Cambridge (1978)
[2]P.R. Rouse, J. Chem. Phys. **21**, 1272 (1953)
[3]M. Doi, S.F. Edwards, *The theory of polymer dynamics*, Clarendon Oxford (1986)
[4]P.G. de Gennes, Physics (USA) **3**, 37 (1967)
[5]D. Richter, B. Ewen, B. Farago, T. Wagner, Phys. Rev. Lett. **62**, 2140 (1989)
[6]A. Wischnewski, M. Monkenbusch, L. Willner, D. Richter, B. Farago, G. Kali, Phys. Rev. Lett. **90**, 058302 (2003)
[7]H. Montes, M. Monkenbusch, L. Willner, S. Rathgeber, L.J. Fetter, D. Richter, J. Chem. Phys. 110, 10188 (1999)
[8]J.D. Ferry, *Viscoelastic properties of polymers*, John Wiley & Sons, NewYork (1980)
[9]L.J. Fetters, D.J. Lohse, D. Richter, T.A. Witten, A. Zirkel, Macromolecules **27**, 4639 (1994)
[10] S. Hoffmann, L. Willner, D. Richter, A. Arbe, J. Colmenero, B. Farago, Phys. Rev. Lett. **85**, 772 (2000)
[11] G.D. Smith, W. Paul, M. Monkenbusch, L. Willner, D. Richter, X.H. Qiu, M.D. Ediger Macromolecules **32**, 8857 (1999)
[12] D. Richter, R. Butera, L.J. Fetters, J.S. Huang, B. Farago, B. Ewen, Macromolecules **25**, 6156 (1992)
[13] D. Richter, B. Farago, R. Butera, L.J. Fetters, J.S. Huang, B. Ewen, Macromolecules **26**, 795 (1993)
[14] D. Richter, B. Ewen, B. Farago, T. Wagner, Phys. Rev. Lett. **62**, 2140 (1989)
[15] M. Doi, S.F. Edwards, J. Chem. Soc. Farad Trans. **274**, 1789 and 1802; **275**, 38 (1978)
[16] S.F. Edwards, J.M.V. Grant, J. Phys. A **6**, 1169 (1993)
[17] P.G. de Gennes, *Scaling concepts in polymer physics*, Cornell University Press, Ithaca (1979)
[18] P.G. de Gennes, J. Chem. Phys. **55**, 572 (1971)
[19] N. Fatkullin, R. Kimmich, Phys. Rev. E **52**, 3273 (1995)
[20] D. Richter, A. Baumgärtner, K. Binder, B. Ewen, J.B. Hayter, Phys. Rev. Lett. **47**, 109 (1981)
[21] P.G. de Gennes, J. Phys. (Paris) **42**, 735 (1981)
[22] G.J. Ronca, J. Chem. Phys. **79**, 79 (1983)
[23] J. des Cloiseaux, Macromolecules **23**, 3992 (1990)
[24] K.S. Schweitzer, J. Chem. Phys. **91**, 5802, 5822 (1989)
[25] D. Richter, L. Willner, A. Zirkel, B. Farago, L.J. Fetters, J.S. Huang, Phys. Rev. Lett. **71**, 4158 (1993)
[26] D. Richter, L. Willner, A. Zirkel, B. Farago, L.J. Fetters, J.S. Huang, Macromolecules **27**, 7437 (1994)
[27] A. Wischnewski, M. Monkenbusch, L. Willner, D. Richter, A.E. Likhtmann, T.C.B. McLeish, B. Farago, Phys. Rev. Lett. **88**, 058301 (2002)
[28] P. Schleger, B. Farago, A. Kollmar, C. Lartigue, D. Richter, Phys. Rev. Lett. **81**, 124 (1998)
[29] R.H. Colby, L.J. Fetters, W.W. Graesseley, Macromolecules **20**, 2226 (1987)
[30] A.E. Likhtmann, T.C.B. McLeish, to be published
[31] N. Clark, T.C.B. McLeish, Macromolecules **26**, 5264 (1993)
[32] T.C.B. McLeish, Advances in Physics **51**, 1379 (2002)
[33] S.T. Millner, T.C.B. McLeish, Phys. Rev. Lett. **81**, 725 (1998)

[34] J. Käs, H. Strey, E. Sackmann, Nature **368**, 226 (1994)
[35] B.H. Zimm, J. Chem. Phys. **24**, 269 (1956)
[36] B. Ewen, D. Richter, Advances in Polymer Science **134**, 1 (1997)
[37] E. Dubois Violette, P.G. de Gennes, Physics (USA) **3**, 181 (1967)
[38] D. Richter, K. Binder, B. Ewen, B. Stühn, J. Phys. Chem. **88**, 6618 (1984)
[39] T. Csiba, G. Jannink, D. Durand, R. Papoular, A. Lopp, L. Aurray, F. Bowe, J.P. Cotton, R. Borsali, J. Phys. II **1**, 381 (1991)
[40] L. Harnau, R.G. Winkler, P. Reinecker, J. Chem. Phys. **102**, 7750 (1995) and **104**, 6355 (1996)
[41] G. Allegra, F. Ganazzoli, Macromolecules **14**, 1110 (1981), Advances in Chemical Physics **75**, 265 (1989)
[42] D. Richter, M. Monkenbusch, J. Allgaier, A. Arbe, J. Colmenero, B. Farago, Y. Cheol-Bae, R. Faust, J. Chem. Phys. **111**, 6107 (1999)
[43] D. Richter, M. Monkenbusch, W. Pyckhout-Hinzten, A. Arbe, J. Colmenero, J. Chem. Phys. **113**, 11398 (2000)
[44] A. Arbe, M. Monkenbusch, J. Stellbrink, D. Richter, B. Farago, K. Almdahl, R. Faust, Macromolecules **34**, 1281 (2001)
[45] A.A. Jones, R.P. Lubianez, M.A. Hanson, S.L. Shostak, J. Polym. Sci. Polym. Phys. Ed. **16**, 1685 (1978)
[46] U.W. Suter, E. Saiz, P.J. Flory, Macromolecules **16**, 317 (1983)
[47] W. Doster, S. Cusack, W. Petry, Nature **337**, 754 (1989)
[48] M. Ferrand, J. Dianoux, W. Petry, G. Zaccai, Proc. Nat. Acad. Sci. (USA) **90**, 9668 (1993)
[49] S. Longeville, W. Doster, G. Kali, Chem. Phys. **292**, 413 (2003)

C 3 Polyelectrolytes and Semiflexible Polymers

R. G. Winkler
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Poisson-Boltzmann Equation – Debye-Hückel Potential 4

3 Conformational Properties of Polyelectrolytes in Solution: Mean Field Approach 6
3.1 Polyelectrolyte at infinite dilution 7
3.2 Polyelectrolyte at finite dilution 7

4 Conformational Properties of Polyelectrolytes in Solution: Counterion Condensation 8

5 Polyelectrolyte Aggregates 12

6 Conformational Properties of Semiflexible Polymers 15
6.1 Electrostatic Persistence Length 16
6.2 Gaussian Semiflexible Polymer Model 17
6.3 Stretching DNA 18

7 Dynamical Properties of Semiflexible Polymers 20
7.1 Chain Dynamics in Free-Draining Limit 20
7.2 Hydrodynamic Interaction 23
7.3 Dynamics of Partially Stretched DNA Molecules 28

8 Conclusions 30

A Appendices 32
A.1 Solution of the Linearized Poisson-Boltzmann Equation 32
A.2 Electrostatic Persistence Length in the Rod Limit 33
A.3 Equation of Motion for Semiflexible Polymers 34
A.4 Mean Square Displacement 36

1 Introduction

Macromolecules constitute an integral part of biological systems. Proteins, for example, carry the major part of the mechanical and chemical functions of cells. Desoxyribonucleic acid (DNA) is the molecular basis of heredity and is constructed of a double helix held together by hydrogen bonds between purine and pyrimidine bases which are bound to two chains containing alternate links of deoxyribose and phosphate (cf. Fig. 1). The extraordinary mechanical and dynamical properties of eukaryotic cells are determined by a three dimensional assembly of protein fibers, the cytoskeleton. A major contribution to this remarkable properties is due to actin filaments, which consist of two strands of globular molecules called actin (cf. Fig. 2), and proteins that crosslink them [1,2].

A key aspect of DNA structure is its large length-to-width ratio. The diameter of the molecule is approximately 2 nm, whereas typical contour lengths range from micrometers in simple viruses to meters in humans. Thus, its structural and dynamical properties can easily by analyzed by optical methods. From a theoretical point of view, this makes DNA an ideal candidate to verify polymer models.

Another of DNA's important structural properties is that the molecule is highly charged under standard, that is, physiological, pH conditions, i.e., DNA is a *polyelectrolyte*. Polyelectrolytes are macromolecules build of monomers with ionizable groups. When the molecules are dissolved in a polar solvent such as water, the ion pairs dissociate. The electrostatic charges of one sign remain localized on the chain whereas the large number of oppositely charged counterions are scattered in the solution. With respect to DNA, a hydrogen atom dissociates from the phosphate group. With an effective density of one fundamental (negative) charge every 0.17 nm of its length, DNA is about as highly charged as a linear molecule can be. The charges make the DNA molecules soluble in water. Indeed, they need to be charged to avoid precipitation and phase separation at the high concentrations that characterize them in vivo. Understanding how nature controls the electrostatic interaction between biomolecules and biostructures is a fascinating challenge.

Aside from biological molecules there is a broad spectrum of colloidal and synthetic macromolecular charged systems. Technologically polyelectrolytes are used in a wide-range of applications such as dewatering agents, drug reduction agents, as additives in detergents and cosmetics, in gels, in superabsorbants etc. [3–8]. Many outstanding scientists payed attention to such charged systems. Among them were Michael Faraday, Peter Debye, and Ernst Hückel.

Polyelectrolytes, and in particular DNA, self-organize in tightly packed structures. An particular example is chromatin, the dense complex of DNA and histone proteins occupying the nuclei of plant and animal cells (see Fig. 3). Moreover, DNA molecules condense into toroids and rods in the presence of multivalent cations [9] or polyamines [10]. Also complexes are formed with cationic polymers [11, 12] and cationic dendrimers [13, 14]. Most of the synthetic cationic agents forming complexes and aggregates[1] are developed for potential use as DNA vectors in novel gene therapies.

The above examples emphasize the importance of the charges of polyelectrolyte molecules. The long-range character of the Coulomb interaction among the various charges is the origin of a rich variety of properties of such systems which are not found in neutral polymers. It is important to keep in mind that the total system is electrically neutral. For every charge on a polyelectrolyte there is a so called counterion present in the solution. The more or less freely

[1] Such complexes are known as *self-assembled complexes for gene delivery.*

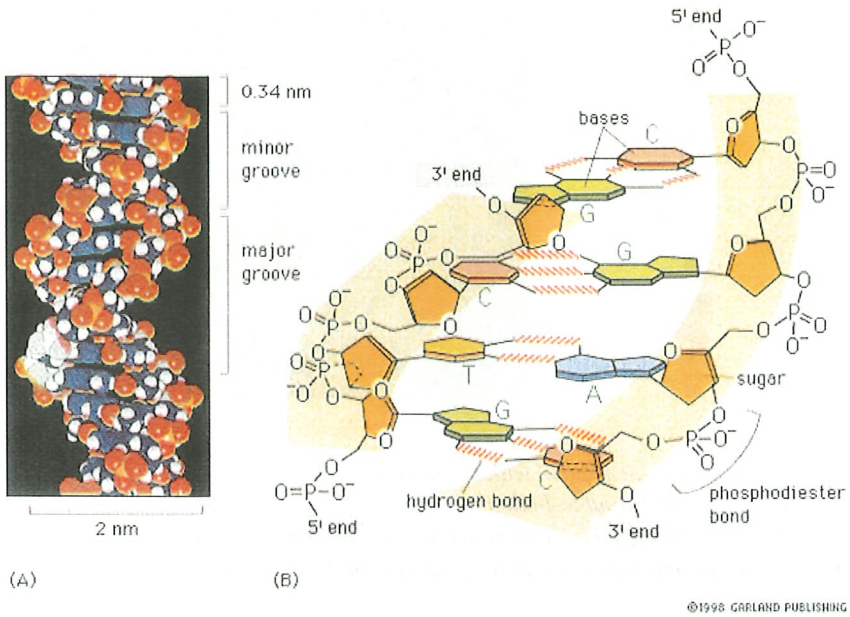


Fig. 1: *DNA is a particular example for a polyelectrolyte. It is about as highly charged as a linear molecule can be. In a polar solvent (water) the hydrogen atoms (not shown) dissociate leaving negatively charged oxygens behind. The hydrogen atoms themselves are floating in the solution. (The figure has been taken from Ref. [15].)*

floating counterions strongly affect the charge-charge interactions among the polyelectrolytes. It is the Coulomb interaction which renders polyelectrolyte systems difficult to tread theoretically. However, understanding of biological systems requires to unravel the features of charge interaction in molecular systems.

For certain properties of DNA molecules, the charges need not to be considered explicitly. Instead, a DNA molecule can be considered as a semiflexible polymer taking the intramolecular Coulomb interaction into account by a persistence length ($l_p \approx 50$ nm). This concept has proven very useful to understand the deformation behavior and the dynamics of single DNA molecules. The cytoskeleton polymers (actin filaments, microtubules, and intermediate filaments) are semiflexible polymers too [1, 2, 16, 17]. Their persistence lengths are on the order of their contour lengths. In contrast to fully flexible linear macromolecules, the statistical mechanical properties and the dynamics of semiflexible chain molecules is not determined by the conformational entropy only, but depends crucially on the bending stiffness. This leads to particular physical phenomena and is fundamental for the biological functions of these molecules.

In the present contribution various fundamental properties of polyelectrolytes and semiflexible chains are discussed. In the first part polyelectrolytes are considered, i.e., the explicit influence of charges on interactions and structural properties of linear macromolecules is discussed. A particular aspect is counterion condensation which is elementary for the understanding of DNA aggregation. In the second part a model for semiflexible chains is introduced. The basic features of the model are outlined and various structural and dynamical properties of DNA molecules

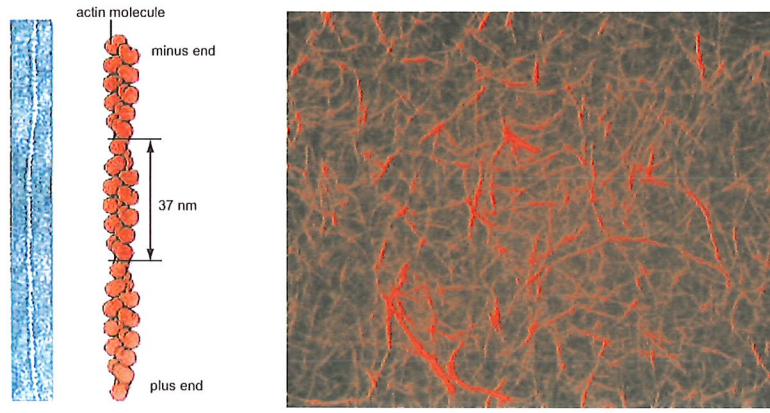


Fig. 2: *G-Actin is a globular protein which polymerizes to form F-actin. Electron micrographs (left) indicate the parallel double-helical structure of the filaments. As seen in the electron micrograph, the actin filament is a weakly bending rodlike molecule (semiflexible chain). Right: Filaments of actin stained with fluorescently-tagged antibodies [18].*

are extracted by comparison between the theoretical model and experimental results.

2 Poisson-Boltzmann Equation – Debye-Hückel Potential

The dissociated ions (counterions) of a polyelectrolyte are freely floating in a solution. Hence, the inter- and intramolecular charge-charge interactions of the polymers are given by an effective potential that is different from the bare Coulomb potential. To gain insight into the effective interaction between two charged particles in a solution, we neglect chain connectivity for a moment. Instead, we consider a solution of charged point particles, N_+ of positive charge and N_- of negative charge. The total electrostatic energy V of such a system is given by

$$V = \frac{1}{2}\sum_{\alpha=1}^{2}\sum_{i=1}^{N_\alpha} z_\alpha e\phi(\boldsymbol{r}_i^\alpha)\,, \tag{1}$$

$$\phi(\boldsymbol{r}) = \sum_{\gamma=1}^{2}\sum_{j=1}^{N_\gamma} \frac{z_\gamma e}{\epsilon|\boldsymbol{r}-\boldsymbol{r}_j^\gamma|}\,. \tag{2}$$

The indices $\alpha\,,\gamma \in \{1,2\}$ indicate the two different species of particles[2]. ϕ is the electrostatic potential, e the elementary charge, and $z_\alpha \in \mathbb{Z}$ is the valency of the particle of species α. The solvent is considered a continuous background with dielectric constant ϵ. Next, we consider the canonical average of ϕ when we fix a particle of kind α at $\boldsymbol{r}'$. Using the properties of a δ-function, we obtain

$$\langle\phi(\boldsymbol{r},\boldsymbol{r}')\rangle_\alpha = \frac{V}{N_\alpha}\left\langle\sum_{i=1}^{N_\alpha}\phi(\boldsymbol{r})\delta(\boldsymbol{r}'-\boldsymbol{r}_i^\alpha)\right\rangle\,, \tag{3}$$

[2]For $\alpha=\gamma$ the term with $i=j$ is left out from the summation.

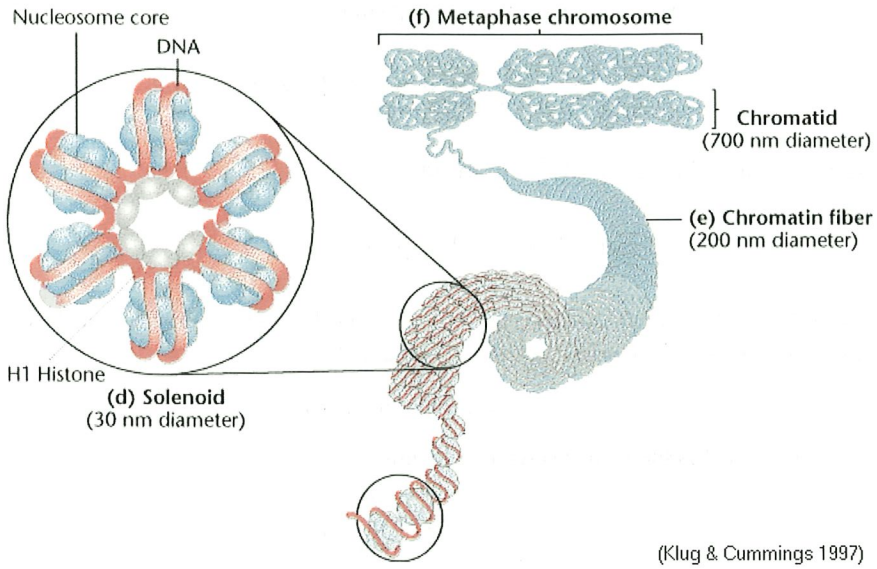


Fig. 3: *In eukaryotic cells, DNA and histone proteins form complexes which are known as chromatin [16]. The basic building unit of chromatin is the nucleosome. The nucleosome itself is an aggregate consisting of eight histone proteins (positively charged) and 146 associated base pairs of DNA. (The figure has been taken from Ref. [19].)*

where

$$\langle \ldots \rangle = \frac{1}{Z} \int \ldots \exp\left(-\beta V\right)\, d^{3N}x \,. \tag{4}$$

Z is the partition function and N the total number of particles. $\beta = 1/k_B T$, where k_B denotes the Boltzmann factor and T the temperature. By taking the Laplacian (Δ) of Eq. (3) we find

$$\Delta \left\langle \phi(\boldsymbol{r}, \boldsymbol{r}') \right\rangle_\alpha = -\frac{4\pi V}{N_\alpha \epsilon} \left\langle \sum_{\gamma=1}^{2} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\gamma} z_\gamma e \delta(\boldsymbol{r} - \boldsymbol{r}_j^\gamma) \delta(\boldsymbol{r}' - \boldsymbol{r}_i^\alpha) \right\rangle = -\frac{4\pi}{\epsilon} \left\langle \rho(\boldsymbol{r}, \boldsymbol{r}') \right\rangle_\alpha \,. \tag{5}$$

$\left\langle \rho(\boldsymbol{r}, \boldsymbol{r}') \right\rangle_\alpha$ is the mean charge density at $\boldsymbol{r}$ with a particle of species α fixed at $\boldsymbol{r}'$. Considering a homogeneous system[3], we can formally express $\rho(\boldsymbol{r}, \boldsymbol{r}')$ by the potential of mean force $w_{\alpha\gamma}(\boldsymbol{r})$ via

$$\left\langle \rho(\boldsymbol{r}) \right\rangle_\alpha = \sum_{\gamma=1}^{2} \varrho_\gamma z_\gamma e \exp\left(-\beta w_{\alpha\gamma}(\boldsymbol{r})\right) \,. \tag{6}$$

$\varrho_\alpha = N_\alpha / V$ is the density of species α. Equations (5) and (6) are exact so far. Following Debye and Hückel, we make the following approximations to obtain an analytical solution of equation (5) [20]:

- The potential of mean force is replaced by

$$w_{\alpha\gamma}(\boldsymbol{r}) \approx z_\gamma e \left\langle \phi(\boldsymbol{r}) \right\rangle_\alpha := z_\gamma e \bar{\phi}_\alpha(\boldsymbol{r}) \,. \tag{7}$$

[3]For a homogeneous system, $\rho(\boldsymbol{r}, \boldsymbol{r}')$ is a function of the difference $\boldsymbol{r} - \boldsymbol{r}'$ only.

We then arrive at the **Poisson-Boltzmann** equation

$$\Delta\bar{\phi}_\alpha(\boldsymbol{r}) = -\frac{4\pi}{\epsilon}\sum_{\gamma=1}^{2}\varrho_\gamma z_\gamma e \exp\left(-\beta z_\gamma e\bar{\phi}_\alpha(\boldsymbol{r})\right) . \tag{8}$$

- The right-hand side of Eq. (8) is linearized:

$$\sum_\gamma \varrho_\gamma z_\gamma e \exp\left(-\beta z_\gamma e\bar{\phi}_\alpha(\boldsymbol{r})\right) \approx \sum_{\gamma=1}^{2}\varrho_\gamma z_\gamma e - \beta\sum_{\gamma=1}^{2}\varrho_\gamma (z_\gamma e)^2\bar{\phi}_\alpha(\boldsymbol{r}) .$$

Since the total system is neutral, the first term on the right-hand side vanishes and the linearized Poisson-Boltzmann equation is obtained

$$\Delta\bar{\phi}_\alpha(\boldsymbol{r}) = \kappa^2\bar{\phi}_\alpha(\boldsymbol{r}) \tag{9}$$

where $l_D = \kappa^{-1}$ is the Debye screening length

$$\kappa^2 = 4\pi l_B\sum_{\gamma=1}^{2} z_\gamma^2\varrho_\gamma \tag{10}$$

and l_B the Bjerrum length[4]

$$l_B = \frac{e^2}{\epsilon k_B T} . \tag{11}$$

With the appropriate boundary conditions for a sphere of diameter l in infinite space, the solution of equation (9) is given by (see Appendix A.1)

$$\bar{\phi}_\alpha(r) = \frac{z_\alpha e}{\epsilon}\frac{e^{-\kappa(r-l)}}{r(1+\kappa l)} , \quad r > l , \tag{12}$$

i.e., the presence of the various ions leads to a screened Coulomb interaction between two charges (**Debye-Hückel** potential). For point particles, i.e., $l = 0$, the potential reduces to $\bar{\phi}_\alpha(r) = z_\alpha e e^{-\kappa r}/r\epsilon$. In section 3 we will use the Debye-Hückel potential to gain insight into the conformational properties of polyelectrolyte chains.

3 Conformational Properties of Polyelectrolytes in Solution: Mean Field Approach

The investigation of the conformational properties of flexible or semiflexible polyelectrolyte molecules requires the calculation of the partition function. Unfortunately, this is not possible even for the Debye-Hückel interaction which captures the counterions in the screening length κ^{-1}, i.e., the counterions are no longer considered explicitly. The non-linear character of this potential renders the strict analytical calculation of the partition function impossible. The coupling of different length scales in charged systems, however, will lead to phenomena which are not present for uncharged chains. Thus, it is important to elucidate the dependence of the characteristic chain dimensions on quantities like chain length, polymer concentration, and interaction strength, even if we have to resort to (severe) approximations.

[4]For water at room temperature: $\epsilon \approx 78.5$ and $l_B \approx 7.14$Å.

3.1 Polyelectrolyte at infinite dilution

The effect of the long-range nature of the Coulomb interaction on chain conformations is most pronounced for an isolated polymer. Such a situation is reached at infinite dilution.
The effect of the repulsive interaction between the like-charged monomers on chain conformations can by understood by a Flory-type calculation [7,21]. It relies on simplifying assumptions but provides a simple physical picture. For illustration, we consider a Gaussian flexible chain composed of N beads each carrying a charge e. In a mean-filed approach, the Flory-type free energy is given by

$$F = k_B T \frac{R^2}{Nl^2} + k_B T l_B \frac{N^2}{R} \,. \tag{13}$$

The first term is the contribution of the Gaussian chain with the mean squared end-to-end distance $\langle r_e^2 \rangle = Nl^2$ [22]. The second term is the electrostatic energy of the N charges. R is a typical length scale of the system. Minimization with respect to R yields

$$R \sim N(l_B l^2)^{1/3} \,. \tag{14}$$

Thus, the size of a chain increases linearly with the chain length. This is quit different form the behavior of a Gaussian chain ($R = \sqrt{N}l$) or a chain in good solvent (self avoiding walk; $R \sim lN^\nu$, with $\nu \approx 0.6$ [21]). The long-range Coulomb interaction leads to an increase of typical chain dimensions proportional to the chain length. This however, does not imply that the chain is fully extended. A local flexibility may still be present but R nevertheless scales with N.

3.2 Polyelectrolyte at finite dilution

At finite densities the Coulomb interaction among the monomers is screened by the counterions. The characteristic chain dimension will then depend on the screening length. In the presence of salt, e.g., sodium and chloride ions, the screening length κ^{-1} not only depends on the counterion concentration but also on the salt concentration. It is then given by $\kappa^2 = 4\pi l_B \left(z_c^2 \varrho_c + 2\varrho_s\right)$, where z_c is the counterion valance, ϱ_c the counterion density, and ϱ_s the density of the monovalent salt. As an example, the concentration of NaCl in biological systems is 0.1 mol/L.
The conformational properties of flexible polyelectrolytes in the presence of salt have been investigated for over half a century without reaching a consensus. Various method have been applied ranging from scaling arguments, variational methods, perturbation calculations, and field-theoretic renormalization group analysis to computer simulations [23]. The latter method is the most reliable because computer simulations provide a (numerical) exact solution of a model system. From the point of few of analytical theory, the perturbation approach by Edwards and Singh proved were useful [24–26].
Qualitatively, the observed behavior can be understood by the following considerations. Because of screening, the Coulomb interaction will be relevant for length scales Δs smaller then the Debye screening length: $\Delta s < l_D = 1/\kappa$ (see Fig. 4). For larger length scales ($\Delta s > l_D$), the interaction among the monomers is screened and the conformation of a polymer is determined by the short-range Debye-Hückel interaction. The statistical properties on this length scale are identical to those of a neutral chain with excluded volume interaction [22], i.e., $R \sim N^\nu$, with $\nu \approx 0.6$. In general, we will find a transition from the dependence $R \sim N$ to $R \sim N^\nu$ when we consider larger and large length scales or when the chain length is increased. The crossover length Δs_c between the two regimes is given by $\Delta s_c \sim 1/\kappa$.

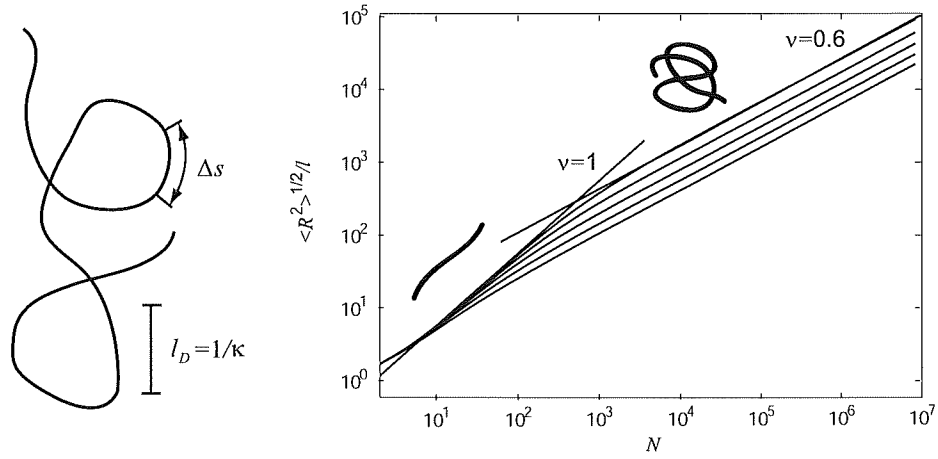


Fig. 4: *Conformational properties of flexible polymers for various screening lengths* $l_D = 1/\kappa$. *The left sketch illustrates various length scales. The right plot displays the mean square end-to-end distances of discrete chains of bond length* l *as a function of chain length and for the* κ *values* 0.05, 0.1, 0.2, 0.4, *and* 0.8 *(from top to bottom). For short chains the dependence* $R \sim N$ *is obtained, whereas for large* N $R \sim N^{0.6}$ *[26]. The crossover between the two regimes clearly depends on* κ.

Figure 4 displays the discussed behavior in a more quantitative manner. The various curves follow from the Edwards-Singh perturbation theory approach. Details of the calculation can be found in Ref. [26]. We clearly see the transition from a Coulomb interaction dominated regime to the regime of short range interaction.

Considering more sophisticated theoretical approaches, like the Polymer-Reference-Interaction-Site model (PRISM) [27, 28], which can take into account counterions and salt-ions explicitly, it is clear that the Debye-Hückel description applies for interaction strength $l_B/l < 1$ only[5] [29–31]. As is already obvious from the calculations of Sec. 2, the interaction among equally charged particles is always repulsive within the Poisson-Boltzmann approach. However, for large interaction strengths correlations among the counterions are important and the counterions need to be taken into account explicitly.

4 Conformational Properties of Polyelectrolytes in Solution: Counterion Condensation

Correlations among counterions give rise to interesting effects and novel structures. A particular aspect, which requires the explicit consideration of the counterions, is know as counterion condensation.

The terminology counterion condensation was already introduced in the 60s of the last century [32,33] and refers to drastic changes in the behavior of counterions when the interaction strength exceeds a critical value.

[5] We will denote the ration l_B/l as interaction strength.

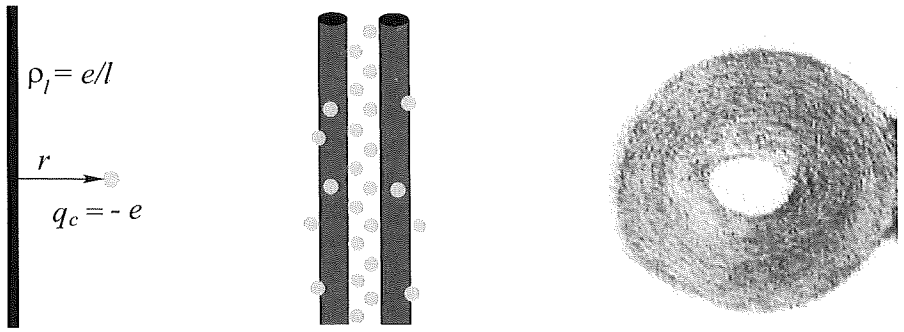


Fig. 5: *Counterion in front of a line charge (left). Counterion arrangement in the low temperature limit on two charged rods (middle). DNA molecule condensed into a torus by adsorption of tetravalent counterions (spermine) (right) [34].*

For illustrative purposes we consider an infinitely long and infinitely thin charged rod, i.e., an infinite line charge (cf. Fig. 5). The linear charge density of the rod is $\rho_l = e/l$. The potential of the line charge is obtained from the Poisson equation with the charge density $\rho_R = \rho_l \delta(x)\delta(y)$, where the rod is located in the origin of the reference system and is pointing along the z-axis. Because of the translational symmetry along the z-axis, it is sufficient to solve the two-dimensional Poisson equation, which yields the result

$$\phi_R(r) = -\frac{2e}{l\epsilon} \ln r \,, \tag{15}$$

where $r = \sqrt{x^2 + y^2}$. The partition function of a single charge in front of the line charge is given by

$$Z_R \sim \int_0^{\rho_0} \exp\left(-(l_B/l) \ln r\right) 2\pi r \, dr = \int_0^{\rho_0} r^{1-2l_B/l} \, dr \,. \tag{16}$$

(The distance ρ_0 is chosen in such a way that other counterions are exclude, i.e., a very dilute system is considered and interactions among the counterions are neglected.) The integral of Eq. (16) diverges at the lower limit for $1 - 2l_B/l \leq 1$ or

$$l_B \geq l \,. \tag{17}$$

For $l_B/l = 1$ is the Coulomb energy of two charges at a distance l equal to the thermal energy $k_B T$. The physical interpretation of the divergence is that systems characterized by such values are unstable. Sufficiently many counterions will condense on the line charge to reduce the interaction strength $l_B/l > 1$ to a value just below one [33]. This is called counterion condensation. The critical value $l_B = l$ is obtained for an infinitely thin line charge at any concentration. For a rod of finite thickness Eq. (17) is valid only in the limit of zero concentration [33].

For a planar geometry of fixed charge, the counterions are always condensed, independent of the surface charge density. This is a consequence of the linearly increasing electrostatic energy with distance for the one-dimensional situation [34].

To determine a critical interaction strength l_B/l for a system at finite density and flexible chains is difficult. Valuable insight into the phenomenon of counterion condensation is obtained by

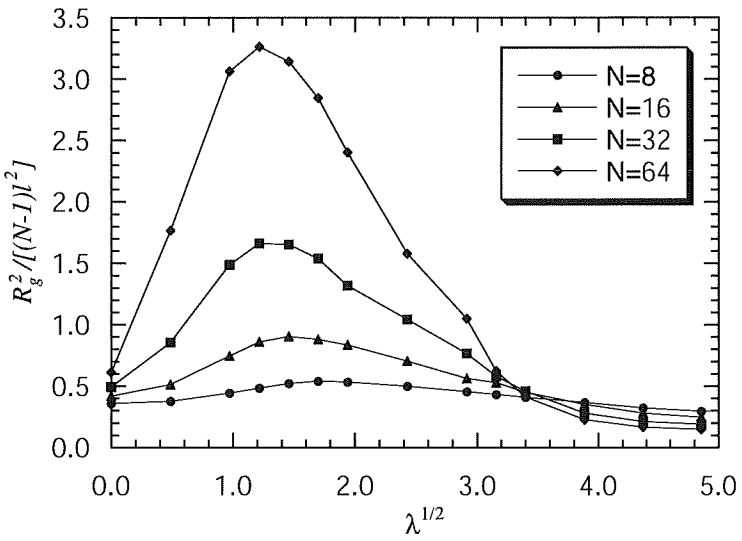


Fig. 6: *Squared radius of gyration divided by the number of bonds as a function of the square root of the interaction strength $\lambda = 1.2 l_B/l$ for various chain lengths N [36].*

computer simulations [35,36] (molecular dynamics simulations, Monte Carlo simulations). By solving the Langevin equation

$$m\ddot{\boldsymbol{r}} = \boldsymbol{F} - m\gamma\dot{\boldsymbol{r}} + \boldsymbol{\Gamma} \tag{18}$$

for a system of point particles of mass m, we studied the conformational properties and the counterion distribution of individual flexible polyelectrolyte chains in the presence of counterions [36]. The force $\boldsymbol{F}$ comprises contributions from the bond length potential (harmonic potential), intramolecular as well as inter-particle excluded volume interactions (Lennard-Jones potential), and the Coulomb interaction. The stochastic force $\boldsymbol{\Gamma}$ is assumed to be stationary, Markovian, and Gaussian (white noise) , i.e.,

$$\langle \boldsymbol{\Gamma}(t) \rangle = 0 , \tag{19}$$

$$\langle \boldsymbol{\Gamma}(t)\boldsymbol{\Gamma}(t') \rangle = 6k_BTm\gamma\delta(t-t') , \tag{20}$$

where γ is the friction constant [37] of the background solvent. This type of equation keeps the temperature fixed rather than the energy and produces canonical time averages. The system is placed in a cubic box and periodic boundary conditions are applied. The long-range Coulomb interaction in the infinite periodic system is taken into account by the Ewald sum [38]. The size of the periodic system is adjusted to yield the same density for any considered chain length.
Figure 6 displays the radius of gyration as a function of the interaction strength and for various chain lengths. The interaction strength $\lambda/1.2 = l_B/l = e^2/(k_BT\epsilon l)$, where l is the bond length, can be changed by changing the temperature, the dielectric constant, or the valency of the counterions. For zero charges, the conformations of a chain in good solvent are found. Here only the short-range excluded volume interaction is relevant. For non-zero charges the chains elongate with increasing interaction strength. The maximum extension is reached close to the

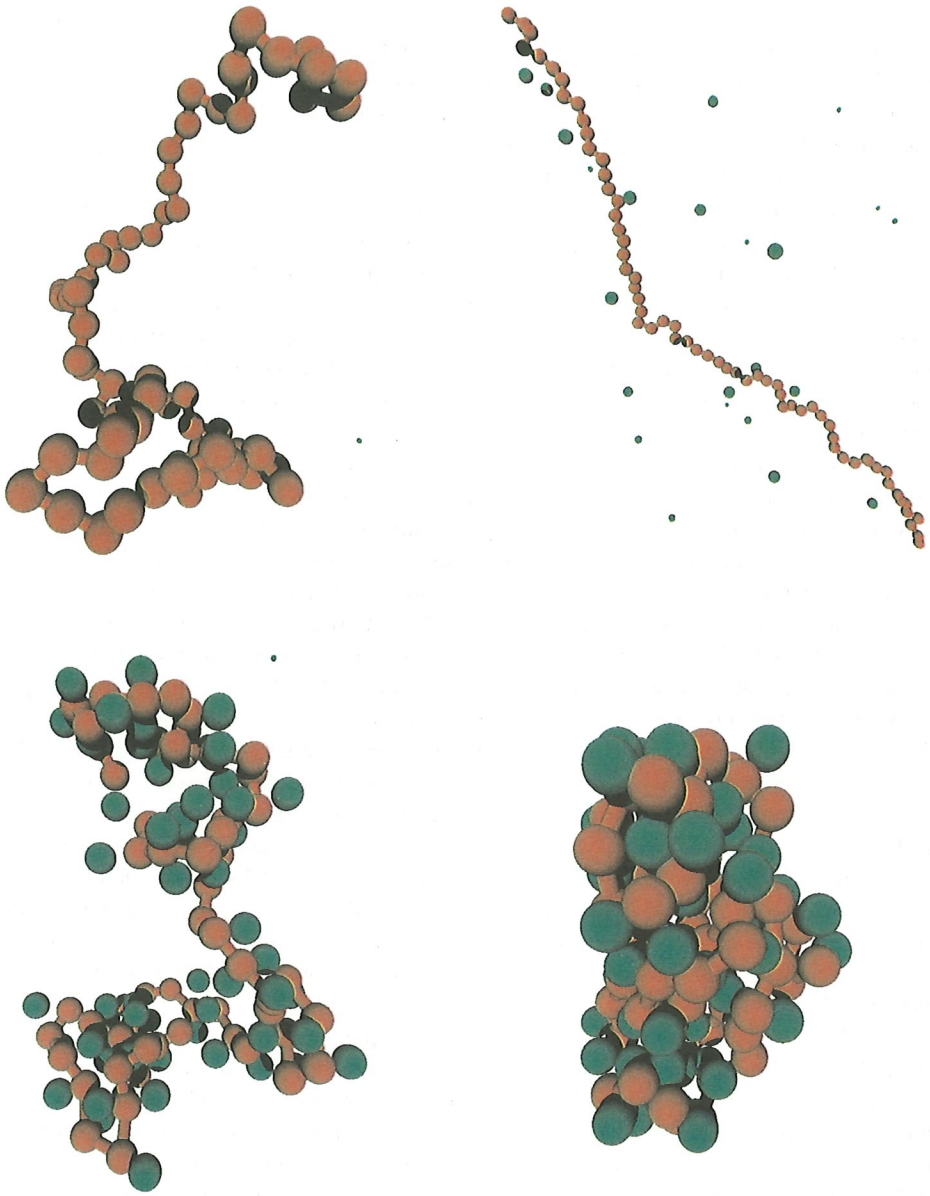


Fig. 7: *Snapshots of polyelectrolyte macromolecules for the interaction strengths* $l_B/l = 0$ *(top left),* $l_B/l = 1.3$ *(top right),* $l_B/l = 9.7$ *(bottom left), and* $l_B/l = 19.7$ *(bottom right). Counterions are colored green [36].*

interaction strength where Manning condensation occurs. For the simulated systems the critical Bjerrum length is in the vicinity of $l_B/l \approx 1.2$. For larger l_B/l the ions condense on the chain resulting in a decrease of chain dimension by screening of the bare Coulomb interaction. For

very large interaction strengths $l_B/l \gtrsim 10$ ion pairs are formed, which leads to an attractive intramolecular interaction and a collapse of the whole structure into a dense globular state (see Fig. 7). For the considered system, the radius of gyration in the collapsed state is even smaller than the one of a neutral chain in good solvent. Simulations with other counterion-monomer radii ratios show hat this does not hold for all ratios. Instead, we find flexible ring-like structures for counterions much smaller than a monomer.

Figure 7 depicts snapshots of chain conformations at various interaction strengths. For zero charges, the conformations correspond to those of chains in a good solvent (top left snapshot of Fig. 7). The top right snapshot corresponds to a chain close to its maximum extension. In the bottom left snapshot almost all counterions are condensed on the chain leading to flexible chain behavior. Finally, at large interaction strengths compact globular structures are obtained. The condensation of counterions leads to ion-pair formation. The resulting dipole moments of the various pairs are the origin of an intramolecular attraction and the formation of compact structures.

As a consequence of counterion condensation, equally charged (rodlike) macroions attract each other in solution. This follows from the PRISM approach [30, 31], when calculating an effective potential among monomers. Computer simulations of two charged rods yield similar results [39]. Considering two parallel uniformly charged rods, the authors find a short range attractive mean force for multivalent ions and a sufficiently large charge density of a rod. From the simulations the distribution of the condensed counterions is inferred. At low temperatures (ground state) the origin of the attraction is the formation of two interlocking structures, resembling Wigner crystals, arranged around the two cylinders (see Fig. 5). In the low temperature limit, electrostatic attraction is always encountered, independent of the valency of the counterions. This simply reflects a general feature of the electrostatics of charged particles. With increasing temperature, the two Wigner crystals lose long-range positional order, but short-range order survives up to high temperatures. The attraction at finite temperature may be attributed to this residual positional order, which is correlated for small rod diameters [39]. The general observations of the simulations agree with the results of the PRISM approach. However, in the considered isotropic and homogeneous solutions rods are not parallel in general. Although there are correlations among the counterions in the solution, I do not expect even short-range order. In contrast, the presence of attractive forces in solutions suggests that already liquid-like order is sufficient for attraction.

The DNA torus shown in Figure 5 provides an example for DNA compaction in the presence of multivalent counterions. In a highly dilute aqueous solution, the multivalent counterions condense on the DNA molecule. Due to its stiffness, the molecule folds into a tightly packed, circumferentially wound torus. Its average radius is about 50 nm, and the distance between the axes of neighboring, parallel portions of the molecule is only slightly large than its diameter.

So far only charge effects were addressed. But, the conformational properties of polyelectrolytes depend on the solvent quality also. Scaling arguments provide a rich phase diagram for the various combinations of interaction strengths, polymer concentrations, and excluded volume interactions. More details on this issue can be found in Refs. [40–42] and references therein.

5 Polyelectrolyte Aggregates

The negative charge of a DNA molecule is essential for its biological function. As discussed in the last section, DNA forms aggregates resembling the DNA packing in viruses in the presence

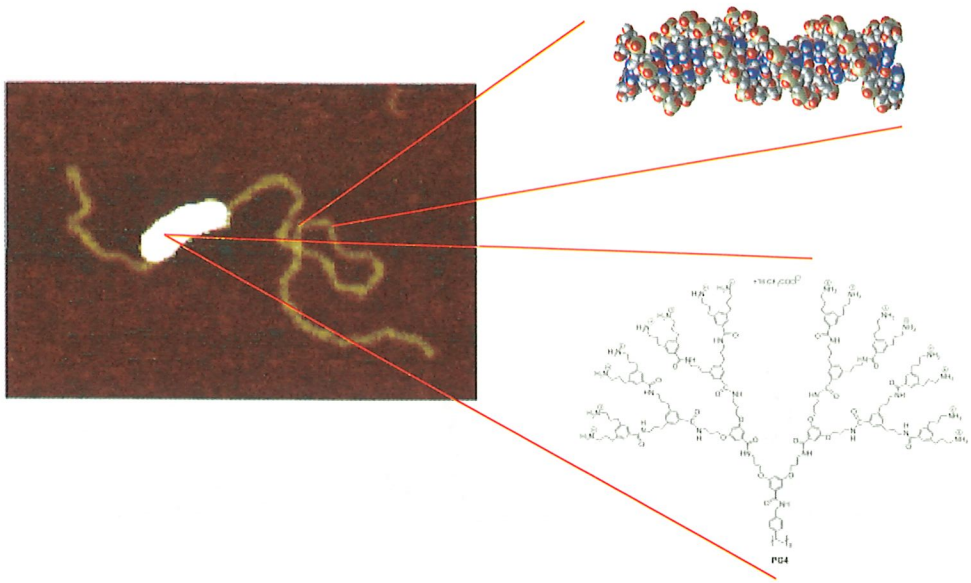


Fig. 8: *High resolution SFM image of a DNA-dendronized polymer complex (white spot) deposited onto mica [14]. The brown line indicates the DNA molecule.*

of multivalent counterions. In viruses DNA is just stored, whereas in eukaryotic cells a much more versatile method is required to store the up to 2 m long molecule and to provide excess for a large number of proteins (gene regulatory proteins, transcription factors, RNA polymerases etc.) at the same time. This is achieved by chromatin, a complex of DNA with histone proteins (cf. Fig. 3) [43]. X-ray crystallography has shown that in a nucleosome 146 bp DNA wrap in 1.65 turns around the histone octamers. The nucleosomes possess a net negative charge because the negative charge of the wrapped DNA is significantly larger (about 20%) than the total positive charge of the histone octamer [14, 34].

Charge reversal or neutralization of DNA is of special interest for the delivery of genes to a living cell for the purpose of gene therapy. Since both, bare DNA and a cell surface are negatively charged, DNA does not approach the cell surface. The idea is to screen the DNA charge such that the resulting complex is neutral or positively charged.

Complexes formed with various cationic molecules have been studied. Polyamines like spermidine and spermine constitute a group of cell components and even though their functions are mostly unknown, it is believed that they are important in the regulation of cell proliferation and differentiation. Moreover, it is widely accepted that they are helpers on DNA packing. Due to their positive charge, they induce DNA compaction, aggregation, and precipitation [10, 14, 44]. The resulting structures, however, have not be resolved on a molecular level. Also in complexes formed with cationic polymers [11, 12] and cationic dendrimers [13, 14], the molecular structure remains unclear, while X-ray diffraction on DNA-lipid complexes (lipoplexes) reveals multilamellar structures [45].

Despite the importance of DNA condensation, we are just beginning to unravel the mechanisms

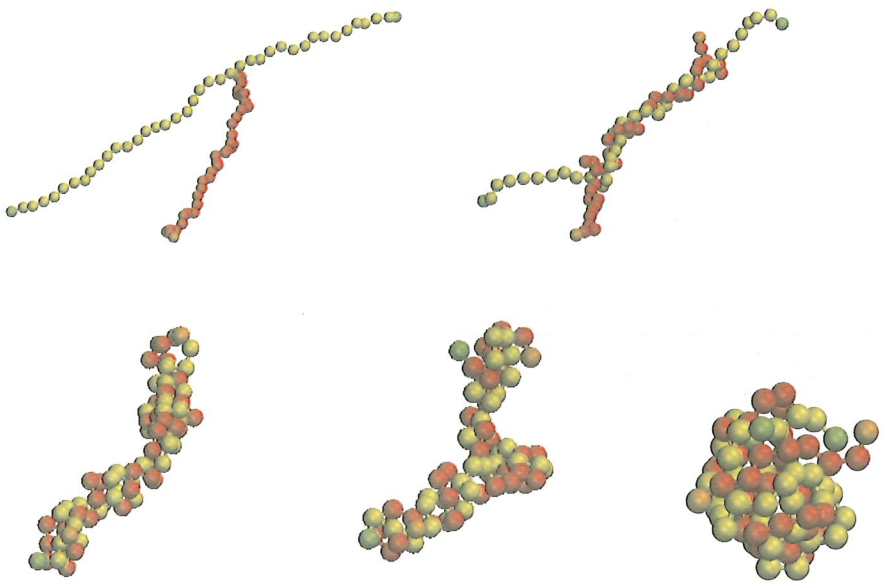


Fig. 9: *Snapshots of two oppositely charged polyelectrolyte macromolecules at successive times [52]. Initially (top left), the system is in a non-equilibrium state. With time, the two chains aggregate in a helical structure and eventually collapse into a dense globule (bottom right). The interaction strength is* $l_B/l = 6.7$ *and the number of beats is* $N = 40$.

leading to particular aggregates. Computer simulations are an important tool in this endeavor, because correlations among the various charges are captured in the numerical solutions for the model systems.

Complexes of polyelectrolytes and oppositely charged spherical colloids [46–49] have been studied intensively, since they resemble nucleosome like structures. But also systems of several oppositely charged chains have been considered [50–55]. Figure 8 shows an example of an DNA-cationic dendrimer complex precipitated onto a coated mica surfaces. The cylindrically shaped dendrimer molecule is approximately 100 nm long. Every monomer carries 16 positive charges, thus, the molecules is highly charged. The complexes were adsorbed on mica which was coated by a positively charged polymer and then imaged by scanning force microscopy [14]. From the analysis of the images, the authors of Ref. [14] conclude that DNA wraps around the cylinder in a helical manner.

To illustrate certain structural features of polyelectrolyte aggregates, Fig. 9 shows snapshots of a computer simulation of two oppositely charged flexible chains of the same length [52, 56]. The simulation method and the forces are described in Sec. 4. The chain length for the particular system of Fig. 9 is $N = 40$ and the interaction strength[6] is $l_B/l = 6.7$. Due to the strong intramolecular interaction, the molecules are in a stretched conformation initially (cf. Fig. 6). The separated molecules attract each other and move towards each other. When they

[6]For DNA, the interaction strength is $l_B/l = 4.2$.

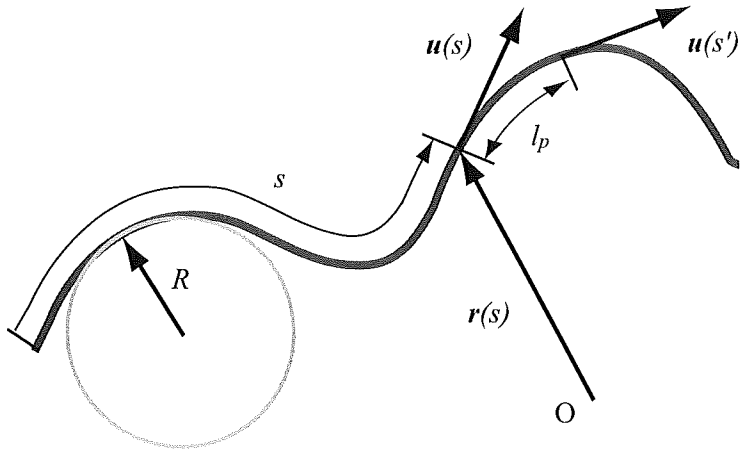


Fig. 10: *Sketch of a continuous linear polymer.*

are close, they start to wrap around each other in a helical fashion. This process continuous until essentially all monomers found a matching counterpart. If one of the molecules would be rigid, we would see a molecule wrapped around a cylinder in a helical manner. This corresponds to the DNA-dendrimer complex of Fig. 8. Hence, the simulations support the conclusion that the DNA wraps around the dendrimer in helical fashion. Since both molecules are flexible, they start to form a super structure and finally collapse into a dense globule. Such a globular structure is typical for neutral systems in an aggregated state. Simulations of a long polyelectrolyte chain interacting with many (not to) short oppositely charged chains (polyamines) yield similar globules [44].

6 Conformational Properties of Semiflexible Polymers

Under physiological conditions, a DNA molecule assumes the form of a disordered coil. Only if two parts come within 1 nm, the strongly repel. The conformational and dynamical properties of such a DNA molecule can be described by a semiflexible chain model. There is a broad spectrum of other semiflexible polymers, like actin filaments, microtubules (they are part of the cytoskeleton), and fd-viruses to name just a few [16, 17]. A characteristic feature of semiflexible polymers is their persistence length l_p, which is essential for their biological functions. The persistence length is a characteristic lengths scale and is defined as the length at which the mean tangent-vector correlation function along the chain contour has decayed to $1/e$, i.e., $\langle \boldsymbol{u}(s)\boldsymbol{u}(s')\rangle = \exp(-|s-s'|/l_p)$ (cf. Fig. 10). In contrast to flexible polymers, the structural and dynamical properties of such molecules are not only determined by entropy, but crucially depend on the bending stiffness, described by the persistence length. The persistence length of a DNA molecules is approximately 50 nm, whereas the one of the cytoskeleton polymers is on the order of their contour lengths.

6.1 Electrostatic Persistence Length

The persistence length of a charged polymer has two contributions: The first one is due to steric restrictions (l_0) and the second one is due to electrostatic interactions (l_e). In general, the electrostatic contribution cannot simply be calculated, since it requires the calculation of the partition function. However, in the limit of a rodlike polymer, an analytical expression can be derived.

The calculation is most easily done within the Kratky-Porod wormlike chain model [57]. The Kratky-Porod model describes the polymer chain as a continuous space curve $\boldsymbol{r}(s)$ of length L, i.e., $0 \leq s \leq L$. The tangent vector $\boldsymbol{u}(s) = \partial \boldsymbol{r}(s)/\partial s$ satisfies the condition $|\boldsymbol{u}(s)| = 1$, independent of the position along the chain contour. In addition, the local curvature is given by $1/R = |\partial \boldsymbol{u}(s)/\partial s|$. This curvature can be associated with a bending energy. For $R \to \infty$ (straight rod), the bending energy is zero and increases with decreasing R. Thus, the bending energy can be defined as

$$\beta V_B = \frac{\epsilon_p}{2} \int_0^L \frac{1}{R^2(s)} ds = \frac{\epsilon_p}{2} \int_0^L \left(\frac{\partial \boldsymbol{u}(s)}{\partial s} \right)^2 ds \,. \tag{21}$$

ϵ_p depends on the persistence length. Explicitly, $\epsilon_p = l_p$ in three dimensional space. Hence, the partition function of the Kratky-Porod wormlike chain is given by

$$Z = \int \exp\left(-\frac{l_p}{2} \int_0^L \left(\frac{\partial \boldsymbol{u}(s)}{\partial s} \right)^2 ds \right) \mathcal{D}^3 x \,, \tag{22}$$

where the path integral has to be determined taking the constraint $|\boldsymbol{u}(s)| = 1$ into account. ($\mathcal{D}^3 x = \lim_{N\to\infty} d^3x_1 \ldots d^3x_N$, $d^3x_i = dx_i dy_i dz_i$) Alternatively, Z can be derived starting from a discrete chain model taking bending restrictions into account by constraints [58]. The partition function (22) accounts for the conformational space only. Since we consider a canonical ensemble, the potential energy and the kinetic energy are independent. The total partition function is given by $Z_{tot} = Z Z_{kin}$, where $Z_{kin} = \int \exp(-\beta \int \boldsymbol{p}^2/(2\varrho)\, ds) \mathcal{D}^3 p$ is the kinetic contribution, $\boldsymbol{p}(s)$ is the momentum and ϱ the linear mass density. Since we are interested in conformational properties only, Z_{kin} is a constant which cancels in any calculation of averages and thus can be neglected in the following discussions.

To derive the electrostatic persistence length, the intramolecular Coulomb interaction is described by the Debye-Hückel energy

$$\beta V_{DH} = \frac{l_B}{l^2} \int_0^L \int_{s'}^L \frac{e^{-\kappa |\boldsymbol{r}(s) - \boldsymbol{r}(s')|}}{|\boldsymbol{r}(s) - \boldsymbol{r}(s')|} \, ds \, ds' \,, \tag{23}$$

where e/l is the linear charge density. We now assume that the rodlike polymer is weakly bending. The partition function is then obtained by a perturbation calculation with a rod as reference conformation. The details of the calculations are presented in Appendix A.2. The result is a new partition function of the form of Eq. (22) with the effective persistence length[7]

$$l_p = l_0 + \frac{1}{4} \frac{l_B}{\kappa^2 l^2} \,. \tag{24}$$

This expression has originally be derived by Odijk [59] and Skolnik and Fixman [60]. A particular feature of l_e is its quadratic dependence on the inverse Debye screening length.

This consideration shows that electrostatic interactions can be captured by an effective persistence length in certain applications.

[7] This expression is valid for an infinitely long chain. For finite chains there are L dependent correction terms.

6.2 Gaussian Semiflexible Polymer Model

The Gaussian chain model is the minimal model for flexible polymers [22]. Similarly, a minimal model for semiflexible chains can be derived. The basic features of this model are: (i) Gaussian segments, i.e., the distribution function of the distance between two points along the chain contour is Gaussian and (ii) bending restrictions are taken into account. The advantage of such a model over the Kratky-Porod wormlike chain model is its analytical tractability, in particular in studies of the dynamics of semiflexible polymers. In the following section various aspects of the Gaussian semiflexible chain model will be discussed.

The distribution and partition function of a Gaussian semiflexible chain can systematically be derived by the maximum entropy principle [58,61]. For a continuous chain the partition function reads

$$Z = \int \exp\left(-\nu \int_0^L \left(\frac{\partial \boldsymbol{r}(s)}{\partial s}\right)^2 ds - \nu_0 \left[\left(\frac{\partial \boldsymbol{r}(0)}{\partial s}\right)^2 + \left(\frac{\partial \boldsymbol{r}(L)}{\partial s}\right)^2\right] - \frac{\epsilon_p}{2} \int_0^L \left(\frac{\partial^2 \boldsymbol{r}(s)}{\partial s^2}\right)^2 ds\right) \mathcal{D}^3 x \,. \quad (25)$$

The Lagrangian multipliers ν, ν_0, and ϵ_p are given by

$$\nu = \frac{3p}{2}\,, \qquad \nu_0 = \frac{3}{4}\,, \qquad \epsilon_p = \frac{3}{4p}\,, \qquad p = \frac{1}{2l_p}\,. \quad (26)$$

Here, p is an abbreviation which depends on the persistence length. The terms with the first derivative in $\boldsymbol{r}$ capture the chain flexibility, i.e., they account for the chain entropy. The term with the second derivative is the bending energy, similar to the bending energy of the Kratky-Porod wormlike polymer (21). The chain ends behave different from the rest of the chain because of the break down of the translational symmetry, which is captured in the terms with ν_0.

The major difference between the Kratky-Porod model and the Gaussian semiflexible chain model is the constraint for the tangent vector. In the Gaussian model the constraint $\boldsymbol{u}^2(s) = 1$ is relaxed to $\langle \boldsymbol{u}^2(s) \rangle = 1$. Hence, the Gaussian model allows for fluctuations of the contour length. Naturally, this leads to differences in the statistical properties of the two models, which are, however, minor for an adequate treatment of the Gaussian model [58, 62].

As an example, Eqs. (27) – (30) present various averages obtained by the Gaussian semiflexible chain model:

$$\langle \boldsymbol{r}_e^2 \rangle = \frac{L}{p} - \frac{1}{2p^2}\left(1 - e^{-2pL}\right)\,, \quad (27)$$

$$\left\langle (\boldsymbol{r}(s) - \boldsymbol{r}(s'))^2 \right\rangle = \frac{|s-s'|}{p} - \frac{1}{2p^2}\left(1 - e^{-2p|s-s'|}\right)\,, \quad (28)$$

$$\langle \boldsymbol{r}_g^2 \rangle = \frac{L}{6p} - \frac{1}{4p^2} + \frac{1}{4p^3 L} - \frac{1}{8p^4 L^2}\left(1 - e^{-2pL}\right)\,, \quad (29)$$

$$\langle \boldsymbol{u}(s)\boldsymbol{u}(s') \rangle = e^{-2p|s-s'|}\,. \quad (30)$$

In particular, the mean square end-to-end distance ($\boldsymbol{r}_e$), the radius of gyration[8] ($\boldsymbol{r}_g$), and the correlation function of the tangent vectors are identical to those obtained by the Kratky-Porod

[8] For the definition see [22].

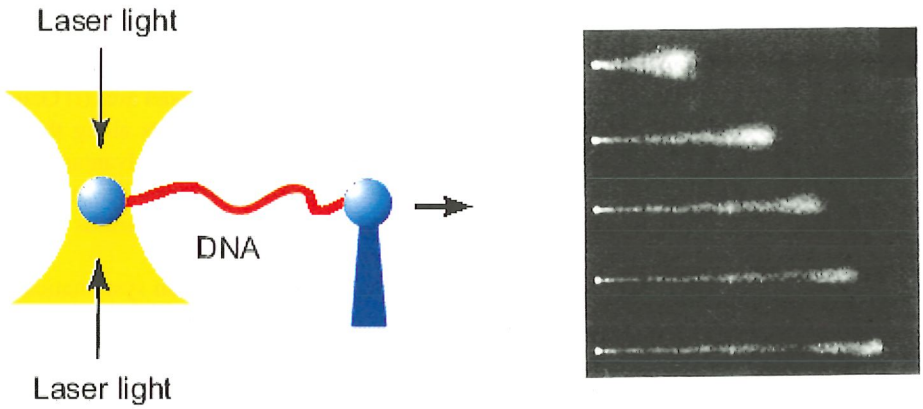


Fig. 11: *Schematic diagram for DNA stretching (left). The DNA is attached to beads that are fixed in space by a laser beam (tweezer) and a pipet. Alternatively the right bead could be positioned by a magnetic bead or a second tweezer. Right: Time-averaged images of tethered DNA molecules in a constant flow field (left to right) [65]. The flow velocity increases from top to bottom.*

model. The mean square end-to-end distance and the mean square radius of gyration assume the following values in the limit of a flexible and a rodlike polymer, respectively,

$$\langle \boldsymbol{r}_e^2 \rangle = \begin{cases} \frac{L}{p}, & pL \gg 1 \\ L^2, & pL \ll 1 \end{cases}, \qquad \langle \boldsymbol{r}_g^2 \rangle = \begin{cases} \frac{L}{6p}, & pL \gg 1 \\ \frac{L^2}{12}, & pL \ll 1 \end{cases}. \tag{31}$$

Finally, the joint probability distribution $\Psi(\boldsymbol{r}(s), \boldsymbol{r}(s'))$ of two points $\boldsymbol{r}(s)$ and $\boldsymbol{r}(s')$ is given by

$$\Psi(\boldsymbol{r}(s) - \boldsymbol{r}(s')) = \left(\frac{3}{2\pi\sigma(s-s')} \right)^{3/2} \exp\left(-\frac{3(\boldsymbol{r}(s) - \boldsymbol{r}(s'))^2}{2\sigma(s-s')} \right), \tag{32}$$

with $\sigma(s-s') = \langle (\boldsymbol{r}(s) - \boldsymbol{r}(s'))^2 \rangle$ (cf. Eq. 28).

6.3 Stretching DNA

Thanks to huge technical advances in manipulation of the structure of double-helix DNA, it has become feasible to measure the force-extension curve of single long DNAs. There are various ways to apply a force to a polymer. Typically, the DNA ends are attached to μm-size beads. The position of a bead can be controlled by an optical trap, a pipet, or a magnetic field, if a bead is magnetic [63]. Alternatively, external fields, like flow fields or electric fields, can be applied [64, 65] (see Fig. 11).

Theoretical investigations of the stress-strain behavior of polymer systems have a long tradition, particular in the field of rubber elasticity. These studies are based on freely-jointed chain models [22] ignoring the bending energy, which turned out to be a serious omission (see Fig. 12). The problem of a wormlike chain (Kratky-Porod model) under tension was first addressed in

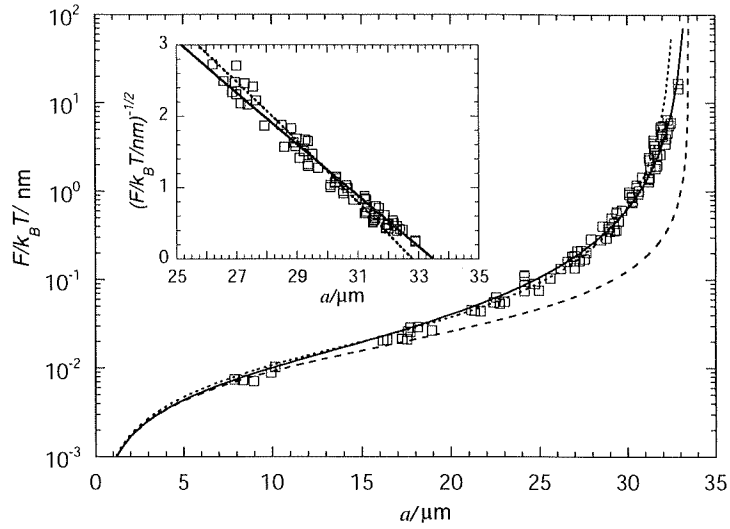


Fig. 12: *Force-extension relation of DNA molecules. The squares are experimental values by Smith et al. [63]. The solid line is a fit of the force-extension curve of the Gaussian semiflexible chain. The logarithmic fit yields the chain length* $L = 33.5$ *μm and the persistence length* $l_p = 53.5$ *nm. The dotted line is calculated using the interpolation formula derived in Ref. [67] for the Kratky-Porod model with the parameters* $l_p = 53$ *nm and* $L = 32.8$ *μm. The dashed line is calculated for a Gaussian flexible chain [61]. The insert shows* $F^{-1/2}$ *versus* a.

Ref. [66] and a full theoretical description of that model was provided in Ref. [67] a few years ago.

In this section, the force-extension relation is discussed for a Gaussian semiflexible polymer. For this purpose, the end point $\boldsymbol{r}_0 = \boldsymbol{r}(s = 0)$ is fixed at the origin of the reference frame, i.e., $\boldsymbol{r}_0 = 0$, and a constant force is applied to the other chain end[9]. The latter is equivalent to the constraint [62] $\langle \boldsymbol{r}(L) \rangle = \boldsymbol{a}$. Applying the maximum entropy principle, the partition function has now to be calculated taking this additional constraint into account. As a result, the Lagrangian multiplier ν depends on the end-to-end distance a, whereas the other parameters are unchanged. For sufficiently flexible polymers ($pL > 1$) we find the following analytical expression for the force-extension relation

$$F/k_BT = 2\nu\frac{a}{L} = \frac{3pa}{L\left(1 - \frac{a^2}{L^2}\right)^2}. \tag{33}$$

For small extensions a/L, the force increases linearly with a. For $a \to L$, the force diverges due to the finite length of the polymer. The force-extension relation (33) deviates significantly from the one of a freely jointed chain [22]. For large forces, Eq. (12) yields $L - a \sim F^{-1/2}$, whereas for a freely jointed chain $L - a \sim F^{-1}$ is obtained [62,67]. This clearly indicates that the bending energy affects the conformational properties even of flexible polymers.

[9] Alternatively, the chain end could be fixed at $\boldsymbol{a}$. However, the force-extension relation for rather flexible chains does not depend on this difference [62].

Figure 12 shows the force-extension relation of DNA molecules. The theoretical curve describes the experimental data by Smith et al. [63] on B-DNA very well. A least square fit of F to the experimental data yields the persistence length $l_p = 1/(2p) = 53.5$ nm and the chain length $L = 33.5$ μm, respectively, and corresponds to $pL \approx 313$. These values agree with those originally obtained by fitting the force-extension relation of the Kratky-Porod wormlike chain [67]: $l_p = 53nm$, $L = 32.8$ μm. Taking the number of base pairs of the experimental DNA (97004 bp) and the length for one base pair (1 bp = 0.34 nm) the chain length is $L = 33$ μm. The lengths obtained by the fits are in very good agreement with this value. The force-extension relation of a flexible chain (without bending energy) is significantly different from the one of a semiflexible chain. This clearly reveals that the bending stiffness has to be taken into account even for seemingly flexible polymers.

A central quantity for the characterization of the conformational properties of polymers is the distribution function $\Psi(\boldsymbol{r}_L)$ of the end-to-end vector $\boldsymbol{r}_L$. The explicit calculation yields the expression ($pL > 1$)

$$\Psi(\boldsymbol{r}_L) = N_c \left(1 - \frac{\boldsymbol{r}_L^2}{L^2}\right)^{-3/2} \left(2 - \frac{\boldsymbol{r}_L^2}{L^2}\right)^{-3} \exp\left(-\frac{3pL}{2\,(1 - \boldsymbol{r}_L^2/L^2)}\right) , \tag{34}$$

where N_c denotes the normalization constant, in three dimensions. This expression describes Monte Carlo simulation data for the Kratky-Porod model excellently [62, 68].

7 Dynamical Properties of Semiflexible Polymers

Not only equilibrium properties of polymers but also their dynamics is affected by the presence of chain stiffness. Naturally, the molecular dynamics is essential for the function of a biopolymer and hence it is important to understand the influence of stiffness on the dynamics of a semiflexible macromolecule. Examples are the transport of molecules, e.g. DNA, in veins by blood flow, diffusion through the cell cytoskeleton, or conformational transitions of proteins.

The successful description of the dynamics of flexible polymer chains by Rouse (Rouse model) and Zimm (Zimm model) [69] for systems without and with hydrodynamic interactions, respectively, is closely related with the Gaussian chain. For the quadratic Hamiltonian of a Gaussian chain a linear equation of motion is obtained which can by solved analytically. The corresponding extension to semiflexible chains has been formulated in Refs. [71, 72] for systems with and without hydrodynamic interaction.

7.1 Chain Dynamics in Free-Draining Limit

In the free-draining limit, no hydrodynamic interaction is present. The equation of motion (Langevin equation) for the mass point $\boldsymbol{r}(s,t)$ is obtained from the Lagrangian with the kinetic energy and the exponent of the partition function (25), which is identified with the potential energy, and reads (see Appendix A.3)

$$\gamma \frac{\partial}{\partial t}\boldsymbol{r}(s,t) + \epsilon_p k_B T \frac{\partial^4}{\partial s^4}\boldsymbol{r}(s,t) - 2\nu k_B T \frac{\partial^2}{\partial s^2}\boldsymbol{r}(s,t) = \boldsymbol{\Gamma}(s,t) , \tag{35}$$

with the boundary conditions

$$\left[2\nu\frac{\partial}{\partial s}\boldsymbol{r}(s,t) - \epsilon_p\frac{\partial^3}{\partial s^3}\boldsymbol{r}(s,t)\right]_{\pm L/2} = 0\,, \tag{36}$$

$$\left[2\nu_0\frac{\partial}{\partial s}\boldsymbol{r}(s,t) \pm \epsilon_p\frac{\partial^3}{\partial s^3}\boldsymbol{r}(s,t)\right]_{\pm L/2} = 0\,. \tag{37}$$

(Note, in the analysis of the chain dynamics the definition range of s is $-L/2 \leq s \leq L/2$.) The polymer molecule is much larger than a fluid particle and it is by far slower than the fluid. Hence, the inertia term in the equations of motion has been neglected. This is similar to the behavior of a colloidal particle in solution. The interaction with the solvent is taken into account by the stochastic force $\boldsymbol{\Gamma}(s,t)$ and a friction force with the friction constant per length γ. The two forces are related by the dissipation-fluctuation theorem (Eq. 39). Due to the separation of time scales between the fluid and the polymer, $\boldsymbol{\Gamma}$ can be approximated by a stationary, Markovian, and Gaussian process (white noise). The correlations are then given by

$$\langle\boldsymbol{\Gamma}(s,t)\rangle = 0\,, \tag{38}$$

$$\langle\Gamma_\alpha(s,t)\Gamma_{\alpha'}(s',t')\rangle = 2\gamma k_B T\delta_{\alpha\alpha'}\delta(s-s')\delta(t-t')\,. \tag{39}$$

Equation (36) is a fourth-order, linear partial differential equation which can be solved by means of a normal mode analysis. The corresponding eigenfunctions are solutions of the eigenvalue equation

$$\epsilon_p k_B T\frac{d^4}{ds^4}\psi_n(s) - 2\nu k_B T\frac{d^2}{ds^2}\psi_n(s) = \xi_n\psi_n(s)\,, \tag{40}$$

and are given by

$$\psi_0 = \sqrt{\frac{1}{L}}\,, \tag{41}$$

$$\psi_n(s) = \sqrt{\frac{c_n}{L}}\left(\zeta_n\frac{\sin\zeta_n s}{\cos\zeta_n L/2} + \zeta_n'\frac{\sinh\zeta_n' s}{\cosh\zeta_n' L/2}\right)\,, \forall\, n \text{ odd}\,, \tag{42}$$

$$\psi_n(s) = \sqrt{\frac{c_n}{L}}\left(-\zeta_n\frac{\cos\zeta_n s}{\sin\zeta_n L/2} + \zeta_n'\frac{\cosh\zeta_n' s}{\sinh\zeta_n' L/2}\right)\,, \forall\, n \text{ even}\,. \tag{43}$$

The c_ns follow from the normalization condition. The wave numbers ζ_n and ζ_n', where $\zeta_n'^2 - \zeta_n^2 = 4p^2$, are determined by the boundary conditions (36) and (37). ψ_0 describes the translational motion of the whole molecule.

The expansion of the position vector and the stochastic force in terms of the eigenfunctions $\psi_n(s)$,

$$\boldsymbol{r}(s,t) = \sum_{n=0}^{\infty}\boldsymbol{\chi}_n(t)\psi_n(s)\,, \qquad \boldsymbol{\Gamma}(s,t) = \sum_{n=0}^{\infty}\boldsymbol{\Gamma}_n(t)\psi_n(s)\,, \tag{44}$$

yields the following equation of motion for the amplitudes $\boldsymbol{\chi}_n(t)$:

$$\gamma\frac{\partial}{\partial t}\boldsymbol{\chi}_n(t) + \frac{\gamma}{\tau_n}\boldsymbol{\chi}_n(t) = \boldsymbol{\Gamma}_n(t)\,, \tag{45}$$

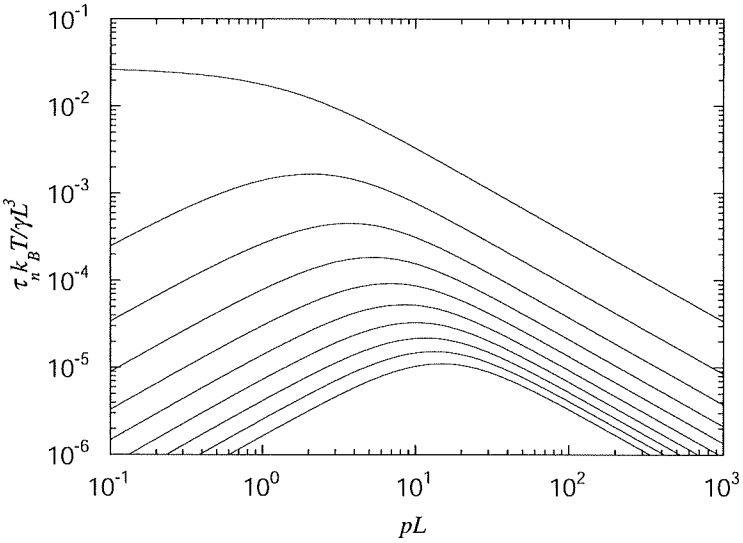


Fig. 13: *Relaxation times τ_n ($n = 1-10$, top to bottom) in the free-draining limit as function of pL. For $pL \gg \zeta_n L/2$, the relaxation times are proportional to L^2 and vary with n^{-2}, whereas in the limit $pL \ll \zeta_n L/2$ and $n > 1$ the τ_ns are proportional to L^4 and exhibit a dependence $(2n-1)^{-4}$ on the mode number. The rotational relaxation time τ_1 is proportional to L^3 in the rod limit.*

with $\gamma/\tau_n = \xi_n$. The stationary state solution of this equation is

$$\boldsymbol{\chi}_n(t) = \frac{1}{\gamma} \int_{-\infty}^{t} e^{-(t-t')/\tau_n} \, \boldsymbol{\Gamma}_n(t') \, dt' \, . \tag{46}$$

The relaxation times are related to the wave numbers ζ_n according to

$$\tau_n = \frac{4\gamma p}{3k_B T(\zeta^4 + 4\zeta^2 p^2)} \, , n > 0 \, , \quad \frac{1}{\tau_0} = 0 \, . \tag{47}$$

The time correlation of appropriate polymer properties is often required when the above results are applied to experimental measurements. Many of these correlation functions can simply be obtained from the correlation functions of the amplitudes $\boldsymbol{\chi}_n$. A straightforward calculation yields

$$\langle \boldsymbol{\chi}_n(t) \boldsymbol{\chi}_m(0) \rangle = \frac{k_B T}{\pi\eta} \tau_n \delta_{nm} \, e^{-t/\tau_n} \, , \quad \forall \, n, \, m \neq 0 \, . \tag{48}$$

As is evident for a linear differential equation, the correlation functions decay exponentially with the characteristic time scale τ_n.
Figure 13 displays the first ten normalized relaxation times $\tau_n k_B T/(\gamma L^3)$. Three regimes can be distinguished for $n > 1$. In the limit of $pL \gg \zeta_n L/2$

$$\tau_n = \frac{\gamma L^2}{3pn^2\pi^2 k_B T} \tag{49}$$

(stretching modes), i.e., the relaxation times are proportional to L^2 and vary like n^{-2}, representing the well-known Rouse relaxation times [69]. The dynamics of this model is discussed in Ref. [70]. For $pL \ll \zeta_n L/2$

$$\tau_n = \frac{64\gamma p L^4}{3\pi^4(2n-1)^4 k_B T} \tag{50}$$

(bending modes), hence, the relaxation times are proportional to L^4 and exhibit the dependence $(2n-1)^{-4}$ on the mode number, making the set of relaxation times much wider separated in this limit. In the intermediate regime both bending and stretching contribute to the relaxation times. As is obvious from Fig. 13, the maximum of $\tau_n(pL)$ shifts continuously to larger pL values with increasing mode number n. This simply reflects the fact that any real polymer looks increasingly rigid on smaller and smaller length scales. Therefore, even for a flexible chain the fast modes, representing the local motion of a chain on short length scale, are rather bending than stretching modes. The first relaxation time τ_1, which is often called rotational relaxation time, exhibits for $pL \gg 1$ the typical Rouse behavior, whereas for $pL \ll 1$ the dependence $\tau_1 = \gamma L^3/(48 k_B T)$ is found [71].

The mean square displacement $\Phi_s(t) := \left\langle (\boldsymbol{r}(s,t) - \boldsymbol{r}(s,0))^2 \right\rangle$ of a particle is an important measure to characterize its dynamics. As shown in Appendix A.4, the mean square displacement significantly depends on chain stiffness. For short times $t/\tau_1 \ll 1$, a flexible chain displays a square root dependence[10] $\langle \Phi_s(t) \rangle_s \sim \sqrt{t}$, whereas a semiflexible chain exhibits the dependence $\langle \Phi_s(t) \rangle_s \sim t^{3/4}$ in the center of mass reference frame.

The free-draining limit applies to polymer chains (shorter than the entanglement length) in a melt [70,73]. For, e.g., DNA in solution, however, hydrodynamic interactions have to be taken into account.

7.2 Hydrodynamic Interaction

To describe the dynamics of macromolecules in a dilute solution the interaction mediated by the motion of the fluid, which is called the hydrodynamic interaction, has to be taken into account. The motion of the solvent fluid around a part of the macromolecule is caused by the moving macromolecule itself and affects the velocities of its other parts. This hydrodynamic interaction can be taken into account in the equations of motion by the well known Oseen tensor [69,75]. However, this tensor has been shown to lead to unphysical behavior of transport properties, e.g., negative translational diffusion coefficients. Rotne and Prager [74] have studied finite beads models to derive a correction to the Oseen tensor which overcomes its deficiencies [75]. The Rotne-Prager tensor reads

$$\begin{aligned} \mathbf{Q}(\boldsymbol{r}(s),\boldsymbol{r}(s')) \;=\; & \frac{\Theta(|\Delta\boldsymbol{r}(s,s')| - d)}{8\pi\eta\Delta r^3(s,s')} \left(\mathbf{I}\Delta r^2(s,s') + \Delta\boldsymbol{r}(s,s') : \Delta\boldsymbol{r}(s,s') \right. \\ & \left. + \frac{d^2}{2\Delta r^2(s,s')} \left[\frac{1}{3}\mathbf{I}\Delta r^2(s,s') - \Delta\boldsymbol{r}(s,s') : \Delta\boldsymbol{r}(s,s') \right] \right) , \end{aligned} \tag{51}$$

with $\Delta\boldsymbol{r}(s,s') := \boldsymbol{r}(s,t) - \boldsymbol{r}(s',t)$, $\mathbf{I}$ the unit matrix, and $\Delta\boldsymbol{r}(s,s') : \Delta\boldsymbol{r}(s,s')$ a 3×3 matrix composed of the components of the vectors $\Delta\boldsymbol{r}(s,s')$. In this equation, d designates the Stokes diameter of a bead, for which the friction constant is $\gamma d = 3\pi\eta d$ and η is the solvent viscosity. Due to the Heaviside step function $\Theta(|\Delta\boldsymbol{r}(s,s')| - d)$ only physically obtainable

[10] To obtain a s independent mean square displacement, the average over the whole chain is considered.

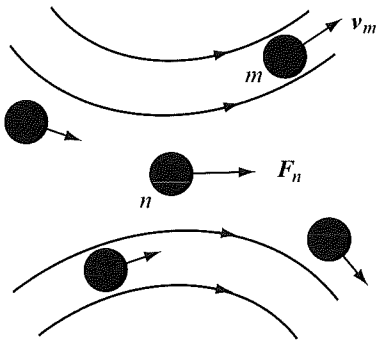


Fig. 14: *The hydrodynamic interaction. The force acting on the particle n creates a velocity field and causes the motion of the other particles [69].*

configurations of the chain, for which $\Delta r(s,s') = |\Delta \boldsymbol{r}(s,s')| > d$ are considered. Moreover, the self-interaction is excluded, since the friction of a bead is taken into account separately. Translated to our continuous polymer model, d is the thickness of the molecule. With the local friction, the total hydrodynamic tensor is given by [69]

$$\mathbf{H}(\boldsymbol{r}(s),\boldsymbol{r}(s')) = \frac{\mathbf{I}}{3\pi\eta}\delta(s-s') + \mathbf{Q}(\boldsymbol{r}(s),\boldsymbol{r}(s'))\,. \tag{52}$$

Because $\mathbf{H}(\boldsymbol{r}(s),\boldsymbol{r}(s'))$ is a nonlinear function of $\boldsymbol{r}(s)$, the Langevin equation is quit difficult to handle. To simplify the analysis, Zimm [69] introduced the preaveraging approximation, which replaces $\mathbf{H}(\boldsymbol{r}(s),\boldsymbol{r}(s'))$ by its average,

$$\mathbf{H}(\boldsymbol{r}(s),\boldsymbol{r}(s')) \to \mathbf{h}(s,s') = \langle \mathbf{H}(\boldsymbol{r}(s),\boldsymbol{r}(s'))\rangle = \int \mathbf{H}(\boldsymbol{r}(s),\boldsymbol{r}(s'))\,\Psi(\{\boldsymbol{r}(s)\})\,\mathcal{D}^3x\,. \tag{53}$$

Since we consider systems at equilibrium and $\mathbf{H}$ depends on two points only, $\Psi(\{\boldsymbol{r}(s)\})$ is equal to the joint probability distribution function $\Psi(\boldsymbol{r}(s)-\boldsymbol{r}(s'))$ (32). Using this distribution function, a straightforward calculation yields $\mathbf{h}(s,s') = \mathbf{I}\,[\delta(s-s')/3\pi\eta + Q(s-s')]$ and

$$Q(s-s') = \frac{\Theta(|s-s'|-d)}{3\pi\eta}\sqrt{\frac{3}{2\pi\sigma(s-s')}}\exp\left(-\frac{3d^2}{2\sigma(s-s')}\right)\,, \tag{54}$$

where the Θ function is introduced to exclude self-interaction. In general, the interactions among the points of the chain are nonlocal since $Q(s-s')$ decreases quit slowly as a function of $(s-s')$ compared to $\delta(s-s')$.
The equation of motion of the polymer chain including hydrodynamic interactions is then given by the Langevin equation [72]

$$\frac{\partial}{\partial t}\boldsymbol{r}(s,t) = \int_{-L/2}^{L/2} \mathbf{h}(s,s')\left(2\nu k_BT\frac{\partial^2}{\partial s'^2}\boldsymbol{r}(s',t) - \epsilon_p k_BT\frac{\partial^4}{\partial s'^4}\boldsymbol{r}(s',t) + \boldsymbol{\Gamma}(s',t)\right)ds'\,. \tag{55}$$

The equation of motion in the free-draining limit is obtained by neglecting the contribution $Q(s-s')$ of the hydrodynamic interaction. To solve equation (55), we insert the eigenfunction

expansion (44), which yields

$$\frac{\partial}{\partial t}\chi_n(t) = \sum_{m=0}^{\infty} \left(Q_{nm} + \delta_{nm}/(3\pi\eta)\right) \left(-\frac{3\pi\eta}{\tau_m}\chi_m(t) + \Gamma_m(t)\right) . \tag{56}$$

The matrix elements Q_{nm} are given by $Q_{nm} = \int_{-L/2}^{L/2}\int_{-L/2}^{L/2} \psi_n(s)Q(s-s')\psi_m(s')\,ds\,ds'$ with the eigenfunctions (41)-(43). A numerical calculation shows that the interaction matrix Q_{nm} is almost diagonal over the whole range of the flexibility parameter pL. Hence, except for the first mode ($n = 1$) in the limit of large stiffness, the interaction matrix is well approximated by

$$Q_{nm} \approx \sqrt{\frac{2}{3\pi^3}}\frac{\delta_{nm}}{\eta L}\int_d^L \frac{(L-s)}{\sqrt{\sigma(s)}} \exp\left(-\frac{3d^2}{2\sigma(s)}\right)\cos\zeta_n s\; ds\,. \tag{57}$$

Hence, the amplitudes $\chi_n(t)$ (in the stationary state) read

$$\chi_n(t) = \frac{\tau_n}{3\pi\eta\tilde{\tau}_n}\int_{-\infty}^{t} e^{-(t-t')/\tilde{\tau}_n}\Gamma_n(t')\,dt'\,, \tag{58}$$

with the relaxation times (cf. Eq. (47))

$$\tilde{\tau}_n = \frac{\tau_n}{1+3\pi\eta Q_{nn}}\,, n > 0\,. \tag{59}$$

Similar to Eq. (48), the time correlation functions are obtained as

$$\langle\chi_n(t)\chi_m(0)\rangle = \frac{k_BT}{\pi\eta}\tau_n\delta_{nm}\,e^{-t/\tilde{\tau}_n}\,, \quad \forall\, n,\, m \neq 0\,. \tag{60}$$

The correlation functions now decay exponentially with the characteristic time scale $\tilde{\tau}_n$.
The hydrodynamic interaction strongly affects the relaxation times, in particular those of flexible polymers. Evaluation of Q_{nn} yields in the appropriate limits

$$\tilde{\tau}_n = \frac{\eta}{\sqrt{3\pi}k_BT}\left(\frac{L}{pn}\right)^{3/2}, \text{ for } \zeta_nL \ll 2pL\,, \tag{61}$$

$$\tilde{\tau}_n \approx \frac{64\eta p}{\sqrt{3\pi^5}k_BT\ln(\Lambda L/d)}\left(\frac{L}{2n-1}\right)^4, \text{ for } L/d > \zeta_nL \gg 2pL\,. \tag{62}$$

The dimensionless quantity $\Lambda L/d$ is a measure of the hydrodynamic aspect ratio of a macromolecule. Comparing Eq. (61) with the corresponding relaxation times of a free-draining system, we find a significantly different dependence on chain length and mode number. This dependence has indeed be conformed experimentally in polymer solutions [69]. With increasing chain stiffness, the hydrodynamic interaction becomes less and less important for the relaxation times. Similar, since ζ_n is an increasing function of mode number n, relaxation times with large mode number are less influenced by hydrodynamic interaction than those with small mode numbers.
The modified relaxation times (61) for flexible polymers naturally affect their mean square displacement. Instead of a square root dependence (cf. Eq. (103)) we now find $\langle\Phi_s(t)\rangle_s \sim t^{2/3}$. The time dependence for semiflexible polymers, however, is essentially unchanged, since the influence of the hydrodynamic interaction on chain dynamics is much smaller for such chains.

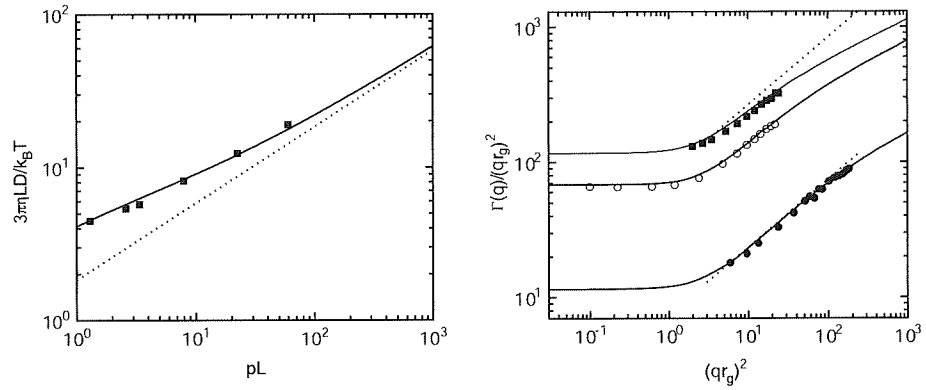


Fig. 15: *Diffusion constants for various DNA fragments from dynamic light scattering (DLS) measurements (squares) [76–78] (left). The solid line follows from Eq. (63). The dotted line corresponds to the Zimm prediction with the radius of gyration of a wormlike chain as input. Right: Decay rates calculated via Eq. (65) together with experimental results for DNA molecules: pDS1 DNA (filled squares) ($pL = 10$) [80], ColE1 DNA (open circles) ($pL = 22.4$) [77], $\phi 29$ DNA (filled circles) ($pL = 60.1$) [78]. The dotted lines represent the relation $\Gamma(\boldsymbol{q}) \sim q^3$.*

The center of mass diffusion constant D follows from the time correlation function for the translational mode $n = 0$ as

$$D = \lim_{t\to\infty} \frac{1}{6Lt} \left\langle (\boldsymbol{\chi}_0(t) - \boldsymbol{\chi}_0(0))^2 \right\rangle = \frac{k_B T}{3\pi\eta L} \left(1 + \frac{\sqrt{6}}{\sqrt{\pi} L} \int_d^L \frac{L-s}{\sqrt{\sigma(s)}} \exp\left(-\frac{3d^2}{2\sigma(s)} \right) ds \right) . \quad (63)$$

Figure 15 presents measured diffusion constants of various DNA molecules [76–78] together with the predictions of Eq. (63) [79]. We like to emphasize that there is no adjustable parameter in the theoretical model. All parameters are taken from experiments. The close agreement between theory and experiment suggests that chain stiffness has even to be taken into account for large scale properties like the translational diffusion coefficient in the case of semiflexible macromolecules in solution.

Insight into the dynamical behavior of macromolecules can be gained by scattering experiments. Since DNA molecules are rather large, their dynamics is accessible by dynamic light scattering (DLS). DLS measures the intramolecular dynamic structure factor (cf. [81]) which is defined as

$$S(\boldsymbol{q}, t) = \frac{1}{L^2} \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \left\langle \exp\left(-i\boldsymbol{q} \left[\boldsymbol{r}(s,t) - \boldsymbol{r}(s',t) \right] \right) \right\rangle \, ds \, ds' \quad (64)$$

for the above continuous chain model, where $\boldsymbol{q}$ is the scattering vector. Calculation of the average exploiting the correlations (60) yields

$$\begin{aligned} S(\boldsymbol{q}, t) \;=\; & \frac{e^{-q^2 D t}}{L^2} \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \exp\left(-\frac{q^2}{6} \sigma(s - s') \right) \\ & \times \exp\left(-\frac{q^2 k_B T}{3\pi\eta} \sum_{n=1}^{\infty} \tau_n \psi_n(s) \psi_n(s') \left(1 - \exp\left(-\frac{t}{\tilde{\tau}_n} \right) \right) \right) ds \, ds' . \end{aligned} \quad (65)$$

For very small scattering vectors ($qr_g \ll 1$, large length scales) is the decay rate of the dynamic intramolecular structure factor much larger than $1/\tilde{\tau}_1$. Thus, the diffusion of the whole molecule is measured: $S(\boldsymbol{q},t) = \exp\left(-Dq^2t\right)$[11]. For scattering vectors $qr_g > 1$, internal modes become essential. The following approximate expression [72] for S is obtained in this scattering vector regime, when the fact that the integrand of Eq. (60) has a sharp peak at $s = s'$ is taken into account

$$S(\boldsymbol{q},t) = \frac{2}{L}\int_0^L \exp\left(-\frac{q^2}{6}\sigma(s) - \frac{q^2 k_B T}{3\pi\eta L}\sum_{n=1}^{\infty}\tau_n \cos\zeta_n s\left(1-\exp\left(-\frac{t}{\tilde{\tau}_n}\right)\right)\right) ds\,. \quad (66)$$

The contribution of the translational motion can be neglected, because q^2Dt is very small on a time scale comparable with the decay rate of the dynamic structure factor at large scattering vectors. In order to study the individual contributions of bending and stretching modes, we consider the limiting cases in which all internal modes are either stretching or bending modes. Using the expression (61) for the relaxation times in the limit of flexible chains and replacing the sum over n in Eq. (66) by an integral, the dynamic structure reads

$$S(\boldsymbol{q},t) \approx S(\boldsymbol{q},0)\exp\left(-1.35\left(\Gamma(q)t\right)^{2/3}\right)\,, \quad \Gamma(q) = \frac{k_B T q^3}{6\pi\eta}\,, \quad (67)$$

for $k_B T q^3/(6\pi\eta) \gg 1$. The decay rate Γ is proportional to q^3 and independent of chain length and persistence length. Therefore, the scattering vector regime in with relation (67) applies is called universal regime.

Assuming that all relaxation times with mode number $n > 1$ are bending modes (62) an analogous calculation yields in the long-time limit

$$S(\boldsymbol{q},t) \approx S(\boldsymbol{q},0)\exp\left(-0.41\left(\Gamma(q)t\right)^{3/4}\right)\,, \quad \Gamma(q) = \frac{\sqrt{6}q^{8/3}p^{1/3}k_B T\ln(\Lambda L/d)}{3\pi^{3/2}\eta}\,. \quad (68)$$

In this limit, the decay rate Γ is proportional to $q^{8/3}$ and depends on the persistence length and the hydrodynamic ration $\Lambda L/d$.

In general, stretching as well as bending modes will contribute to the decay of the dynamic structure factor. From the analytical expressions given above, it is obvious that chain stiffness influences the functional shape $S(\boldsymbol{q},t)$. Generally, the dynamic structure factor cannot be represented by a single exponential function. In some limiting cases a stretched exponential can be used approximately.

Experimentally the initial decay rate $\Omega(\boldsymbol{q}) = -\lim_{t\to 0} d\ln(S(\boldsymbol{r},t))/dt$ is often used to characterize the decay of the dynamic structure factor. Alternatively, a decay rate $\Gamma(\boldsymbol{q})$ can be obtained from the condition

$$S(\boldsymbol{q}, t = \Gamma^{-1}(\mathrm{q}))/S(\boldsymbol{q},0) = e^{-1}\,. \quad (69)$$

As shown in Ref. [79], the dynamic structure factor is a rather universal function for $t\Gamma(\boldsymbol{q})$. The decay rates of various DNA molecules in solution are presented in Fig. 15 together with the decay rate calculated from the dynamic structure factor (65). In addition, the dotted line corresponds to the relation $\Gamma(\boldsymbol{q}) \sim q^3$. The figure exhibits excellent agreement between the theoretical results and the experimental measurements. Three scattering vector regimes can

[11] $S(\boldsymbol{q},0) \approx 1$ in this limit.

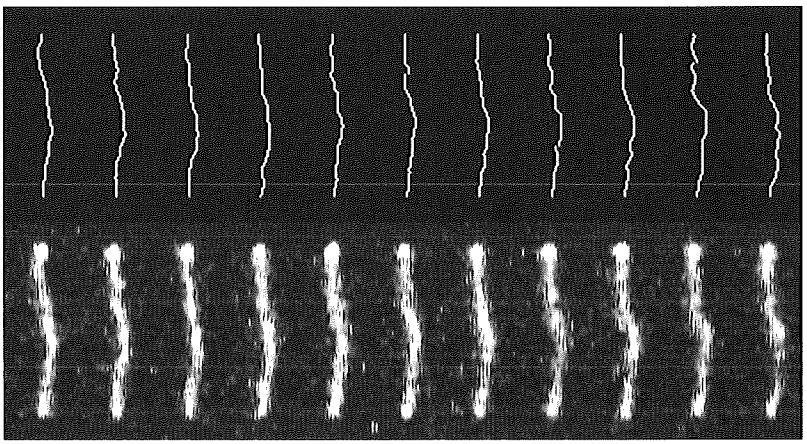


Fig. 16: *Video images of single molecules of DNA held at a fixed extension of* 12 μm ($L = 20$ μm). *The bottom column shows a series of successive raw video images, whereas the column on top shows the corresponding parameterizations used in the data analysis [82].*

be distinguished. In the small scattering vector regime $qr_g < 1$, the translational motion of the whole molecule dominates. The decay rate is given by $\Gamma = q^2 D$, as is evident from the result for COlEl DNA. In the intermediate scattering vector regime $1/r_g < q < p$, the decay rate is approximately proportional to q^3 for ColEl DNA and $\phi 29$ DNA. For pDS1 DNA, there exists no intermediate scattering vector regime, because the gap between the small scattering vector regime, with dominating translational motion, and the large scattering vector regime, $q > p$, in which molecular stiffness is important, is small in comparison with more flexible macromolecules. The decay rate is rather proportional to $q^{8/3}$ for $q > p$ of the pDS1 DNA molecule. The crossover from the intermediate regime to the large scattering vector regime is visible for $\phi 29$ DNA.

There is a broad spectrum of other quantities which could be discussed in the context of semi-flexible chains. Some more results are presented in Refs. [71–73, 79].

7.3 Dynamics of Partially Stretched DNA Molecules

To study the dynamics of polymers in solution, we linearized the equations of motion by preaveraging the hydrodynamic tensor. This allows us to solve the equations of motion in terms of a normal mode analysis. The normal mode approach has been tested experimentally by measuring the fluctuations of single partially stretched DNA molecules [82]. For this purpose, λ-Phage DNA was attached to streptavidin-coated beads and stained by a dye. The two beads were fixed by optical tweezers (cf. Fig. 11). Fluorescence images were recorded with an image-intensified CCD camera. An example of a sequence of such images is displayed in Fig. 16. The parameterized data were used in a normal mode analysis. The authors of Ref. [82] find that the motion of DNA can be described by linearly independent normal modes. However, the observed dependence of the measured relaxation times on the mode number for various extension could not be explained by the Zimm model.

A qualitative and quantitative understanding of the experimental data requires the incorporation

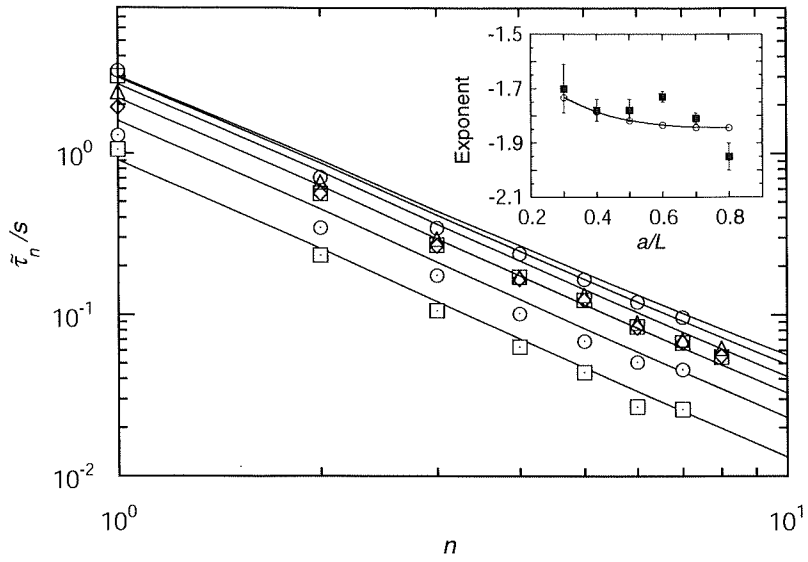


Fig. 17: *Relaxation times of stretched DNA molecules for the extensions* $a/L = 0.3$, 0.4, 0.5, 0.6, 0.7, *and* 0.8 *(from top to bottom) [83]. The experimental data are taken from Ref. [82]. The insert shows the exponents* ϑ *of the relation* $\tau_n \sim n^{\vartheta}$ *corresponding to the various curves.*

of stretching in the theoretical description [83]. Hence, the preaveraging of the hydrodynamic tensor has to be performed with a distribution function for a stretched chain.

To account for the experimental geometry, we consider a semiflexible polymer with one end fixed at the origin of the reference frame and the other one at a point $\boldsymbol{a} = (0, 0, a)^T$. The calculation of the distribution function $\Psi(\Delta r(s, s'))$ is somewhat lengthy, but can be performed analytically. The final result is a Gaussian with a deformation dependent (a) mean and width. The preaveraged hydrodynamic tensor can also be calculated analytically. The explicit expressions are presented in Ref. [83]. Due to the stretching of the chain, $\mathbf{Q}(s, s')$ becomes asymmetric, with $Q_{xx} = Q_{yy}$ (transverse to the stretching direction).

To solve the equation of motion (55) for fixed ends, we use the eigenfunction expansion $\boldsymbol{r}(s, t) = \sum_{n=1}^{\infty} \boldsymbol{\chi}_n(t)\psi_n(s) + \boldsymbol{a}s/L$. The boundary conditions ($\psi_n(0) = \psi_n(L) = 0$) are satisfied for the eigenfunctions $\psi_n(s) = \sqrt{2/L}\sin(n\pi s/L)$. The equations of motion are then similar to those of Eq. (56), but ν is now deformation dependent and has to be replaced by the value of Eq. (33) [$\nu = 3p/(2\mu^2) = 3p/(2[1 - (a/L)^2]^2)$].

Since the DNA molecule is rather long (20 μm) and hence flexible the relaxation times τ_n are well approximated by the Rouse relaxation times (cf. Sec. 7.1). The relaxation times transverse ($x-$axis) to the stretching direction including hydrodynamic interactions are the given by

$$\tilde{\tau}_n^x = \frac{\tau_R \mu^2}{n^2[1 + 3\pi\eta Q_{nn}^x]}, \tag{70}$$

with the Rouse relaxation time $\tau_R = \eta L^2/(\pi p k_B T)$. The relaxation times are anisotropic as a consequence of stretching.

The numerically calculated relaxation times for the extensions a/L between 0.3 and 0.8 are

shown in Fig. 17 together with the experimental data (symbols). The necessary parameters are taken from experiments [82]: $L = 20$ μm, $d = 2.5$ nm, $1/2p = 50$ nm, $\eta = 0.2$ P, and $T = 300$ K. The parameters imply $pL = 200$, i.e., the DNA molecules are rather flexible. The Rouse relaxation time is then $\tau_R = 62$ s. The agreement between the experimentally measured relaxation times and the theoretical values is excellent. The theoretical data cover the same $\tilde{\tau}_n$ regime as the experimental values for $n = 1$ as well as for $n = 8$. The theoretical data follow rather precisely a straight line. Thus, the n dependence of the relaxation times is well described by a power law for small n values [82]. The insert shows the exponents (ϑ) of the straight lines $\tilde{\tau}_n \sim n^{\vartheta}$ fitted to the theoretical lines (solid line) and the experimental data (symbols). The exponents agree with each other within a few percent error.

To achieve quantitative agreement between the theoretical and experimental data, it is necessary to describe a DNA molecule by a semiflexible chain model. This is reflected in the factor μ^2 in Eq. (70). For a Gaussian chain, only the factor μ would appear. The factor μ^2 is however necessary to quantitatively cover the experimentally observed relaxation time range. This conclusion is consisted with the discussion of the force-extension curve of Sec. 6.3.

Within certain assumptions, an analytical approximation for the relaxation times is obtained. In particular, we find for $a/L > 0.5$

$$\tilde{\tau}_n^x = \frac{\tau_R \mu^2}{n^2} \left[1 + \frac{3L}{4a} \left(\ln \frac{L}{n\pi d} + \frac{1}{n} - C \right) \right]^{-1} , \tag{71}$$

where $C = 0.577215\ldots$ is Euler's constant [83]. The relaxation times exhibit in addition to the Rouse factor n^{-2} a logarithmic dependence on the mode number. The logarithmic term cannot be ignored. For moderate chain lengths, its contribution to the mode dependence of the relaxation times is on the order of the excluded volume interaction for slight deformed chains.

8 Conclusions

A understanding of the structural, dynamical, and functional aspects of biological systems will only be possible by an interdisciplinary effort. Biologists, physicists, chemists, and scientist form other fields will have to bring in their knowledge. The current contribution shows that many existing theoretical models can be applied to biological systems and help to unravel their characteristics. On the other hand, biology provides samples which allow us to test theoretical models by direct visualization.

Various aspects of polyelectrolytes in solution and semiflexible chain have been discussed in this lecture. A mean field approach yields an effective repulsive interaction potential (Debye-Hückel potential) among equally charged particles. Beyond a certain interaction strength, however, this approach breaks down on small length scales next to a charge and counterions have to be taken into account explicitly. The condensation of the counterions leads to correlations among charges and ultimately gives rise to an attractive interaction among equally charged objects. Such effects are responsible for DNA compaction. More sophisticated structures appear for complex cationic structures. Examples are nucleosomes and chromatin as well as DNA–cationic dendrimer complexes which may serve as vectors in gene therapy.

For certain conformational and dynamical properties of biopolymers, counterions can be taken into account by an effective persistence length resembling a semiflexible polymer. DNA at physiological conditions is a particular example. There is a broad spectrum of other semiflexible polymers in biological systems. Among them are the constituents of the cytoskeleton, in

particular actin filaments. Stretching experiments of DNA molecules clearly reveal the semiflexible character of these molecules. Similar conclusions can be drawn from the analysis of DNA dynamics. Short unconstraint DNAs follow the predictions for the dynamics of semiflexible polymers. Long DNAs display the dynamics of flexible polymers on sufficiently large length scales. Measurements of the relaxation dynamics of single stretched DNA molecules support the usefulness of the normal mode concept for the analysis of chain dynamics. Moreover, a quantitative agreement between experimental measurements and theoretical data once more requires a description in terms of a semiflexible chain model.
There are many more aspects of the dynamics of biopolymers which have not been addressed in this lecture. In particular the dynamics of DNA in flow fields has received considerable attention in the last decade. Flow properties are important for the transport of biopolymers in blood flow or through the cell cytoskeleton. The interested reader will find further aspects in references [65, 84–86] on experimental facts and in Refs. [87, 88] on theoretical investigations.

In summary, the investigation of the structural and dynamical properties of polyelectrolytes and semiflexible polymers is a challenging and exciting task. The understanding of various biological functions relays on our knowledge of such systems.

A Appendices

A.1 Solution of the Linearized Poisson-Boltzmann Equation

The linearized Poisson-Boltzmann equation for spherical particles of diameter l is given by

$$\Delta\bar{\phi}_\alpha(\boldsymbol{r}) = \begin{cases} \kappa^2\bar{\phi}_\alpha(\boldsymbol{r})\,, & r > l \\ 0\,, & 0 < r < l \end{cases}\,, \tag{72}$$

where $r = |\boldsymbol{r}|$. The boundary conditions are that $\bar{\phi}_\alpha(r)$ must vanish at infinity, that $\bar{\phi}_\alpha(r)$ be continuous across the surface specified by $r = l$, and that ϵ times the normal derivative of $\bar{\phi}_\alpha(r)$ be continuous across the surface $r = l$. The last condition comes from the fact that the normal component of the displacement vector $\boldsymbol{D}$ is continuous across any charge free surface. Fixing particle α at the origin of the reference system and introducing spherical coordinates gives

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\bar{\phi}_\alpha(r)}{dr}\right) = \begin{cases} \kappa^2\bar{\phi}_\alpha(r)\,, & r > l \\ 0\,, & 0 < r < l \end{cases}\,. \tag{73}$$

To solve the equation for $r > l$, the function $\chi(r) = r\bar{\phi}_\alpha(r)$ is introduced, which yields the linear differential equation

$$\frac{d^2\chi}{dr^2} - \kappa^2\chi = 0\,, \quad r > l\,.$$

The solution of this equation is easily found. Applying the boundary condition $\bar{\phi}_\alpha(r \to \infty) = 0$, the solution for $\bar{\phi}_\alpha$ is

$$\bar{\phi}_\alpha(r) = \frac{A}{r}e^{-\kappa r}\,, \quad r > l\,. \tag{74}$$

The solution for $0 < r < l$ follows directly from Eq. (73) and is given by

$$\bar{\phi}_\alpha(r) = \frac{B}{r} + C\,, \quad 0 < r < l\,. \tag{75}$$

To evaluate the constant B Gauss' law is applied. Since the charge density within the sphere with radius $r < l$ is $\rho = z_\alpha e\delta(\boldsymbol{r})$, integration of the Maxwell equation $\nabla\boldsymbol{D} = 4\pi\rho$ yields $B = z_\alpha e/\epsilon$ and hence

$$\bar{\phi}_\alpha(r) = \frac{z_\alpha e}{\epsilon r} + C\,, \quad 0 < r < l\,. \tag{76}$$

The conditions that $\bar{\phi}_\alpha(r)$ and $\epsilon d\bar{\phi}_\alpha(r)/dr$ have to be continuous at $r = l$ determine the remaining constants A and C:

$$\begin{aligned} A &= \frac{z_\alpha e e^{\kappa l}}{\epsilon(1+\kappa l)}\,, \\ C &= -\frac{z_\alpha e\kappa}{\epsilon(1+\kappa l)}\,. \end{aligned}$$

Thus, the total solution is

$$\bar{\phi}_\alpha(r) = \begin{cases} \dfrac{z_\alpha e e^{-\kappa(r-l)}}{\epsilon r(1+\kappa l)}\,, & r > l \\ \dfrac{z_\alpha e}{\epsilon r} - \dfrac{z_\alpha e\kappa}{\epsilon(1+\kappa l)}\,, & 0 < r < l \end{cases}\,. \tag{77}$$

(78)

A.2 Electrostatic Persistence Length in the Rod Limit

The intramolecular Coulomb interaction of the continuous charge distribution along an infinitely long polymer in the presence of counterions is given by the Debye-Hückel potential

$$\beta V_{DH} = \frac{l_B}{l^2} \int_0^\infty \int_{s'}^\infty \frac{e^{-\kappa|\boldsymbol{r}(s)-\boldsymbol{r}(s')|}}{|\boldsymbol{r}(s)-\boldsymbol{r}(s')|}\, ds\, ds' \,. \tag{79}$$

No rearrangement or fluctuations of the charges are allowed. With the tangent vector, the difference vector between two points along the chain contour is expressed by

$$\boldsymbol{r}(s) - \boldsymbol{r}(s') = \int_{s'}^{s} \boldsymbol{u}(s'')\, ds'' \,. \tag{80}$$

If a weakly bending rod is assumed, $\boldsymbol{u}$ is well approximated by the following Taylor expansion

$$\boldsymbol{u}(s'') \approx \boldsymbol{u}(s') + \frac{\partial \boldsymbol{u}}{\partial s'}(s''-s') + \frac{1}{2}\frac{\partial^2 \boldsymbol{u}}{\partial s'^2}(s''-s')^2 \,. \tag{81}$$

Substitution of Eq. (81) into Eq. (80) yields

$$\boldsymbol{r}(s) - \boldsymbol{r}(s') \approx \boldsymbol{u}(s')(s-s') + \frac{1}{2}\frac{\partial \boldsymbol{u}}{\partial s'}(s-s')^2 + \frac{1}{6}\frac{\partial^2 \boldsymbol{u}}{\partial s'^2}(s-s')^3 \,, \tag{82}$$

and

$$\frac{e^{-\kappa|\boldsymbol{r}(s)-\boldsymbol{r}(s')|}}{|\boldsymbol{r}(s)-\boldsymbol{r}(s')|} \approx \frac{e^{-\kappa|s-s'|}}{|s-s'|} + \frac{1}{24}\left(\frac{\partial \boldsymbol{u}}{\partial s'}\right)^2 e^{-\kappa|s-s'|}\left[|s-s'| + \kappa(s-s')^2\right] \,. \tag{83}$$

The constraint of a unit tangent vector implies $\boldsymbol{u}\partial\boldsymbol{u}/\partial s = 0$ and $(\partial \boldsymbol{u}/\partial s)^2 = -\boldsymbol{u}\partial^2\boldsymbol{u}/\partial s^2$, which has been taken into account to derive Eq. (83). The first term on the right-hand side of Eq. (83) accounts for the Debye-Hückel energy of a straight rod. It is independent of any particular conformation of the chain. Hence, the term yields a constant contribution to the partition function and vanishes in any calculation of thermal averages. Thus, we will neglect this term in the following. Substitution of the rest of Eq. (83) into Eq. (79) and evaluation of one of the integrals yields

$$\beta V_{DH} = \frac{l_B}{8\kappa^2 l^2} \int_0^\infty \left(\frac{\partial \boldsymbol{u}}{\partial s}\right)^2 ds \,. \tag{84}$$

Collecting the various energy terms we find

$$\begin{aligned} &\frac{l_0}{2}\int_0^\infty \left(\frac{\partial \boldsymbol{u}}{\partial s}\right)^2 ds + \frac{l_B}{l^2}\int_0^\infty \int_{s'}^\infty \left(\frac{e^{-\kappa|\boldsymbol{r}(s)-\boldsymbol{r}(s')|}}{|\boldsymbol{r}(s)-\boldsymbol{r}(s')|} - \frac{e^{-\kappa|s-s'|}}{|s-s'|}\right) ds\, ds' \\ &\quad \approx \frac{l_0}{2}\int_0^\infty \left(\frac{\partial \boldsymbol{u}}{\partial s}\right)^2 ds + \frac{l_B}{8\kappa^2 l^2}\int_0^\infty \left(\frac{\partial \boldsymbol{u}}{\partial s}\right)^2 ds = \frac{l_p}{2}\int_0^\infty \left(\frac{\partial \boldsymbol{u}}{\partial s}\right)^2 ds \,. \end{aligned} \tag{85}$$

The (screened) Coulomb interaction yields a contribution to the bending stiffness of the rodlike molecule. The effective persistence length l_p is given by [59,60]

$$l_p = l_0 + \frac{1}{4}\frac{l_B}{\kappa^2 l^2} \,. \tag{86}$$

A.3 Equation of Motion for Semiflexible Polymers

The kinetic energy of a linear sequence of mass points ($i = 0 \ldots N$) transforms in the continuum limit into

$$T_{kin} = \frac{1}{2}\sum_{i=0}^{N} m\left(\frac{d\boldsymbol{r}_i}{dt}\right)^2 \longrightarrow \frac{1}{2}\int_{-L/2}^{L/2} \varrho\left(\frac{\partial \boldsymbol{r}(s,t)}{\partial t}\right)^2 ds\,, \tag{87}$$

where $\varrho = \lim_{l\to 0} m/l$ is the linear mass density. Identifying the exponent of Eq. (25) as potential energy V_s, the following Langevin function ($\mathcal{L} = T_{kin} - V_s$) is obtained

$$\mathcal{L}[\boldsymbol{r}(s),\dot{\boldsymbol{r}}(s)] = \frac{\varrho}{2}\int_{-L/2}^{L/2}\left(\frac{\partial \boldsymbol{r}(s,t)}{\partial t}\right)^2 ds - k_BT\nu\int_{-L/2}^{L/2}\left(\frac{\partial \boldsymbol{r}(s,t)}{\partial s}\right)^2 ds \tag{88}$$

$$-\frac{\epsilon_p k_BT}{2}\int_{-L/2}^{L/2}\left(\frac{\partial^2 \boldsymbol{r}(s,t)}{\partial s^2}\right)^2 ds - k_BT\nu_0\left[\left(\frac{\partial \boldsymbol{r}(s,t)}{\partial s}\right)^2_{s=L/2} + \left(\frac{\partial \boldsymbol{r}(s,t)}{\partial s}\right)^2_{s=-L/2}\right].$$

The equation of motion follows from Lagrange's equation of the second kind

$$\frac{d}{dt}\frac{\delta}{\delta\dot{\boldsymbol{r}}}\mathcal{L}[\boldsymbol{r}(s),\dot{\boldsymbol{r}}(s)] - \frac{\delta}{\delta \boldsymbol{r}}\mathcal{L}[\boldsymbol{r}(s),\dot{\boldsymbol{r}}(s)] = 0\,, \tag{89}$$

where $\delta/\delta\boldsymbol{r}$ denotes the functional derivative and $\dot{\boldsymbol{r}} = \partial \boldsymbol{r}/\partial t$. To calculated the functional derivative, the rule

$$\frac{\delta \boldsymbol{r}(s,t)}{\delta \boldsymbol{r}(s',t)} = \delta(s-s') \tag{90}$$

is required and the fact that derivatives with respect to time and the contour coordinate (s) can be exchanged with the functional derivative. With

$$\frac{d}{dt}\frac{\delta\mathcal{L}}{\delta\dot{\boldsymbol{r}}(s,t)} = \varrho\frac{d}{dt}\int_{-L/2}^{L/2}\dot{\boldsymbol{r}}(s',t)\frac{\delta\dot{\boldsymbol{r}}(s',t)}{\delta\dot{\boldsymbol{r}}(s,t)}\,ds' = \varrho\frac{d}{dt}\int_{-L/2}^{L/2}\dot{\boldsymbol{r}}(s',t)\delta(s'-s)\,ds' = \varrho\frac{\partial^2 \boldsymbol{r}(s,t)}{\partial t^2}\,, \tag{91}$$

$$-\frac{\delta\mathcal{L}}{\delta \boldsymbol{r}(s,t)} = 2k_BT\nu\int_{-L/2}^{L/2}\boldsymbol{r}'(s',t)\frac{\delta \boldsymbol{r}'(s',t)}{\delta \boldsymbol{r}(s,t)}\,ds' + k_BT\epsilon_p\int_{-L/2}^{L/2}\boldsymbol{r}''(s',t)\frac{\delta \boldsymbol{r}''(s',t)}{\delta \boldsymbol{r}(s,t)}\,ds'$$

$$+2k_BT\nu_0\left(\boldsymbol{r}'(s',t)\frac{\delta \boldsymbol{r}'(s',t)}{\delta \boldsymbol{r}(s,t)}\bigg|_{L/2} + \boldsymbol{r}'(s',t)\frac{\delta \boldsymbol{r}'(s',t)}{\delta \boldsymbol{r}(s,t)}\bigg|_{-L/2}\right), \tag{92}$$

here $\boldsymbol{r}'(s',t) = \partial \boldsymbol{r}(s',t)/\partial s'$ and $\boldsymbol{r}''(s',t) = \partial^2 \boldsymbol{r}(s',t)/\partial s'^2$, and

$$\frac{\delta}{\delta \boldsymbol{r}(s,t)}\frac{\partial \boldsymbol{r}(s',t)}{\partial s'} = \frac{\partial\delta(s-s')}{\partial s'} \tag{93}$$

$$\frac{\delta}{\delta \boldsymbol{r}(s,t)}\frac{\partial^2 \boldsymbol{r}(s',t)}{\partial s'^2} = \frac{\partial^2\delta(s-s')}{\partial s'^2} \tag{94}$$

as well as partial integration

$$
\begin{aligned}
-\frac{\delta\mathcal{L}}{\delta \boldsymbol{r}(s,t)} &= -2k_BT\nu \int_{-L/2}^{L/2} \frac{\partial^2 \boldsymbol{r}(s',t)}{\partial s'^2}\delta(s-s')\,ds' + 2k_BT\nu \frac{\partial \boldsymbol{r}(s',t)}{\partial s'}\delta(s-s')\Big|_{-L/2}^{L/2} \\
&\quad + k_BT\epsilon_p \int_{-L/2}^{L/2} \frac{\partial^4 \boldsymbol{r}(s',t)}{\partial s'^4}\delta(s-s')\,ds' \\
&\quad - k_BT\epsilon_p \frac{\partial^3 \boldsymbol{r}(s',t)}{\partial s'^3}\delta(s-s')\Big|_{-L/2}^{L/2} + k_BT\epsilon_p \frac{\partial^2 \boldsymbol{r}(s',t)}{\partial s'^2}\frac{\partial \delta(s-s')}{\partial s'}\Big|_{-L/2}^{L/2} \\
&\quad + 2k_BT\nu_0 \left(\boldsymbol{r}'(s',t)\frac{\delta \boldsymbol{r}'(s',t)}{\delta \boldsymbol{r}(s,t)}\Big|_{L/2} + \boldsymbol{r}'(s',t)\frac{\delta \boldsymbol{r}'(s',t)}{\delta \boldsymbol{r}(s,t)}\Big|_{-L/2} \right) ,
\end{aligned} \tag{95}
$$

the following equation of motion and boundary conditions are obtained:

$$
\varrho\frac{\partial^2 \boldsymbol{r}(s,t)}{\partial t^2} - 2k_BT\nu\frac{\partial^2 \boldsymbol{r}(s,t)}{\partial s^2} + k_BT\epsilon_p\frac{\partial^4 \boldsymbol{r}(s,t)}{\partial s^4} = 0\,, \tag{96}
$$

$$
\left(2\nu\frac{\partial \boldsymbol{r}(s,t)}{\partial s} - \epsilon_p\frac{\partial^3 \boldsymbol{r}(s,t)}{\partial s^3}\right)_{-L/2}^{L/2} = 0\,, \tag{97}
$$

$$
\left(2\nu_0\frac{\partial \boldsymbol{r}(s,t)}{\partial s} \pm \epsilon_p\frac{\partial^2 \boldsymbol{r}(s,t)}{\partial s^2}\right)_{-L/2}^{L/2} = 0\,. \tag{98}
$$

The same set of equations is obtained in the continuum limit of a discrete semiflexible chain model. Adding a stochastic and a friction force, which account for the interaction with a background medium, and neglecting the inertia term ($\sim \varrho$) yields Eq. (35).

A.4 Mean Square Displacement

The mean square displacement $\Phi_s(t)$ of a point $\boldsymbol{r}(s,t)$ is defined as

$$\Phi_s(t) := \left\langle (\boldsymbol{r}(s,t) - \boldsymbol{r}(s,0))^2 \right\rangle \,. \tag{99}$$

Using the eigenfunction expansion (44) and the orthogonality of the eigenfunctions (41) – (43)

$$\Phi_s(t) = \frac{1}{L}\left\langle (\chi_0(t) - \chi_0(0))^2 \right\rangle + \left\langle \left(\sum_{n=1}^{\infty} \chi_n(t)\psi_n(s) - \sum_{n=1}^{\infty} \chi_n(0)\psi_n(s) \right)^2 \right\rangle \,. \tag{100}$$

Inserting the center-of-mass mean square displacement $\left\langle (\chi_0(t) - \chi_0(0))^2 \right\rangle / L$ and the correlation functions (48) yields

$$\Phi_s(t) = 6D_0 t + 6D_0 L \sum_{n=1}^{\infty} \tau_n \psi_n(s)^2 \left(1 - e^{-t/\tau_n}\right) \,, \tag{101}$$

with $D_0 = k_B T/\gamma L$. Averaging over all s yields

$$\langle \Phi_s(t) \rangle_s = \frac{1}{L} \int_{-L/2}^{L/2} \Phi_s(t)\, ds = 6D_0 t + 6D_0 \sum_{n=1}^{\infty} \tau_n \left(1 - e^{-t/\tau_n}\right) \,. \tag{102}$$

An approximate expression for $\langle \Phi_s(t) \rangle_s$ is obtained in the limit $t/\tau_1 \ll 1$, when the sum over n is replaced by an integral and the relaxation times (49) and (50) are used:

$$\langle \Phi_s(t) \rangle_s = \begin{cases} 6D_0 t + \sqrt{\dfrac{k_B T}{3\gamma\pi p}}\sqrt{t}\,, & pL \gg 1 \\ 12D_0 t + 2.5\left(\dfrac{k_B T \sqrt[3]{p}}{\gamma} t\right)^{3/4}\,, & pL \ll 1 \end{cases} \,. \tag{103}$$

In the rod limit $pL \to 0$ the mean square displacement is determined by τ_1 and

$$\langle \Phi_s(t) \rangle_s = 12 D_0 t \,. \tag{104}$$

References

[1] L. Eichinger, B. Köppel, A. Noegle, M. Schleicher, M. Schliwa, K. Weijer, W. Witke, and P. Janmey, Biophys. J. **70**, 1054 (1996)

[2] P. Janmey, *Cell Membranes and the Cytoskeleton*, Volume 1A of *Handbook of Biological Physics*, Chapter 17, 1054 (North Holland, Amsterdam, 1995)

[3] C. Tanford, *Physical Chemistry of Macromolecules* (Wiley, New York, 1961)

[4] M. Mandel, *Polyelectrolytes* (Ridel, Dordrecht, 1988)

[5] M. Hara Ed., *Polyelectrolytes: Science and Technology* (Marcel Dekker, New York, 1993)

[6] S. Förster and M. Schmid, Adv. Polym. Sci. **120**, 51 (1995)

[7] J.-L. Barrat and J.-F. Joanny, Adv. Chem. Phys. **XCIV**, 1 (1996)

[8] *Macro-ion Characterization: From Dilute Solutions to Complex Fluids*, edited by K. Schmitz (American Chemical Soceity, Washington, DC, 1994)

[9] N. V. Hud, K. H. Downing, and R. Balhorn, Proc. Natl. Acad. Sci. U.S.A. **92**, 3581 (1995)

[10] R. Golan, L. I. Pietrasanta, W. Hsieh, and H. G. Hansma, Biochemistry **38**, 14069 (1999)

[11] H. G. Hansma, R. Golan, W. Hsieh, C. P. Lollo, P. Mullen-Ley, and D. Kwoh, Nucleic Acid Res. **26**, 2481 (1998)

[12] D. D. Dunlap, A. Maggi, M. R. Soria, and L. Monaco, Nucleic Acid Res. **25**, 3095 (1997)

[13] W. Chen, N. J. Turro, and D. A. Tomalia, Langmuir **16**, 15 (1999)

[14] I. Gössl, L. Shu, A. D. Schlüter, and J. P. Rabe, J. Am. Chem. Soc. **124**, 6860 (2002)

[15] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology* (Garland Publishing, New York, 1998)

[16] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of The Cell* (Garland Publishing, New York, 1989)

[17] E. Frey, K. Kroy, J. Wilhelm, and E. Sackmann, in: *Dynamical Networks in Physics and Biology*, edited by G. Forgacs and D. Beysens (Springer Verlag, Berlin, 1998)

[18] See: http://www.dartmouth.edu/~langford/Research.html

[19] W. S. Klug and M. R. Cummings, *Concepts of Genetics* (Prentice Hall, Upper Saddle River NJ, 1997)

[20] D. A. McQuarrie, *Statistical Mechanics* (Harper and Row, New York, 1979)

[21] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell Univerity Press, Ithaca, NY, 1979)

[22] Frielinghaus, chapter **C1** of this book

[23] M. Ullner, J. Phys. Chem. B **107**, 8097 (2003)

[24] S. F. Edwards and P. Singh, J. Chem. Soc., Faraday Trans. 2 **75**, 1001 (1979)

[25] K. Gosh, G. A. Carri, and M. Muthukumar, J. Chem. Phys. **115**, 4367 (2001)

[26] T. Hofmann, R. G. Winkler, and P. Reineker, J. Chem. Phys. **118**, 6624 (2003)

[27] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic Press, London, 1986)

[28] K. S. Schweizer and J. G. Curro, Adv. Polym. Sci. **116**, 319 (1995)

[29] C.-Y. Shew and A. Yethiraj, J. Chem. Phys. **106**, 5706 (1997)

[30] T. Hofmann, R. G. Winkler, and P. Reineker, J. Chem. Phys. **114**, 10181 (2001)

[31] T. Hofmann, R. G. Winkler, and P. Reineker, J. Chem. Phys. **119**, 2406 (2003)

[32] T. Onishi, N. Imai, and F. Oosawa, J. Phys. Soc. Japan **15**, 896 (1960)

[33] G. S. Manning, J. Chem. Phys. **51**, 924 (1969)

[34] W. M. Gelbart, R. F. Bruinsma, P. A. Pincus, and V. A. Parsegian, Physics Today **53**, 38 (2000)

[35] M. Stevens and K. Kremer, J. Chem. Phys. **103**, 1669 (1995)

[36] R. G. Winkler, M. Gold, and P. Reineker, Phys. Rev. Lett. **80**, 3731 (1998)

[37] H. Risken, *The Fokker-Planck Equation* (Springer, Berlin, 1984)

[38] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987)

[39] N. Grønbech-Jensen, R. J. Mashl, R. F. Bruinsma, and W. M. Gelbert, Phys. Rev. Lett. **78**, 2477 (1997)

[40] H. Schiessel, Macromolecules **32**, 5673 (1999)

[41] A. V. Lyulin, B. Dünweg, O. V. Borisov, and A. A. Darinskii, Macromolecules **32**, 3264 (1999)

[42] H. J. Limbach, C. Holm, and K. Kremer, Europhys. Lett. **60**, 566 (2002)

[43] H. Schiessel, J. Phys.: Condens. Matter **15**, 699 (2003)

[44] R. S. Dias, A. A. C. C. Pais, M. G. Miguel, and B. Lindman, J. Chem. Phys. **119**, 8150 (2003)

[45] J. O. Rädler, I. Koltover, T. Salditt, and C. R. Safinya, Science **275**, 810 (1997)

[46] T. Wallin T and P. Linse, J. Phys. Chem **100**, 17873 (1996)

[47] C. Y. Kong and M. Muthukumar, J. Chem. Phys. **109**, 1522 (1998)

[48] R. Messina, C. Holm, and K. Kremer, Phys. Rev. E **64**, 021405 (2001)

[49] K.-K. Kunze and R. R. Netz, Phys. Rev. Lett. **85**, 4389 (2000)

[50] D. Srivastava and M. Muthukumar, Macromolecules **27**, 1461 (1994)

[51] J. B. Imbert, J. M. Victor, N. Tsunekawa, and Y. Hiwatari, Phys. Lett. A **258**, 92 (1999)

[52] R. G. Winkler, M. O. Steinhauser, and P. Reineker, Phys. Rev. E **66**, 021802 (2002)

[53] K.-K. Kunze and R. R. Netz, Europhys. Lett. **58**, 299 (2002)

[54] Y. Hayashi, M. Ullner, and P. Linse, J. Phys. Chem. B **107**, 8198 (2003)

[55] R. Messina, J. Chem. Phys. **119**, 8133 (2003)

[56] R. G. Winkler, NJP **6**, 11 (2004)

[57] O. Kratky and G. Porod, Recl. Trav. Chim. Pays-Bas. **68**, 1106 (1949)

[58] R. G. Winkler, P. Reineker, and L. Harnau, J. Chem. Phys. **101**, 8119 (1994)

[59] T. Odijk, J. Polym. Sci., Polym. Phys. Ed. **15**, 477 (1977)

[60] J. Skolnick and M. Fixman, Macromolecules **10**, 944 (1977)

[61] R. G. Winkler and P. Reineker, Macromolecules **25**, 6891 (1992)

[62] R. G. Winkler, J. Chem. Phys. **118**, 2919 (2003)

[63] S. B. Smith, L. Finzi, and C. Bustamante, Science **258**, 1122 (1992)

[64] J. M. Schurr and S. B. Smith, Biopolymers **29**, 1161 (1990)

[65] T. T. Perkins, D. E. Smith, R. G. Larson, and S. Chu, Science **268**, 83 (1995)

[66] M. Fixman and J. Kovac, J. Chem. Phys. **58**, 1564 (1973)

[67] J. F. Marko and E. D Siggia, Macromolecules **28**, 8759 (1995)

[68] J. Wilhelm and E. Frey, Phys. Rev. Lett. **77**, 2581 (1996)

[69] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, Oxford, 1986)

[70] D. Richter, chapter **C2** of this book

[71] L. Harnau, R. G. Winkler, and P. Reineker, J. Chem. Phys. **101**, 7750 (1995)

[72] L. Harnau, R. G. Winkler, and P. Reineker, J. Chem. Phys. **104**, 6355 (1996)

[73] L. Harnau, R. G. Winkler, and P. Reineker, Europhys. Lett. **45**, 488 (1999)

[74] J. Rotne and S. Prager, J. Chem. Phys. **50**, 4831 (1969)

[75] J. K. G. Dhont, chapter **A2** of this book

[76] S. S. Sorlie and R. Pecora, Macromolecules **23**, 487 (1990)

[77] K. Soda and A. Wada, Biophys. Chem. **20**, 185 (1984)

[78] J. C. Thomas, S. A. Allison, J. M. Schurr, and R. D. Holder, Biopolymers **19**, 1451 (1980)

[79] R. G. Winkler, L. Harnau, and P. Reineker, Macromol. Theory Simul. **6**, 1007 (1997)

[80] J. Seils and T. Dorfmüller, Biopolymers **31**, 813 (1991)

[81] W. Schweika, chapter **B1** of this book

[82] S. R. Quake, H. Babcock, and S. Chu, Nature **388**, 151 (1997)

[83] R. G. Winkler, Phys. Rev. Lett. **82**, 1843 (1999)

[84] D. E. Smith, H. P. Babcock, and S. Chu, Science **283**, 1724 (1999)

[85] P. LeDuc, C. Haber, G. Bao, and D. Wirtz, Nature **399**, 564 (1999)

[86] D. E. Smith and S. Chu, Science **281**, 1335 (1998)

[87] T. Hofmann, R. G. Winkler, and P. Reineker, Phys. Rev. E **61**, 2840 (2000)

[88] R. Rzehak, D. Kienle, T. Kaeakatsu, and W. Zimmermann, Euro. Phys. J. E **2**, 3 (2000)

C 4 Protein-polymer mixtures

R. Tuinier
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 Proteins as colloids or atoms **4**
2.1 Colloid-atom analogy 4
2.2 Colloid-protein analogy 5
2.3 Atom-protein analogy: phase behavior 7

3 Colloid-polymer mixtures **11**
3.1 Introduction to depletion interaction 11
3.2 Pair interaction between colloids due to polymers 13
3.3 Phase behavior of polymer-colloid mixtures 16

4 Phase behavior of protein solutions **22**
4.1 Protein-polysaccharide mixtures 22
4.2 The role of polymer chains in crystallization 24

5 Summary **27**

A Appendix: Generalized Gibbs adsorption equation **28**

B Appendix: Free volume fraction α from scaled particle theory **29**

1 Introduction

Many biological systems in our daily life such as food and our cells consist of liquid dispersions crowded with biopolymers [1],[2],[3],[4],[5]. In food products the high concentration of biopolymers may lead to a phase instability that is usually unwanted. Producers of food products would like to be able to understand and especially to predict when instability takes place.

In blood, large concentrations of red blood cells (RBCs) and blood serum proteins are present [5]. The concentration of blood serum proteins is of the order of 75g/L. Due to illnesses the serum protein may increase leading to an enhanced attraction between the RBCs. More attraction between colloidal particles (like RBCs) increases the sedimentation rate, which explains why the RBC sedimentation rate has been and is being used to detect illnesses [5].

The cytoplasm of living cells is also crowded with biopolymers and proteins [6]. It has even been conjectured that the small, localized phases in cells, with distinct sharp interfaces, are due to crowding-induced phase separation [1].

Why is this biomacromolecular crowding related to attraction and phase instability? In order to get an understanding of the interactions and phase stability in those crowded systems let us focus on what happens when proteins and polymers are mixed. A general scheme of what happens when globular proteins (indicated as the spheres) are mixed with polymer chains (indicated as the coils), as based on the extensive work over the last few decades by Tolstoguzov [2], is schematically depicted in Fig. 1.

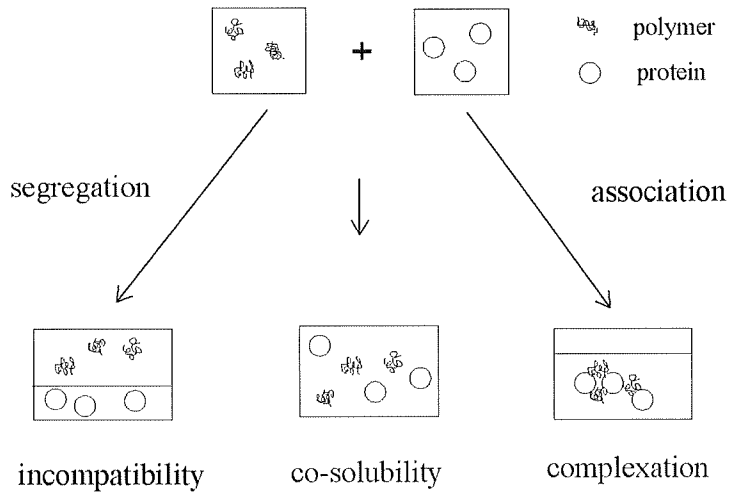


Fig. 1: *Schematic overview of the possible states of mixed polymer-globular protein mixtures. The proteins are simplified as spherical particles, which sometimes is a reasonable approximation for globular proteins. The polymer chains are indicated as spherical fluffy coils, representing flexible polymers.*

When protein particles are mixed with polymers the interaction between them is either segregative (polymer and protein repel each other and are denoted as incompatible) or associative (there is attraction between protein and polymer) [2],[3],[4]. In case of associative interaction, oppositely charged (bio)polymers form complexes due to electrostatic interaction [4]. If the complexes are sufficiently attractive, they may concentrate in liquid coacervate drops, leading to a phase separation of the mixture: one phase is enriched in both biopolymers and the other phase mainly contains solvent. The complexes may also lead to an instability including the formation of solid precipitates.

If the polymers are uncharged or have a similar charge as the proteins, they generally tend to segregate. The picture in Fig. 2 is an example of such a demixed system containing a polysaccharide mixed with whey proteins in an aqueous 0.10 M NaCl solution. In the case of segregative interaction, a phase separation takes place with the (bio)polymers concentrated in separate phases. In such phase separated biopolymer-protein mixtures one can actually measure the interfacial tension between the phases, which is of the order of 10^{1} μN/m [7].

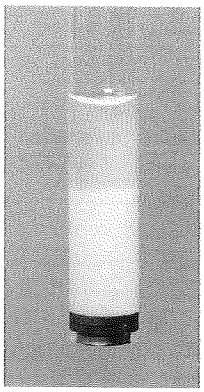


Fig. 2: *A demixed protein-polysaccharide mixture of aggregated whey protein particles (with a radius of 30 nm) and exocellular polysaccharides with a radius of gyration of 86 nm. The phase that is enriched in proteins is the turbid lower phase while the polysaccharides are concentrated in the upper phase. The coexisting phases are separated by a sharp interface.*

When a polymer interacts segregatively with a protein particle, a reduction of the polymer density near the protein particle is found. This reduction of polymer segments, due to a loss of conformational entropy of the chains near an interface, is termed depletion [8], which will be explained in more detail in section 3.

This chapter focuses on the segregative or depletion interaction in mixtures of polymers and proteins. In order to give a theoretical description of the phase stability of such mixtures we will discuss the protein-colloid-atom analogy in section 2. In section 3 a simple model for colloid-polymer mixtures will be addressed that allows to calculate the pair interaction and phase behavior . The theory described in section 3 will be used in section 4 to compare with the phase behavior of protein solutions. First it will be compared with the liquid-liquid phase separation in protein-polysaccharide mixtures. Secondly, the role of adding polymers in protein crystallization is discussed, which is important in the analysis of the 3D-structure of proteins, making it possible to better understand its biological function.

2 Proteins as colloids or atoms

In order to make progress and describe polymer-protein mixtures theoretically it is important to simplify the protein particles. At a first glance it may seem that such quite complicated structures as proteins are too complex to compare with colloids. Of course their chemistry is more involved: the 3D structures of proteins are very diverse as are their resulting biological functions. Colloids usually consist of one or a few simple chemical compounds, whereas proteins consist of a large number of different amino acids. For a full description of the pair interaction between two protein particles one needs detailed information on the Van der Waals attraction, electrostatic repulsion, hydrophobic interaction, hydration layers etc. However, under several conditions the physical behavior of a dispersion of proteins is not very different from that of colloids, as will be shown in sections 2.2 and 2.3. Such an approach may be criticized because nature is simplified, but is useful because it allows predicting certain properties, such as the phase behavior . But let us first focus on the colloid-atom analogy, which has helped to better understand and describe the physics of colloidal dispersions and will later be of use to understand the physics of protein dispersions.

2.1 Colloid-atom analogy

Already in early times it was recognized that colloidal particles suspended in a low-molecular-weight solvent behave as simple fluids. From the work of Van der Waals it followed that the model structure of a simple liquid is that of a hard sphere system. In 1910 Perrin [9] microscopically studied the motion of colloidal particles and verified that they act as giant molecules. Perrin settled the issue of the colloid-atom analogy with simple experiments on Brownian motion [10]. A few decades later, the statistical mechanics formulation for treating colloids as atoms was given. Still, no theory was yet available predicting for instance the pressure of atomic liquids in terms of pair interactions. With the onset of using computers to simulate fluids it became possible to obtain statistical averages that were sufficiently reliable to give representative bulk properties, based on a given pair interaction so (approximate) statistical mechanical theories could be tested. Previously, theories could only be validated directly with experiments. By studying systems with model pair interactions, say of the Lennard-Jones type, it became clear that the structure of a concentrated fluid is primarily determined by the hard repulsive part, with the softer attractive part as a perturbation. This resulted in the important finding that *the model structure* of a simple liquid and a stable colloidal fluid is that of a collection of hard spheres. Nowadays, the theories can be tested also experimentally in a much better way due to the development of synthesized homodisperse model colloids.

Thus the formalism connecting structure and thermodynamic properties of a colloidal dispersion is identical to that of a one-component gas or fluid. Therefore, we may use the (equilibrium) structural and thermodynamic properties derived for a hard sphere gas to describe the properties of a colloidal dispersion. As an illustration we discuss a scattering experiment of a colloidal dispersion approaching an instability point, see for instance Bódnar *et al.* [11]. The intensity of scattered radiation can be attributed to local fluctuations in the colloidal particle concentration. When the intensity increases it implies that the concentration fluctuations increase and extend over larger regions of space. Progressing further, the spatial concentration fluctuations lead to a complete separation of spatial regions with different particle concentrations and (mostly) different mass densities, i.e., phase separation is taking place. These phenomena are comparable to what happens in simple atomic or molecular fluids

near an instability point [12]. Colloidal dispersions thus mimic atomic fluids in many aspects in which the colloidal particles act as giant atoms. A difference between atomic systems and colloidal dispersions that is experimentally relevant is the scattering light by these systems. The relevant length scale in colloidal dispersions is on the order of the wavelength of visible light. Therefore a colloidal liquid can be recognized as a turbid fluid (roughly speaking for instance very white for a dispersion with particles of radii larger than, say, 100 nm, and bluish if the colloidal particles have a radius of about 30 nm) and light scattering can be used (for not too high particle concentrations; if the turbidity exceeds 1 cm^{-1} multiple scattering techniques are required) for the analysis of the particle size and structure of the dispersion. For atomic fluids one is restricted to X-Rays or neutrons with a much smaller wavelength.

2.2 Colloid-protein analogy

As to illustrate the analogy between some dispersions of protein particles and colloidal dispersions we give a few examples of experimental results on protein dispersions and its interpretation with theory that is commonly used to describe colloidal dispersions of hard spheres. As a first example we give small-angle neutron scattering results (SANS) on polydisperse (about 10% in size) β-caseins [13]. The protein β-caseins is one of the caseins present in large quantities in milk. In H_2O or D_2O, β-caseins form spherical micelles. The structure factor *S(K)* of a concentration series of β-caseins as a function of the wave vector *K* is plotted in Fig. 3. Theoretically one may use the Percus-Yevick theory to describe the structure factor of a hard sphere dispersion [14]. This has been extended to include size polydispersity [15] and the resulting predictions for the structure factors for the three β-casein volume fractions used are plotted in Fig. 3. The agreement between the data points and the theoretical prediction (no adjustable parameters) is quite reasonable taking into account the complexity of the protein particles.

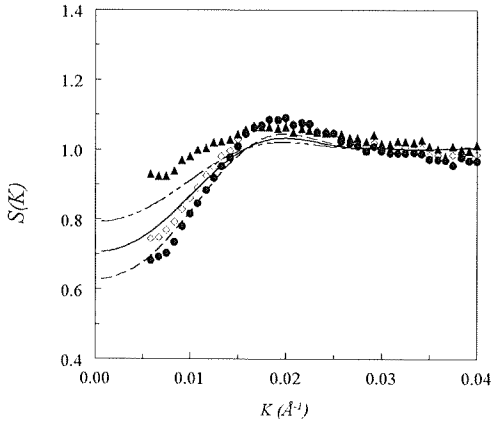


Fig. 3: *Structure factors of a dispersion of β-caseins (spherical micelles with a radius of gyration: 8 ± 1 nm) dispersed in D_2O at 298 K for three concentrations: 10 g/L (closed triangles), 15 g/L (open diamonds) and 20 g/L (closed circles) corresponding to volume fractions ϕ=0.035, 0.052 and 0.07 (specific volume of β-caseins is 3.5 mL/g). SANS data are indicated as the data points. Curves are the results using the polydisperse Percus-Yevick theory with a polydispersity of 10%. Redrawn from de Kruif et al. [13].*

As a second example that illustrates that protein dispersions sometimes behave as colloidal dispersions we address the measured diffusion coefficients of casein micelles. Milk contains a volume fraction $\phi \sim 0.11$ of casein micelles, which are polydisperse association colloids consisting of various caseins linked together by calcium phosphate (present in the continuous phase of milk) and stabilized by κ-caseins on the outside of the micelles. The κ-caseins act as polymer brushes that are well-known stabilizers for colloids. The casein micelles are rather stable; they are hardly influenced by changing the temperature or increasing the salt concentration. The diffusion coefficient D of the casein micelles in their natural solvent (= the continuous phase of skim milk, milk without fat) was measured using dynamic light scattering by de Kruif [16]. For small scattering angles the focus is on large length scales where the collective diffusion is probed. Up to first order in volume fraction dependence the short-time collective diffusion (see ref. [17] for an extensive explanation and discussion on diffusion of colloidal particles) of a dispersion of hard spheres reads:

$$\frac{D}{D_0} = 1 + 1.45\phi + \ldots \tag{1}$$

here D_0 is the diffusion coefficient at infinite dilution. The collective diffusion coefficient thus (initially) increases with increasing volume fraction ϕ : the gradient in the concentration fluctuations increases with ϕ. For large scattering angles one probes the short-time self-diffusion, which for hard spheres can be written as [17]:

$$\frac{D}{D_0} = 1 - 1.83\phi + \ldots \tag{2}$$

On short length scales the diffusion thus decreases when the particle concentration of hard spheres increases: the particles hinder one another in their motions. In Fig. 4 results for both collective as well as self-diffusion are plotted for casein micelles (symbols) [16]. The predictions of equations (1) (dashed line) and (2) (full line) are also plotted and the agreement with the experimental data is quite good. Also for the viscosity of casein micelle dispersions de Kruif and co-workers have shown that it follows the hard sphere behavior.

From the above it becomes clear that spherical repulsive proteins behave as hard spheres up to a certain degree. It should be realized that details in the interactions might lead to deviations from hard sphere behavior with increasing concentration. Nevertheless, it is promising that a simple description in terms of hard spheres allows to describe such a complicated system as a protein solution.

Under conditions of attractions between the proteins the situation is more complex. Piazza [18] established a simple model of protein interparticle interaction near crystallization conditions. He supposed the interaction could by described as the adhesive hard sphere potential. This is a hard sphere potential plus a very short-ranged attraction. The measured osmotic compressibility of lysozyme could be described very well using this model that has been used in the past to describe colloidal model dispersions with short-ranged attractions. The analogy between colloids and proteins thus may hold under various conditions for the static and dynamic physical properties of proteins dispersions. What is of special interest is of course the phase behavior of protein dispersions.

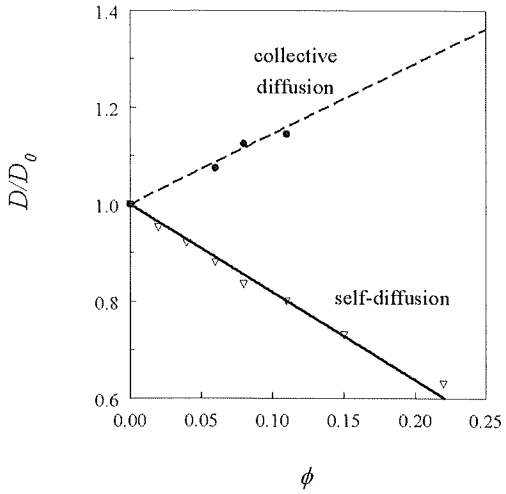


Fig. 4: *Collective (filled circles) and self-diffusion (open triangles) of casein micelles as a function of their volume fraction in the continuous phase of skim milk, redrawn from de Kruif, [16]. The theoretical predictions for hard spheres are plotted for collective (dashed line) and self-diffusion (full line).*

2.3 Atom-protein analogy: phase behavior

At the end of the 19th century Van der Waals [19] presented a theory that demonstrates how a collection of gas molecules with excluded volume plus an attraction between them becomes unstable below a certain temperature. Below that critical temperature an increase of the pressure may lead to a condensation of the gaseous phase into a liquid state. Although the gaseous state has the largest entropy, the liquid configurations are favored by their low energy, as first described correctly by Van der Waals. It is also known, however, that for high concentrations a collection of atoms may undergo a fluid-to-solid transition at high temperatures, where the gaseous or fluid state is stable. On the basis of their excluded volume one may first approximate the atoms as hard spheres. Computer simulations [20] have demonstrated that a fluid-solid transition occurs in a collection of hard spheres. Due to the fact that the hard spheres are less stuck and still have some motional freedom inside an ordered lattice of spheres (the solid) as compared to a concentrated fluid where the spheres have a tendency to jam, the entropy loss due to crystallization thus seems to be overcompensated by the motional entropy gain that a phase transition takes place. It is thus unnecessary to have an attractive force for a phase transition. What happens when there is also attraction between the atoms as occurs in practice?

The pair interaction between two atoms $u(r)$, with centre-to-centre distance r, can be well described for many atomic fluids using the (generalized) Lennard-Jones potential:

$$u(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{2n} - \left(\frac{\sigma}{r}\right)^{n}\right] \qquad (3)$$

For the n=6 (the 'classical' Lennard-Jones potential) and n=11 conditions we plotted the Lennard-Jones potential in Fig. 5. Increasing n leads to shorter range of attraction between the particles.

It is interesting to study the resulting phase diagrams in systems of many particles interacting through equation (3). This can be done using computer simulation methods. We represent the phase diagrams for three cases A) (Fig. 6), B) and C) (Fig. 7). As a reference point we will first show the result for the hard sphere interaction $u(r) = \infty$ for $r \leq \sigma$ and $u(r) = 0$ for $r > \sigma$, i.e. the hard sphere case A) for which there is a temperature independent fluid-solid phase transition. The fluid coexists at $\phi = 0.494$ with the solid phase at $\phi = 0.545$ as is known from computer simulation results [20], see Fig. 6.

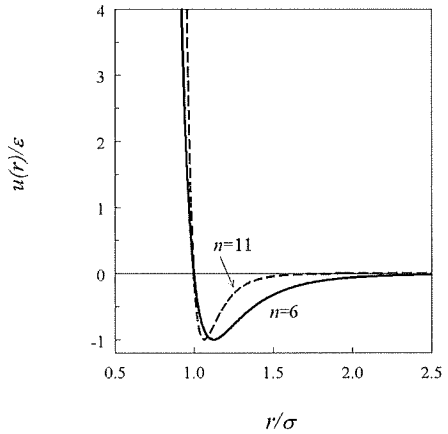


Fig. 5: *The generalized Lennard-Jones interaction potential following equation (3) for n = 6 and n = 11. By changing n the range of attraction is modified.*

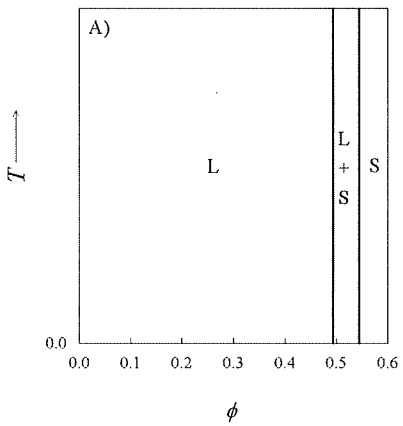


Fig. 6: *Phase diagram of a pure hard sphere dispersion.*

There is no gas-liquid coexistence for hard sphere fluids. For the Lennard-Jones-cases n = 6 and 11 Vliegenthart *et al.* [21] computed the phase diagram using Gibbs Ensemble method simulations. Some of their results are redrawn in Fig. 7 B) and C). For n=6 there is a wide region of gas-liquid (G+L) coexistence, with a critical temperature T_c. At large volume fractions, there is a liquid-solid (L+S) coexistence that meets the liquid branch of the gas-liquid coexistence at the triple temperature T_t where gas, liquid and solid coexist. Below T_t the solid coexists with a gas phase. For a shorter range of attraction (n=11) the topology of the phase diagram changes significantly as compared to n=6, see Fig. 7 C) compared to B). The fluid phase is stable for n=6, but lies just below the fluid-solid coexistence for n=11. There is mainly a (stable) fluid-solid transition. The phase transition is shifted and is now located under the fluid-solid transition. For shorter ranges of attraction, n>11, the gas-liquid coexistence curve shifts further downwards with respect to the fluid-solid curve [21]. This 'hidden' gas-liquid transition becomes metastable for shorter-ranged attractions as compared to the n=6 case, attaining a flatter shape. For n=12 and n=18, the fluid-to-solid coexistence dominates and the metastable gas-liquid coexistence lies very far below the F-S binodal [21].

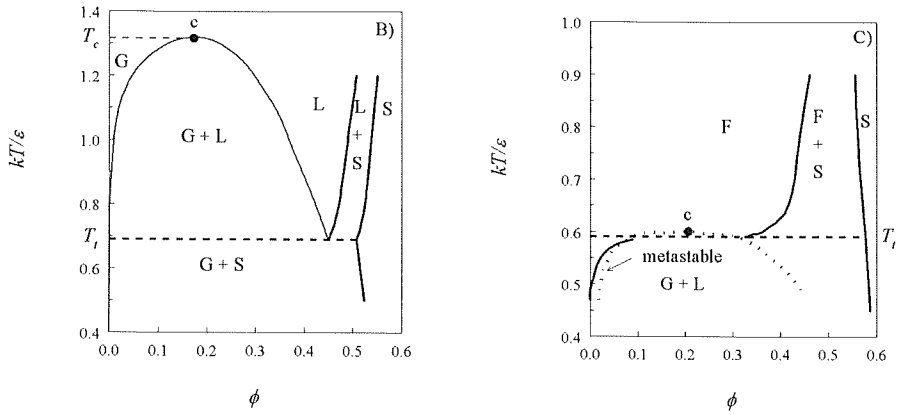


Fig. 7: *Phase diagram for B) Lennard-Jones pair interaction with n=6 and C) Lennard-Jones pair interaction with n = 11. The results are redrawn from Vliegenthart et al. [21].*

Due to the analogy between colloidal and atomic fluids these results also have implications for the behavior of colloidal suspensions. When the range of attraction is much smaller than the particle diameter, one may not expect a stable gas-liquid transition. Indeed when phase separation occurs in colloidal dispersions, the condensed phase is usually 'solid-like', i.e., there is no 'liquid-liquid' phase separation. In recent years it is being accepted that the hidden, metastable gas-liquid phase, although not stable, may still influence the behavior of non-equilibrium phenomena when the dispersion is suddenly quenched into the metastable region, e.g., by sudden cooling or adding destabilizing components like (concentrated) salt, non-adsorbing polymers, etc. In colloidal dispersions one finds three types of phases: the colloidal gas, colloidal liquid and colloidal solid (or crystal) phases. The continuum in these phase states is now the solvent, which can be viewed upon as a background.

Let us now compare the Lennard-Jones phase diagrams in Fig. 7 with that of an aqueous solution of proteins. As an example we present the phase diagram of a solution of the eye lens protein γ_{II}-crystallin at pH 7 in 100 mM Na-phosphate buffer [22], see Fig. 8. The critical temperature T_c is 278.4 K and the critical protein volume fraction ϕ_c=0.21. At high temperatures there only exists a fluid-to-crystal (protein crystals are the solid S-phase) transition. At lower temperature the so-called 'solubility curve', which is the binodal curve of the fluid-solid (F-S) coexistence region, shifts towards lower protein volume fractions. Significantly below the F-S curve a metastable gas-liquid G-L region is found, including a critical point. This phase diagram closely resembles that of a Lennard-Jones-phase diagram for large n. In fact, the case of n=11 in Fig. 7 (C) is much closer to the protein phase diagram as compared to Fig. 7 (B). As many protein dispersions that can be crystallized, of which lysozyme is the most well-known, are comparable to the phase diagram in Fig. 8, it appears that the range of attraction (say ~20 to 30% of their size) in salty aqueous protein dispersions is quite fixed. Due to the similarity of the phase diagrams of colloids and proteins it is justified to treat protein solutions as spheres with a hard sphere repulsion combined with a short-ranged, temperature-dependent attraction. For a colloid/protein analogy in an associative protein-polymer mixture, see Weinbreck *et al.* [23].

It is of course interesting to study what happens to these systems when polymers are added to protein dispersions because they are often mixed together in practice and since polymers are well-known crystallization promoters. Therefore, in the next section we will address the theory of polymer-colloid mixtures both on the level of the pair interaction as well as on the level of the phase behavior.

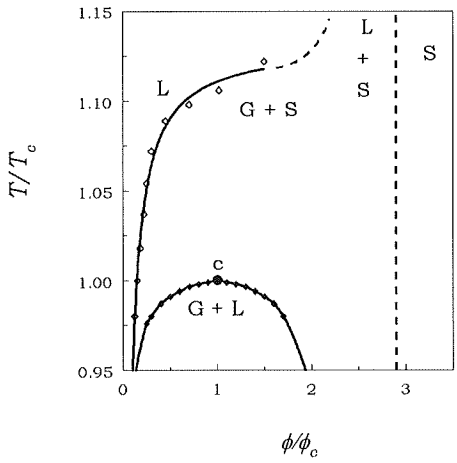


Fig. 8: *Phase diagram of the lens protein γ_{II}-crystallin in 100 mM Na-phosphate buffer at pH 7. Redrawn from Benedek and co-workers [22]. The data points were determined experimentally from the phase boundaries. The curves are a guide to the eye.*

3 Colloid-polymer mixtures

3.1 Introduction to depletion interaction

In 1954 Asakura and Oosawa showed that two plates immersed in a solution of ideal non-adsorbing polymers, where depletion takes place at the plate surfaces, attract one another [24]. Using statistical mechanics they derived an expression for the polymer density (the number of chains per volume) between the plates ρ_{in}. When the distance between the plates h decreases, ρ_{in} becomes smaller than ρ, the concentration outside the plates. Consequently, this concentration difference leads to an expression for the pressure difference outside and between the plates. The resulting interaction potential $W(h)$, plotted in Fig. 9, demonstrates that there is a significant attraction between the plates due to the depletion of ideal polymer chains. The contact potential gives the depletion thickness per plate, see, which is close to the polymer's radius of gyration R_g.

More relevant is of course the attraction between two spheres as depicted in Fig. 10 due to the non-adsorbing polymer chains, which is used to explain the mechanism responsible for the attractive depletion interaction. There will be a layer close to the sphere surface where the concentration of polymer centers will be zero, denoted as the depletion layer, as indicated by the (dashed) layers around the spheres in Fig. 10. Outside this layer the polymer concentration equals the bulk polymer concentration. The concentration gradient due to this depletion layer results in an osmotic pressure gradient. For a single sphere this osmotic pressure gradient is isotropic (Fig. 10 (a)). However, if the depletion layers overlap the osmotic pressure becomes anisotropic, leading to a net osmotic force, as indicated by the arrows in Fig. 10 (b), that pushes the particles together.

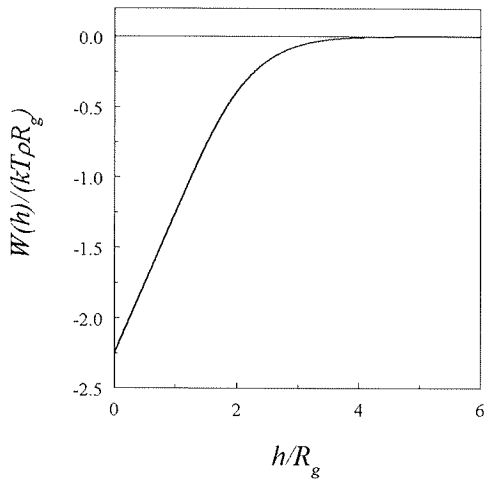


Fig. 9: *Interaction potential W(h) between two hard plates in a solution of ideal polymer chains. The interaction potential at contact $W(0)/kT = -4\rho R_g/\sqrt{\pi}$; the depletion thickness per plate is $2R_g/\sqrt{\pi} \approx 1.13R_g$ [24],[8].*

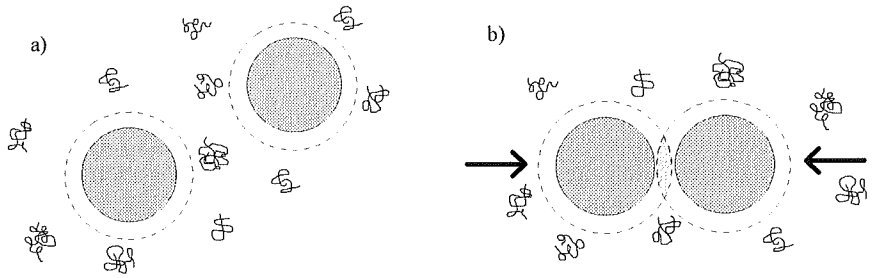


Fig. 10: *Schematic picture of two colloidal spheres in a polymer solution with non-adsorbing polymers. The depletion layers are indicated by the short dashes. When there is no overlap (a) the osmotic pressure on the spheres due to the polymer solution is isotropic. For overlapping depletion layers (b) the osmotic pressure on the spheres is unbalanced; the excess pressure is indicated by the arrows.*

In that case, for ideal (the polymer segments do not 'feel' other segments) polymers, the free energy of interaction equals the overlap volume, which is indicated by the hatched region, times osmotic pressure, as will be shown in section 3.1. The resulting attraction induces a phase transition upon exceeding a certain concentration of colloidal spheres and polymers. A colloid-enriched phase is then in equilibrium with a polymer-enriched phase.

Protein-polymer mixtures in a common solvent are quite involved systems as compared to for instance a gas of atoms. Proteins are usually dissolved in an aqueous salt solution, which is quite complex as such. A common procedure in soft condensed matter physics is to consider the solvent as a continuous background. This can be rationalized by realizing that the solvent molecules are much smaller than colloidal particles. Although protein particles often are at the lower limit of the colloidal range (1 nm to 1 μm), the solvent molecules are still small to such a degree that they may be regarded as background as to describe the thermodynamic properties of the protein solution. The solvent, however, indirectly mediates the forces between the protein particles. As we have seen in section 2, we may sometimes describe proteins in solutions as colloidal particles. This suggests that the proteins in aqueous salt solutions may be described as colloidal particles in a background solvent.

Adding polymer to the solvent-protein system makes the situation more complex. Especially polymer chains in solution have many degrees of freedom. A classical simplification of the polymer-colloid mixture is to replace the polymer chain with an effective sphere. Asakura and Oosawa [25] suggested replacing a polymer chain with a hard sphere in the limit of small polymer concentrations. Vrij [26] proposed using a freely overlapping sphere (FOS) to describe the polymer chain. The FOS is a hard sphere for the colloidal hard sphere but is freely permeable for other FOSs and thus mimics polymer chains under theta-conditions. The pair interaction scheme Vrij [26] suggested reads:

$$
\begin{aligned}
&W_{c\text{-}c}(r) = \infty && r < \sigma_c \\
&W_{c\text{-}c}(r) = 0 && r > \sigma_c \\
&W_{c\text{-}FOS}(r) = \infty && r < (1/2)(\sigma_c + \sigma_{FOS}) \\
&W_{c\text{-}FOS}(r) = 0 && r > (1/2)(\sigma_c + \sigma_{FOS}) \\
&W_{FOS\text{-}FOS}(r) = 0 && \text{all } r
\end{aligned}
\tag{4}
$$

Here $W(r)$ is the pair interaction potential, c refers to the colloid, σ_c is the hard sphere diameter (=2R, where R is the sphere's diameter), σ_{FOS} is the FOS diameter ($a_{FOS} = \sigma_{FOS}/2$ is the FOS radius) and r is the distance between the particle centers, see Fig. 11. This model will be used in section 3.2 to derive expressions for the effective pair interaction between the colloidal hard spheres and to arrive at a prediction for the phase behavior in section 3.3.

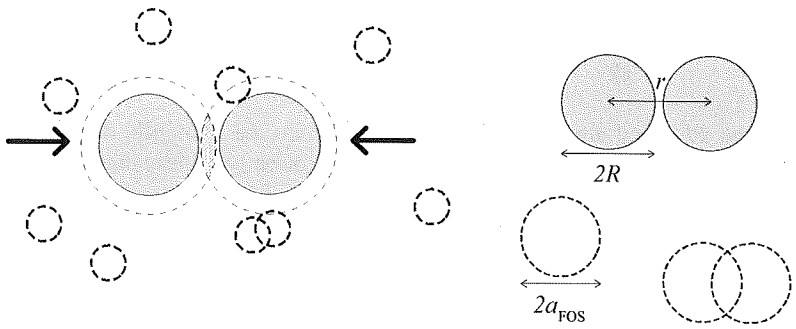


Fig. 11: *Left: Two hard spheres (filled particles) in a dispersion of freely overlapping spheres (FOSs). Due to the (non-adsorbing) FOSs there exists a depletion layer around each hard sphere (dashed layer around the hard spheres) with a thickness equal to the a_{FOS}. When the depletion layers overlap the FOSs push the hard spheres together. Right: indication of the relevant length scales: r, the centre-to-centre distance between the hard sphere centers, R, hard sphere radius, and a_{FOS}, the FOS-radius.*

3.2 Pair interaction between colloids due to polymers

The effective interaction between two colloidal particles can be calculated thermodynamically using the generalized Gibbs adsorption equation [27] (for a derivation, see Appendix A):

$$
-\left(\frac{\partial W(h)}{\partial \mu_q}\right)_h = N_q(h) - N_q(h=\infty)
$$

or

$$
W(h) = -\int_{-\infty}^{\mu_q} d\mu_q' \left[N_q(h) - N_q(h=\infty)\right] \tag{5}
$$

that relates the interaction potential to the excess amount of particles when two particles are brought together at a distance h In case of two spheres with a centre-to-centre distance r

(see Fig. 11): $h = r-2R$. The quantity N_q is the number of species of type q. In case of FOS particles as depleting agents:

$$d\mu_q = d\mu_{\mathrm{FOS}} = kTd \ln \rho_{\mathrm{FOS}} \tag{6}$$

where ρ_{FOS} is the number of FOSs per volume. So, for two spheres in a dispersion of FOSs and noting that $N_{FOS}(h)$–N_{FOS} (h=∞) is an excess quantity, so $N_{FOS}(h)$–$N_{FOS}(h=\infty) = \Gamma_{FOS}(h)$–$\Gamma_{FOS}(h=\infty)$, where Γ_{FOS} is the adsorbed amount of FOS particles:

$$\frac{W(h)}{kT} = -\int_0^{\rho_{\mathrm{FOS}}} d\rho_{\mathrm{FOS}}{}' \left[\frac{\Gamma_{\mathrm{FOS}}(h) - \Gamma_{\mathrm{FOS}}(h=\infty)}{\rho_{\mathrm{FOS}}{}'} \right] \tag{7}$$

Since Γ_{FOS} is proportional to ρ_{FOS} it follows:

$$\frac{W(h)}{kT} = \Gamma_{\mathrm{FOS}}(h=\infty) - \Gamma_{\mathrm{FOS}}(h) \tag{8}$$

The difference in the adsorbed amount of FOS particles when the two colloidal spheres are separated a distance h compared to the adsorbed amount around two individual spheres (the r.h.s. of equation (8)) is exactly ρ_{FOS} times the volume of the hatched region or the overlap volume of depletion layers in Fig. 11. In other words for (ideal, non-self-interacting) FOS particles equation (8) can be rewritten as:

$$W(h) = -\rho_{\mathrm{FOS}} kT\, V_{\mathrm{overlap}} \tag{9}$$

Since $\Pi_{\mathrm{FOS}}=\rho_{\mathrm{FOS}}\, kT$ is the (ideal) osmotic pressure it says that the interaction is just minus the overlap volume V_{overlap} times the osmotic pressure Π_{FOS}, as was suggested by Asakura and Oosawa [25] and Vrij [26]. Above we have derived this expression straightforwardly from thermodynamics. Simple geometry helps us to compute the overlap volume:

$$V_{\mathrm{overlap}} = \frac{2}{3}\pi a_{\mathrm{FOS}}^3 \left(1 - \frac{h}{2a_{\mathrm{FOS}}}\right)^2 \left(2 + \frac{3R}{a_{\mathrm{FOS}}} + \frac{h}{2a_{\mathrm{FOS}}}\right) \tag{10}$$

valid for $0<h<2(R+a_{\mathrm{FOS}})$. Combining equations (9) and (10) using $\rho_{\mathrm{FOS}} = 3\phi_{\mathrm{FOS}}/4\pi a^3{}_{\mathrm{FOS}}$ yields:

$$\frac{W(h)}{kT} = \begin{cases} -\phi_{\mathrm{FOS}} \left(\dfrac{3R}{2a_{\mathrm{FOS}}} + 1 + \dfrac{h}{4a_{\mathrm{FOS}}} \right) \left(1 - \dfrac{h}{2a_{\mathrm{FOS}}} \right)^2 & for\ \ 0 \le h \le 2\left(R + a_{\mathrm{FOS}}\right) \\ 0 & for\ \ h > 2\left(R + a_{\mathrm{FOS}}\right) \end{cases} \tag{11}$$

A few illustrative results for $W(h)$ are plotted in Fig. 12. At fixed size ratio a_{FOS}/R an increase of the relative concentration of FOSs, ϕ_{FOS} (which is unity at the overlap concentration: the polymer or FOS concentration where the system is exactly fully filled with FOSs or polymer coils, respectively), leads to a proportional increase of the magnitude of the attraction between

the hard spheres. An increase of the size ratio a_{FOS}/R at fixed volume fraction of FOSs leads to a decrease of the contact potential and an increase of the range of attraction.

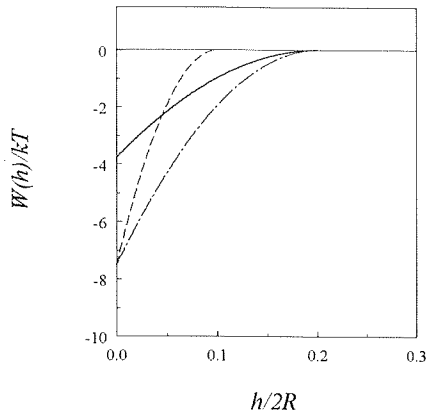


Fig. 12: *Results for the depletion interaction potential W(h) between two hard spheres in a sea of FOSs as a function of the distance between the spheres following equation (11). Dashed curve: $\phi_{FOS}=0.5$, $q=a_{FOS}/R=0.1$, full curve: $\phi_{FOS}=0.5$, $q=a_{FOS}/R=0.2$; Dot-dashed curve: $\phi_{FOS}=1$, $q=a_{FOS}/R=0.2$.*

When applying the FOS-hard colloidal sphere model one should assign a certain value to the size of a FOS particle as to link polymer chains to FOSs. Often the FOS radius is taken equal to the polymer's radius of gyration R_g. Formally, it should be equal to the relevant depletion layer thickness around the spheres. In case of very large spheres the depletion thickness near a flat wall could be taken. The depletion thickness can be calculated from the segment density profile [8]. For ideal chains the polymer segment density near a hard non-adsorbing wall was calculated by Eisenriegler [28]. A calculation of the depletion thickness from this profile gives $a_{FOS} = 2R_g/\sqrt{(\pi)}$, which is indeed very close to R_g. For more details on the polymer density near interfaces, see [29].

In practice one should keep the following factors in mind [30]:

1) the depletion thickness will become polymer density-dependent near and above the overlap concentration; the depletion thickness behaves as the correlation length that scales as $(\rho/\rho^*)^{-3/4}$ for polymer densities ρ larger than the overlap density ρ^* under good solvent conditions.
2) a long polymer chain can easily wrap around a small colloidal sphere; the depletion thickness reduced to values much smaller than the radius of gyration or correlation length for large size ratios q.
3) the osmotic pressure will deviate from Van't Hoff's law near and above the overlap concentration, especially when the solvent quality is better.

For the moment we will only stick to the simple FOS model and use Van't Hoff's law and take $a_{FOS} = R_g$. This will be a reasonable approximation for $\rho/\rho^* << 1$ and is insightful for our aim.

Very recently the depletion potential between two spheres in a polymer solution has been measured using the scanning optical tweezer technique [31]. By trapping two silica spheres

with a diameter of 1.25 μm the interaction was measured between those spheres in a solution of DNA molecules with a radius of gyration of 500 nm. In Fig. 13 their results are reproduced for concentrations of 140 μg/mL (filled squares) and 190 μg/mL (open circles). The corresponding relative densities of the DNA chains ρ/ρ^* (= ϕ_{FOS}) are 1.46 and 1.98.

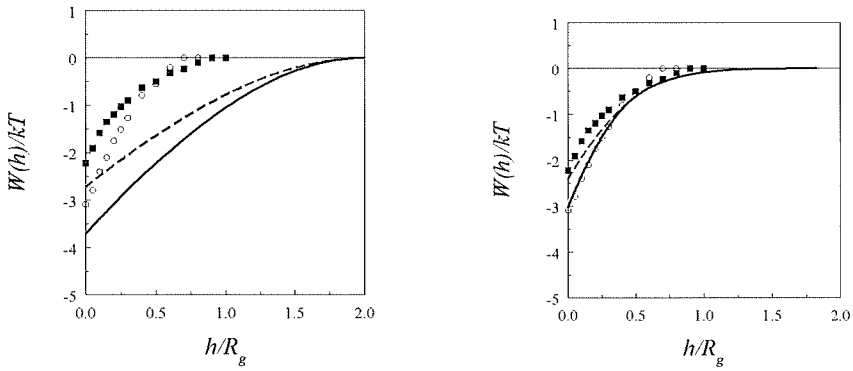


Fig. 13: *Comparison of the interaction potential between two silica spheres, with a diameter of 1.25 μm in a non-adsorbing solution with DNA molecules with a radius of gyration of 500 nm as measured by Verma et al. [31] using optical tweezers. Filled squares represented 140 μg/mL and the open circles 190 μg/mL DNA, corresponding to (relative) polymer concentrations ρ/ρ^* of 1.46 and 1.98, respectively). The results of the FOS theory are given by the curves (left) and the results for polymers with excluded volume interaction [32] are given by the curves in the right plot.*

In the left plot of Fig. 13 the hard sphere plus FOS results (equation (11)) for $a_{FOS}=R_g=500$ nm and $R=625$ nm, so $q=a_{FOS}/R=0.8$, and $\phi_{FOS}=1.46$ (dashed) and 1.98 (full curve) are plotted. The order of magnitude is reasonable but the range of attraction is overestimated by the theory. This is due to the assumption: $a_{FOS}=R_g$. For polymer concentrations above the overlap concentration ρ/ρ^* one should take into account the non-ideality of the polymer solution. Taking into account the excluded volume interactions [32], which is more involved theoretically, gives the results in the right plot of Fig. 13. The agreement with the experiment is now much better. Still, the very simple FOS model gives a reasonable order of magnitude prediction for the interaction.

3.3 Phase behavior of polymer-colloid mixtures

Osmotic equilibrium and free volume theory

Although the pair interaction between two colloidal hard spheres due to non-adsorbing polymer chains as we have encountered in the previous section is interesting and insightful as such, the resulting phase behavior is of primary interest for biological systems. First, we simplify the systems we are interested in – globular spherical-like proteins mixed with biopolymer chains in a salty aqueous solution – into a three-component system of spherical colloids, polymer chains and a simple solvent. Again, we treat the aqueous salt solution as background as is common in colloid physics. This means that we have simplified the system to an effective two-component system. Next we simplify the polymers as in the previous

section by replacing them with freely overlapping spheres (FOSs). Still a system of hard spheres plus FOSs is not easy to fully describe thermodynamically. A useful method for an approximate thermodynamic description is the so-called osmotic equilibrium theory. It is based upon placing the system of interest in equilibrium with a reservoir that only contains one of the components. Commonly, one puts the component that 'perturbs' or that is smaller in the reservoir, see Fig. 14. For our situation of interest, we consider a system of hard spheres plus FOSs, held in equilibrium with a dispersion of FOSs. Since we would like to fix the temperature, the volume and the chemical potential, so energy- and mass transport are possible, we work in the grand canonical ensemble. The colloidal hard spheres are however fixed in the system (they can not pass the membrane), so we have to consider the semi-grand canonical potential $\Omega(N_c,V,T,\mu_{FOS})$:

$$\begin{aligned}\Omega(N_c,V,T,\mu_{FOS}) &= F(N_c,V,T)+\int_{-\infty}^{\mu_{FOS}^R}\frac{\partial\Omega(N_c,V,T,\mu_{FOS}^{R\,\prime})}{\partial\mu_{FOS}^{R\,\prime}}d\mu_{FOS}^{R\,\prime}\\ &= F(N_c,V,T)-\int_{-\infty}^{\mu_{FOS}^R}N_{FOS}d\mu_{FOS}^{R\,\prime}\end{aligned} \tag{12}$$

Here N_c is the number of colloidal particles in the system volume V. The free energy of the pure hard sphere dispersion is given by $F(N_c,V,T)$.

The second term on the r.h.s. of equation (12) is the perturbation due to adding FOSs. Note the analogy with the generalized Gibbs adsorption equation (5); equation (11) is its 'many-body' analogue. If no FOSs are present $\mu_{FOS}=-\infty$ and $\Omega = F$. Due to the osmotic equilibrium $\mu_{FOS} = \mu^R_{FOS}$. The number of FOSs in the system N_{FOS} depends on μ^R_{FOS}. In order to make progress we use the free volume theory. It is based upon the assumption [33]:

$$N_{FOS} = \rho^R_{FOS} V_{free} \tag{13}$$

where V_{free} is the ensemble-averaged free volume for FOSs in the system under the assumption that the colloid-colloid correlations are undistorted. A free volume fraction α can be defined:

$$\alpha = \frac{V_{free}}{V} \tag{14}$$

which can be calculated from scaled particle theory (SPT), see [33].
For the conditions in Fig. 14, the SPT result, see Appendix B for a derivation, equals:

$$\alpha = (1-\phi)e^{-\left[A\left(\frac{\phi}{1-\phi}\right)+B\left(\frac{\phi}{1-\phi}\right)^2+C\left(\frac{\phi}{1-\phi}\right)^3\right]} \tag{15}$$

with $A=3q+3q^2+q^3$, $B=4.5q^2+3q^3$ and $C=3q^3$, where $q=a_{FOS}/R$ [33]. Equations (13) and (14) yield $N_{FOS}=\alpha V\rho^R_{FOS}$.

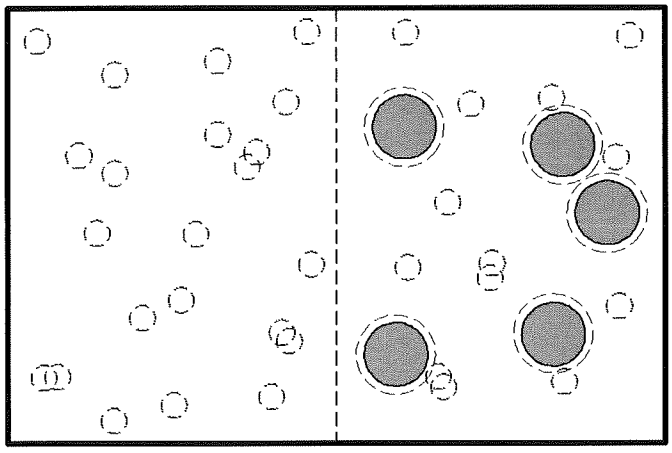


Fig. 14: *Schematic picture of a system containing hard spheres mixed with FOSs held in osmotic equilibrium with a reservoir that only contains FOSs following [33]. System and reservoir are connected through a semi-permeable membrane that is permeable for the FOSs but impermeable for the hard spheres.*

For 'ideal' FOSs we can use again equation (6) for the chemical potential. So, finally equation (12) can be rewritten as:

$$\Omega(N_c,V,T,\mu_{FOS}) = F(N_c,V,T) - \alpha V\, kT\, \rho^{R}_{FOS} \tag{16}$$

Note that the second term of the r.h.s. simply equals $-\Pi V_{free}$. Again there is an analogy with the simple pair potential picture expressed by equation (9); the overlap volume in Fig. 11 is a 'free volume' for the FOSs. It is useful to normalize expression (16) using $\omega = \Omega v_c/kTV$ and (hence) $f = Fv_c/kTV$, where v_c is the volume of a colloidal hard sphere. Using $\rho_{FOS} = 3\phi_{FOS}/4\pi a^3_{FOS}$, equation (16) now becomes:

$$\omega = f - \alpha v_c \rho^{R}_{FOS} = f - \alpha q^{-3} \phi^{R}_{FOS} \tag{17}$$

Note that the relative concentration of FOS particles, ϕ_{FOS}, can exceed unity.

Predicted phase behavior

Now depending on the functional form of ω the system may phase separate into two (or more) phases. In Fig. 15 an osmotic equilibrium is sketched of a phase separated colloidal fluid plus FOSs system, demixed into colloid-enriched (A) and colloid-diluted (B) phases. The binodal points, i.e. the coexistence points, can be computed from imposing the coexistence conditions:

$$\Pi_f = \Pi_s \quad \text{and} \quad \mu_f = \mu_s \tag{18}$$

for the colloidal fluid (f)-to-solid (s) transition. In case the fluid becomes unstable the condition is:

$$\Pi_g = \Pi_l \quad \text{and} \quad \mu_g = \mu_l \tag{19}$$

for the colloidal gas (g)-liquid (l) transition. The coexistence points can be computed using equation (17) by making use of the thermodynamic relationships:

$$\frac{\mu}{kT} = \frac{1}{kT}\left(\frac{\partial \Omega}{\partial N_c}\right)_V = \frac{v_c}{kTV}\left(\frac{\partial \Omega}{\partial \phi}\right)_V = \left(\frac{\partial \omega}{\partial \phi}\right)_V \tag{20}$$

and

$$\frac{\Pi v_c}{kT} = -\frac{v_c}{kT}\left(\frac{\partial \Omega}{\partial V}\right)_{N_c} = -\left(\frac{\partial (\omega V)}{\partial V}\right)_{N_c} = \phi\left(\frac{\partial \omega}{\partial \phi}\right)_{N_c} - \omega \tag{21}$$

valid for fixed polymer concentration in the reservoir.

In case of hard sphere dispersions the Carnahan-Starling equation of state [34] gives an accurate description for f of the fluid phase:

$$f = \phi \ln \phi + \frac{4\phi^2 - 3\phi^3}{(1-\phi)^2} - \phi \tag{22}$$

For the solid phase we use a semi-empirical equation of state of an FCC solid [35],[30],[33]:

$$f = 2.1306\phi + 3\ln\left[\frac{\phi}{1-\phi/\phi_c}\right] \tag{23}$$

In absence of polymers ϕ_{FOS} =0 and using equations (22) and (23) leads to a fluid-solid coexistence containing a fluid phase with a volume fraction of 0.492 and a solid phase with a volume fraction of 0.542, which is actually quite close to the computer simulation results [20] of 0.494 and 0.545, respectively.

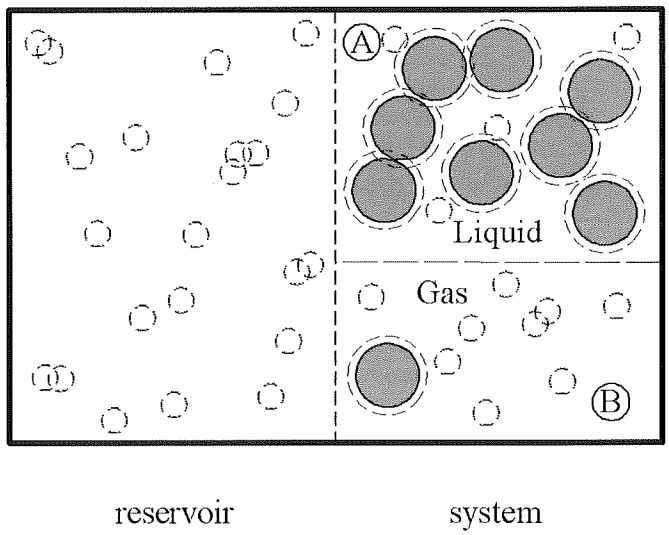


Fig. 15: *Schematic picture of the reservoir-system equilibrium as in Fig. 14 but now for a phase-separated system. In this example A) is the colloidal liquid phase and B) the colloidal gas phase.*

In Fig. 16 we give some examples of calculated phase diagrams for q=0.25 and q=0.6. In the upper left graph the relative FOS concentration in the reservoir is plotted versus the colloid volume fraction for q=0.25. At ϕ_{FOS} =0, the fluid-solid transition at ϕ_c=0.492 and 0.542 is recovered. Increasing the amount of FOSs leads to a broadening of the fluid-solid coexistence. Phase separation takes place significantly below the overlap concentration of FOSs (ϕ_{FOS}=1). By using equation (15) for the free volume fraction $\alpha=V/V_{free}=\phi_{FOS}/\phi^R_{FOS}$, it is possible to calculate the FOS concentration in the reservoir. The resulting binodal curve is plotted in the lower left graph in Fig. 16. The indicated tie-lines (a tie-line connects two binodal points) demonstrate the partitioning of the FOSs over the two phases. At lower colloid volume fractions there is more space for the FOSs, so more free volume and α will be larger as compared to the higher volume fractions in the colloidal solid phase, where α is very small. The fluid phase does not become unstable below the binodal curve; there is only a *metastable* gas-liquid transition, which will be addressed in section 4.2.

In the upper right plot of Fig. 16 the result for the binodal in the ϕ^R_{FOS} vs ϕ representation is given for q=0.6. Again, the fluid-solid coexistence is found at ϕ_{FOS} =0, but the broadening of this coexistence is much less as compared to the low q-case. For this size ratio the fluid becomes unstable and phase separates into a gas-liquid equilibrium below the fluid-solid equilibrium curve. A critical point is found at $\phi=\phi_c$ = 0.188. Above ϕ^R_{FOS}=0.49 the fluid is

unstable. At ϕ^R_{FOS} =1.33 there is a triple line where three phases, colloidal gas, liquid and solid phases coexist. For larger values of ϕ^R_{FOS} there is a gas-solid coexistence with a very diluted colloidal gas phase. Again using α the phase diagram in the ϕ_{FOS} (system concentration) versus ϕ representation can be calculated. The phase diagram is dominated by the gas-liquid coexistence while it is noted that liquid-solid, gas-liquid-solid and gas-solid coexistence takes place at large colloid volume fractions and large FOS concentrations. In section 4.1 we compare the results for colloidal gas-liquid (or so-called 'liquid-liquid' phase separation) with experimental data on protein-polymer mixtures that phase separate into two liquid phases.

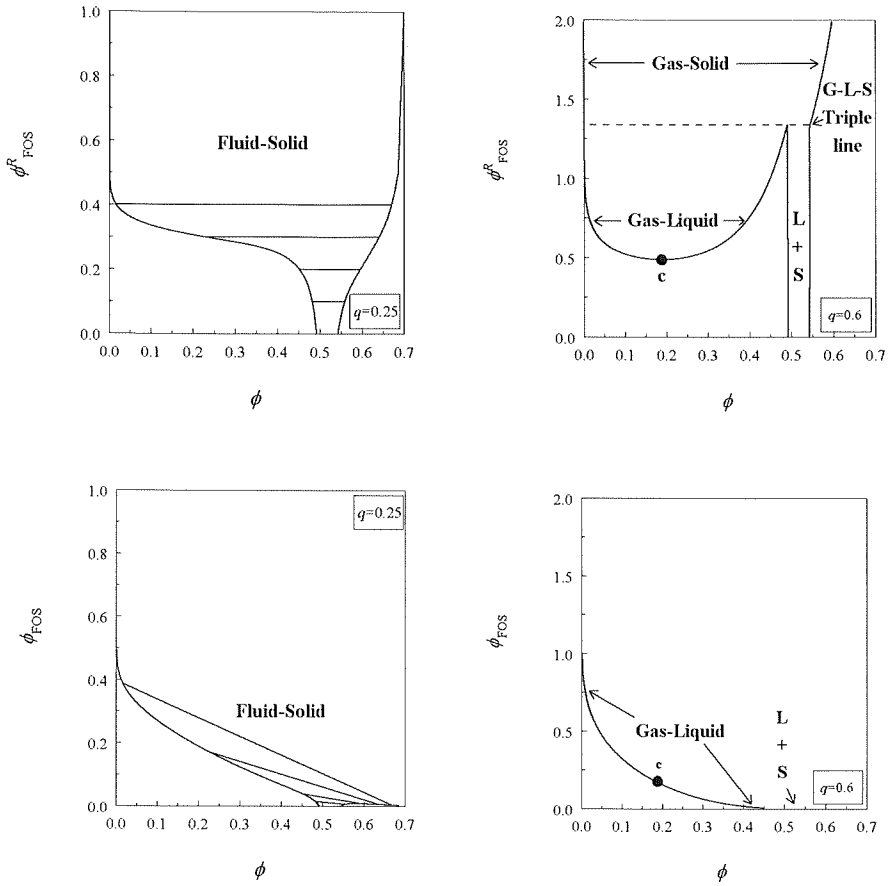


Fig. 16: *Phase diagrams for FOS-hard sphere mixtures calculated using the osmotic equilibrium theory. Left: results for q=0.25 with FOSs reservoir (upper) and system (lower) volume fractions versus the colloid volume fraction. Right: results for q=0.6 with FOSs reservoir (upper) and system (lower) volume fractions versus the colloid volume fraction.*

4 Phase behavior of protein solutions

4.1 Protein-polysaccharide mixtures

As addressed in the introduction, in many protein-polysaccharide mixtures the depletion interaction between the protein particles due to the polysaccharides leads to a phase separation above a certain concentration of polysaccharides at fixed protein concentration. Here we give some phase diagrams of such systems and compare them with the phase diagrams as calculated with a mixture of hard spheres and freely overlapping spheres (FOSs).

Firstly, we redraw some liquid-liquid phase lines of a phase-separated system of casein micelles mixed with guar gum [36] in Fig. 16 (left), where the guar concentration is plotted versus the casein micelle volume fraction. The polysaccharide guar gum, produced by the plant *Cyamopsis tetragonoloba*, is a food thickener and belongs to the group of galactomannans, of which tara gum and locust bean gum are also well-known. The guar gum solution was mixed with casein micelles. Casein micelles are association colloids (see section 2.2). By preparing skim milk it is possible to produce a stable suspension with casein micelles as the main component. Skim milk can be regarded as a suspension of casein micelles with a volume fraction of ~ 10%. The large amount of other but very small (< 5 nm) components (salts, lactose and whey proteins) is considered to belong to the continuous phase since mixtures of guar and skim milk without casein micelles are stable in all proportions [36]. Guar can easily be depolymerized by chemical hydrolysis. This hydrolysis occurs at low pH and high temperatures and can be used to prepare a series of guar gums varying in molar mass. So, a series of guar gums was prepared by degradation in that way [36].

In Fig. 17 (left) the phase lines are plotted for casein micelles mixed with native guar gum (molar mass M= 1260 kg/mol, radius of gyration R_g=109 nm), guar gum heated for 10 minutes at pH 1.5 (M= 564 kg/mol, R_g=73 nm) and heated for 30 minutes at pH 1.5 (M= 184 kg/mol, R_g=35 nm).

Mixing guar gum with casein micelles led to a phase separation [36]. This phase separation originates from a depletion interaction leading to an effective attraction between the casein micelles by non-adsorbing guar. As can be observed, the polymer concentration at the phase boundary increases with decreasing size of the guar gum polymer chains.

In Fig. 17 we also (right graph) plot the result from the osmotic equilibrium theory for FOSs plus hard spheres (see section 3.3) for $a_{FOS}=R_g$, so for a_{FOS}/R=1.09 (native), 0.73 (10 mins) and 0.35 (30 mins) (the radius of the casein micelles R is 100 nm). The trend in the results as well as the order of magnitude for the guar concentration required to induce phase separation agrees with the experimental results: a larger value for R_g/R gives a phase transition at smaller polymer/FOS concentrations. Taking into account that guar behaves as a polymer in a good solvent in aqueous solutions, the location of the phase boundary is predicted reasonably well. When the guar gum chains are made shorter (degradation), polysaccharides are somewhat less effective as a thickener. At the same time the phase boundary shifts to higher guar concentrations when the chains are shorter, so that a higher guar concentration can be used without inducing a phase separation. So, preventing a phase separation and increasing the viscosity, which is the role of thickeners such as guar in food dispersions, are opposite constraints.

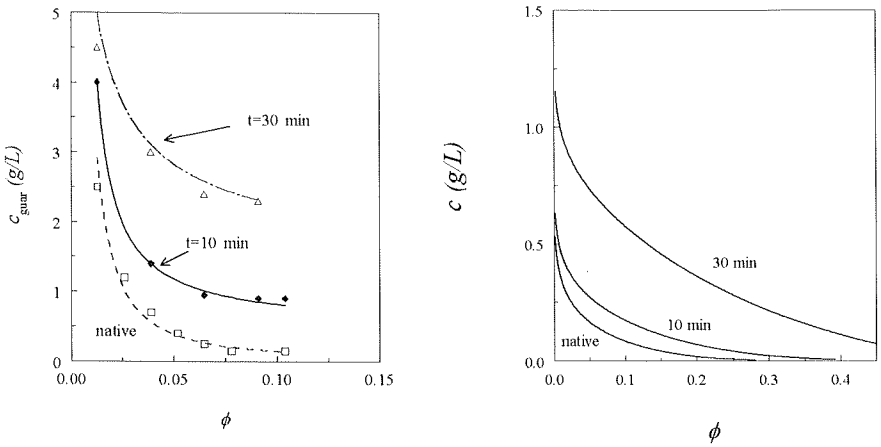


Fig. 17: *Experimental phase boundaries of guar gum-casein micelle mixtures [36] as derived from visual observation for three heating times of three guar samples as indicated in the plot (left). Right: calculated phase boundaries from the free volume theory (see section 3.3). The curves were calculated from the radii of gyration and molar masses given in the text and a radius of the casein micelles of 100 nm.*

Next, we compare the results of a mixture of casein micelles and an exocellular polysaccharide with the free volume theory. The exocellular polysaccharide was produced by a lactic acid bacterium and has a radius of gyration of 86 nm and a number-averaged molar mass of 1470 kg/mol [37]. In an aqueous solution with 0.10 M salt this polymer behaves as in a good solvent. The limiting concentrations establishing the phase diagram of the mixture of casein micelles and the exocellular polysaccharide (EPS) [38] are redrawn as the filled diamonds in Fig. 18. The dashed curve is the result for hard spheres plus FOSs for q=0.86. This result is reasonably close to the experimental result. The experimental critical colloid volume fraction is expected to be located at a colloid volume fraction above 0.13 since at all investigated casein micelle concentrations (up to a volume fraction of 0.13) the volume of the casein micelle rich phase was always smaller than that of the EPS-rich phase. The theoretical critical point is indicated by the dot. We have also plotted the osmotic equilibrium theory result for polymers in the excluded volume limit mixed with hard spheres for q=0.86 (for details see [30],[32]) as the full curve. It follows that the measured phase line lies in between the results for fully interacting and non-interacting polymer chains. It follows that in practice many systems will phase separate in the range between the predictions for the ideal and excluded volume limit.

This could be expected since the effective Flory-Huggins parameter χ is still close to 0.5 (0.48 for dextran) for polysaccharides in water, whereas χ=0 for polymer chains in the full excluded volume limit. Still, the FOS model gives a good first order of magnitude prediction of the phase diagrams of polysaccharides plus (spherical) protein mixtures. It is meanwhile realized that the FOS model is very simplified as compared to real polymer chains.

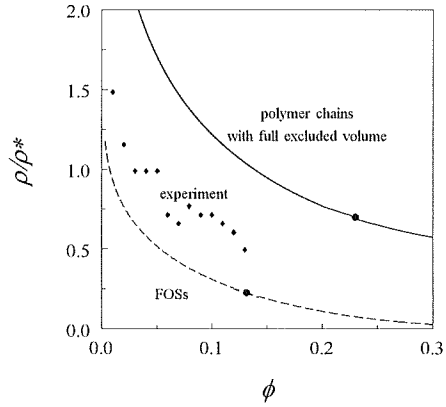


Fig. 18: *Phase diagram of an exocellular polysaccharide- casein micelle mixture with q=0.86 [38] in comparison with the free volume predictions for FOSs (dashed curves; $\rho/\rho^* = \phi_{FOS}$) and polymer chains in the excluded volume limit [30] (full curves).*

4.2 The role of polymer chains in crystallization

Adding non-adsorbing polymers to protein solutions enhances crystallization. For instance Polson *et al.* [39] reported this already several decades ago. Obtaining good quality crystals is essential for the determination of the protein structure using X-ray diffraction. Knowledge of the protein structure helps to elucidate the biochemical role of the protein. Until the 1960s only a few protein crystal forms were reported in scientific Journals per year. In the 1970s and 80s this number has increased strongly, while since the 90s more than hundred protein crystal forms are being published annually [40]. Attempts to induce crystallization using precipitating agents such as organic solvents and adding salt were commonly used until, say, 1980. Later it was realized how efficient adding polymers as precipitating agents is, so it became more popular. Flexible polymer chains, such as polyethylene glycol (PEG) are known to be effective crystallization promoters in protein solutions of, for instance, lysozyme [41] and leghemoglobin [42]. Scattering studies show that adding non-adsorbing polymers to protein dispersions induces an effective attraction between the protein molecules that can be described as a depletion interaction [43], which is the additional attraction between the proteins that enhances crystallization.

In order to understand why the polymers help crystallization it is useful to investigate the conditions under which protein crystallization is enhanced. Helpful in that sense is the molecular dynamics computer simulation study of Ten Wolde and Frenkel [44]. They studied the crystallization of a collection of spherical particles interacting through a generalized Lennard-Jones-like pair interaction. The phase diagram of their system is qualitatively comparable to the phase diagram in Fig. 19 (left): a fluid-solid coexistence with a 'hidden' gas-liquid coexistence curve inside the unstable region. From the simulation results it appeared that good crystals only form when a quench into the metastable gas-liquid region is made that lies near the metastable critical point, indicated by the dot in Fig. 19. The pathway for the nucleation of crystals changes dramatically in the neighborhood of the metastable

critical point, mainly due to a significantly lowered free-energy barrier [44]. Far from the critical point a quench into the unstable gas-solid region leads to a phase separation with random 'solid' aggregates instead of crystals due to the fact that the energy barrier for nucleation is too high. The current picture is that the mixture first phase separates into two coexisting colloidal fluid phases ('gas-liquid'), out of which the crystals grow in the concentrated liquid phase. The optimum phase diagram is thus a diagram with a 'hidden' gas-liquid curve *just below* the solubility curve or fluid-solid line. In that situation one easily quenches into the metastable gas-liquid region, preventing the possibility of the formation of random aggregates. A phase diagram that should be close to optimum crystallization conditions, according to the criteria given above, is plotted in Fig. 19 (left). Also in Fig. 19 we have plotted the phase diagram for a hard sphere/FOS-mixture, but with the inverse of the FOS reservoir concentration versus the colloid volume fraction (compare with the upper left graph of Fig. 16) for q = 0.25 (right). This plot qualitatively closely resembles the result for optimum crystallization conditions, the inverse FOS or polymer concentration playing the role of the temperature in the Lennard-Jones system or protein solution.

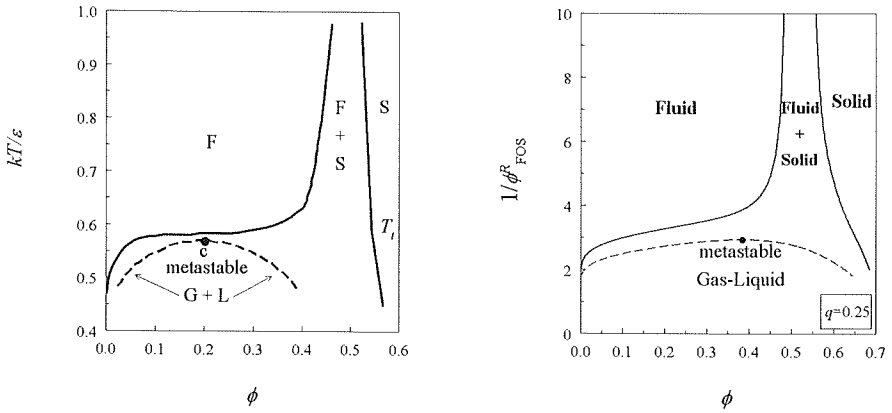


Fig. 19: *Sketch of the phase diagram of a Lennard-Jones fluid close to optimum crystallization conditions (left) and the phase diagram of a FOS-colloid mixture with q = 0.25 where the inverse FOS concentration in the reservoir is plotted versus the colloid volume fraction (right). Note the analogy between the diagrams.*

This analogy indicates that by adding polymers one could shift the phase diagram in the direct of optimum crystallization conditions. For instance, if the metastable gas-liquid coexistence would be located very far below the solubility curve, adding polymers would then shift the gas-liquid region much closer to the solubility curve.

As to illustrate this concept of the role of non-adsorbing polymer chains on the protein solution phase behavior, we will discuss the results of adding the polymer PEG to a solution with the protein apoferritin, studied by Tanaka and Ataka [45]. Apoferritin is an iron storage protein consisting of 24 subunits. The effective radius is about 8 nm and the molar mass of apoferritin is 440 kg/mol. Inside the molecule there are cavities for iron ions, see also the schematic picture in Fig. 20.

It is not easy to crystallize a solution of apoferritins by adding salt ions. Traditionally, a well-defined scale, known as the 'Hofmeister series', is used as a measure for the efficiency of precipitating proteins. Except when using Cd-ions, a solution of apoferritin cannot be crystallized in the common manner with the usual salt ions as precipitating agents. Adding PEG however does make it possible to induce crystallization.

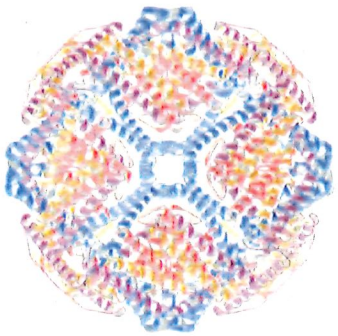


Fig. 20: *Schematic picture of the structure of the amino acid groups in the protein apoferritin. Note the different subunits.*

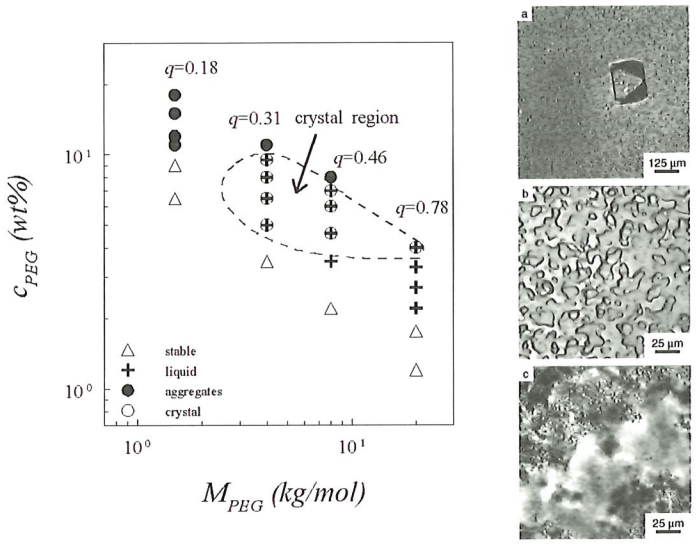


Fig. 21: *Phase behavior of apoferritin with PEG The apoferritin concentration was kept constant at 54 g/L and the molar mass and concentration of PEG was varied as indicated in the diagram. Results are redrawn from [45]. The micrographs (right) represent the various kinds of unstable solutions that were found: A) crystals, B) liquid domains and C) random aggregates. Micrographs are copyright (2002) by The Journal of Chemical Physics from ref. [45].*

In Fig. 21 the redrawn experimental data were obtained from visual and microscopic evaluation of PEG-apoferritin mixtures in aqueous 0.6 M NaCl solutions. The concentration of apoferritin was fixed at 54 g/L. The PEG concentration and molar mass were varied.

Four molar masses M (and radii of gyrations R_g) of the PEGs were used: 1.5 (1.4 nm), 4.0 (2.5 nm), 8.0 (3.7 nm) and 20 (6.2 nm) kg/mol, giving $q = a_{FOS}/R = R_g/R = 0.18$, 0.31, 0.46, and 0.78. Four kinds of situations were observed after mixing PEG with apoferritins [45]. For sufficiently small concentrations (depending on M, hence q), the mixture was stable (triangles in Fig. 21). Above a certain concentration a phase transition took place. At $q = 0.18$ random aggregates (filled circles) were found, which is typical for a protein solution undergoing a fluid-to-solid transition, but where there are no proper conditions for the creation of good quality crystals. This was also found for the highest concentrations at $q = 0.31$ and 0.46. For $q = 0.31$, 0.46 and 0.78 there was a region where liquid domains (crosses) were formed, indicative for a 'gas-liquid' phase transition, usually referred to as 'liquid-liquid' phase separation. For $q = 0.78$ liquid domains were found in the entire unstable regime. Finally, good quality crystals (open circles) were formed at $q = 0.31$ and 0.46 for not too high PEG concentrations.

It thus follows that adding PEG indeed allows providing conditions for good crystallization within a specific range of protein-polymer size ratios and polymer concentrations. In Fig. 19 the result for a polymer-colloid mixture with $q = 0.25$ was given. For a colloid-polymer mixture this appears to be close to the 'good crystallization' condition: the critical point is close to the 'solubility' fluid-solid curve but lies still within the unstable region (a metastable 'liquid-liquid' phase separation precedes crystallization). For larger size ratios up to $q = 0.3$ (above which there is a triple point and there is a stable gas-liquid transition) the FOS-hard sphere model predicts that the critical point moves even closer to the solubility curve. For smaller values of q, the liquid-liquid region shifts to temperatures far below the solubility curve, making it harder to crystallize. This explains why it is harder to obtain good crystals with $q = 0.18$ in the protein-PEG mixture.
Some deviations are expected between FOS-colloid and protein-polymer mixtures since a FOS particle is of course different from a PEG polymer molecule and one should insert polymer-polymer interactions (which may shift the triple point to values larger than q) to give a better description of the effects of real polymer chains. Further, the equation of state of the protein solution may deviate, depending on the conditions, from the equation of state of the dispersion of hard sphere colloids.

5 Summary

The segregative interaction and resulting phase behavior in aqueous protein-polymer mixtures has been addressed. It is argued that the physical properties of protein dispersions can be described using colloid physics theories as is shown for the diffusion, scattering and phase behavior of protein dispersions. That allows to treat a (bio)polymer-protein mixture, where the particles interact segregatively, as a colloid-polymer mixture where the polymers are depleted from the colloid surfaces. In its simplest form the depletion induced pair interaction between colloids mediated by polymers, as well as the resulting phase behavior, are presented and compared to measurements on biological systems. The analogy between protein solution and colloidal or atomic phase behavior helps to understand protein crystallization, and the role of non-adsorbing polymers as protein crystallization promoters.

F. Weinbreck (NIZO food research, The Netherlands), S. Wiegand and G.A. Vliegenthart (both at Jülich) are acknowledged for useful suggestions and comments.

A Appendix: Generalized Gibbs adsorption equation

In case of adsorption of species (depletion = negative adsorption) it is useful to define the grand potential J:

$$J = F - \sum_i \mu_i N_i \tag{A1}$$

where F is the Helmholtz free energy, μ_i is the chemical potential of species i, and N_i is the number of particles i. Now we fix temperature and volume, so changing the distance between the colloidal particles h is related to a change in the Helmholtz free energy given by:

$$dF = -Kdh + \sum_i \mu_i dN_i \tag{A2}$$

where K is the force required to move the particles. Combining equations (A1) and (A2) relates the grand potential J to the force between the particles:

$$dJ = -Kdh - \sum_i N_i d\mu_i \tag{A3}$$

In case of one species q adsorbs or is depleted, cross-differentiation leads to:

$$\left(\frac{\partial K}{\partial \mu_q}\right)_h = \left(\frac{\partial N_q}{\partial h}\right)_{\mu_q} \tag{A4}$$

Classical mechanics teaches us that the interaction potential is equal to integration over the force-distance curve, hence:

$$K = -\left(\frac{\partial W}{\partial h}\right)_{\mu_q} \tag{A5}$$

Combining equations (A4) and (A5) yields the generalized Gibbs adsorption equation:

$$-\left(\frac{\partial W(h)}{\partial \mu_q}\right)_h = N_q(h) - N_q(h=\infty) \tag{A6}$$

that relates the interaction potential to the excess amount of particles when two particles are brought together at a distance h.

B Appendix: Free volume fraction α from scaled particle theory

The chemical potential of a FOS particle can be written as:

$$\mu_{\mathrm{FOS}} = \mu_{\mathrm{FOS}}{}^{*} + kT \ln \frac{N_{\mathrm{FOS}}}{V} + w = \mu_{\mathrm{FOS}}{}^{*} + kT \ln \frac{N_{\mathrm{FOS}}}{V_{\mathrm{free}}} \tag{B1}$$

where $\mu_{\mathrm{FOS}}{}^{*}$ is a standard constant chemical potential ($kT \ln \lambda^3$ for hard spheres, where λ is the de Broglie wavelength) and w is the work required to bring a FOS particle from the reservoir containing only FOSs to the system with hard spheres. Using equations (B1) and (14) it follows that the free volume fraction α is related to the work w through:

$$w = -kT \ln \alpha \tag{B2}$$

For very small FOSs the free volume fraction is just unity minus the sum of the overall hard sphere volumes plus FOS depletion layers:

$$\alpha_{\text{small FOSs}} = 1 - \frac{4\pi}{3} \rho_c (R + a_{\mathrm{FOS}})^3 \tag{B3}$$

so one may write:

$$w_{\text{small FOSs}} = -kT \ln \left[1 - \frac{4\pi}{3} \rho_c (R + a_{\mathrm{FOS}})^3 \right] \tag{B4}$$

where ρ_c is the number of colloids per volume N_c/V ($\phi = 4\pi R^3 \rho_c/3$). In the opposite limit of very large FOSs the work to place a FOS particle in the system is just the volume of a FOS times the osmotic pressure due to the colloidal hard spheres Π_c:

$$w_{\text{large FOSs}} = \frac{4\pi}{3} a_{\mathrm{FOS}}^3 \Pi_c \tag{B5}$$

One may now propose a series expansion for w in x:

$$w(x) = w(x=0) + x \left(\frac{\partial w}{\partial x} \right)_{x=0} + \frac{1}{2} x^2 \left(\frac{\partial w}{\partial x} \right)^2_{x=0} + \frac{4\pi}{3} \left(x a_{\mathrm{FOS}} \right)^3 \Pi_c \tag{B6}$$

where x is a scaling variable of the size a_{FOS}. The last term is from the large FOS limit of equation (B5). By realizing that equation (B4) can be written as:

$$w(x) = -kT \ln \left[1 - \frac{4\pi}{3} \rho_c (R + x a_{FOS})^3 \right] \tag{B7}$$

the first three terms on the r.h.s. in equation (B6) can be calculated. The pressure of the hard sphere fluid in the fourth term can be calculated using the Percus-Yevick equation of state (which is consistent within the SPT approach):

$$\frac{\Pi_c}{\rho_c kT} = \frac{1+\phi+\phi^2}{(1-\phi)^3} = \frac{1}{1-\phi} + \frac{3\phi}{(1-\phi)^2} + \frac{3\phi^2}{(1-\phi)^3} \tag{B8}$$

Then one finds from equation (B6) using equations (B7) and (B8):

$$\frac{w}{kT} = -\ln[1-\phi] + \frac{\phi}{1-\phi}A + \frac{\phi^2}{(1-\phi)^2}B + \frac{\phi^3}{(1-\phi)^3}C \tag{B9}$$

With A, B and C defined below equation (15). Substitution in equation (B2) then yields equation (15).

References

[1]H. Walter, D.E. Brooks, *FEBS Lett.* **361** (1995) 135.
[2]V.B. Tolstoguzov, *Food Hydrocoll.* **4** (1991) 429.
[3]J.-L. Doublier, C. Garnier, C. Renard, C. Sanchez, *Curr. Opin. Coll. Int. Sci.* **5** (2000) 184.
[4]S.L. Turgeon, M. Beaulieu, C. Schmitt, C. Sanchez, *Curr. Opin. Colloid Interface Sci.*, **8** (2003) 401.
[5]J.Janzen, D.E. Brooks, *Clin. Hemorheol.* **9** (1989) 695, S.E. Bedell, B.T. Booker, *Am. J. Med.* **78** (1985) 1001; B. Neu, H.J. Meiselman, *Langmuir* **17** (2001) 7973.
[6]A.P. Minton, *Biophys. Chem.* **57** (1995) 1.
[7]E. Scholten, R. Tuinier, R.H. Tromp, H.N.W. Lekkerkerker, *Langmuir* **18** (2002) 2234.
[8]R. Tuinier, J. Rieger, C.G. de Kruif, *Adv. Colloid Interface Sci.* **103** (2003) 1.
[9]J. Perrin, *Die Atome*, Steinkopff, Dresden, 1914.
[10] J.K.G. Dhont, chapter A.2 in this book.
[11] I. Bodnár, J.K.G. Dhont, H.N.W. Lekkerkerker, *J. Phys. Chem.* **100** (1996) 19614.
[12] J.-P. Hansen, I. R. McDonald, *Theory of Simple Liquids*, Academic Press, San Diego, 1986.
[13] C.G. de Kruif, R. Tuinier, C. Holt, P.A. Timmins, P.A., H.S. Rollema, *Langmuir* **18** (2002) 4885.
[14] G. Nägele, chapter B.16, in: 'Soft Matter: complex materials on mesoscopic scale,' 33rd IFF winter school, J.K.G. Dhont, G. Gompper, D. Richter, Eds. (ISBN 3-89336-297-5) 2002.
[15] A.Vrij, *J. Colloid Interface Sci.* **90** (1982) 110.
[16] C.G. de Kruif, *Langmuir* **8** (1992) 2932.
[17] J.K.G. Dhont, Chapter 6 in : *An introduction to Dynamics of Colloids*, Elsevier, Amsterdam, 1996.
[18] R. Piazza, *Phys. Rev. E* **58** (1998) R2733.
[19] for an explanation of the Van der Waals model, see e.g. B. Widom, *Science* **157** (1967) 375.
[20] B.J. Alder, T.E. Wainwright, *J. Chem. Phys.* **27** (1957) 1208; W.G. Hoover, F.H. Ree, *J. Chem. Phys.* **49** (1968) 3609.
[21] G.A. Vliegenthart, J.F. Lodge, H.N.W. Lekkerkerker, *Physica A*, **263** (1999) 378.
[22] M.L. Broide , C.R. Berland, J. Pande, O. Ogun, G.B. Benedek, *Proc. Natl. Acad. Sci. USA* **88** (1991) 5660; C.R. Berland, G.M. Thurston, M. Kondo, M.L. Broide, J. Pande, Ö. Ögun, G.B. Benedek, *Proc. Natl. Acad. Sci. USA* **89** (1992) 1214; N. Asherie, A. Lomakin, G.B. Benedek, *Phys. Rev. Lett.* **77** (1996) 4832.
[23] F. Weinbreck, R. de Vries, P. Schrooyen, C.G. de Kruif, *Biomacromolecules* **4** (2003) 293.
[24] S. Asakura, F. Oosawa, *J. Chem. Phys.* **22** (1954) 1255.
[25] S. Asakura, F. Oosawa, *J. Polym. Sci.* **33** (1958) 183.
[26] A. Vrij, *Pure Appl. Chem.* **48** (1976) 471; H. de Hek, A. Vrij, *J. Colloid Interface Sci.* **84** (1981) 409.
[27] D.G. Hall, *J. Chem. Soc. Faraday Trans.* **68** (1972) 2169.
[28] E. Eisenriegler, *J. Chem. Phys.* **79** (1983) 1052.
[29] E. Eisenriegler, chapter B.10, in: 'Soft Matter: complex materials on mesoscopic scale,' 33rd IFF winter school, J.K.G. Dhont, G. Gompper, D. Richter, Eds. (ISBN 3-89336-297-5) 2002.
[30] D.G.A.L. Aarts, R. Tuinier, H.N.W. Lekkerkerker, *J. Phys.: Condens. Matter* **14** (2002)

7551.
[31] R. Verma, J.C. Crocker, T,C, Lubensky, A.G. Yodh, *Macromolecules* **33** (2000) 177.
[32] R. Tuinier, D.G.A.L. Aarts, H.H. Wensink, H.N.W. Lekkerkerker, *Phys. Chem. Chem. Phys.* **5** (2003) 3707.
[33] H.N.W. Lekkerkerker, *Colloids Surf.* **51** (1990) 419; H.N.W. Lekkerkerker, W.C.K. Poon, P.N. Pusey, A. Stroobants, P.B. Warren, *Europhys. Lett.* **20** (1992) 559
[34] N.F. Carnahan, K.E. Starling, *J. Chem. Phys.* **51** (1969) 635.
[35] W.W. Wood, *J. Chem. Phys.* **20** (1952) 1334.
[36] R. Tuinier, E. ten Grotenhuis, C.G. de Kruif, *Food Hydrocoll.* **14** (2000) 1.
[37] R. Tuinier, P. Zoon, C. Olieman, M.A. Cohen Stuart, G.J. Fleer, C.G. de Kruif, *Biopolymers* **49** (1999) 1.
[38] R. Tuinier, C.G. de Kruif, *J. Chem. Phys.* **110** (1999) 9296.
[39] A. Polson, G.M. Potgieter, J.F. Largier, J.F. Joubert, G.E.F. Mears,, *Biochim. Biophys. Acta* **82** (1964) 463.
[40] A. Roussel, L. Serre, M. Frey, J.C. Fontecillacamps *J. Cryst. Growth* **106** (1990) 405.
[41] A.M. Kulkarni, A.P. Chatterjee, K.S. Schweizer, C.F. Zukoski, *J. Chem. Phys.* **113** (2000) 9863.
[42] Ya. D. Devedzhiev, N. N. Kudryavtseva, L. I. Borodenko, O.S. Genina, B.P. Atanasov, *Sov. Phys. Crystallogr.* **35** (1990) 292.
[43] R. Tuinier, J.K.G. Dhont, C.G. de Kruif, *Langmuir* **16** (2000) 1497; A. Tardieu, S. Finet, F. Bonneté, *J. Cryst. Growth* **232** (2001) 1; S. Finet, A. Tardieu, *J. Cryst. Growth* **232** (2001) 40.
[44] P.R. ten Wolde, D. Frenkel, *Science* **277** (1997) 1975; *Theor. Chem. Acc.* **101** (1999) 205.
[45] S. Tanaka, M. Ataka, *J. Chem. Phys.* **117** (2002) 3504.

C 5 The protein glass transition

Vandana Kurkal and Jeremy C. Smith
Interdisciplinary Center for Scientific Computing
Universität Heidelberg, Heidelberg

Contents

1 Introduction 2

2 Glassy behaviour of proteins 2

3 Detecting the dynamical transition 3
3.1 Mössbauer spectroscopy 3
3.2 X-ray diffraction 4
3.3 Neutron Scattering 4
3.4 Molecular Dynamics (MD) Simulation 5

4 Dynamical transition and protein function 5

5 Dynamical transition and neutron frequency windows 7

6 Effect of solvent on dynamical transition 7
6.1 The role of protein-water hydrogen bond 7
6.2 Translational hydration water dynamics drives dynamical transition 9

7 Principal components of the protein dynamical transition 12
7.1 Principal component analysis 12
7.2 Free energy landscapes 13
7.3 Harmonic, Quasi-harmonic and multiminima modes 13

8 Conclusions 15

1 Introduction

Proteins exhibit a dynamical transition at $\approx$ 150 K - 220 K involving deviation at higher temperatures from linear dependence on temperature of the average mean-square displacement, $\langle u^2 \rangle$ of the protein atoms. Several techniques have demonstrated the existence of this transition, including Mössbauer spectroscopy [1–3], neutron scattering [4–13] and X-ray crystallography. [14–20] For atleast some proteins, there exists a correlation between the dynamical transition and protein function. For example myoglobin [21], elastase [22], ribonuclease [23] and bacteriorhodopsin [9] are inactive below the transition temperature. This suggests that the motions activated at the transition may be of functional importance. Hence, a deeper understanding of the relationship between dynamics and activity of proteins can be gleaned by characterising the dynamical transition.

Proteins behave like glasses at low temperature. The glassy nature of proteins is described in the next section followed by a description of various experimental and theoretical work on the protein glass transition. The remainder of the chapter deals with the relationship between the dynamical transition and protein function/activity, the role of solvent in the dynamical transition and a principal component analysis of computer simulations of the dynamical transition in proteins.

2 Glassy behaviour of proteins

The liquid-to-glass transition occurs when lowering the temperature results in system becoming trapped in local minima on the energy landscape. The liquid loses its ability to flow [24] and appears as a "structurally disordered solid" [25]. A similar phenomenon takes place in proteins. The atoms are trapped in local minima below the glass transition but possess enough kinetic energy to overcome barriers so as to explore the surrounding local minima above the dynamical transition temperature [24–27].

Mode Coupling theory (MCT) distinguishes two kinds of motions in glass-forming liquids: a very fast secondary β relaxation processes due to local rearrangements involving low energy barriers and a more co-operative α relaxation process corresponding to slow diffusive motions [24, 27–29]. The β processes follow the standard Arrhenius temperature dependence whereas the α processes are strongly hindered below the transition temperature, the relaxation following a stretched exponential behaviour [26, 30].

A key characteristic feature observed in glass-forming substances, that is absent in proteins, is a well-defined glass transition temperature, T_g which is the temperature at which there is a jump in the heat capacity [26] C_p. However, proteins do seem to conform to MCT, usually a sign of fragile glass-forming systems, and the jump in C_p may be gradually spread over a range of temperatures. This may be due to the presence of several α processes that are initiated at different temperatures.

The experimental and theoretical evidence for the glass transition in proteins suggests that hydrated proteins as single molecules may possess the complexity necessary for cooperative dynamics comparable to that present in simple glass-forming systems [26]. Experimental and theoretical studies on protein dynamical transition will be dealt in with some detail in the next section.

3 Detecting the dynamical transition

The dynamical transition has been observed using variety of spectroscopic and scattering experiments [1–20] as well as molecular dynamics (MD) simulations [31–40]. These methods will now be discussed. Experimental techniques detecting the dynamical transition measure different quantities. For example Mössbauer spectroscopy [1–3], detects the motion of a single probe nucleus whereas neutron scattering [4–13] results from the dynamics of more globally distributed probes. Moreover, the timescales to which the measurements are sensitive also vary. Despite these differences, it has been observed with several techniques that there exists a dynamical transition in proteins in the temperature range $\approx$ 180 - 220 K.

3.1 Mössbauer spectroscopy

In Mössbauer spectroscopy, the Mössbauer nucleus introduces a time threshold that filters out dynamical processes which are slower than the lifetime of the nucleus [41]. Dynamical processes with relaxation times close to the lifetime of the Mössbauer nucleus broaden the narrow absorption lines and the time sensitivity of the shape, and the intensity of the Mössbauer spectra is more pronounced just above the time threshold. Dynamical processes with relaxation times which are in the time window corresponding to the lifetime of the nucleus give rise to additional broad spectral lines. These can be detected only if they originate from a distinct dynamic processes [42,43].

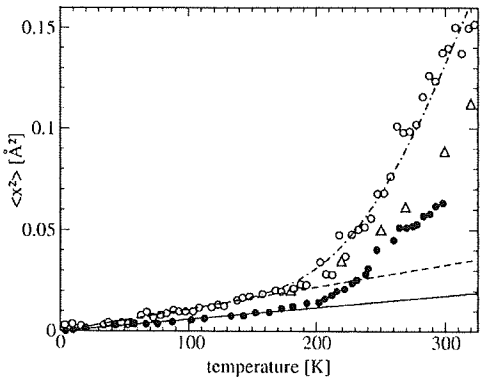


Fig. 1 : *Mean square displacement from incoherent neutron scattering (from Ref. [6]);* ○ *: time resolution 80 ps;* △ *: time resolution 12 ps.* • *: Mean square displacements from Mössbauer absorption spectroscopy on Myoglobin crystals (from Ref. [7]).*

Mössbauer spectra of proteins show a characteristic temperature dependence. The intensity of the absorption lines dramatically decreases [1, 42, 44–46] and their width increases [45] with increasing temperature. $\langle u^2 \rangle$ of a particular nucleus can be obtained from Mössbauer spectroscopy by using the relation connecting Lamb-Mössbauer factor (f) and the $\langle u^2 \rangle$

$$f = exp[-k^2 \frac{\langle u^2 \rangle}{3}]$$

where $k = 2 * \pi/\lambda$ and $\lambda = 0.86$ Å for $^{5}7$Fe.

A transition in $\langle u^2 \rangle$ of the iron site of myoglobin at around 180 - 200 K was reported [7, 47]. Mössbauer spectroscopic data [7] are shown as filled circles in Fig. 1. As can be seen, the $\langle u^2 \rangle$ exhibits a linear behaviour below $\approx$ 200 K, and then shows a continuous increase in the slope with increasing temperature.

3.2 X-ray diffraction

X-Ray crystallographic refinement [14] of proteins shows that there is a spread of electron density around their average structure [14]. This is quantified by the Debye-Waller factors of individual atoms which are determined by the mean-square fluctuations. A dynamical transition in Ribonuclease A was has been claimed to be seen at 180 - 200 K using this effect [15]. X-ray diffraction study on crystalline myoglobin [48] did not result in an inflexion in the $\langle u^2 \rangle$ at the transition temperature. The $\langle u^2 \rangle$ obtained from crystallographic thermal factors represents a static distribution due to the coherent nature of X-ray diffraction and is thus very different from the $\langle u^2 \rangle$ values obtained from incoherent spectroscopies that measure time correlations [48]. Hence, if throughout the temperature range examined, the same set of conformations are occupied, then the dynamic transition will not be observed in X-ray diffraction, but still can be observed in the incoherent spectroscopic experiments. The linear temperature-dependence of $\langle u^2 \rangle$ in crystallographic myoglobin would then correspond to all substates being equally occupied at all temperatures. On the other hand, if there exists a possibility that some substates become vacant at low temperatures, then a dynamical transition should also be observed using X-ray diffraction. Also, for a dynamical transition to occur in the X-ray diffraction studies, abrupt changes in the average positions should occur at the transition temperature. When the temperature is increased, co-ordinate shifts may be observed leading to the reduction of the population of the position occupied at lower temperatures.
A study of human lysozyme was performed using the normal mode refinement method [49] and a clear inflexion in $\langle u^2 \rangle$ as a function of temperature was demonstrated. It was suggested that the presence of the dynamical transition might be related to the formation of cubic ice in the large solvent channels of the crystal. In contrast, myoglobin shows no apparent solvent channel. Thus, the difference in size of the solvent channel might explain why there exists a difference in the behaviour of the temperature dependence of the crystal diffraction between lysozyme and myoglobin.
Recent studies on ligand rebinding kinetics in myoglobin [19, 20], with a combination of crystallographic refinement and Fourier transform infrared spectroscopy showed that below 180 K, photodissociated ligands migrate to specific sites within an internal cavity of an essentially immobilized frozen protein from which they subsequently rebind by thermally-activated barrier crossing. On photodissociation above 180 K ligands can escape from these sites aided by protein fluctuations that transiently open exit channels.

3.3 Neutron Scattering

Neutron scattering is one of the major tools for studying dynamical transition. Neutron scattering experiments probe specifically the motion of hydrogen atoms as these have an anomolously large incoherent neutron-scattering cross-section. Since hydrogen atoms are abundant and uniformly distributed in proteins, this method offers a global view of protein dynamics. The first neutron scattering experiment reporting dynamical transition behaviour was carried out on hydrated myoglobin powders [8]. This work investigated the atomic motion in hydrated myo-

globin over a range of 4-350 K and on the time scale of 0.1-100 ps. At temperatures below 180 K, myoglobin was found to behave as a harmonic solid. Above this temperature, extra motions were seen which were of a non-vibrational nature. These were attributed to jumps between states of different energies *via* barrier crossing. Later, the presence of dynamical transition was reported for bacteriorhodopsin at $\approx$ 230 K [9], at 200 K in α amylase [11], and at 220 K for enzyme xylanase [12]. An example of neutron scattering data showing a dynamical transition is shown in Fig. 1 [6]. This transition is very similar to that observed in Mössbauer spectroscopy . Both the Mössbauer spectrum and the neutron scattering data show a linear temperature dependence below $\approx$ 200 K followed by an anharmonic behaviour with increase in temperature. On increasing the temperature, an extra decrease in the neutron elastic intensity occurs accompanied by the appearance of quasielastic scattering. Quasielastic scattering is a broad peak centered on the elastic peak and is due to diffusive motions such as barrier crossing or continuous diffusion.

3.4 Molecular Dynamics (MD) Simulation

Present day computers have made possible the use of MD simulations on large systems such as proteins. A simulation of ferrocytochrome c exhibited a dynamical transition at $\approx$ 220 K [34]. Subsequent calculations showed the presence of a dynamical transition also in MD simulations of myoglobin [35, 36]. Later, MD simulations of myoglobin were used to calculate and decompose quasielastic neutron scattering. This resulted in the attribution of the nonvibrational component of the scattering to liquid-like diffusive motions of the side chains, rather than conformational transitions.

A wide variety of research progressed in the direction of dealing with questions concerning the role of solvent in the dynamical transition and what motions are excited. These questions will be addressed and discussed in the later sections.

4 Dynamical transition and protein function

Protein function is dependent on protein flexibility. As, during the dynamical transition, there is a significant increase in flexibility, a loss of function might be expected below the transition temperature. A number of studies have indeed shown that at least some proteins cease to function below the dynamical transition. For example, the dynamics of membrane protein chromatophores from *Rhodospirillum rubrum* were investigated using the Mössbauer effect [50] after incorporation of ^{57}Fe and were shown to undergo a transition at $\approx$ 170 K [2]. Interestingly, the temperature dependence of the efficiency of the photoinduced electron transfer was also observed to undergo a change at $\approx$170 K. The function of light-driven proton pump protein, bacteriorhodopsin also exhibits a strong correlation with dynamical transition phenomena [9, 10, 51].

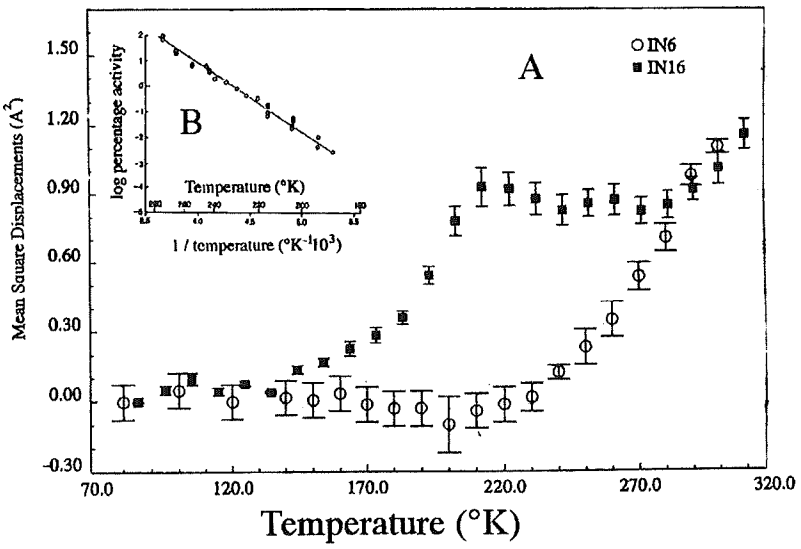


Fig. 2 : *Variation of $\langle u^2 \rangle$ for glutamate dehydrogenase in a cryosolution as a function of temperature measured using the instrument IN6 and IN16 at ILL, Grenoble. From Ref. [13].*

Neutron scattering experiments have been performed to examine the temperature dependence of motions in a cryosolution of the enzyme glutamate dehydrogenase (GDH) [52] and compare these with the enzyme activity [13]. The results showed that activity remains below the measured picosecond-timescale dynamical transition at ≈220 K, and there was no significant deviation of the activity from Arrhenius behaviour down to 190 K. These results indicate that there exists a range of temperature (190-220 K) at which the enzyme rate-limiting step does not require, and is not affected by anharmonic motions taking place on the picosecond timescale.
Further dynamic neutron scattering experiments were performed on the GDH solution with two instruments : IN6 and IN16 at Laue-Langevin reactor in Grenoble, of different energy resolution, permitting the separate determination of $\langle u^2 \rangle$ on the sub- $\approx$ 100 ps and sub- $\approx$ 5 ns time scales, respectively [13]. The resulting elastic incoherent intensities obtained on these instruments were used to plot $\langle u^2 \rangle$ vs temperature as shown in Fig. 2. Upward inflection of the $\langle u^2 \rangle$ indicates the onset of anharmonic dynamics as the temperature is increased. As can be seen, there exists a clear timescale dependence of dynamical transition. The onset of anharmonic motion occurs at $\approx$ 140 K on IN16 and $\approx$ 220 K on IN6. In addition to the dynamical transition at $\approx$ 140 K, there are inflexions in the IN16 $\langle u^2 \rangle$ at $\approx$ 185 K, $\approx$ 210 K, and $\approx$ 280 K. If a dynamical transition is defined as an inflexion in $\langle u^2 \rangle$, then the IN16 profile demonstrates four dynamical transitions in the sample. The three higher-temperature transitions do not correspond to the transitions from anharmonic to harmonic behaviour, but rather to modifications of the anharmonic behaviour itself. The lowest temperature at which the transition from harmonic to anharmonic behaviour takes place is heavily dependent on the time window of the instrument used to study the dynamics. Comparison with the temperature profile of the activity of the enzyme in the same solvent (shown as an inset in Fig. 2) reveals that the dynamical transition at

220K has no effect on the activity of GDH solutions.

5 Dynamical transition and neutron frequency windows

Detecting the dynamical transition on a neutron scattering instrument depends on the relation between the timescale of the relaxation processes responsible for the increased $\langle u^2 \rangle$ and the energy resolution of the neutron scattering spectrometer. As seen in the case of GDH solutions [13], the dynamical transition temperature in the same solution depends strongly on the timescale of the motions observed and the transition temperature shifts from $\approx$220 K to $\approx$150K with improvement in the energy resolution. This can qualitatively be explained in terms of temperature-dependent timescales of internal protein motions. An increase in temperature results in faster extra fluctuations that move into the accessible timescale window of the spectrometer used.

Two possible scenarios exist for the existance of dynamical transition in proteins [53]. A change with temperature of the long-time probability distribution of the single atom displacements forms the first, "equilibrium" scenario. Here, the characteristic relaxation frequencies of the dynamics are all within the energy resolution of the instrument used, and the dynamical transition, in principle, can then lead to a characterization of energy levels occupied by different conformational substates. In the second scenario, there exists no change in the time-converged atomic probability distribution with temperature. This is called the 'frequency-window' scenario. This would correspond, for example, to the systems possessing multiple minima with the minima having the same energy. In this case, temperature-dependent apparent dynamical transition behavior will be observed when the relaxation frequencies of the dynamics responsible for the mean-square displacement are too slow to be detected by the finite energy resolution instrument. If the frequencies increase with temperature such that they fall into the frequency window of the instrument, then the dynamical transition is observed.

For systems where the frequency window scenario dominates, the dynamical transition provides information about the timescales of the motion that fall into the energy window of the instrument. For the dynamics involving barrier crossing, the dynamical transition analysis provides information regarding the barriers on the timescales accessed.

6 Effect of solvent on dynamical transition

At physiological temperatures dry proteins are, in comparison to hydrated ones, relatively rigid and may well be inactive. A number of studies have probed the relationship between protein hydration and dynamical transition [9, 54–60]. These suggest that solvent plays a major role.

6.1 The role of protein-water hydrogen bond

The non-linear increase in the mean-square displacements observed for the heme iron in myoglobin crystals [1] was interpreted as a glass transition caused by the protein hydration shell [61]. Various experimental and MD simulation studies, have demonstrated that a change in the environmental parameters induces non-harmonic behaviour in protein dynamical properties [10, 26, 51]. In analogy with supercooled liquids, the onset of additional anharmonic or diffusive motions at the dynamical transition temperature is associated with structural relaxation or α processes [62]. Both the transition temperature and the amplitudes are sensitive to

the solvent environment of the protein, with amplitude decreasing upon dehydration [9,63] and the transition temperature increasing with increasing solvent viscosity [64, 65]. These observations were interpreted [63] using a simple model that stresses the role of polar interactions between the solvent and the protein in the dynamical transition. This model is based on the assumption that there exists just two states for each hydrogen bond namely, open and closed. Breaking a bond requires an energy ΔH. The open/closed state population of hydrogen is then given by the Boltzmann factor $\ell = exp(-\Delta G/RT)$ where $\Delta G = \Delta H - T\Delta S$. Displacement of non-polar hydrogen atoms in the side-chains is associated with this process and it is assumed that the formation of a bond is very fast. This model leads to an exponential increase in the amplitudes of motion as a function of temperature, whereas the relaxation rate, dominated by the fast bond formation remains fixed. The dynamical transition would then be due to fast bond fluctuations.

MD simulations have shown that there exists a correlation between dynamical transitions in a protein and a glass-like transition in the surrounding solvent [34]. It was shown that room-temperature atomic displacements of a protein can be significantly suppressed by cooling the solvent or by fixing the solvent coordinates. [66] Simulation evidence also suggests that at the dynamical transition the average protein:water hydrogen bond lifetime decreases and the number of different water molecules taking part in protein water hydrogen bond increases [67]. It was concluded that the dynamical transition is connected to rearrangement of the protein-water hydrogen bond network.

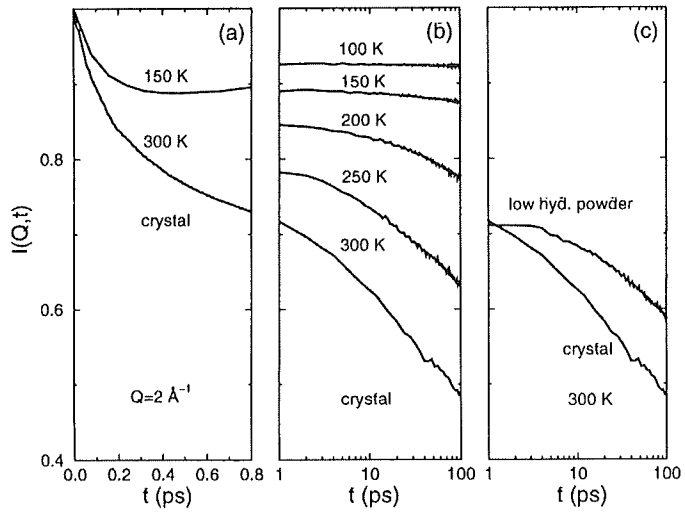


Fig. 3 : *Temperature dependence of the incoherent intermediate scattering functions for non-exchangeable hydrogen atoms computed from MD simulations of the RNase crystal : (a) 1 ps time scale, (b) 100 ps time scale, (c) Comparison of the intermediate scattering functions in the crystal and dehydrated powder at 300 K. From Ref. [38].*

Further simulations investigating the role of the hydrogen-bond network on the dynamical transition provided a detailed analysis of protein-water hydrogen bonds that distinguishes between

fast and slow hydrogen bond dynamics [38]. It has concluded that the additional anharmonic protein motions activated above the transition temperature arise from the slow relaxation of the water hydrogen bond network.

Effects of temperature and hydration on protein structural relaxation are shown in Fig. 3. Here, the incoherent intermediate neutron scattering function, $I(Q,t)$ is plotted as a function of time. $I(Q,t)$ initially displays a fast decay on the timescale at both 150 K (low temperature) and 300 K (high temperature) in the hydrated crystal. On the 100 ps time scale, $I(Q,t)$ below 200 K is almost flat, due the onset of structural arrest below the dynamical transition temperature (180 K). Above the dynamical transition temperature, the $I(Q,t)$ displays a secondary decay due to the onset of structural relaxation. The fast initial decay was suggested to be reminiscent of the β processes and the secondary decay as that of α processes as in glasses [38].

6.2 Translational hydration water dynamics drives dynamical transition

The observed solvent couplings lead to the question of whether the dynamical transition in a solvated protein is controlled by the solvent or whether the intrinsic anharmonicity of the protein dynamics suffices. To investigate these issues, dual heat bath MD simulations were performed in which the protein and solvent were held at different temperatures [40].

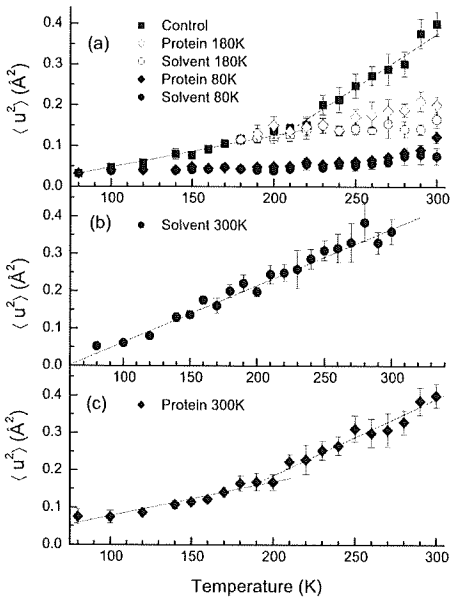


Fig. 4 : *Mean square fluctuations $\langle u^2 \rangle$ of protein non hydrogen atoms for different sets of simulations. (a) Control set with protein and solvent at same temperature, protein held at 80 K, solvent held at 80 K, protein held at 180 K, solvent held at 180 K. (b) Solvent held at 300 K. (c) Protein held at 300 K. From Ref. [40].*

Dual heatbath simulations had previously demonstrated that cold solvent strongly inhibits protein fluctuations whereas hot solvent increases them. [66] Varying the temperature of one component while keeping the other temperature constant dissociates changes with temperature of features of the protein energy landscape from those inherent to solvent. This enables the driving force behind the protein transition to be identified.
The protein fluctuations calculated from the control set of simulations (in which in each simulation, the protein and the solvent are at the same temperature), together with those obtained by fixing the temperature of one component below the dynamical transition while varying the temperature of the other, are shown in the Fig. 4a. In the control set, the experimental dynamical transition is reproduced, with non-linearity starting at $\approx$ 220 K. It is also seen that fixing the temperature of the solvent at 80 K and 180 K suppresses the dynamical transition, and $\langle u^2 \rangle$ of the protein increases linearly with temperature indicating that the low-temperature solvent cages the protein dynamics. It is also evident from Fig. 4a that holding the protein temperature constant at 80 K or 180 K and varying the solvent temperature also abolishes dynamical transition. Figs. 4b and 4c show the effect of holding one component above the transition temperature while varying the temperature of the other. Holding the solvent temperature at 300 K leads to increased protein fluctuations at most temperatures relative to the other simulation sets. However, there is again no clear deviation from linearity showing that there is no dynamical transition behaviour. In contrast, fixing the protein at 300 K and varying the solvent temperature recovers dynamical behaviour in the protein at a slightly lower temperature than in the control set. The above results indicate that the only effects due to changes with temperature in the sampled region of the protein energy landscape appear when the solvent is fixed at a temperature of 300 K. The absence of a dynamical transition with the solvent fixed at 300 K indicates that these changes do not control the transition. However, when the protein is held at 300 K, variations with temperature in the sampled solvent landscape trigger the protein transition.

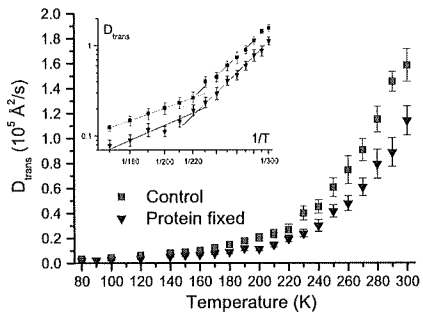


Fig. 5 : *Translational diffusion constant D_{trans} for sets of simulations where the protein and solvent are at the same temperature and for the proteins held fixed. Inset: the same data plotted as log D_{trans} versus $\frac{1}{T}$. Straight line fits below and above 120 K are also shown. From Ref. [40].*

The question as to which properties of the solvent is responsible for inducing protein transition is addressed by examining the translational diffusion constant and the dipole rotational correlation terms for the water molecules (Figs. 5, 6 and 7). A qualitative transition in the temperature dependence of the water translational diffusion constant at the dynamical transition is shown

in Fig. 5. If the diffusion behaves like an activated process *i.e.*, $D_{trans} \propto exp(-\alpha/T)$, then a straight line behaviour is expected as shown as an inset in Fig. 5. What is observed are two regimes of linear behaviour below and above transition involving a lowering of the effective activation energy for water translational diffusion above the transition. This observation remains true even when the protein atoms are fixed, thereby showing that the water transition is inherent to the solvent shell and is independent of the protein dynamics. The behaviour of D_{trans} vs T in Fig. 5 clearly shows that, above the transition, the translational diffusion increases rapidly with temperature.

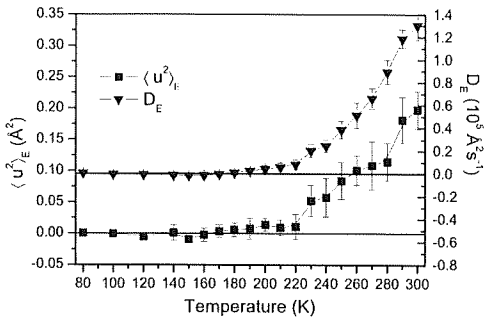


Fig. 6 : *Excess mean square fluctuation, $\langle u^2 \rangle_E$ and excess water translational diffusion constant D_E versus temperature. $\langle u^2 \rangle_E$ is defined as $\langle u^2 \rangle - \langle u^2 \rangle_H$ where $\langle u^2 \rangle_H$ is the linear part of $\langle u^2 \rangle$ obtained by fitting to the data below 220 K. D_E is calculated from $D_{trans} - D_H$ where D_H is the linear part of D_{trans} obtained by fitting t the data below 220 K. All data calculated from the control simulations. From Ref. [40].*

A comparison plot of the excess water translation diffusion constant over and above the effective diffusion constant for harmonic motion and the excess mean square fluctuation of the protein is given in Fig. 6. The two quantities are seen to vary almost identically with temperature, indicating again that water translational diffusion drives the protein dynamical transition. A plot of dipole correlation terms as a function of temperature (Fig. 7) indicates that the water dipole rotational autocorrelation times undergo no change through the dynamical transition when the protein is fixed. Hence, the water rotational reorientation is decoupled from the transition in the translational diffusion seen in the simulation. The observations discussed above provide an answer to the question of whether the solvent translational diffusion is responsible for the change in slope of $\langle u^2 \rangle$ versus temperature in the protein. It is clear that the dynamical transition of a protein when hydrated is driven by changes with temperature in the solvent dynamics. The solvent transition drives dynamical transition behaviour primarily in the side chain atoms of the external protein regions. *i.e.*, those closest to the solvent.

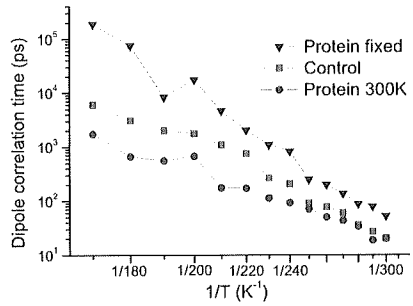


Fig. 7 : *Dipole autocorrelation time on a logarithmic scale versus* $\frac{1}{T}$. *From Ref. [40].*

This dovetails with the "radially-softening" description of protein dynamics, based on MD simulation analysis of quasi elastic neutron scattering data, in which it was shown that the average dynamical properties of a protein at 300 K vary smoothly with increasing distance from the protein core, involving a gradual increase of the diffusive amplitudes and a narrowing and shift to shorter times of the distribution of diffusive relaxation processes. [68]

7 Principal components of the protein dynamical transition

In order to further investigate the motions involved in protein dynamical transition, MD trajectories of myoglobin were subjected to Principal Component Analysis (PCA) [69]. It was observed that the dynamics activated at the transition can be described by a very small number of components, pointing to an underlying simplicity in the process involved. The modes activated were found to be collective and distributed over the protein molecule.

7.1 Principal component analysis

PCA is a convenient method for representing the conformational space explored in an MD trajectory [70, 71]. It determines the essential motions present in the simulation: the principal component modes. The set of principal components is the solution to the eigenvalue problem of the second-moment matrix, A, of the square-root mass-weighted internal atomic displacements. The elements of the second moment matrix A_{ij}, are given by

$$A_{ij} = \sqrt{m_i m_j} \times \langle (r_i(t) - {r_i}^m)(r_i(t) - {r_i}^m) \rangle$$

where r_i is the position, ${r_i}^m$, the mean position, and m_i the mass of atom i, and the average is taken over the time frames of the trajectory. The diagonalization of A yields the eigenvector w_k. *i.e.* the principal components and their associated eigenvalues, χ_k. χ_k is the mass-weighted mean-square fluctuation along w_k and can be used to calculate the atom averaged mass-weighted atomic fluctuation of the protein.

7.2 Free energy landscapes

The free energy landscapes of the low frequency principal components are shown in Fig. 8. Figs. 8a, 8b and 8c show the free energy profiles along the modes 1, 2 and 5 respectively. These energies are given by $\mu_k(q) = -k_B T ln P_k(q)$ where k_B is the Boltzmann constant, T, the temperature and $P_k(q)$ being the probability of finding the protein at position q along mode k.

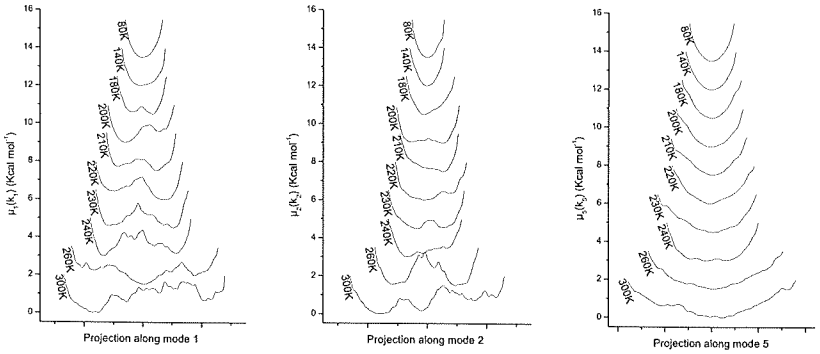


Fig. 8 : *Free energy profiles of (a) the 1st ; (b) the 2nd and (c) the 5th-lowest principal component modes. From Ref. [69].*

As can be seen from Fig. 8a, for mode 1, the profile is approximately quadratic below 180 K. The onset of dynamical transition is characterised by the appearance at 180 K in mode 1 of double-well behaviour with a free energy barrier of $\approx k_B T$. Above $\approx$ 240 K, the mode 1 profile is highly rugged. Mode 2 is approximately harmonic for T $\approx$ 200 K, above which multiple minima start appearing. Mode 5, as can be seen from Fig. 8c, is an example of a mode that is harmonic at low temperatures and becomes quasiharmonic above. These graphical representation of modes over a temperature range 180 - 300 K shows that these principal components involve collective dynamics distributed over most of the myoglobin molecule. In mode 1, the transition between the free energy minima at 200 K comprises the relative motion of two large rigid body blocks of $\approx$ 900 atoms.

7.3 Harmonic, Quasi-harmonic and multiminima modes

The anharmonicity factor (ρ_k), the measure of the deviation of mode k from harmonicity. To determine this, a coefficient c_{kl}, projecting the normal mode vector space onto that of the principal component can be calculated using the relation $c_{kl} = (w_k \cdot w_l^{NM}$ where w_k is the eigenvector and w_l^{NM} are the normal modes. The harmonic variance along the kth principal component mode χ_k^H, is then given by $\chi_k^H = \sum_l c_{kl} \chi_l^{NM}$, and the anharmonicity factor is defined as $\rho_k = \sqrt{\frac{\chi_k}{\chi_k^H}}$. The standard deviation (σ), is defined as $\sigma^2 = 1000 \cdot \overline{(P(q) = G(q))^2}$ where $P(q)$ is the probability distribution and $G(q)$ is the Gaussian fit.

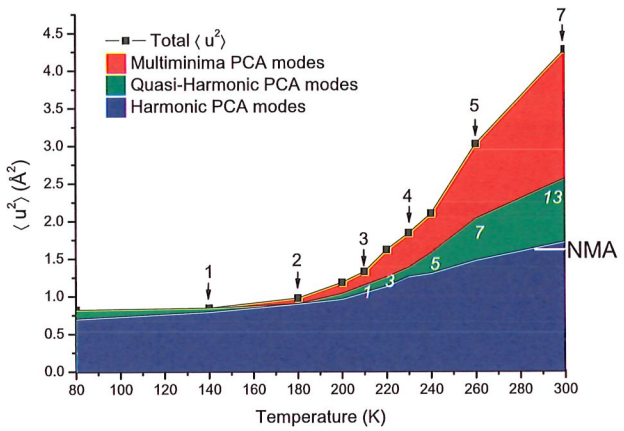


Fig. 9 : *Decomposition of the mass-weighted protein* $\langle u^2 \rangle$ *into contributions arising from the harmonic, quasi-harmonic and multiminimum classes of principal component mode. Vertical arrows indicate number of multiminimum modes and in italics are the number of quasiharmonic modes at certain temperatures. From Ref. [69].*

The anharmonicity factor and the standard deviation allows the classification of the principal components into three types.

1. Harmonic modes : Those which exhibit approximately Gaussian probability distributions ($\sigma < 1$) for which the anharmonicity is low ($\rho \leq 2$).

2. Quasiharmonic modes : those which exhibit approximately Gaussian probability distributions ($\sigma < 1$) but posess considerable anharmonicity ($\rho > 2$).

3. Multiminimum modes which cannot be well fitted by a Gaussian (σ ¿ 1) and are also anharmonic ($\rho > 2$).

The contributions made by these principal components to the mean square displacements ($\langle u^2 \rangle$) are plotted in Fig. 9. The increase in $\langle u^2 \rangle$ signaling the incipient phase of the transition ($\approx$ 180 - 210 K) is seen to arise from a very small number of principal components. By 210 K, only four of the PCA modes are not harmonic and among these three have multiminimum behaviour and one is quasiharmonic. At 210 K, 75% of the increase over the linear $\langle u^2 \rangle$ is due to multiminimum dynamics. At 300 K, only 0.3% of the total number of modes in the protein deviated from harmonic behaviour. 70% of the excess $\langle u^2 \rangle$ at 300 K arises from seven multiminimum modes with the other 30% originating from 13 quasiharmonic principal components [69].
These studies provide a description of the subnanosecond protein dynamics activated at the transition. The results indicate that the solvent:surface interaction propagates to the interior of the protein *via* collective dynamics that can be described by a very small number of principal components. The modes initiating the transition are global and distributed over the protein. The quasiharmonic components correspond to softened effective harmonic motion. The largest

displacements, however, arise from the activation of barrier crossing, *i.e.*, multiminimum components. In the mode with the largest contribution to the onset of the transition in the present system the barrier crossing involves blocks of supersecondary structural elements moving relative to each other.

8 Conclusions

The protein dynamical transition involves the onset of diffusional dynamics at a temperature range of $\approx$ 180 - 220 K. It is now established that solvent plays a major role in the dynamical transition, *via* relaxation of the water hydrogen bond network and solvent molecule translational displacement. The solvent transition drives the protein dynamical transition behavior primarily in the side-chain atoms of the external protein regions *i.e.*, those closest to the solvent. The modes initiating the dynamical transition are global and distributed over the protein. The larger displacements involved are found to arise from the activity of barrier crossing, multiminimum principal components. The rigid-body motion concerned qualitatively resembles structural changes seen in proteins in different functional states. [72] However, it is yet unclear as to what extent motions activated in the dynamical transition describe functional protein dynamics. Certain proteins are found to cease functioning below the dynamical transition. Enzyme activity is dependent on dynamics at the active site. The understanding of physical properties of proteins such as dynamical transition should allow further understanding of protein equilibrium fluctuations and their relation to function.

References

[1] F. Parak and H. Formanek, Acta. Crystallogr. **27**, 573 (1971).

[2] F. Parak *et al.*, FEBS Lett. **117**, 368 (1980).

[3] E. W. Knapp, S. F. Fischer, and F. Parak, J. Am. Chem. Soc. **86**, 5042 (1982).

[4] H. D. Middendorf, Ann. Rev. Biophys. Bioeng. **13**, 425 (1984).

[5] S. Cusack, Comm. molec. cell. Biophys **3**, 243 (1986).

[6] T. Kleinert *et al.*, Biochemistry **37**, 717 (1998).

[7] F. Parak, Rep. Prog. Phys. **66**, 103 (2003).

[8] W. Doster, S. Cussack, and W. Petry, Nature **337**, 754 (1989).

[9] M. Ferrand, A. J. Dianoux, W. Petry, and G. Zaccai, Proc. Natl. Acad. Sci. USA **90**, 9668 (1990).

[10] J. Fitter, R. E. Lechner, and N. A. Dencher, Biophys. J **73**, 2126 (1997).

[11] J. Fitter, Biophys. J **76**, 1034 (1999).

[12] V. Reat, R. D. M. F. J. L. Finney, R. M. Daniel, and J. C. Smith, Proc. Natl. Acad. Sci. USA **97**, 9961 (2000).

[13] R. M. Daniel *et al.*, Biophys. J. **77**, 2184 (1999).

[14] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou, Nature **280**, 558 (1979).

[15] R. F. Tilton, J. C. Dewan, and G. A. Petsko, Biochemistry **31**, 2469 (1992).

[16] H. Frauenfelder, F. Parak, and R. D. Young, Annu. Rev. Biophys. Chem **17**, 451 (1988).

[17] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).

[18] H. Frauenfelder and D. T. Leeson, Nat. Struct. Biol. **5**, 757 (1998).

[19] B. H. McMohon, J. D. Muller, C. A. Wraight, and G. U. Nienhaus, Biophys. J **74**, 2567 (1998).

[20] A. Ostermann, R. Waschipky, F. G. Parak, and G. U. Nienhaus, Nature **404**, 205 (2000).

[21] R. H. Austin *et al.*, Biochemistry **14**, 5355 (1975).

[22] B. F. Rasmussen, A. M. Stock, D. Ringe, and G. A. Petsko, Nature **357**, 423 (1992).

[23] X. Ding, B. F. Rasmussen, G. A. Petsko, and D. Ringe, Biochemistry **33**, 9285 (1994).

[24] C. A. Angell, Science **267**, 1924 (1995).

[25] I. E. Iben *et al.*, Phys. Rev. Lett **62**, 1916 (1989).

[26] J. L. Green, J. Fan, and C. A. Angell, J. Phys. Chem **98**, 13780 (1994).

[27] F. H. Stillinger, Science **267**, 1935 (1995).

[28] A. P. Sokolov *et al.*, J. Non-Cryst. Solids **172**, 138 (1994).

[29] B. Frick and D. Richter, Science **267**, 1939 (1995).

[30] R. Abseher, H. Schreiber, and O. Steinhauser, Proteins **25**, 366 (1996).

[31] P. J. Steinbach, R. J. Loncharich, and B. Brooks, Proc. Natl. Acad. Sci. USA **90**, 9135 (1993).

[32] S. Melchionna, M. Falconi, and A. Desideri, J. Chem. Phys **108**, 6033 (1998).

[33] J. A. Hayward and J. C. Smith, Biophys. J. **82**, 1216 (2002).

[34] C. F. Wong, C. Zheng, and J. A. McCammon, Chem. Phys. Lett. **154**, 151 (1989).

[35] J. Smith, K. Kuczera, and M. Karplus, Proc. Natl. Acad. Sci USA **87**, 1601 (1990).

[36] R. J. Loncharich and B. R. Brooks, J. Mol. Biol **215**, 439 (1990).

[37] P. J. Steinbach and B. R. Brooks, Proc. Natl. Acad. Sci USA **93**, 55 (1996).

[38] M. Tarek and D. J. Tobias, Phys. Rev. Lett. **88**, 138101 (2002).

[39] A. R. Bizzarri and S. Cannistraro, J. Phys. Chem. B **106**, 6617 (2002).

[40] A. L. Tournier and J. C. Smith, Biophys. J. **85**, 1871 (2003).

[41] R. L. Mössbauer, J. of Phys. **C6**, 11 (1976).

[42] S. G. Cohen, E. R. Bauminger, I. Nowik, and S. Ofer, Phys. Rev. Lett. **46**, 1244 (1981).

[43] K. H. Mayo, F. Parak, and R. L. Mössbauer, Phys. Rev. Lett. **82**, 468 (1981).

[44] H. Keller and P. G. Debrunner, Phys. Rev. Lett **45**, 68 (1980).

[45] F. Parak *et al.*, FEBS Lett **117**, 368 (1980).

[46] F. Parak, E. N. Frolov, R. L. Mössbauer, and V. I. Goldanskii, J. Mol. Biol. **145**, 825 (1981).

[47] F. Parak, E. W. Knapp, and D. Kucheida, J. Mol. Biol **161**, 177 (1982).

[48] S. H. Chong *et al.*, Eur. Biophys. J **30**, 319 (2001).

[49] Y. Joti, M. Nakasako, A. Kidera, and N. Go, Biol. Crystall **D58**, 1421 (2002).

[50] E. N. Frolov *et al.*, Biophys. J. **19**, 253 (1991).

[51] U. Lehnert *et al.*, Biophys. J. **75**, 1945 (1998).

[52] R. M. Daniel and H. H. Petach, Biochem. J. **305**, 17 (1996).

[53] T. Becker *et al.*, submitted.

[54] G. Careri, A. Giansanti, and J. A. Rupley, Proc. Natl. Acad. Sci. **83**, 6810 (1986).

[55] F. Demmel, W. Doster, W. Petry, and A. Schulte, Biophys. J **26**, 327 (1997).

[56] J. Fitter, R. E. L. anf G. Büldt, and N. A. Dencher, Proc. Natl. Acad. Sci. **93**, 7600 (1996).

[57] Y. U. F. Krupyanskii *et al.*, Z. Naturforsch **37**, 57 (1982).

[58] J. Partridge, P. R. Dennison, B. D. Moore, and P. J. Halling, Biochim. Biophys. Acta **1386**, 79 (1998).

[59] J. Perez, J. M. Zanotti, and D. Durand, Biophys. J. **77**, 454 (1999).

[60] J. C. Smith, S. Cusack, P. L. Poole, and J. L. Finney, J. Biomol. Struct. Dyn. **93**, 2974 (1990).

[61] T. Bachleitner, Biophys. J. **50**, 213 (1986).

[62] W. Doster, S. Cusack, and W. Petry, Phys. Rev. Lett. **65**, 1080 (1990).

[63] W. Doster and M. Settles, *Hydration Processes in Biology* (IOS Press, Amsterdam, ADDRESS, 1998).

[64] H. Lichtenegger, Biophys. J. **76**, 414 (1999).

[65] A. M. Tsai, D. A. Neumann, and L. N. Bell, Biophys. J. **79**, 2728 (2000).

[66] D. Viktup, D. Ringe, G. A. Petsko, and M. Karplus, Natur. Struct. Biol. **7**, 34 (2000).

[67] C. Arcangeli, A. R. Bizzarri, and S. Cannistraro, Chem. Phys. Lett. **291**, 7 (1998).

[68] S. A. Dellerue, J. Petrescu, J. C. Smith, and M. C. Bellissent-Funel, Biophys. J. **81**, 1666 (2001).

[69] A. Tournier and J. C. Smith, Phys. Rev. Lett. **91**, 208106 (2003).

[70] M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).

[71] M. Karplus, S. Hayward, and N. Go, Proteins **33**, 496 (1998).

[72] M.Gerstein *et al.*, J. Mol. Biol. **234**, 357 (1993).

C 6 Protein folding: Theory

P. Grassberger
John von Neumann - Institut für Computing
Forschungszentrum Jülich GmbH

Contents

1 Introduction **2**

2 The computational physicist's approach **3**
2.1 Force fields 3
2.2 MC versus MD 4
2.3 Other ground state search algorithms 5
2.4 Coarse-grained models 6
2.5 Ground state searches in the HP model 7

3 The computer scientist's approach **9**

4 The statistical physicist's questions **10**
4.1 Are proteins just random heteropolymers? 10
4.2 How many qualitatively different folds can exist? 10
4.3 Is there a molecular basis for neutral evolution? 11
4.4 How deterministic is protein folding? 12

5 Outlook **12**

1 Introduction

Today molecular biology is heavily dependent on massive computations. Projects like the deciphering of the human genome would not be possible without fast computers, but the same is true also for many other tasks like establishing phylogenetic trees, finding similar genes in different organisms, understanding large metabolic networks, and last but not least, protein folding. Indeed, understanding how proteins, characterized by their sequence of amino acids, find their way to fold precisely into that ("native") shape in which they perform their biological function, is considered by many as the most difficult problem. And it is not by chance that the biggest supercomputer ever built, "blue gene", which is due to be finished by the year 2005 and which will perform with 1 petaflop (10^{15} flops; about 30 times more than the most powerful present supercomputer), will be entirely be devoted to protein folding.

Proteins are linear (i.e., unbranched) polymers, built from the 20 naturally occurring amino acids. Their lengths vary considerably, from about 20 amino acids up to several thousands. Most typical single-domain proteins have between 100 and 500 amino acids. Each amino acid consists of the backbone (which is the same for all amino acids) and a side group. The side groups vary in size (the smallest, glycine, has just a single H atom, while others have up to 9 heavy atoms plus about the same number of H atoms), in electrical charge (some are positive charged, others are negative or neutral), and in hydrophobicity. Hydrophobicity essentially depends on the electric dipole moment of the side group. Polar amino acids are hydrophilic (want to be in contact with water), while non-polar ones are hydrophobic [1, 2].

These differences between amino acids are responsible for the specific shapes into which proteins fold. One distinguishes between primary, secondary, tertiary, and quaternary structure. The primary structure is the amino acid sequence itself and the set of all covalent bonds (e.g. disulfide bridges). The secondary structure describes local structures which consist mostly of α-helices, β-sheets, and "loops" connecting them. The tertiary structure describes the global folding topology, unless the protein is composed of several chains or of several distinct domains. Their arrangement is called the quaternary structure.

It was established by Anfinsen in the sixties [3] that proteins fold in vitro spontaneously and reversibly. By changing temperature or pH value one can make the protein unfold, and by reestablishing the original conditions it will fold again into the native state. This means that there is no machinery needed for folding, and that the native state is entirely determined by the amino acid sequence. Actually, this is not quite true since large proteins need so-called chaperones for their proper folding, but these are rather unspecific molecules which essentially prevent misfoldings by temporarily shielding the protein from reactive surfaces.

Experimental methods to determine the structure of a given protein are costly and in general not easy. This is in contrast to determining DNA and thus also amino acid sequences which today is very fast and can be done routinely. Therefore the number of known genes (= amino acid sequences) is much bigger today than the number of known structures, and this gap is widening more and more. Thus it would be a big achievement if one could predict reliably the structure from the sequence. This *protein fold prediction* problem is the first of a series of related problems. Another is to find the *pathway* along which the folding proceeds, and yet another is the *inverse fold prediction* or *proton design* problem: Finding a sequence which would lead to a given 3-d structure.

In the following I will mostly be concerned with the fold prediction problem. For this, the most successful methods today are not based on finding a ground state of some hamiltonian, but by means of data base searches using heuristics and exploiting analogies with proteins whose

structures are already known. Such methods will be discussed later. In the next section we will concentrate on methods based on an explicit hamiltonian. Such methods are called " ab initio" in the protein folding literature, even if the hamiltonian is not fundamental but is heuristic.

2 The computational physicist's approach

2.1 Force fields

For a theoretical physicist, it might seem that predicting the native structure is straightforward. One just writes down the Schrödinger equation and finds its ground state. But things are not so simple. First of all, one is not interested in the ground state only (which would be relevant at zero absolute temperature), but the native state is vaguely described as the state of lowest *free energy*. This involves an integration over "microscopic" degrees of freedom, keeping only the "relevant" structural aspects fixed. Actually, while the free energy of a protein is well defined, its " free energy landscape" is a much more fuzzy concept. The reason is that in the former one integrates over *all* thermal fluctuations, while one imagines some of the degrees of freedom kept fixed in the latter. Nevertheless it is often convenient to use this concept of a free energy landscape. But in the following we shall sometimes disregard the above subtlety and use the words "energy landscape" and "free energy landscape" synonymously.
A bigger problem is that a medium size protein with 400 amino acids, say, contains about 10^4 valence electrons. Solving such a Schrödinger equation truly ab initio is at present out of question. Finally, to make life not too easy, nature has built most proteins such that native and misfolded states have rather small energy differences – while the individual positive and negative contributions to the energy are much larger. Therefore, even very small errors will have rather drastic consequences.
This holds not only for calculations starting from the Schrödinger equation, but also for all calculations using phenomenological potentials. There are a number of such potentials (or "force fields") in the literature, the most popular ones include CHARMM [4], AMBER [5], ECEPP [6], and GROMOS [7]. They typically contain Lennard-Jones interactions, electrostatic interactions, and torsion angle dependent energy terms. One crucial distinction is between force fields which assume rigid valence bonds and fixed bond angles (ECEPP) and force fields which allow for flexible bonds (AMBER, GROMOS). The former is of course a rather drastic simplification as far as the dynamics is concerned, although it should have much less effects on structural properties. Actually, when one wants to use molecular dynamics (MD), then rigid bonds and bond angles are quite a nuisance, since the dynamics of rigid objects connected by flexible joints is quite messy [1]. It is much less of a problem in Monte Carlo (MC) simulations. On the other hand, assuming bonds to be rigid eliminates the fastest motion in a protein. Thus one can use much larger time steps if the oscillations in bond length and bond angles are frozen. Another aspect concerns the simulation of the protein's environment. Most proteins are in the cell immersed in water. Exceptions are of course membrane proteins and proteins which act as parts of some larger structures. Simulating the water (or membrane, or entire structure) can be very costly. On the other hand, it is in either case very wrong just to forget about the environment (although it is done very often). This is particularly so for globular proteins surrounded by water, since water is not just neutral. Its polarizability plays a crucial role. Therefore a number of approximate treatments have been proposed. One possibility is to use a distance

[1] The same problem appears also in robotics, and indeed the mathematics needed in both fields are rather similar.

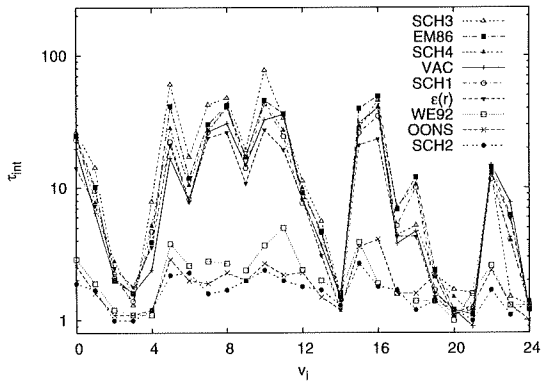


Fig. 1: *Integrated autocorrelation times at $T = 300\,K$ of Met-Enkephalin whose configuration is described only by its 24 torsion angles. The variable v_0 is the total energy, variables v_i with $i = 1, \ldots, 24$ are the angles. From Ref.[14], where also the acronyms for the different solvation schemes are explained.*

dependent electrostatic permittivity [8, 9]. A more refined alternative is the solvent-accessible area approach [10, 11, 12, 13, 14]. In this approach one assumes that each type of atom which is exposed to the solvent (water) contributes to the energy a term which is proportional to its accessible surface and to a phenomenological "solvation parameter". There are a large number of solvation parameter sets in the literature, and it is not clear which ones are the best. Also, it is not clear whether different solvation parameter sets, even if they give the same energies in equilibrium, behave dynamically in the same way. Indications are that this is indeed not true, and that different sets show dramatically different relaxation times [14], see Fig. 1.

2.2 MC versus MD

There are two main basic strategies for simulating a classical dynamical system with many degrees of freedom: Molecular dynamics and Monte Carlo (MC) algorithms. In the former one directly simulates the actual motion, while one makes shortcuts in the latter such that equilibrium (Gibbs-Boltzmann) distributions are sampled correctly, without claiming to have realistic dynamics. If one wants to follow the actual folding path of a protein, one must therefore, if one wants to be realistic, use MD. This is one reason why most packages used by chemists to simulate proteins and other biomolecules use MD [5, 7].

But MD is slow. The promise of MC methods is that one can move more efficiently in phase space, if one chooses suitable moves. There exist mainly two basic MC strategies [15]: Metropolis (or 'Metropolis-Hasting', or 'Markov chain') and sequential sampling. In the former one starts with one configuration and proposes moves which *deform* it. These moves might increase or decrease the energy. If the energy is decreased, the move is accepted, otherwise it is rejected with a finite probability. There are a vast number of extensions of this basic strategy. One is hybrid MC [16] which is indeed in between MC and MD. Essentially, the proposed moves in hybrid MC are MD moves with rather large time steps. Due to the latter, the simulation would be very imprecise if these moves were accepted unconditionally. By using a Metropolis type accept/reject step, one corrects for this and obtains a formally exact method.

Other methods which are still within the realm of the Markov chain strategy are multicanonical [17], parallel tempering [18], and flat histogram [19] methods. Each of these methods modifies in one way or another the original hamiltonian so that one is less held up by barriers and local minima. These changes are then of course compensated by suitable re-weighings so that one finally samples indeed from the correct Boltzmann distribution. In the multicanonical method one essentially adds a term which makes the density of states flat in energy, and which has to be determined in previous auxiliary simulations. In the flat histogram method such a term is built during the simulation itself. In parallel tempering one makes k simultaneous simulations at different temperatures $T_1 \ldots T_k$. Periodically, one proposes to switch systems between two neighbouring temperatures and accepts this switch according to a Metropolis type criterion. For a detailed discussion of these and several related methods see [20]. Results obtained in this way are quite impressive [21, 22, 23]. Many of these tricks can also be used for MD simulations. One looses then of course the realism of the dynamics, but one can still keep some of its advantages (mainly if water is simulated explicitly; see below). A package where some of these advanced methods are built in or can be added relatively easily is SMMP [24].

Sequential MC methods are very different. Instead of deforming a fully assembled polymer, they proceed by *building* the protein *from scratch*, by adding amino acids one after the other (why they are also called " growth algorithms"). If not done carefully, the growth will lead to structures which most likely are very bad energetically. To avoid this, one uses two tricks: First of all, one lets the protein grow with a bias such that new monomers are added in energetically favourable places. This has of course to be accompanied by re-weighing. Each configuration gets a weight which is inversely proportional to the probability with which it is chosen. If the bias is optimal, this "Rosenbluth weight" [25] cancels exactly the Boltzmann weight, leading thereby to perfect importance sampling. In practice this will never happen. Thus the combined Rosenbluth and Boltzmann weights will show large fluctuations, leading to inefficiencies. At this point comes the second trick, i.e. re-sampling. This means that configurations with too low weight are killed, while configurations with too high weight are cloned. There are many different implementation of this idea [15, 26, 27], my own favourite is the pruned-enriched Rosenbluth Method (PERM) which uses backtracking by means of recursive function calls, and where the killing/cloning thresholds are learned and updated during the simulation [28]. Recently, it was shown that the idea of multicanonical resp. flat histogram sampling can also be used for PERM [29, 30].

One particular problem for all MC methods is the ambient water. In vacuo, even a large unphysical move where a part of the protein is either displaced or grown has still a large chance to be accepted. This is no longer true when the space around the protein is filled with solvent. In MD, the solvent and the protein would move in a concerted way such that there is no conflict, but all existing MC algorithms (except maybe hybrid MC) are severely affected by this. This is the main reason why in practice most realistic simulations are done with MD and not with MC, and this is also the main reason why one should find better ways to mimic the effect of the solvent without simulating it explicitly.

2.3 Other ground state search algorithms

Although a protein *in vivo* is not in its ground state, one can to a very good approximation consider fold prediction as an optimization problem where the hamiltonian is the cost function. There are a number of optimization algorithms applicable to this, all of which have in common that they do not give the correct Boltzmann distribution away from $T = 0$.

One of the best known such methods are genetic algorithms (GA). It seems that GA have been applied mostly to highly simplified models (see next subsection), but not so much to realistic models [31, 32]. As we shall discuss later, they have not been very successful when compared to multicanonical simulations or to PERM. This is maybe not very surprising, since GA do not use much specific knowledge about the system. On the other extreme are methods which are very specifically constructed for the highly simplified HP-model (see next subsection) and give for it exact ground states, if they converge at all [33, 34].
Among the many methods designed for ground state searches of more realistic protein models, we mention just two: The *stochastic tunneling* method of [35], and the *conformational space annealing* method of [36]. Stochastic tunneling deforms the potential surface so that high barriers are leveled down and the trajectory can 'tunnel' more easily through the barrier. In conformational space annealing one builds a library of configurations which are selected such as to have low energy but at the same time be well separated in state space. Less 'valuable' configurations are repeatedly replaced by better ones, i.e. by new configurations which have lower energy but are not much closer to other configurations in the library. As the search for lower energies goes on, one releases slowly the constraint of large separation, so that one searches with increasing intensity in better and better, but also smaller and smaller, regions.

2.4 Coarse-grained models

We have already discussed two simplifications, namely implicit water and rigid bonds. But since the resulting models are still much too complicated, people have gone far beyond that. In particular, the following models have been studied in detail:

- United-atom models. A first approximation might consist in disregarding H atoms. More radically, one can replace the backbone and/or the side group of each amino acid by a single fictitious particle. Maybe the most sophisticated and best studied model in this line is UNRES [37].

- Lattice models. The most radical and also most widely made choice is a simple cubic (simple square in 2 dimensions) lattice, although this is not the most realistic. Finer lattices were used e.g. in [38]. But one does not have to go to very fine lattices, already using an fcc lattice should be better than an sc lattice [34]. Notice that the model of [39], although formally off-lattice, is also closely related to an fcc-based lattice model.

- In a lattice model it is most natural to us , i.e. two amino acids interact if and only if they are on neighbouring lattice sites, and the interaction strength depends only on their types [40, 41]. But contact interactions can also be used off-lattice. The total energy is written as a sum over non-bonded neighbouring amino acid pairs. If the i-th amino acid is of type $a_i \in \{1, 2, \ldots 20\}$, we have

$$H = \sum_{1 \le i < j \le N} \Delta_{ij} \epsilon(a_i, a_j), \tag{1}$$

  where $\Delta_{ij} = 1$ if monomers i and j are neighbours, and $\Delta_{ij} = 0$ else. Although this model has been quite successful, it has been shown [42] that such a model has very strict limitations. Taking the actual native configuration of crambin from the proton data bank, It was proven exactly in [42] that there cannot exist any set $\{\epsilon(a, a')\}$ such that the actual native configuration of crambin has a lower energy than all other configurations.

Table 1: *Lowest energies of the ten 3-dimensional HP sequences of Unger et al.[54]. The second column shows the lowest energies found by the GA used in [54]. The third column shows the results found with an improved GA in [31]. The fourth column shows the results of nPERM, while the last column gives exact lower bounds on the ground state energy.*

sequence Nr.	Ref[54]	Ref[31]	nPERM	lower bounds
1	-27	-27	-32	-33
2	-29	-30	-38	-41
3	-35	-38	-45	-49
4	-34	-34	-42	-45
5	-32	-36	-43	-48
6	-29	-31	-35	-37
7	-20	-25	-28	-32
8	-29	-34	-38	-43
9	-32	-33	-41	-45
10	-24	-26	-31	-35

- In spite of this, people have indeed gone even further and reduced the number of amino acid species. The most radical model in this respect is the HP model of K. Dill and collaborators [43, 44, 45]. There it is assumed that the main driving force is hydrophobicity, and thus only two classes of amino acids are distinguished: hydrophobic ones (H) and hydrophilic or *polar* ones (P). The interaction between them is described by Eq.(1) with $\epsilon(PP) = \epsilon(HP) = 0, \epsilon(PP) = -1$. It was indeed shown in [46] that hydrophobicity plays the main role in the Miyazawa-Jernigan [40] contact potentials, although there it is not assumed that it is a simple zero/one feature.

2.5 Ground state searches in the HP model

It is known that finding the ground state of a lattice heteropolymer is an NP-complete problem [47]. This implies that there cannot be any algorithm which finds the ground state of any polymer with N monomers in a time polynomial in N, although checking that a proposed ground state is indeed correct is polynomial. Thus protein fold prediction in a simple lattice model like the HP model is basically as difficult as any other difficult combinatorial problem. In addition, also the protein design problem, i.e. the problem of finding an amino sequence which folds into a given configuration, is NP-hard [48]. It is therefore tempting to study the HP model not so much for its implication for real proteins but as a test bed for heuristic combinatorial algorithms. The rest of this subsection is to be viewed from this perspective.

Maybe the first stochastic algorithm used for this problem was simulated annealing with Metropolis MC and with local moves. Using ten random HP sequences of length $N = 48$ as test cases to be folded on the sc lattice, this approach failed quite miserably [49]: Only for one of the sequences the known ground state energy was hit. In contrast, the PERM algorithm as implemented in [50] found all ten ground state energies within reasonable CPU time. Similar results were obtained also for other test cases, but it was found that multicanonical and other generalized ensemble methods did equally well [51, 52]. A substantial further improvement was obtained with " new PERM" (nPERM) [53], which brought the CPU times for all ten sequences

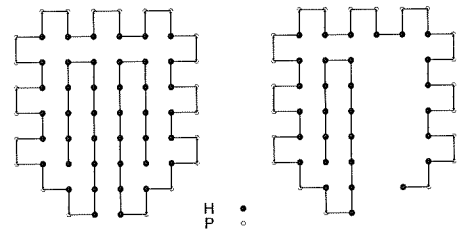


Fig. 2: *Left side: ground state configuration of a $N = 64$ chain in 2D from [55]. Other states with the same energy differ in the detailed folding of the tails in the interior, but have identical outer shapes. Right side: when about 3/4 of the chain is grown, one has to pass through a very unstable configuration which is stabilized only later, when the core is finished.*

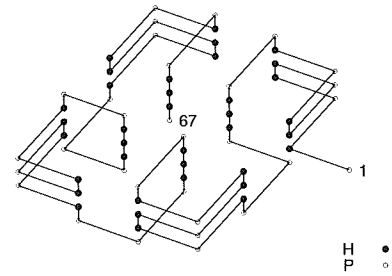


Fig. 3: *Ground state configuration ($E = -56$) of the $N = 67$ HP sequence given in [33]. It forms a structure resembling an α/β barrel. When starting at monomer #67 (β sheet end), nPERM could find it easily, but not when starting from monomer #1.*

down to $<$ 1h on a slow work station.

The essential improvement in nPERM is that good (high weight) configurations are not simply cloned as in old PERM, but that two *different* configurations are made when the weight is estimated to become too large. More precisely, if there exist $\geq k$ different possibilities to continue the growth of a configuration, and if the current weight is sufficiently high, then these k continuations are chosen. The number of off-spring made is never larger than the number of different continuations. The main consequence is that we avoid thereby the 'lack of diversification' which also plagues genetic algorithms when the optimum is searched to greedily.

Genetic algorithms are quite popular in the computer science community, and there exists a similar set of ten HP sequences of length $N = 64$ which had been used as test bed in two previous analyses. With nPERM we could find lower energies for *each* sequence, see Table 1.

Chain growth algorithms usually do not perform so well if there are important long distance contacts (i.e., if the distance *along the chain* between two amino acids in spatial contact is large). This is particularly so when one arm has to grow a large part without getting support

from the rest of the chain, because the region between the two is filled out later [53]. Two examples are given in Figs. 2 and 3. But long range contacts decrease the speed of folding also in real proteins [56]. It is not clear whether the coincidence between these two observations is an accident or is more profound. After all, the chain growth algorithm was not meant to mimic the actual folding process. On the other hand one could imagine that, even if the chains are not grown during the folding (as they are in PERM), the set of contacts has to grow in a similar way, explaining thereby the similarity.

3 The computer scientist's approach

As we said already in the introduction, today the physics-based approach described in the preceding section is less successful for protein fold prediction than heuristic approaches that do not use any explicit energy function or any other dynamical concepts. Instead, these approaches use heuristics derived from proteins with known folds and analogies with proteins with similar sequences. Progress in fold prediction is monitored by the bi-annual CASP competitions [57]. In all these competitions so far such methods have done much better than the ab initio methods described in the previous section.

Secondary structure prediction is particularly easy. Alpha helices and β- sheets are stabilized by hydrogen bonds. Since different amino acids differ also in their tendency to form hydrogen bonds, one can rather robustly predict the likelihood for these secondary structures. Another useful information is that prolin is more stiff than other amino acids because of a second valence bond between the backbone and the side chain, and cannot therefore be in the interior of an α-helix or of a β-sheet. All this and similar information can used either in a hidden Markov model or in a feed-forward neural network. There are a large number of such tools available on the web, see e.g. [58] and the citations given there.

After this is done, one has to assemble the secondary structure elements to obtain the tertiary structure. What helps here is the fact that hydrophobic amino acids must (except for a membrane protein) be buried in the interior, while polar ones should be on the surface. Whenever this is not sufficient, one can use ' homologous proteins' (i.e., proteins evolved historically from the same ancestor), provided their structure is already known. If on takes a typical protein and changes 10 percent of all its amino acids in a random way, one can be pretty sure that the protein will no longer fold into the same shape. But homologous proteins often fold into the same shape, even though less than 40 percent of their amino acids are the same. This remarkable feature is dictated by the requirement that a mutation should not disrupt the production of proteins fulfilling the same function, if the mutant line is to survive. We will discuss this in more detail in the next section, but here it is sufficient to say that homologies are usually a good indicator for structural similarity, even if the sequence similarity is low.

Another method to predict tertiary structures is *threading* [58]. Using the properties of the individual amino acids, one can predict the 'optimal' environment of any particular stretch of the protein. On the other hand, one can for each protein in a data base (whether it is homologous to the protein at hand or not) just write down the environment along its backbone. Threading essentially means that one pretends to substitute the local amino sequence (but not its environment formed by the other parts!) of the data base protein by the local amino acid sequence of the protein to be analyzed. Thus one aligns the *sequence* to be analyzed with a large number of *structures*. Only if the alignment is good, i.e. if the protein fits fits into the environment, one can assume that the actual fold is similar.

4 The statistical physicist's questions

Up to now we have not yet touched a true *theory* of protein folding. We have used the tools provided by theoretical physicists essentially like engineers. Biologists, medical doctors, and pharmaceutical industry are all mainly interested in understanding how particular proteins fold or how one has to construct an amino acid sequence such that it folds into a particular way. In general they are not very interested in universal features or in explaining *why* something happens as it does – except, of course, when this would also help them in their practical needs. Physicists are notoriously different in this respect. To them it usually seems obvious that universal behavior is the more interesting the less it is tied to specific features, just because then it is more universal. Critical phenomena are the best example for this approach, but it transpires also through all the other facets of theoretical physics.

4.1 Are proteins just random heteropolymers?

When statistical physicists entered the field of protein science, it seemed very natural to assume that proteins are, as regards their folding, just random. Of course their catalytic properties (if they are enzymes) are not random, but it was not clear why this should have any effect on their folding statistics.

Maybe the most important result of studies of grossly simplified models is that we now know that this cannot be true. The reason is simply that randomly chosen amino acid sequences would, in the vast majority of cases, not lead to good folders. Instead they would have highly degenerate ground states (on lattice) or a huge amount of metastable states (in the continuum). Roughly spoken, in a random heteropolymer there are two 'transitions' (which are of course no real phase transitions, since the systems are finite) [59]. One is a collapse transition which leads to a compact globule whose internal structure is still 'fluid' and does not know about the native configuration, the other is a sort of 'freezing' transition which then fixes the native internal structure.

Imagine now that the collapse transition happens already at a much higher temperature $T = T_c$ than the freezing transition T_f. All the way down from T_c to T_f the globule will become more and more dense and stiff. There might not be a real thermodynamic phase transition in between, but there might happen something like a glass transition. Thus when the real freezing point is reached, the globule will already be 'frozen' for all practical needs, and it will remain in a metastable state without reaching the true ground state.

If, however, $T_c \approx T_f$ (or, rather, $(T_c - T_f)/T_c \ll 1$), then the globule is still sufficiently mobile at $T_f + \epsilon$ so that the ground state is reached. Roughly spoken, real proteins follow the second scenario (otherwise we could hardly exist), while random heteropolymers follow – in all simple models – the first one. Otherwise said, there was a selection pressure during evolution towards polymers which are good folders.

Indeed, by simulating this evolution on the computer it has been possible to "breed" good folders [60, 61].

4.2 How many qualitatively different folds can exist?

First a remarkable experimental fact: In spite of the millions of different proteins which exits in the entire living world, it is estimated that there exist only 1000 qualitatively different folds [62]. This means that even proteins which have nothing in common must often fold into roughly

the same shape. Why is this so? A priori one might have guessed that there should be as many different folds as there are sequences.
To shed light on this problem, Wingreen *et al.* [63] studied a variety of simplified models, from off-lattice models to the extremely much simplified 'purely hydrophobic' (PH) model. In all of them, they found that different configurations have very different *designability*. A configuration is designable, if there is at least one amino acid sequence for which it is the unique ground state. Many configurations are not designable at all (i.e., for every sequence there exists another configuration with the same or even a lower energy). Others are highly designable, i.e. there are very many sequences for which it is the unique ground state. The spectrum of designabilities is much wider than expected, and this can also be rather easily be explained in the simplified models.
Notice that this result is in some sense 'dual' to the observation made in the previous subsection. There it was shown that good folders are sparse in the set of *sequences*, now we see that also designability is sparsely distributed in the set of *configurations*. The number of most designable configurations is very small.
Could it be that the small number of qualitatively different folds found in real proteins is connected to this, i.e. that naturally occurring proteins have to be highly designable? In the following subsection we will argue that this is indeed the case.

4.3 Is there a molecular basis for neutral evolution?

The observation of the last subsection has an important consequence for evolution. Assume that an organism would use a protein whose configuration just barely desingable. Thus there are only a few sequences which would fold into this shape. Any mutation would lead to a high risk that the mutated sequence no longer folds properly, and the organism would be in trouble. On the other hand, if the protein were highly designable, there would be a much larger chance that the mutation is harmless. Therefore there is an evolutionary advantage to use highly designable protein configurations.
Indeed, this agrees nicely with our previous observation that the sequence similarity of homologous sequences is usually much lower than the similarity of their folds. It also agrees with the theory of neutral evolution [64] which stipulates, based partially on the paleontological record, that most mutations which are not immediately eliminated are neutral, i.e. have neither good nor bad *immediate* effects. But they can become important later, by adding genetic diversity which can be crucial if the environment changes.
To test these ideas, the authors of [65] proceeded as follows: They started from a random sequence in the protein data bank (PDB), i.e. from a sequence which is known to be a good folder and whose native state is also known. Then they made a number of point mutations and checked whether the mutated sequences still fold into the same configuration. For this they used the contact potential of [41] and threading against a large set of structures in the PDB. Many mutations will no longer fold properly, but for those which do so, they can repeat the procedure: make a second point mutation, keep those which still fold correctly, make another mutation to them, etc.
In this way they obtained a set of sequences, connected by paths of single point mutations, which all fold into the same configuration. Surprisingly (or not), this set is both sparse and dense at the same time. It is sparse because it of course contains only a tiny fraction of all sequences. It is dense in the sense that any two randomly picked sequences taken from the set seem to be completely uncorrelated statistically. Thus the set seems uniformly distributed in

sequence space. Basically the same phenomenon is true also for RNA [66].

4.4 How deterministic is protein folding?

Actually, this question includes two aspects: How unique is the folding path, when a protein folds correctly? And how often does it not fold correctly?
Traditionally, biochemists had assumed, in analogy with more simple chemical reactions, that there is a unique folding path passing through a well defined transition state (a saddle in the energy landscape). It is now known that this is not the case. The accepted picture of the (free) energy landscape is that of a " funnel" [67], with a more or less monotonic overall decrease of the energy and with only shallow and smooth secondary minima, as the native state is approached. This should hold at least for good folders, while there are much more and much deeper local minima in randomly picked heteropolymers, leading to a much more rugged landscape. In any case, there is no unique folding path, and the "transition state" is rather an ensemble of very diverse microstates.
Similarly, traditional biochemists seem to have underestimated for a long time the probability that proteins might fold incorrectly [68]. It is now known that many diseases, including Creutzfeldt-Jakob and Alzheimer, are due to misfoldings. Most common is that misfolded proteins, instead of burying their hydrophobic parts inside, display part of them on their surface which them gives rise to agglomeration. To prevent this, there exist chaperones which keep the folding proteins away from other hydrophobic surfaces. But the fact that this is not always successful, and that misfolding is quite common also in healthy cells, is suggested by a quite elaborate machinery which breaks up and destroy unwanted proteins [69]. Usually, it is assumed that these are mostly well-folded proteins which are no longer needed, but it might well be that these are also, to a large part, misfolded ones [68].

5 Outlook

A landmark paper where the notion of the 'protein folding problem' was coined, was Levinthal's paper of 1968 [70] where he showed that the number of possible protein conformations is much too big for a random unbiased search. But this so-called *Levinthal's paradox* is paradox only if one forgets about all physical sciences since the last 2000 years. Of course no dynamics is an unbiased search. No stone you drop from your hand would fall down, if there were no force pulling it, and if it had to make a search among *all* other possible positions, in addition to the one in front of your feet.
Basically, I believe that a 'protein folding problem' in the sense of a clear-cut problem which can be *solved* does not exist. We understand perfectly well why, in principle, proteins fold. We just don't know most of the many details, and simulating the folding is still far too slow. Future computers, together with better algorithms, will allow us to make steady progress. As we make this progress, we will also be able to identify some of the more universal and theoretically deeper aspects which finally will lead to a real theory of protein folding worth its name.

References

[1] J.D. Watson *et al.*, *Molecular biology of the gene* (Benjamin/Cummings, Menlo Park, 1987).

[2] A.L. Lehninger, D.L. Nelson, and M.M. Cox, *Principles of Biochemistry* (Worth Publishers, N.Y. 1993).

[3] C.B. Anfinsen, Science **181**, 223 (1973).

[4] http://www.scripps.edu/brooks/c27docs/Charmm27.Html.

[5] http://amber.scripps.edu

[6] http://www.tc.cornell.edu/Research/Biomed/CompBiologyTools/eceppak/README.html

[7] http://www.igc.ethz.ch/gromos

[8] B. Hingerty, R.H. Richie, T.L. Ferrel, and J. Turner, Biopolymers **25**, 427 (1985).

[9] Y. Okamoto, Biopolymers **34**, 529 (1994).

[10] B. Lee and M. Richards, J. Mol. Biol. **55**, 379 (1971).

[11] C. Chothia, Nature **248**, 338 (1974).

[12] D. Eisenberg, A.D. McLachlan, Nature **319**, 199 (1986).

[13] A.H. Juffer, F. Eisenhaber, S.J. Hubbard, D. Walther, P. Argos, Protein Sci. **4**, 2499 (1995).

[14] B.A. Berg and H.-P. Hsu, cond-mat/0306435 (2003).

[15] J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics (Springer, New York 2001)

[16] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth, Physics Letters B **195**, 216 (1987).

[17] B.A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).

[18] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[19] F. Wang and D.P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[20] Y. Okamoto, cond-mat/0308360 (2003).

[21] U.H.E. Hansmann, J. Chem. Phys. **120**, 417 (2004).

[22] H. Kokubo H and Y. Okamoto, Chem. Phys. Lett. **383**, 397 (2004).

[23] G. Favrin, A. Irbäck, and S. Wallin, Proteins **47**, 99 (2002).

[24] F. Eisenmenger, U.H. Hansmann, S. Hayryan, and C.-K. Hu, Comp. Phys. Commun. **138**, 192 (2001).

[25] M.N. Rosenbluth and A.W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955)

[26] D. Frenkel, G.C.A.M. Mooji, and B. Smit, J. Phys. C **3**, 3035 (1991).

[27] J. Sadanobu and W.A. Goddard, J. Chem. Phys. **106**, 6722 (1997).

[28] P. Grassberger, Phys. Rev. **E56**, 3682 (1997).

[29] M. Bachmann and W. Janke, cond-mat/0304613; cond-mat/0310707 (2003)

[30] T. Prellberg and J. Krawczyk, cond-mat/0312253 (2003).

[31] A.L. Patton, W.F. Punch III, and E.D. Goodman, *A standard GA approach to native protein conformation prediction*, in 6th Proc. Int'l. Conf. on Genetic Algorithms, pp. 574-581 (1995).

[32] R. König and T. Dandekar, BioSystems **50**, 17 (1999); J. Mol. Model. **5**, 317 (1999).

[33] K. Yue and K.A. Dill, Phys. Rev. E **48**, 2267 (1993); Proc. Natl. Acad. Sci. USA **92**, 146 (1995).

[34] R. Backofen and S. Will, *Optimally compact finite sphere packings – hydrophobic cores in the fcc*, Proc. of the 12th Annual Symp. on Combinatorial Pattern Matching (CPM2001), Lecture Notes in Computer Science, 2001, 2089 (2001).

[35] A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102/1 (2003); T. Herges, A. Schug, H. Merlitz, and W. Wenzel, Nanotechnology **14**, 1161 (2003).

[36] J. Lee and H.A. Scheraga, Intl. J. of Quantum Chem. **75**, 255 (1999); J. Lee, A. Liwo, D.R. Ripoll, J. Pillardy, J.A. Saunders, K.D. Gibson, and H.A. Scheraga, Intl. J. of Quantum Chem. **77**, 90 (2000); S.-Y. Kim, S.J. Lee, and J. Lee, J. Chem. Phys. **119**, 10274 (2003).

[37] A. Liwo *et al.*, Protein Sci. **2**, 1715 (1993); J. Comput. Chem. **18**, 849, 874 (1997).

[38] A. Godzik, A. Kolinski, and J. Skolnick, J. Mol. Biol. **227**, 227 (1992); A. Kolinski and J. Skolnick, J. Chem. Phys. **107**, 953 (1997).

[39] D.A. Debe, M.J. Carlson, J. Sadanobu, S.I. Chan, and W.A. Goddard, J. Phys. Chem. B **103**, 3001 (1999).

[40] S. Miyazawa and R.L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[41] U. Bastolla, M. Vendruscolo, and E.-W. Knapp, Proc. Natl. Acad. Sci. USA **97**, 3977 (2000).

[42] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11101 (1998).

[43] K.A. Dill, Biochemistry **24**, 1501 (1985)

[44] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986 (1989).

[45] H.S. Chan and K.A. Dill, J. Chem. Phys. **95**, 3775 (1991).

[46] H. Li, C. Tang, and N.S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997)

[47] B. Berger and T. Leighton, J. of Comput. Biol. **5**, 27 (1998).

[48] N.A. Pierce and E. Winfree, Protein Eng. **15**, 779 (2002).

[49] K. Yue *et al.*, Proc. Natl. Acad. Sci. USA **92**, 325 (1995).

[50] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, Proteins **32**, 52 (1998).

[51] A. Irbäck, "Dynamical-parameter algorithms for protein folding", in *Monte Carlo Approach to Biopolymers and Protein Folding*, eds. P. Grassberger *et al.*, pp. 98-109 (World Scientific, Singapore, 1998)

[52] G. Chikenji, M. Kikuchi, and Y. Iba, Phys. Rev. Lett. **83**, 1886 (1999).

[53] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003).

[54] R. Unger and J. Moult, *A genetic algorithm for 3d protein folding simulations*, in 5th Proc. Int'l. Conf. on Genetic Algorithms, pp. 581-588 (1993).

[55] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).

[56] D. Baker, Nature **405**, 39 (2000).

[57] http://predictioncenter.llnl.gov/casp5/Casp5.html.

[58] G. Steger, *Bioinformatik* (Birkhäuser, Basel 2003).

[59] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[60] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, Folding & Design **1**, 221 (1996).

[61] C. Micheletti, A. Maritan, J.R. Banavar, J. Chem. Phys. **110**, 9730 (1999).

[62] S.E. Brenner, C. Chothia, and T.J.P. Hubbard, Curr. Opin. Struct. Biol. **7**, 369 (1997).

[63] H. Li, C. Tang, and N.S. Wingreen, Proc. Natl. Acad. Sci. USA **95**, 4987 (1998); Proteins **49**, 403 (2002); N.S. Wingreen, H. Li, and C. Tang, Polymer **45**, 699 (2004); J. Miller, C. Zheng, N.S. Wingreen, and C. Tang, Proteins **47**, 506 (2002)..

[64] M. Kimura, *The theory of neutral evolution* (Cambridge Univ. Press, 1983).

[65] U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, J. Mol. Evol. **56**, 243; **57**, 103 (2003).

[66] P. Schuster, W. Fontana, P.F. Stadler, and I.L. Hofacker, Proc. R. Soc. London B **255**, 279 (1994). Proc. Natl. Acad. Sci. USA **92**, 3626 (1995).

[67] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes, Annu. Rev. Phys. Chem. **48**, 539 (1997).

[68] C.M. Dobson, Nature **426**, 884 (2003).

[69] http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Proteasome.html

[70] C. Levinthal, J. Chem. Phys. **65**, 44 (1968).

C 7 Protein Folding: experiment

Thomas Kiefhaber
Biozentrum der Universität Basel
Abteilung Biophysikalische Chemie

Contents

1 Introduction 2

2 Characterization of Protein Folding Transition States 2

2.1 Kinetics of two-state folding 2
2.2 Rate Equilibrium Free Energy Relationships 4
2.3 Rate equilibrium free energy relationships in protein folding 5
2.4 Self-interaction and cross-interaction parameters 7
2.5 Sequential and parallel transition states 9
2.6 Experimental data on transition state movements in protein folding 10
2.7 Conclusions 13

3 Intrachain Diffusion and the Speed Limit for Protein Folding 13

3.1 Using TTET to measure absolute rates of intrachain diffusion 14
3.2 Intrachain contact formation in glycine-rich polypeptides 15
3.3 Scaling of end-to-end diffusion with chain length 17
3.4 Effect of amino acid sequence on intrachain diffusion 19
3.5 Time scales for the earliest steps in protein folding 21

1 Introduction

Experimental work on the mechanism of protein folding has been greatly influenced by Levinthal's famous paper of 1969 [1], in which he pointed out that a polypeptide chain would require an astronomical time to explore at random all possible conformations in order to finally reach the native state. This motivated the search for partially folded intermediates guiding the protein to the native state. Although many proteins have been found to fold through transient intermediates (for reviews see [2-9]), theoretical studies have sparked off a discussion on their role in the folding process. Simplified models for polypeptide chains, in combination with energy landscape theory, suggested that the Levinthal paradox could be resolved by a small energy bias against locally unfavourable conformations [10]. This would reduce folding rates to the experimentally observed time scales. As a common representation of the theoretical work, funnel-like folding landscapes were proposed [11-16] in which the polypeptide chain has a vast number of different possible ways of reaching the native state. Partially folded states observed in these models mainly represent misfolded states trapped in local energy minima rather than obligatory folding intermediates. It was further suggested that the polypeptide chain does not encounter major energy barriers for the actual folding process and that the rate of folding is limited by a region in the folding funnel where the gain in energy cannot compensate for the loss in conformational entropy [12]. Thus, the folding kinetics were proposed to be determined mainly by entropic barriers.

This lecture gives an overview on experimental results on the properties of free energy barriers in protein folding. The first part of the lecture deals with the properties of the free energy barriers with a focus on protein folding transition states, which provide evidence for the importance of intermediates and for the existence of defined transition states. The second part of the lecture will discuss recent results on the dynamics of intrachain diffusion in polypeptide chains, which sets the speed limit for the conformational search during the folding process.

2 Characterization of Protein Folding Transition States

2.1 Kinetics of two-state folding

For simplicity of the kinetics analysis we will in the following discuss the properties of two-state folding reactions. Kinetic analysis of more complex reactions is discussed in ref. [17]. However, the characterization of transition states of complex reactions is done in analogy to the methods discussed below, by analyzing the microscopic rate constants (k_i) for the individual kinetic steps. Experimental investigations have revealed simple two-state folding without detectable intermediates for more than 20 small proteins (for an overview see [18]):

$$U \underset{k_u}{\overset{k_f}{\rightleftharpoons}} N \tag{1}$$

k_f and k_u are the microscopic rate constants for the folding and unfolding reactions, respectively. The single observable macroscopic rate constant (λ) for this mechanism is readily derived as

$$\lambda = k_f + k_u \tag{2}$$

The equilibrium of the reaction is determined by the flows from the native state N to the unfolded state U and vice versa:

$$k_f \cdot [U]_{eq} = k_u \cdot [N]_{eq} \tag{3}$$

and the equilibrium constant (K) for folding is

$$K := \frac{[N]_{eq}}{[U]_{eq}} = \frac{k_f}{k_u} \tag{4}$$

A straightforward parameter to be inferred from two-state behaviour is the free energy for folding (ΔG^0) which is connected by the van't Hoff relation with the equilibrium constant (K)

$$\Delta G^0 = -RT \cdot \ln(K) = -RT \cdot \ln\left(\frac{[N]_{eq}}{[U]_{eq}}\right) = -RT \cdot \ln\left(\frac{k_f}{k_u}\right) \tag{5}$$

Thus, protein stability can be measured in two different ways, either by equilibrium methods or by kinetic measurements. The agreement between the free energies for folding inferred from equilibrium and kinetic measurements is commonly used as a test for two-state folding.
The same formalism can be applied to obtain information on the free energy of activation ($\Delta G^{0\ddagger}$) for a reaction using transition state theory

$$k = k_0 \cdot e^{-\frac{\Delta G^{0\ddagger}}{RT}} \tag{6}$$

k_0 is the pre-exponential factor and represents the maximum rate constant for a reaction in the absence of free energy barriers. Combination of eqs. 4 and 6 allows us to gain information on the free energy of the complete reaction coordinate (see Fig.1).

$$\Delta G^0 = \Delta G_f^{0\ddagger} - \Delta G_u^{0\ddagger} \tag{7}$$

In the original Eyring theory the pre-exponential factor corresponds to the frequency of a single bond vibration

$$k_0 = \kappa \frac{k_B T}{h} \approx 6 \cdot 10^{12}\ s^{-1} \text{ at } 25\ ^{\circ}\text{C} \tag{8}$$

κ represents the transmission coefficient, which is usually assumed to be 1 for reactions in solution. Since protein folding involves formation and breakage of weak intramolecular interactions rather than changes in covalent bonds, the Eyring prefactor will not be useful for the analysis of protein folding reactions. The pre-exponential factor for protein folding reactions will be strongly influenced by intrachain diffusion processes. It is likely around 10^7 to $10^8\ s^{-1}$. This will be discussed in the second part of this lecture.

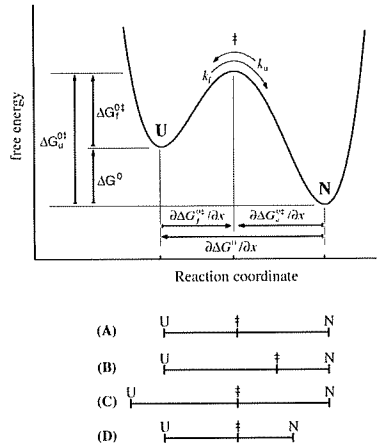


Fig. 1: *Free energy profile for a two-state folding reaction and possible origins of transition state movement (A) Reference reaction (B) Hammond behaviour (C,D) apparent transition state movement due to ground state effects on the unfolded and native protein, respectively.*

2.2 Rate Equilibrium Free Energy Relationships

For the characterization of free energy barriers in protein folding several concepts from physical organic chemistry have been applied. A frequently used approach is the analysis of linear free energy relationships. For many reactions in organic chemistry the changes in activation free energy ($\Delta G^{0\ddagger}$) induced by changes in solvent conditions or by modifications in structure are linearly related to the corresponding changes in equilibrium free energy (ΔG^0) between reactants and products [19] (linear rate-equilibrium free energy relationships; REFERs). A proportionality constant,α_x, can be defined to quantify the energetic sensitivity of the transition state relative to the ground states in respect to a perturbation, ∂x [19].

$$\alpha_x = \frac{\partial \Delta G^{0\ddagger}/\partial x}{\partial \Delta G^0/\partial x} \tag{9}$$

α_x is a measure for the position of the transition state along the reaction coordinate probed by ∂x and it is commonly used to gain information on the structural properties of the transition state. For many reactions the α_x-values are constant over a broad range of $\partial \Delta G^0$ indicating little effect of the perturbations on the structure of the transition state [20]. For other reactions, α_x is sensitive to changes in ΔG^0 which leads to non-linear REFERs and indicates structural changes in the transition state relative to the ground states. It was shown that the characterization of non-linear rate-equilibrium free energy relationships can give valuable information on the mechanism of a reaction and on the properties of the transition barriers since the non-linearities can have different origins: (i) a change in the rate-limiting step on a sequential pathway (Fig.2B) [20], (ii) a change in the mechanism of the reaction due to a switch between parallel pathways [20] (iii) a movement of the position of the transition state along a broad barrier region following Hammond behavior (Fig.2A) [21,22] or (iv) structural changes in the ground state(s) (ground state effects) [23].

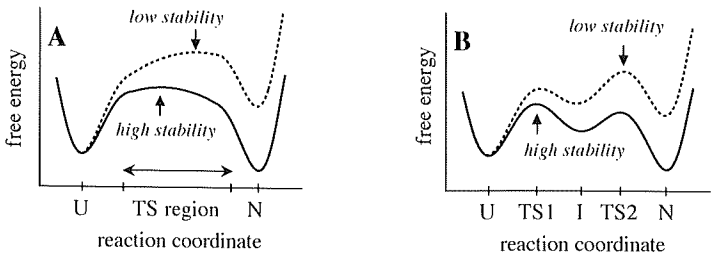


Fig. 2: *Effect of a destabilization of the native state on the free energy barrier for (A) a broad barrier region and (B) a barrier region with two consecutive transition state and a "hidden" high energy intermediate.*

Detailed analyses of non-linear free energy relationships have therefore been widely used to elucidate reaction mechanisms in organic chemistry and the mechanism of complex biological reactions like the catalytic mechanisms of enzymes [24]. In the following we will discuss how linear and non-linear free energy relationships can be applied to characterize the barriers for protein folding reactions.

2.3 Rate equilibrium free energy relationships in protein folding

Rate-equilibrium free energy relationships for protein folding reactions can be derived in various ways. Let us consider the Gibbs fundamental equation

$$\mathrm{d}\Delta G^0 = \Delta V^0 \mathrm{d}p - \Delta S^0 \mathrm{d}T + \Sigma \Delta \mu_i^0 dn_i \tag{10}$$

where ΔG^0, ΔV^0, ΔS^0 and $\Sigma\Delta\mu_i^0$, and are the differences in Gibbs free energy, volume, entropy and chemical potential, respectively, between the unfolded and the native states. If we make use of transition state theory [25,26], the Gibbs equation can be applied to the activation free energy:

$$\mathrm{d}\Delta G^{0\ddagger} = \Delta V^{0\ddagger} \mathrm{d}p - \Delta S^{0\ddagger} \mathrm{d}T + \Sigma \Delta \mu_i^{0\ddagger} dn_i \tag{11}$$

Combining equs. (9–11) shows that protein folding reactions can be perturbed in different ways to gain information on different properties of the transition state. This allows to probe different reaction coordinates for the folding reaction. α_p and α_T-values can be obtained by changes in pressure at constant temperature or changes in temperature at constant pressure, respectively [27,28].

$$\alpha_p = \frac{\partial \Delta G^{0\ddagger} / \partial p}{\partial \Delta G^0 / \partial p} = \frac{\Delta V^{0\ddagger}}{\Delta V^0} \tag{12}$$

$$\alpha_T = \frac{\partial \Delta G^{0\ddagger} / \partial T}{\partial \Delta G^0 / \partial T} = \frac{\Delta S^{0\ddagger}}{\Delta S^0} \tag{13}$$

α_p and α_T give information on the volume and on the entropy of the transition state, respectively [29,30]. From eq. (13) we can further calculate the changes in enthalpy along the reaction coordinate using the Gibbs-Helmholtz equation.

The value of α_T changes with temperature due to the change in molar heat capacity (ΔC_p^0) typically associated with protein folding reactions [31]. This allows the definition of another reaction coordinate [32]

$$\alpha_C = \frac{\Delta C_p^{0\ddagger}}{\Delta C_p^0} \tag{14}$$

α_C allows the characterization of a transition state in terms of its relative solvent exposure, since ΔC_p^0 mainly arises from differences in interactions of the protein with the solvent [33].

The addition of chemical denaturants like urea or GdmCl is the most commonly used perturbation in the study of protein folding reactions. The concentration of these chaotropic compounds generally has a linear effect on both the equilibrium free energy (ΔG^0) [34,35] and on the activation free energies [4] for folding $\Delta G_f^{0\ddagger}$ and unfolding $\Delta G_u^{0\ddagger}$ (Fig.3). Accordingly, for two state folding the logarithm of the single observable apparent rate constant, λ ($\lambda = k_f + k_u$) vs. chemical denaturant concentration yields a V-shaped curve, commonly called Chevron plot (Fig.4C) [36,37]. This plot gives information on the refolding reaction at low denaturant concentrations (the refolding limb) and information on the unfolding reaction at high denaturant concentrations (the unfolding limb).

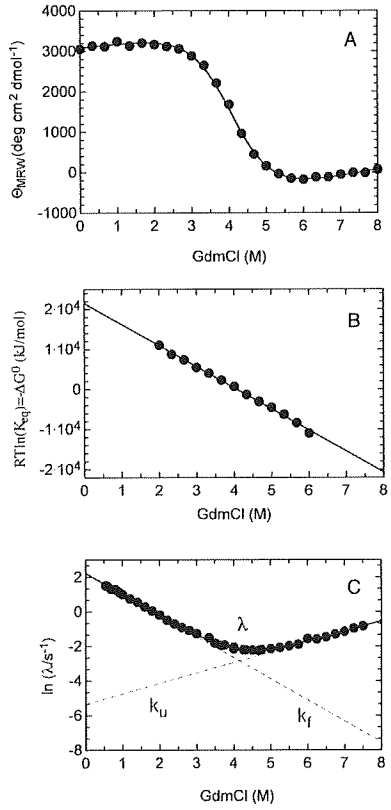


Fig. 3: *(A) GdmCl-induced equilibrium unfolding transition for tendamistat (B) Change in $-\Delta G^0$ with GdmCl concentration calculated using the equilibrium constant determined from the data shown in panel (A). (C) Effect of GdmCl on the apparent folding rate constant (λ). The V-shaped plot is commonly termed "Chevron Plot" and allows calculation of k_f and k_u as indicated (cf. eq. 2)*

The proportionality constants, m_x, are defined as:

$$m_{eq} = \frac{\partial \Delta G^0}{\partial [Denaturant]} \tag{15}$$

$$m_{f,u} = \frac{\partial \Delta G^{0\ddagger}_{f,u}}{\partial [Denaturant]} \tag{16}$$

and a denaturant-induced free energy relationship can be obtained [4]

$$\alpha_D = \frac{\partial \Delta G^{0\ddagger}_f / \partial [Denaturant]}{\partial \Delta G^0 / \partial [Denaturant]} = \frac{m_f}{m_{eq}} \tag{17}$$

α_D reflects the relative sensitivity of the transition state to changes in denaturant concentration. Since the m_{eq}-value was shown to be proportional to the changes in accessible surface area upon unfolding of the protein [38], α_D is interpreted as the relative amount of accessible surface area buried in the transition state and it should correlate with α_C.
Other solvent additives or ligands such as alcohols (2,2,2-trifluoroethanol), polyols (glycerol, sugars), salts (Na_2SO_4, NaCl), D_2O as solvent or proton concentration (change in pH) can also change the stability of a protein and can therefore be used to define the corresponding α-values [27,28].
Equations (12) to (14) and (17) can be considered as medium- or solvent-induced rate-equilibrium free energy relationships. The effect of structural changes introduced by protein engineering are commonly used to obtain a structure-induced free energy relationship [37,39,40]:

$$\alpha_S = \phi_f = \frac{\partial \Delta G^{0\ddagger} / \partial \text{Structure}}{\partial \Delta G^0 / \partial \text{Structure}} \tag{18}$$

α_S, which is commonly called ϕ_f in protein folding, reports on the energetics of the interactions formed by a side chain with the rest of the protein in the transition state relative to the native state. The free energy of the unfolded state serves as a reference. It can be defined for a single residue (ϕ_f^i), for the whole protein (ϕ_f^{prot}) [27,41,42] or for a substructure [27,42,43].
It should be noted that the type of free energy relationships described here are commonly termed "Brønsted plots" in protein folding literature, which is misleading. "Brønsted plots" were originally used to relate the effect of a change in the rate constant of an acid or base-catalyzed reaction to the dissociation constant of the catalyst. The rate-equilibrium relationships considered here are rather Leffler type relationships, since they directly relate the rate constants of a reaction to the equilibrium constants of the *same* reaction.

2.4 Self-interaction and cross-interaction parameters

As discussed above, the observation and characterization of non-linear rate-equilibrium free energy relationships gives information on the properties of the transition barriers for chemical reactions.

A practical way to detect and to analyze non-linearities in free energy relationships was proposed by Jencks and coworkers [22] by defining self-interaction and cross-interaction parameters. A self-interaction parameter (p_x) measures the shift in the position of the transition state along the reaction coordinate with changing ΔG^0 upon a perturbation ∂x.

$$p_x = \frac{\partial \alpha_x}{\partial \Delta G_x^0} = \frac{\partial^2 \Delta G^{0\ddagger}}{(\partial \Delta G_x^0)^2} \tag{19}$$

By definition, a shift in the position of the transition state towards the destabilized state, e.g. as a result of Hammond behaviour or due to sequential barriers, will give a positive p_x-value. For a folding reaction perturbed by addition of denaturant or destabilized by mutations the corresponding self-interaction parameters p_D and p_S are [27,28]

$$p_D = \frac{\partial \alpha_D}{\partial \Delta G_D^0} = \frac{\partial^2 \Delta G^{0\ddagger}}{(\partial \Delta G_D^0)^2} \tag{20}$$

$$p_S = \frac{\partial \phi_f}{\partial \Delta G_S^0} = \frac{\partial^2 \Delta G^{0\ddagger}}{(\partial \Delta G_S^0)^2} \tag{21}$$

The use of self interaction parameters in the analysis of transition state movements is often not sensitive enough, because the energy range of the measurements is too narrow or the curvatures are too small [22]. A more sensitive test for transition state movements is provided by cross-interaction parameters [22], which measure changes in the position α_x of the transition state of a reaction (measured using the perturbation ∂x) caused by a second perturbation ∂y.

$$p_{xy} = \frac{\partial \alpha_x}{\partial \Delta G_y^0} = \frac{\partial^2 \Delta G^{0\ddagger}}{(\partial \Delta G_x^0)(\partial \Delta G_y^0)} = \frac{\partial \alpha_y}{\partial \Delta G_x^0} = p_{yx} \tag{22}$$

For non-linear REFERs the value of α_x will change with the amount of a second perturbation ∂y, resulting in a non-zero p_{xy}-value. A shift in the position of the transition state towards the destabilized state will yield positive p_{xy}-values. Practical examples for cross-interaction parameters are measurements of Chevron plots (cf. Fig.3C) at different temperatures (p_{DT}) or pressures (p_{Dp}). Experimental results will be shown in section 2.6.
In mutational studies the changes in $\Delta G^{0\ddagger}$ and ΔG^0 are usually determined as a function of the denaturant concentration for each mutant. This directly allows the calculation of the denaturant-structure cross interaction parameter (p_{DS}) [27,28].

$$p_{DS} = \frac{\partial \alpha_D}{\partial \Delta G_S^0} = \frac{\partial^2 \Delta G^{0\ddagger}}{(\partial \Delta G_S^0)(\partial \Delta G_D^0)} = \frac{\partial \phi_f}{\partial \Delta G_D^0} = p_{SD} \tag{23}$$

p_{DS} tests whether changes in the stability of the native state caused by mutations ($\partial \Delta G_S^0$) have an effect on the relative solvent exposure of the transition state for folding (α_D).
If the effect of a change in structure ∂S is tested in the wild type and mutant ($\partial S'$) backgrounds, we can calculate a structure-structure cross-interaction parameter.

$$p_{ss'} = \frac{\partial \phi_f^s}{\partial \Delta G_{s'}^0} \tag{24}$$

$p_{ss'}$ indicates whether ϕ_f (=α_s) changes when the protein stability is altered by an additional mutation [27,28,44].

Changes in the length of the reaction coordinate due to the effects of a mutation on the ground states can easily be mistaken for genuine transition state movement [23,27,28]. Both phenomena have very similar effects on the experimentally observed data but they are based on completely different free energy barriers. Figure 1 shows possible effects of a perturbation ∂x on the location of the transition state for a protein folding reaction with the reference conditions given in Fig.1(A). In Fig.1(B) a perturbation leads to a transition state movement along the reaction coordinate relative to both ground states. The length of the reaction coordinate is unchanged, i.e. the structure of both ground states is not affected by the perturbation or they are affected to the same extent. This scenario is in accordance with Hammond behaviour if we assume that the native state is destabilized by the perturbation. Fig.1(C) and (D) show apparent transition state movements caused by ground state effects. In both cases the absolute position and the structure of the transition state remain unchanged but the structure of either the unfolded state (Fig.1C) or of the native state (Fig.1D) changes, which leads to a change in length of the reaction coordinate. In this case an apparent transition state movement will be observed if the position of the transition state is normalized against the length of the reaction coordinate. This demonstrates that the characterization of transition state movements requires the determination of the effects of a perturbation on the rate constants for the forward and backward reaction, on the equilibrium constant and on the length of the reaction coordinate.

2.5 Sequential and parallel transition states

Non-linear REFERs provide a unique tool to characterize complex reaction mechanisms and especially to distinguish between sequential and parallel reaction pathways [20]. If destabilization of the native state leads to a shift in the rate-limiting step to a more native-like transition state (larger α_x) on a sequential pathway a positive p_x value will be observed (downward curvature in the chevron plot; Fig.4A).

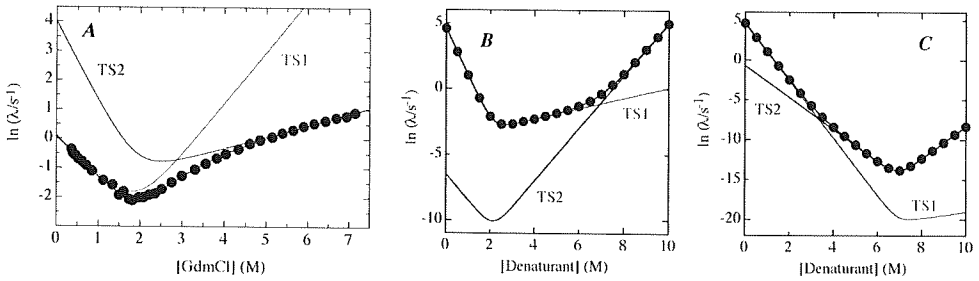


Fig. 4: *Non-linear rate-equilibrium free energy relationships for (A) sequential folding of tendamistat over two consecutive transition states according to the mechanism shown in Fig.2B (taken from ref. [45]) and (B, C) parallel folding pathways (simulated chevron plots taken from ref. [27]).*

Such a behaviour also implies that a high energy intermediate is "hidden" on the free energy landscape, which might not be detectable directly due to its low stability (see Fig.2B). If, in contrast, folding occurs along parallel pathways, the reaction with the most native-like transition state will be most sensitive to a reduced stability of the native-state and an alternative parallel pathway with a more unfolded-like transition state can become rate-limiting with increasing denaturant concentrations. This leads to a decrease in α_x with decreasing protein stability, which will result in a negative p_x-value (Fig.4B). This phenomenon is sometimes referred to as anti-Hammond behaviour [43,46].

2.6 Experimental data on transition state movements in protein folding

The most common way to study protein folding reactions is the variation of the concentrations of a chemical denaturant like GdmCl or urea. For many two-state folders perfectly linear free energy relationships have been reported up to a 30 kJ/mol range of variation in protein stability ($p_D = \partial\alpha_D / \partial\Delta G_D^0 = 0$) see Fig.3C [18,47] suggesting that for those proteins the folding transition states are narrow and robust maxima. For several two-state folders, however, non-linear free energy relationships ($p_D \neq 0$) have been observed without evidence for transiently populated intermediate states. Hammond behavior due to a broad transition state region was originally proposed as origin for these non-linear REFERs (Fig.2A) [48]. However, for tendamistat it was shown that the non-linearity is caused by a denaturant-induced switch between two distinct transition states and a high energy intermediate (Figs.2B and 4A [45]. Analysis of curved chevron plots reported for a large number of different proteins showed that a sequential folding model with consecutive transition states and at least one metastable high energy intermediate is able to explain denaturant-induced non-linear activation free energy relationships in many other apparent two-state folders [49]. This argues for the generality of free energy landscapes with a limited number of sequential transition states in apparent two-state folding.

Structural information on the properties of transition states for several proteins has been obtained from ϕ_f ($=\alpha_S$)-value analysis. The results from these studies give a picture of transition states as distorted native states for the major part of a protein (diffuse transition states) or for large substructures (polarized transitions states). For diffuse transition states the ϕ_f-values are around 0.3 throughout the protein, indicating partial formation of the native set of interactions [42].

Cross-interaction parameters have only rarely been used to characterize protein folding reactions. For tendamistat, the denaturant dependence of the folding reaction was measured at different pressures [29]. This allowed the calculation of the pressure/denaturant cross-interaction parameter (p_{Dp}). Fig.5 shows equilibrium transition curves and chevron plots at pressures between 20 and 1000 bar. Increasing pressure destabilizes tendamistat leading to a shift of the GdmCl-induced unfolding transition to lower GdmCl concentrations (Fig.5A) and indicating that the native state has a larger volume than the unfolded state.

Interestingly, the volume of the transition state exceeds the volume of the native state at high denaturant concentrations as indicated by a decrease in the unfolding rate constant with increasing pressure above 5 M GdmCl (Fig.5B). This scenario is not treated in the original Leffler postulate, which assumes that the properties of the transition state are between those of the reactants and the products. The denaturant/pressure-induced non-linear activation free energy relationships in tendamistat indicate a transition state movement towards a less solvent-exposed structure when the protein is destabilized [29].

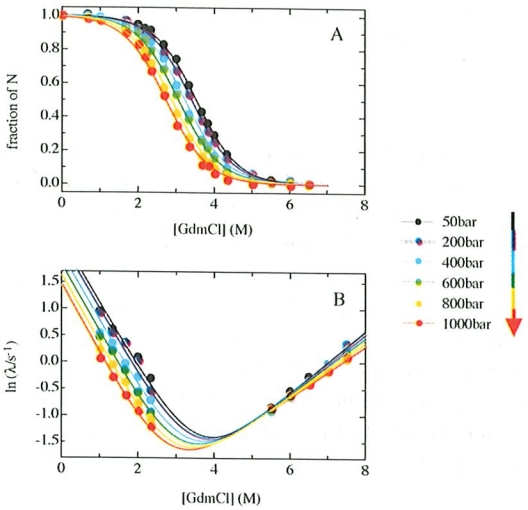


Fig. 5: *Effect of pressure on (A) the GdmCl-induced unfolding transition and (B) the GdmCl-dependence of the apparent rate constant for tendamistat folding at pH2. The data in figures 5A and 5B were fitted globally with* $\partial m_f / \partial p = \partial \Delta V_f^{0\ddagger} / \partial [GdmCl] = 2.5 \pm 0.6$ (cm^3/mol)/M *(data and fits were taken from ref.[29])*

The data were used to calculate the denaturant/pressure cross-interaction parameter

$$p_{Dp} = \frac{\partial \alpha_D}{\partial \Delta G_p^0} = \frac{\partial \alpha_p}{\partial \Delta G_D^0} = p_{pD} \qquad (25)$$

Fig.6 shows the effect of $\partial \Delta G_p^0$ on the α_D-value for the denaturant-pressure dependent folding data shown in figure 5. The α_D-value increases significantly with decreasing stability indicating a positive p_{Dp}-value (Fig.6A). Fig.6B shows that the effect of $\partial \Delta G_p^0$ is only observed in the kinetic m-values while m_{eq} is unchanged due to compensating changes in m_f and m_u. This indicates a true denaturant/pressure-induced transition state movement.

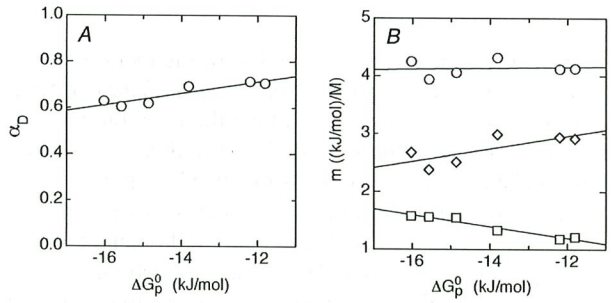


Fig. 6: *Effect of* $\partial \Delta G_p^0$ *on* α_D *(A) and on* m_{eq} *(○)* m_f *(◇) and* m_u *(□) (B) for the pressure/denaturant perturbation of tendamistat folding shown in Fig.5.*

The popularity of the ϕ_f-value analysis of folding transition states provided a wealth of data on α_D, ΔG^0 and $\Delta G_{f,u}^{0\ddagger}$ for many variants of several proteins. This allowed the calculation of denaturant/structure cross-interaction parameters (p_{DS}) to analyze the effect of a change in protein stability (ΔG^0) caused by a structural change on the α_D-value, i.e. on the relative change in solvent accessible surface area in the transition state [27,28].

$$p_{DS} = \frac{\partial \alpha_D}{\partial \Delta G_S^0} = \frac{\partial \phi_f}{\partial \Delta G_D^0} = p_{SD} \tag{26}$$

For several proteins non-zero p_{DS}-values were reported in literature [48,50-52]. For most proteins this change in the position of the transition state could be shown due to ground state effects on the structure of the unfolded state rather than caused by Hammond behavior [28]. As an example, Fig.7 shows the effect of a change in ΔG^0 induced by mutations ($\partial \Delta G_S^0$) on the α_D-values (panel A) and on the m_{eq}-, m_f- and m_u-values (panel B) for an SH3 domain. An increasing α_D-value with decreasing protein stability is observed in accordance with a relative movement of the transition state towards the native state. However, the m_{eq}-value changes significantly upon mutation indicating structural changes it at least one of the ground states. The value of m_f changes by the same amount as m_{eq} while m_u is essentially independent of protein stability (Fig.7B). This indicates that the structure of the transition state does not move relative to the structure of the native state (cf. Fig.2).

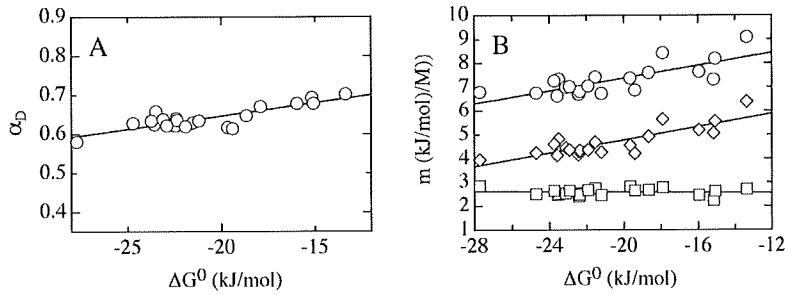


Fig. 7: *Effect of changes in ΔG_S^0 (A) on α_D and (B) on m_{eq} (○), m_f (◇) and m_u (□) for the Sso7d SH3 domain.*

An increase in m_{eq}- and m_f-values point at the disruption of interactions in the unfolded state upon mutation which leads to a less compact unfolded state. These results show that the apparent shift in the position of the transition state along the reaction coordinate seen for the SH3 domain is a ground state effect on the unfolded state rather than Hammond behaviour. Structural changes in the unfolded state can be rationalized in the light of the increasing evidence for residual structure of unfolded proteins. Recently, several high resolution NMR studies have shown that both native-like and non native-like interactions are present in unfolded states of several proteins [53-63] [64]. These interactions may be disrupted by mutations and thus lead to changes in the m_{eq}-values [65]. The analysis of all available data in literature on the effect of mutations on the kinetic and equilibrium m-values shows that most apparent transition state movements upon mutation are due to ground state effects in the unfolded protein. Only for a few proteins weak Hammond behaviour is observed, which is usually accompanied by pronounced ground state effects.

2.7 Conclusions

Self-interaction and cross-interaction parameters are useful tools to detect non-linear rate-equilibrium free energy relationships in protein folding. They can give valuable information on the shape of free energy barriers and thus on the mechanism of protein folding. However, non-linear free energy relationships in protein folding can have different molecular origins like a change in the rate limiting step of the folding reaction (Fig.4B) or Hammond behaviour (Fig.8C,D). As in simple organic reactions structural changes in proteins (e.g. a point mutation) can significantly change the structure of the native state and/or of the unfolded state leading to ground state effects. These effects can, however, be distinguished from genuine transition state movement if the position of the transition state relative to both the native state and the unfolded state is determined. The analysis of transition state movements in a large number of well-characterized proteins using self-interaction and cross-interaction parameters shows that Hammond behaviour is rare in protein folding reactions. This indicates that folding transition states are narrow regions on the free energy landscape which are robust against perturbations. Apparent transition state movements detected by p_{DS} cross-interaction parameters are in most proteins due to ground state effects, most commonly due to structural changes in the unfolded state. This supports the presence of residual structure in unfolded proteins. The analysis of the denaturant-induced REFERs for a large number of proteins further supports sequential folding pathways with obligatory high-energy intermediates as a simple and general explanation for curvatures in chevron plots. This suggests that apparent two-state folding and folding through transiently populated intermediates share similar free energy landscapes with consecutive barriers on sequential pathways. The major difference between two-state and multi-state folding is the relative stability of partially folded intermediates.

3 Intrachain Diffusion and the Speed Limit for Protein Folding

An important issue in protein folding is the rate at which a folding polypeptide chain can explore conformational space to form energetically favourable interactions. This process determines the maximum speed at which a protein can reach its native state. The rate at which unstructured polymer chains can explore conformational space is limited by intrachain diffusion, i.e. by the rate at which two points on a polymer chain can make contact. The knowledge of the rates of intrachain contact formation in polypeptide chains and their dependence on amino acid sequence and chain length is therefore essential for the understanding of the dynamics of the earliest steps in protein folding and for the characterization of the free energy barriers for protein folding reactions. Free diffusion of particles in solutions was treated extensively almost 100 years ago [66-68] but until recently, only little was known about the absolute rates of intrachain diffusion restricted by motions in polypeptides. Numerous theoretical studies have been made to investigate the process of intrachain diffusion in polymers [14,69-72]. Using a first passage time approach Szabo, Schulten & Schulten [70] predicted single exponential behaviour for intrachain contact formation, when the interconversion between individual conformations is fast and when the relative number of conformations allowing end-to-end contact is small (SSS-theory). These studies were not able to give absolute numbers but they predicted that the rate of contact formation scales with $N^{-1.5}$ for intrachain diffusion processes in polymers, where N is the number of polymer building blocks, viz. the number of amino acids in the case of a polypeptide chain.

This value is in agreement with the earlier treatment of a similar process by Jacobsen and Stockmayer [69], who treated end-to-end diffusion as a purely entropic search process.

3.1 Using TTET to measure absolute rates of intrachain diffusion

We have applied the method of triplet-triplet energy transfer (TTET) between thioxanthone and 1-naphthyl alanine (NAla) to measure intrachain contact formation in unstructured polypeptide chains [73,74]. Triplet-triplet energy transfer (TTET) from a triplet donor to a triplet acceptor group involves transfer of two electrons (Dexter mechanism) and requires *van der Waals* contact between donor and acceptor [75]. Efficient electron transfer occurs if the triplet donor has a high extinction coefficient, a high intersystem crossing efficiency and a higher triplet energy than the acceptor.

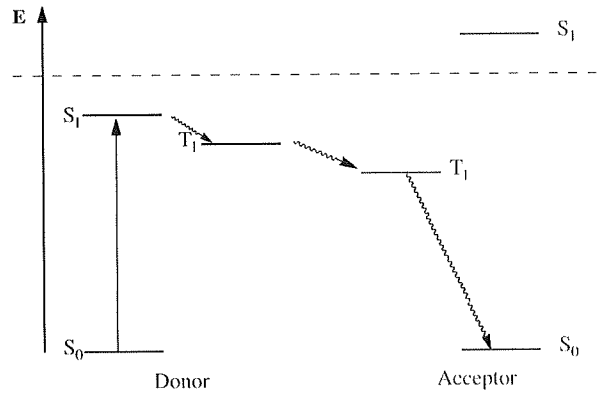


Fig. 8: *Jablonski diagram for triplet-triplet energy transfer(TTET).*

In suitable donor/acceptor pairs the transfer process is diffusion-controlled and can easily be followed by changes in the absorbance of the triplet states. This makes TTET a perfect tool to measure absolute rates of contact formation between two reporter groups. When triplet donor and acceptor groups are attached to a polypeptide chain, the TTET kinetics are determined by the dynamics and by the conformational properties of the chain. In unfolded polypeptides a large number of different conformations are present in equilibrium which gives a wide distribution of donor-acceptor distances. Some polypeptide chains statistically have contact between donor and acceptor at the time of the laser flash. This will lead to instantaneous TTET in these molecules. The remaining molecules have to sample conformational space by intrachain diffusion which will also lead to collisional TTET (Fig.9). The observed transfer kinetics in these molecules reflects the rate at which chain dynamics allows the two labeled points on the polypeptide chain to meet. This assumes that electron transfer is faster than chain diffusion (k_T, $k_{TTET} >> k_c$ and k_{-c}; Fig.9).

We used naphthalene as a triplet acceptor and xanthone as a triplet donor. The triplet states of xanthone are produced by a 4 ns laser flash at 355 nm. Formation of xanthone triplets from the excited singlet state is a rapid process with a time constant ($1/k_T$) of about 3 ps and a high quantum yield for intersystem crossing (~99%).

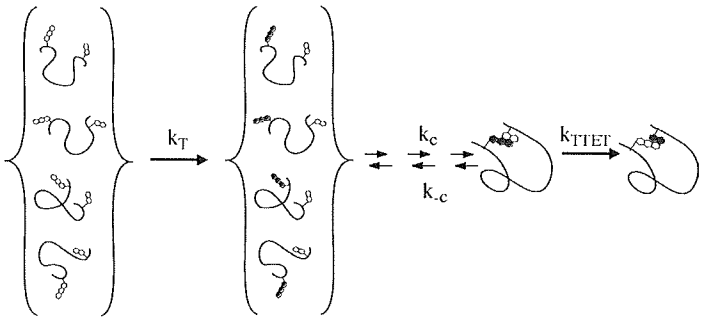
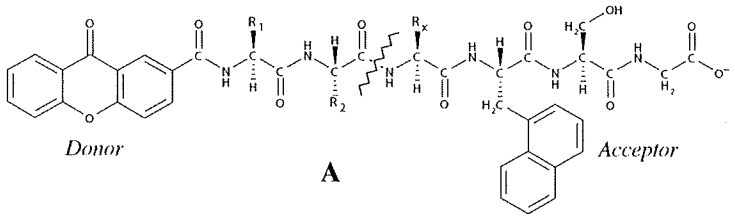


Fig. 9: *Schematic representation of TTET experiments to measure rate constants intrachain diffusion between specific positions in polypeptide chains.*

Xanthone triplets are long lived ($t_{1/2}\approx15$ µs in water), which allows measurements of TTET processes from several nanoseconds (limited by the length of the laser pulse) up to about 20 µs. TTET from xanthone to naphthylacetic acid has a bimolecular rate constant of $k_t=4\cdot10^9$ $M^{-1}s^{-1}$ in water for donor and acceptor free in solution, which is the value expected for a diffusion-controlled bimolecular reaction. Characterization of the photochemistry of the TTET reaction using a 30 fs laserflash setup showed a time constant ($1/k_{TTET}$) around 3 ps. This shows that electron transfer is fast and allows measurements of absolute rate for diffusion processes slower than about 20 ps.

The xanthone group was attached to the N-terminus of various model peptides and the non-natural amino acid 1-naphthyl alanine (NAla) was used to introduce the naphthalene acceptor group near the C-terminus. The length and the sequence of the polypeptide chain between donor and acceptor were varied which gave peptides (A) of the general form



A

where R_1 and R_2 represent any amino acid side chain and *x* indicates the number of amino acids between donor an acceptor. All peptides used in the studies described in the following were unstructured as judged by CD and NMR spectroscopy.

3.2 Intrachain contact formation in glycine-rich polypeptides

To measure intrachain diffusion in flexible polypeptide chains a series of poly(glycine-serine) peptides of different length was synthesized. Fig.10 shows kinetics of intrachain TTET for a Xan-$(Gly\text{-}Ser)_{14}$-NAla-Ser-Gly peptide. The formation of xanthone triplets is reflected by the laser-induced appearance of an absorbance band at 590 nm.

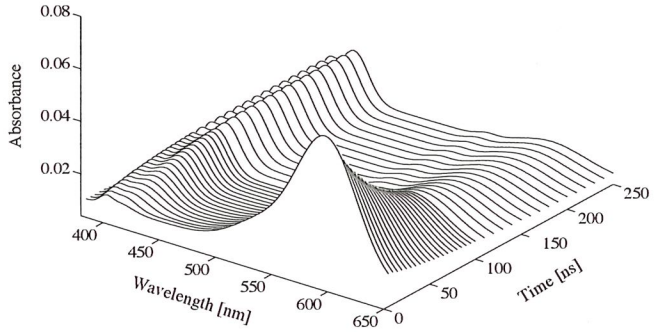


Fig. 10: *Time-dependent change in the absorbance spectrum of a Xan-(Gly-Ser)$_{14}$-NAla-Ser-Gly peptide after a 4 ns laser flash at 355 nm. The decay in the intensity of the xanthone triplet absorbance band around 590 nm is accompanied by a corresponding increase in the naphthalene triplet absorbance band around 420 nm (taken from ref.[74]).*

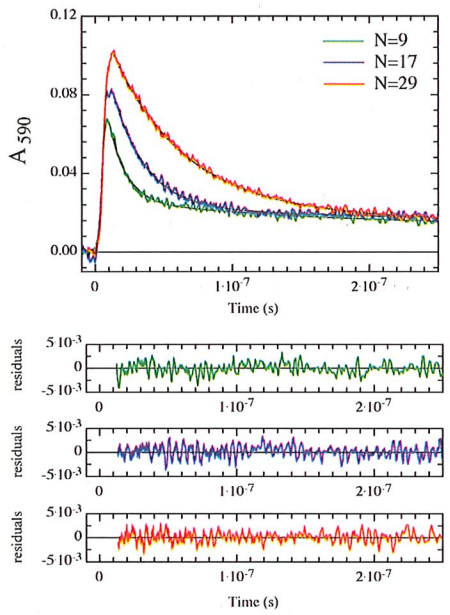


Fig. 11: *Time course of formation and decay of xanthone triplets in peptides of the form Xan-(Gly-Ser)$_n$-NAla-Ser-Gly after a 4 ns laser flash at t=0 and measured by the change in absorbance of the xanthone triplets at 590 nm. Data for different numbers of peptide bonds (N) between donor and acceptor are displayed. Additionally, single exponential fits of the data (solid lines) and the corresponding residuals are displayed (data taken from ref. [74]).*

Decay of the xanthone triplets is accompanied by the appearance of a new absorbance band at 420 nm indicating formation of naphthalene triplets through intramolecular TTET. Monitoring the time-dependent decrease of the xanthone triplet absorbance band at 590 nm (Fig. 11) clearly shows that intrachain contact formation can be described by single exponential kinetics [74], in agreement with predictions from SSS-theory [70]. A fit of the data gives a time constant for contact formation ($\tau=1/k_c$) of 57±3 ns for formation of an *i, i+29* contact (Fig.11).

The observation of single exponential kinetics indicates that (i) the rates of interconversion between the different conformations in the ensemble of random coil states is fast and thus allows the chain to maintain the equilibrium distribution of the ensemble of conformations that has not made contact and that (ii) the number of equilibrium conformations that have donor and acceptor in *van der Waals* contact is small compared to the total number of states [70]. Rapid interconversion between the different conformations in the ensemble of unfolded states is in agreement with sub-nanosecond relaxation kinetics in a cyclic azo-peptide after a photo-induced switch from the *cis* to the *trans* conformation [76]. The TTET kinetics can not be due to intermolecular exchange processes. The peptide concentrations used in the experiments are typically in the 50-100 μM range, which should not allow bimolecular processes with half-times faster than 50 μs, even for the shortest peptides. This was confirmed in experiments using a 1:1 mixture of peptides containing either xanthone or naphthalene both present at the same concentration as in the intramolecular transfer experiments. In these experiments no TTET from xanthone to naphthalene was observed. Also contributions from through-bond transfer processes can be neglected, since this can not occur over distances beyond 8 bonds [77,78]. Even in the shortest peptides used for TTET donor and acceptor were separated by 11 bonds.

3.3 Scaling of end-to-end diffusion with chain length

The scaling of the end-to-end diffusion with chain length was measured by varying the number of amino acids between xanthone and naphthylalanine between 2 and 56 amino acids. This allows measurements of contact formation kinetics for distances ranging from *i, i+3* to *i, i+57*, which covers the range of side chain contacts in small proteins. Fig.11 displays three representative TTET kinetics for peptides of different length. For all polypeptide chains single exponential kinetics for contact formation are observed. Fig.12 shows the effect of increasing chain length on the rate constant for end-to-end diffusion. For long donor-acceptor distances (x>20) the rate of contact formation decreases with $N^{-1.7\pm0.1}$ ($N=x+1$ is the number of peptide bonds between donor and acceptor). This indicates a stronger distance dependence than expected for purely entropy-controlled intrachain diffusion in ideal freely-jointed Gaussian chains, which should scale with $k \sim N^{-1.5}$ (refs. [69,70]) However, Flory already pointed out that excluded volume effects should significantly influence the chain dimensions [79]. Accounting for excluded volume effects in the end-to-end diffusion model of Szabo, Schulten and Schulten [70] leads to $k \sim N^{-1.8}$, which is nearly identical to the value found for the long poly(glycine-serine) chains. This indicates that the dimensions and the dynamics of unfolded polypeptide chains in water are significantly influenced by excluded volume effects, which is in agreement with recent results on conformational properties of polypeptides derived from simplified Ramachandran maps [80]. Another source for the stronger distance dependence observed in our experiments might be an additional small enthalpic barrier, which was observed in the temperature dependence of intrachain diffusion (F.K. & T.K., in preparation). For shorter chains (N<20) the observed simple scaling law breaks down and for very short distances contact formation becomes virtually independent of chain length with a limiting value of $k_0 = 1.9 \cdot 10^8\ s^{-1}$. As pointed out above, the limiting rate for contact formation in short chains is not due to limits of triplet-triplet transfer (k_T in Fig.9), since this process occurs on the sub-nanosecond time scale. Obviously, the intrinsic dynamics of polypeptide chains are limited by different processes for motions over short and over long segments.

This is in agreement with polymer theory which suggests that the properties of short chains are strongly influenced by chain stiffness which leads to a breakdown of theoretically derived scaling laws for ideal chains [81].

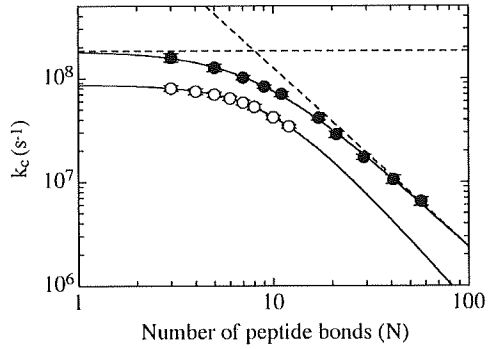


Fig. 12: *Effect of increasing chain length (N) on the rate constant of contact formation in a series of peptides with alternating glycine-serine (●) and with polyserine (○) between donor and acceptor. The solid lines represent fits to the equation* $k_c = \frac{1}{1/k_0 + 1/(k_i \cdot N^m)}$. *The fits give values of* $k_0 = (1.8\pm0.2)\cdot10^8\ s^{-1}$, $k_i = (6.7\pm1.6)\cdot10^9\ s^{-1}$ *and* $m = -1.72\pm0.08$ *for the poly(glycine-serine) series and of* $k_0 = (8.7\pm0.8)\cdot10^7\ s^{-1}$, $k_i = (1.0\pm0.8)\cdot10^{10}\ s^{-1}$ *and* $m = -2.1\pm0.3$ *for polyserine. The dashed lines represent the limiting regimes for contact formation in the poly(glycine-serine) peptides with a length independent upper limit for contact formation of* $k_0 = (1.8\pm0.2)\cdot10^8\ s^{-1}$ *and a length-dependent term of* $k_i \cdot N^m = 6.7\cdot10^9 \cdot N^{1.7}$ *(taken from ref.[74].*

To characterize the dimensions and the stiffness of polymer chains Flory [81] introduced the characteristic ratio (C_n)

$$C_n = \frac{\langle r^2 \rangle_0}{nl^2} \tag{27}$$

for a chain with n segments of length l. The characteristic ratio relates the increase in the mean square end-to-end distance $\langle r^2 \rangle_0$ of an unperturbed polymer chain, i.e. of a real chain without intramolecular interactions, to the change in $\langle r^2 \rangle_0$ of an ideal random-walk or freely jointed chain, which grows with $\langle r^2 \rangle_0 = nl^2$. For short chains C_n increases with chain length due to preferential chain propagation into one direction. For long chains C_n reaches a constant limiting value (C_∞). In this limit $\sqrt{\langle r^2 \rangle_0}$ grows proportional to $n^{1/2}$, like an ideal random-walk chain. However, in a real chain $\sqrt{\langle r^2 \rangle_0}$ increases by a factor of C_∞ more per segment compared to an ideal random-walk chain. The value of C_∞ gives the average number of consecutive chain segments that propagate in the same direction ("statistical segment"). Larger values of C_∞ therefore indicate stiffer chains. For polypeptides chains poly-glycine represents the most flexible chain with $C_\infty = 2$ (refs. [82,83]).

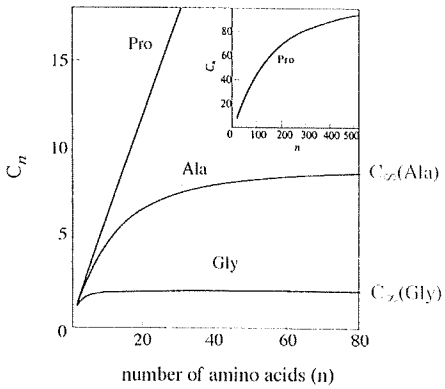


Fig. 13: *Effect of increasing chain length (n) on characteristic ratio (C_n) for polyglycine, polyalanine and polyproline (Figure taken from ref. [81]).*

The 1:1 mixture of glycine and serine used in our experiments is expected to have a C_∞-value of about 3 and C_∞ should be reached for N>10 (ref. [84]). Fig.12 shows, however, that the poly(glycine-serine) chains behave like random chains only over distances longer than 20-30 amino acids when the contact rates scale with $k_c \sim N^{-1.7}$. This indicates increased chain stiffness compared to the predicted value which could be due to specific intrachain hydrogen bonds or van-der-Waals interactions and to excluded volume effects that restrict the number of chain conformations. This model is supported by the measurements of activation energies for the peptides of different length. Diffusion over long distances shows activation enthalpies around kT (2.5 kJ/mol at 25 °C) whereas motions over short distances encounters significant energy barriers around 15 kJ/mol.

3.4 Effect of amino acid sequence on intrachain diffusion

Both theoretical considerations [82-84] and experimental results [85] show that polypeptide chains are especially flexible around glycyl residues. All amino acids except proline (C_∞>100) and glycine (C_∞=2) are predicted to have C_∞ values around 9 and C_∞ should be reached for intrachain distances longer than about 40-50 amino acids [82-84]. This indicates increased chain stiffness and longer mean end-to end distances ($\sqrt{\langle r^2 \rangle_0}$) compared to the poly(glycine-serine) series. To test the effect of chain stiffness on peptide dynamics we measured the rates of contact formation in polyserine peptides [74]. Fig.12 shows the distance dependence of intrachain diffusion in a series of Xan-$(Ser)_x$-NAla-Ser-Gly peptides with x=2-11 (N=3-12). We observed single exponential kinetics for contact formation in all polyserine peptides. For short chains (N<5) contact formation is virtually independent of chain length with a limiting value of $k_0 = 8.7 \cdot 10^7$ s^{-1}. This shows that the local dynamics in polyserine are about 2-fold slower than in the poly(glycine-serine) peptides, which seems to be small compared to the largely different properties expected for the stiffer polyserine chains. A possible origin of this rather small effect is that the decreased flexibility and the longer end-to-end distances in the polyserine chains are compensated by the decreased conformational space available for polyserine compared to poly(glycine-serine) peptides.

For longer polyserine chains contact formation slows down with increasing chain length. The effect of increasing chain length on the rates of contact formation seems to be slightly larger in polyserine compared to the poly(glycine-serine) series ($k \sim N^{-2.1\pm0.2}$). However, due to limitations in peptide synthesis it was not possible to obtain longer peptides, which will be required to get a reliable scaling law for the polyserine peptides.
The kinetics of contact formation over short distances differ only by a factor of 2 for polyserine compared to poly(glycine-serine) (Fig.12). To investigate the effect of other amino acids on local chain dynamics, TTET was measured in short host-guest peptides of the canonical sequence Xan-Ser-Xaa-Ser-NAla-Ser-Gly using the guest amino acids Xaa = Gly, Ser, Ala, Ile, His, Glu, Arg and Pro [74].

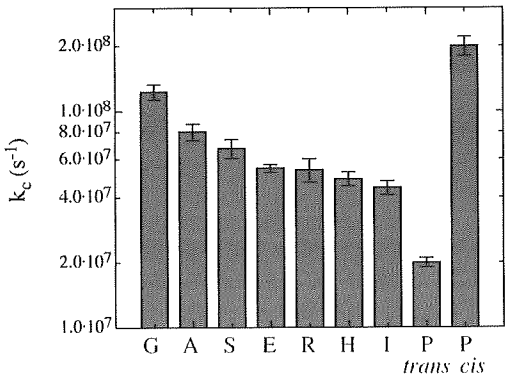


Fig. 14: *Effect of amino acid sequence on local chain dynamics measured in host-guest peptides with the canonical sequence Xan-Ser-Xaa-Ser-NAla-Ser-Gly. The guest amino acid Xaa was varied and the rate constants for the different guest amino acids are displayed (data taken from ref.[74]).*

Fig.14 shows that the amino acid side chain indeed has only little effect on the rates of contact formation. All amino acids except proline and glycine show very similar dynamics. Interestingly, there is a small but significant difference in rate between short side chains (Ala, Ser) and amino acids with longer side chains (Ile, Glu, Arg, His). Obviously, chains that extend beyond the C_β-atom slightly decrease the rates of local chain dynamics, whereas charges do not influence the dynamics. Glycyl and prolyl residues show significantly different dynamics compared to all other amino acids, as expected from their largely different conformational properties [83]. As seen before (Fig.12) glycine accelerates contact formation about 2-fold compared to serine. Proline shows slower and more complex kinetics of contact formation with 2 relaxation times of $1/k_1 = 2.5 \cdot 10^8$ s^{-1} and $1/k_2 = 2 \cdot 10^7$ s^{-1} and respective amplitudes of $A_1 = 80\pm5\%$ and $A_2 = 20\pm5\%$. This essentially reflects the *cis-trans* ratio at the Ser-Pro peptide bond in our host-guest peptide, which has a *cis* content of 16±2% as determined by 1D ^{1}H NMR spectroscopy using the method described by Reimer et al. (ref. [86]). Since the rate of *cis-trans* isomerization is slow (1/k~ 20 s at 22 °C) there is no equilibration between the two isomers on the time scale of the TTET experiments. This allows the measurement of the dynamics of both the *trans* and the *cis* form. The results show that the two isomers significantly differ in their dynamic properties of *i, i+4* contact formation.

3.5 Time scales for the earliest steps in protein folding

The results described above show that the amino acid sequence has only little effect on the local dynamics of polypeptide chains. All amino acids show very similar rates of end-to-end diffusion with time constants between 12 and 20 ns for the formation of *i, i+4* contacts. Polypeptide chains are significantly more flexible around glycine (τ=8 ns) and stiffer around prolyl residues (τ=50 ns for the *trans* isomer). Presumably due to the shorter distances resulting from a *cis* peptide bond the rates of *i, i+4* contact formation are fastest around the *cis* Ser-Pro bond (τ=4 ns). These results allow us to set an upper limit to the rates of formation of the first productive local contacts during protein folding (Fig.15). β-hairpins, which are the most local structures in proteins, are often glycine and proline rich [87]. Gly-Ser is actually one of the most frequent sequences found in hairpin loops. Our results indicate that the time constants for the first steps in the formation of the tightest turns with *i, i+3* contacts are around 5 ns for Gly and *cis* Xaa-Pro (Fig.14). In glycine- and proline-free turns these rates are slowed down to about 10 to 20 ns, depending on the amino acid sequence. Formation of α-helices is most likely initiated by formation of a helical turn, which involves formation of an *i, i+4* interaction [88,89]. Since helices are usually free of glycyl and prolyl residues the initiation can not occur faster than about 12-20 ns (Figs.12 and 14). It should be noted that these rates do not represent the time constants for nucleation of helices and hairpins (Fig.15), which requires formation of more than one specific interaction and most likely encounters additional barriers. Our results rather represent an upper limit for the kinetics of secondary structure formation and can thus be used to calculate the height of the free energy barriers once rates for helix and hairpin nucleation can be measured directly. The measured time constants for intrachain contact formation can, however, directly be used to predict the kinetics of loop formation, which are typically of the size of 6 to 10 amino acids [90]. Contacts at the end of loops form with time constants of about 15 ns for glycine-rich loops and 30-40 ns for stiffer loops.

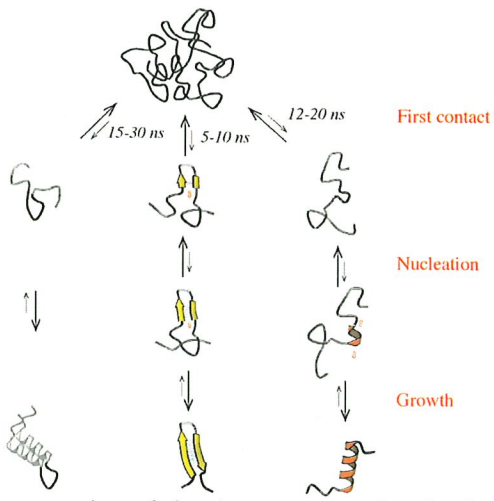


Fig. 15: *Schematic representation of the time constants for the first steps in formation of loops, β-hairpins and α-helices during protein folding derived from the intrachain diffusion rates measured by TTET (taken from ref.[74].*

References

[1] Levinthal, C., J. Chim. Phys. 65, 44 (1968).
[2] Tanford, C., Adv. Prot. Chem. 23, 121 (1968).
[3] Tanford, C., Adv. Prot. Chem 23, 218 (1968).
[4] Tanford, C., Adv. Prot. Chem. 24, 1 (1970).
[5] Baldwin, R.L., Annu. Rev. Biochem. 44, 453 (1975).
[6] Kim, P.S., Baldwin, R.L., Annu. Rev. Biochem. 59, 631 (1990).
[7] Jaenicke, R., Progr. Biophys. Mol. Biol. 49, 117 (1987).
[8] Schmid, F.X., Curr Opin Struct Biol 2, 21 (1992).
[9] Jaenicke, R., Progr. Biophys. Mol. Biol. 71, 155 (1999).
[10] Zwanzig, R., Szabo, A., Bagchi, B., Proc. Natl. Acad. Sci. USA 89, 20 (1992).
[11] Bryngelson, J.D., Wolynes, P.G., Proc. Natl. Acad. Soc. USA 84, 7524 (1987).
[12] Wolynes, P.G., Onuchic, J.N., Thirumalai, D., Science 267, 1619 (1995).
[13] Thirumalai, D., *Theoretical perspectives on in vitro and in vivo protein folding*, in *Statistical Mechanics, Protein Structure, and Protein Substrate Interactions*, S. Doniach, Editor. 1994, Plenum Press. p. 115.
[14] Thirumalai, D., J. Phys. 5, 1457 (1995).
[15] Guo, Z., Thirumalai, D., Biopolymers 36, 83 (1995).
[16] Chan, H.S., Dill, K.A., J. Chem. Physics 100, 9238 (1994).
[17] Kiefhaber, T., *Protein Folding Kinetics*, in *Methods in Molecular Biology, Vol. 40: Protein Stability and Folding Protocols*, B.A. Shirley, Editor. 1995, Humana Press: Totowa, NJ. p. 313.
[18] Jackson, S.E., Folding & Design 3, R81 (1998).
[19] Leffler, J.E., Science 117, 340 (1953).
[20] Leffler, J.E., Grunwald, E., *Rates and equilibria of organic reactions*. 1963, New York: Dover. 458.
[21] Hammond, G.S., J. Am. Chem. Soc. 77, 334 (1955).
[22] Jencks, W.P., Chem. Rev. 85, 511 (1985).
[23] Farcasiu, D., J. Chem. Ed. 52, 76 (1975).
[24] Jencks, W.P., *Catalysis in Chemistry and Enzymology*. 1969, New York: McGraw-Hill Book Company.
[25] Evans, M.G., Polanyi, M., Trans. Faraday Soc. 31, 875 (1935)
[26] Eyring, H., J. Chem. Phys. 3, 107 (1935).
[27] Sánchez, I.E., Kiefhaber, T., Biophys. Chem. 100, 397 (2003).
[28] Sánchez, I.E., Kiefhaber, T., J. Mol. Biol. 327, 867 (2003).
[29] Pappenberger, G., Saudan, C., Becker, M., Merbach, A.E., Kiefhaber, T., Proc. Natl. Acad. Sci. USA 97, 17 (2000)
[30] Jäger, M., Nguyen, H., Crane, J.C., Kelly, J.W., Gruebele, M., J. Mol. Biol. 311, 373 (2001).
[31] Pohl, F.M., FEBS Lett. 65, 293 (1976).
[32] Chen, B.L., Baase, W.A., Schellman, J.A., Biochemistry 28, 691 (1989).
[33] Privalov, P.L., Makhatadze, G.I., J Mol Biol 224, 715 (1992).
[34] Greene, R.F.J., Pace, C.N., J. Biol. Chem. 249, 5388 (1974).
[35] Santoro, M.M., Bolen, D.W., Biochemistry 27, 8063 (1988).
[36] Ikai, A., Fish, W.W., Tanford, C., J. Mol. Biol. 73, 165 (1973).
[37] Matthews, C.R., Meth. Enzymol. 154, 498 (1987).
[38] Myers, J.K., Pace, C.N., Scholtz, J.M., Protein Sci. 4, 2138 (1995).

[39] Goldenberg, D.P., Frieden, R.W., Haack, J.A., Morrison, T.B., Nature 338, 127 (1989).
[40] Fersht, A.R., Matouschek, A., Serrano, L., J. Mol. Biol. 224, 771 (1992).
[41] Itzhaki, L.S., Otzen, D.E., Fersht, A.R., J. Mol. Biol. 254, 260 (1995).
[42] Sánchez, I.E., Kiefhaber, T., J. Mol. Biol. 334, 1077 (2003).
[43] Fersht, A.R., Itzhaki, L.S., elMasry, N.F., Matthews, J.M., Otzen, D.E., Proc. Natl. Acad. Sci. USA 91, 10426 (1994).
[44] Martinez, J.C., Pisabarro, M.T., Serrano, L., Nat. Struct. Biol. 5, 721 (1998).
[45] Bachmann, A., Kiefhaber, T., J. Mol. Biol. 306, 375 (2001).
[46] Matthews, J.M., Fersht, A.R., Biochemistry 34, 6805 (1995).
[47] Schönbrunner, N., Koller, K.-P., Kiefhaber, T., J. Mol. Biol. 268, 526 (1997).
[48] Matouschek, A., Fersht, A.R., Proc. Natl. Acad. Sci. USA 90, 7814 (1993).
[49] Sánchez, I.E., Kiefhaber, T., J. Mol. Biol. 325, 367 (2003).
[50] Matouschek, A., Otzen, D.E., Itzhaki, L., Jackson, S.E., Fersht, A.R., Biochemistry 34, 13656 (1995).
[51] Silow, M., Oliveberg, M., Biochemistry 36, 7633 (1997).
[52] Oliveberg, M., Tan, Y.-J., Silow, M., Fersht, A.R., J. Mol. Biol. 277, 933 (1998).
[53] Evans, P.A., Topping, K.D., Woolfson, D.N., Dobson, C.M., Proteins 9, 248 (1991).
[54] Neri, D., Billeter, M., Wider, G., Wüthrich, K., Science 257, 1559 (1992).
[55] Logan, T.M., Theriault, Y., Fesik, S.W., J. Mol. Biol. 236, 637 (1994).
[56] Smith, C.K., Bu, Z.M., Anderson, K.S., Sturtevant, J.M., Engelman, D.M., Regan, L., Protein Science 5, 2009 (1996).
[57] Sari, N., Alexander, P., Bryan, P.N., Orban, J., Biochemistry 39, 965 (2000).
[58] Wong, K.B., Clarke, J., Bond, C.J., Neira, J.L., Freund, S.M., Fersht, A.R., Daggett, V., J. Mol. Biol. 296, 1257 (2000).
[59] Teilum, K., Kragelund, B.B., Knudsen, J., Poulsen, F.M., J. Mol. Biol. 301, 1307 (2000).
[60] Kortemme, T., Kelly, M.J., Kay, L.E., Forman-Kay, J.D., Serrano, L., J. Mol. Biol. 297, 1217 (2000).
[61] Garcia, P., Serrano, L., Durand, D., Rico, M., Bruix, M., Protein Sci. 10, 1100 (2001).
[62] Choy, W.Y., Forman-Kay, J.D., J. Mol. Biol. 308, 1011 (2001).
[63] Kazmirski, S.L., Wong, K.B., Freund, S.M., Tan, Y.J., Fersht, A.R., Daggett, V., Proc. Natl. Acad. Sci. USA 98, 4349 (2001).
[64] Yi, Q., Scalley-Kim, M.L., Alm, E.J., Baker, D., J. Mol. Biol. 299, 1341 (2000).
[65] Klein-Seetharaman, J., Oikawa, M., Grimshaw, S.B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L.J., Dobson, C.M., Schwalbe, H., Science 295, 1719 (2002).
[66] Einstein, A., Ann. d. Phys. 19, 289 (1906).
[67] von Smoluchowski, M., Ann. d. Phys. 21, 756 (1906).
[68] von Smoluchowski, M., Phys. Z. 17, 557 (1916).
[69] Jacobsen, H., Stockmayer, W.H., J. Phys. Chem. 18, 1600 (1950).
[70] Szabo, A., Schulten, K., Schulten, Z., J. Chem. Phys. 72, 4350 (1985).
[71] de Gennes, P.G., J. Phys. Lett. 46, L639 (1985).
[72] Fixman, M., Farady Discuss. Chem. Soc. 83, 199 (1987).
[73] Bieri, O., Wirz, J., Hellrung, B., Schutkowski, M., Drewello, M., Kiefhaber, T., Proc. Natl. Acad. Sci. USA 96, 9597 (1999).
[74] Krieger, F., Fierz, B., Bieri, O., Drewello, M., Kiefhaber, T., J. Mol. Biol. 332, 265 (2003).
[75] Klessinger, M., Michl, J., *Excited states and photochemistry of organic molecules*. 1995, Weinheim/D: VCH.
[76] Spörlein, S., Carsten, H., Satzger, H., Renner, C., Behrendt, R., Moroder, L., Tavan, P., Zinth, W., Wachtveitl, J., Proc. Natl. Acad. Sci. USA 99, 7998 (2002).

[77] Closs, G.L., Johnson, M.D., Miller, J.R., Piotrowiak, P., J. Am. Chem. Soc. 111, 3751 (1989).
[78] Wagner, P.J., Klán, P., J. Am. Chem. Soc. 121, 9626 (1999).
[79] Flory, P.J., *Principles of polymer chemistry*. 1953, Ithaca: Cornell University Press.
[80] Pappu, R.V., Srinavasan, R., Rose, G.D., Proc. Natl. Acad. Sci. USA 97, 12565 (2000).
[81] Flory, P.J., *Statistical Mechanics of Chain Molecules*. 1969, Munich: Hanser Publishers.
[82] Brant, D.A., Flory, P.J., J. Am. Chem. Soc. 87, 2791 (1965).
[83] Schimmel, P.R., Flory, P.J., Proc. Natl. Acad. Sci. USA 58, 52 (1967).
[84] Miller, W.G., Brant, D.A., Flory, P.J., J. Mol. Biol. 23, 67 (1967).
[85] Schwalbe, H., Fiebig, K.M., Buck, M., Jones, J.A., Grimshaw, S.B., Spencer, A., Glaser, S.J., Smith, L.J., Dobson, C.M., Biochemistry 36, 8977 (1997).
[86] Reimer, U., Scherer, G., Drewello, M., Kruber, S., Schutkowski, M., Fischer, G., J. Mol. Biol. 279, 449 (1998).
[87] Wilmot, C.M., Thornton, J.M., J. Mol. Biol. 203, 221 (1988).
[88] Zimm, B.H., Bragg, J.K., Journ. Chem. Phys. 31, 526 (1959).
[89] Lifson, S., Roig, A., J. Chem. Phys. 34, 1963 (1961).
[90] Leszcynski, J.F., Rose, G.D., Science 234, 849 (1986).

C 8 Motor proteins

Joe Howard
Max Planck Institute for Molecular Cell Biology and Genetics, Dresden

Contents

1 Introduction 2
2 Cell Motility and Motor Proteins 2
3 Motility Assays 4
4 Single-Molecules Assays 6
5 Atomic Structures 6
6 Proteins as Machines 7
7 Chemical Forces 9
8 Effect of Force on Chemical Equilibria 9
9 Effect of Force on the Rates of Chemical Reactions 11
10 Absolute Rate Theories 13
11 Role of Thermal Fluctuations in Motor Reactions 15
12 A Mechanochemical Model for Kinesin 16
13 Conclusions and Outlook 18

1 Introduction

Motor proteins are molecular machines that convert the chemical energy derived from the hydrolysis of ATP into mechanical work used to power cellular motility. In addition to specialized motile cells like muscle fibers and cellular processes like cilia, all eukaryotic cells contain motor proteins (Fig. 1). The reason is that eukaryotic cells are large and their cytosols are crowded with filaments and organelles; as a result, diffusion is too slow to efficiently move material from one part of a cell to another [1]. Instead, the intracellular transport of organelles such as vesicles, mitochondria, and chromosomes is mediated by motor proteins. These proteins include myosins and dyneins that are relatives of the proteins found in the specialized muscle and ciliated cells, as well as members of a third family of motor proteins, the kinesins, which are distantly related to the myosin family.

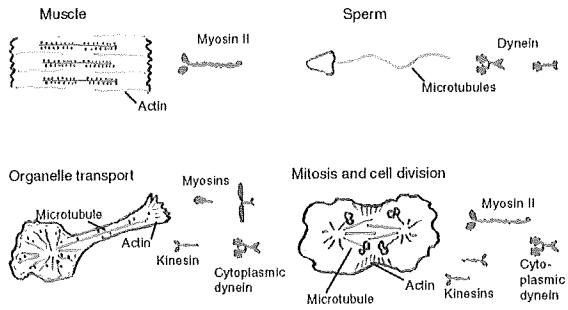


Fig. 1: *Motor proteins and cellular motility*

The focus of this chapter is on how motor proteins work. How do they move? How much fuel do they consume, and with what efficiency? How do chemical reactions generate force? What is the role of thermal fluctuations? These questions are especially fascinating because motor proteins are unusual machines that do what no manmade machines do they convert chemical energy to mechanical energy directly, rather than via an intermediate such as heat or electrical energy. Tremendous insight into this chemomechanical energy transduction process has come from technical developments over the last ten years that allow single protein molecules to be detected and manipulated. The goal of this review is to provide a framework within which to understand these new observations: how do mechanical, thermal, and chemical forces converge as a molecular motor moves along its filamentous track. For background, the reader is directed to *Molecular Biology of the Cell* [2] for an introduction to the biology of cells and molecules, to *Cell Movements* [3] for a broad review of cell motility, and to *Mechanics of Motor Proteins and the Cytoplasm* [4] for more detailed discussion of the mechanics of molecular motors and the cytoskeleton.

2 Cell Motility and Motor Proteins

The study of motor proteins begins with myosin, which drives the contraction of muscle. Myosin was first isolated as a complex with actin filaments by Kühne [5], though it was not until the 1940s that the complex was dissociated into the separate proteins, myosin and actin [6],[7]. The discovery of the myosin crossbridges by H. E. Huxley in 1957 [8], Fig. 2A provided a molecular basis for the contraction of muscle: the bending or rotation of these

crossbridges causes the actin-containing thin filaments to slide relative to the myosin-containing thick filaments, and the sliding of these filaments, in turn, leads to the shortening of the muscle, as had been demonstrated a few years earlier [9],[10].

Since its initial discovery, the crossbridge (also called a head) has proven to be central to the mechanism of cell motility. Dynein, which drives the beating of cilia, was identified in the 1963 [11]. The dynein crossbridges cause the adjacent doublet microtubules to slide with respect to each other (Fig. 2B). Because shear between the microtubules at the base of the cilium (e.g. near the head of the sperm) is prevented by strong linkages, the sliding is converted into bending of the microtubules along the length of the cilium. In this way the sperm undergoes its snake-like propulsion through solution. Kinesin, which moves organelles along microtubules, was purified in 1985 [12],[13]. Attached to the cargo at one end, the crossbridges at the other end of the kinesin molecule walk along the surface of the microtubule (Fig. 2C).

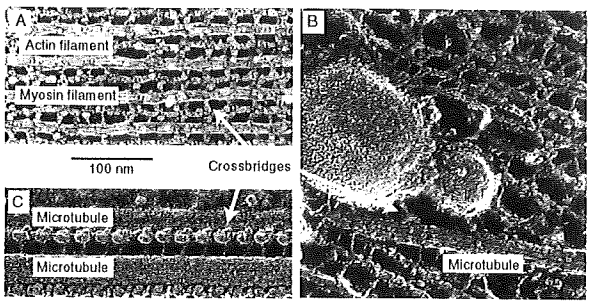


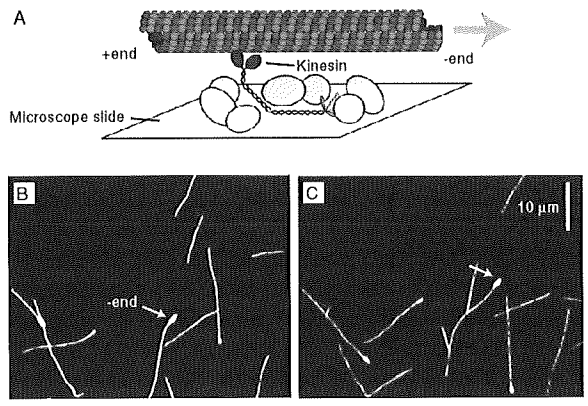
Fig. 2: *Crossbridges formed by the motor domains of motor proteins drive motion along cytoskeletal filaments. A. Muscle. Myosin crossbridges protruding from the thick, myosin-containing filament drive the sliding of the thin, actin-containing filaments. B. Sperm. Dynein crossbridges cause the sliding of adjacent microtubules. C. Kinesin crossbridges walk along microtubules carrying organelles*

In all cases studied in detail, the motion of a motor protein is directed. Actin filaments and microtubules are polar structures made of asymmetric protein subunits, and a given motor always moves towards a particular end of the filament. The myosin, dynein and kinesin families have a large number of members. For example, humans have 33 genes which code for proteins of similar amino acid sequence to the heavy chain of muscle myosin (http://www.gene.ucl.ac.uk/nomenclature/, http://www.mrc-lmb.cam.ac.uk/myosin/myosin.html), and they have 21 dynein heavy-chain genes and 45 kinesin genes [14]; http://www.blocks.fhcrc.org/~kinesin/index.html. Interestingly, different myosins go in different directions along actin filaments and different kinesins go in different directions along microtubules. This is important because the orientation of actin filaments and microtubules in cells is tightly controlled: thus, by using differently directed motors cells are able to move cargoes from one part of the cell to another (and back) in order to organize the cell's internal structure.

3 Motility Assays

The study of motor proteins was revolutionized by the development of in vitro motility assays in which the motility of purified motor proteins along purified cytoskeletal filaments is reconstituted in cell-free conditions. An important milestone in this development was the visualization of fluorescent beads coated with purified myosin moving along actin cables in the cytoplasm of the alga *Nitella* [15]. This was quickly followed by the first completely reconstituted assay in which motor-coated beads were shown to move along oriented filaments made from purified actin that had been bound to the surface of a microscope slide [16]. Though "threads" of actin and myosin had been known to contract in the presence of ATP [17], this contraction was very slow. The significance of the new findings was that they proved that myosin (together with actin) was sufficient to produce movement at rates consistent with the speeds of muscle contraction and cell motility.

There are two geometries used in in vitro motility assays: the gliding assay and the bead assay. In the gliding assay, the motors themselves are fixed to the substrate, and the filaments are observed under a light microscope as they diffuse down from solution, attach to, and glide along the motor-coated surface (Fig. 3) in the presence of ATP (Fig. 4). In the bead assay, filaments are fixed to a substrate, such as a microscope slide, and motors are attached to small plastic or glass beads with typical diameters of 1 μm. The motions can be recorded and the speed measured by tracking the centroid of the bead or the leading edge of the filament (see [18] for detailed methods). There is good overall agreement between the speed of a motor protein in vitro and the speed of the cellular motion that is attributed to the motor (Table 1).



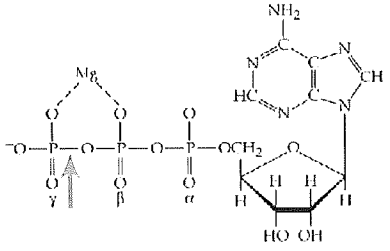


Fig. 3: *Left. A Motility assay in which the motor is bound to a surface and the filament is observed to glide over the surface in the presence of ATP. B, C. Fluorescently labeled microtubules gliding across a kinesin-coated surface (15 seconds between frames). The arrowed microtubule has a bright mark on its slowly growing end (called the minus end): this end leads, showing that kinesin is a minus-end-directed motor.*

Fig. 4: *Right. The hydrolysis of the gamma-phosphate bond (arrowed) of ATP can be summarized by the following reaction:*

$$ATP \leftrightarrow ADP + P_i$$

$$\Delta G = \Delta G_0 - kT \ln \frac{[ATP]_c}{[ADP]_c[P_i]_c} \qquad \Delta G_0 = \ln\left[\frac{[ATP]_{eq}}{[ADP]_{eq}[P_i]_{eq}}\right] = -54 \times 10^{-21}\ \text{J}$$

where the subscript c refers to cellular concentrations and the subscript eq refers to equilibrium concentrations. In cells, the reaction is very far from equilibrium with typical concentrations [ATP] = 1mM, [ADP] = 0.01mM and [P_i] = 1mM; this makes the free energy very large and negative $\Delta G \cong -100 \times 10^{-21}$ J.

Table 1 *Motor speeds in vivo and in vitro*

	Motor	Speed[a] in vivo (nm/s)	Speed[b] in vitro (nm/s)	ATPase[c] (s^{-1})	Function
	Myosins				
1	Myosin IB	?	200	6	Amoeboid motility, hair cell adaptation
2	Myosin II	6,000	8,000	20	Fast skeletal muscle contraction
3	Myosin II	200	250	1.2	Smooth muscle contraction
4	Myosin V	200	350	5	Vesicle transport
5	Myosin VI	ND	-58	0.8	Vesicle transport?
6	Myosin XI	60,000	60,000	ND	Cytoplasmic streaming
	Dyneins				
7	Axonemal	-7,000	-4,500	10	Sperm and cilial motility
8	Cytoplasmic	-1,100	-1,250	2	Retrograde axonal transport, mitosis, transport in flagella
	Kinesins				
9	Conventional	1,800	840	44	Anterograde axonal transport
10	Nkin	800	1,800	78	Transport of secretory vesicles
11	Unc104/KIF	690	1,200	110	Transport of synaptic vesicle precursors and mitochondria
12	Fla10/KinII	2,000	400	ND	Transport in flagella, axons, melanocytes
13	BimC/Eg5	18	60	2	Mitosis and meiosis
14	Ncd	ND	-90	1	Meiosis and mitosis

4 Single-Molecules Assays

The progress of research on motor proteins has gone hand-in-hand with increases in the sensitivity of light microscope techniques. Single protein molecules can now be observed and manipulated. A crucial development was the visualization of individual actin filaments by darkfield microscopy [19]. This was followed by visualization of microtubules by differential interference contrast microscopy [20] and actin filaments by fluorescence microscopy [21]. Further refinement of the motility assays led to detection of movement by single motor molecules [22]. With improved fluorescence sensitivity, it was even possible to image individual fluorescently labeled motors [23] (rather than the much larger filaments), and to watch the motors individually while they move along filaments [24],[25]. The combination of these assays with increasingly sophisticated optical and mechanical techniques such as optical tweezers [26] has allowed measurement of the stepwise movement of motors along their filaments (Fig. 5) and the measurement of the force generated by a single motor protein (Fig. 11).

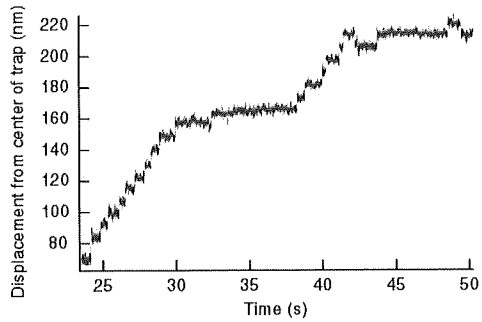


Fig. 5: *A rat kinesin molecule taking 8nm steps along a microtubule at low ATP concentration. (Courtesy of Nick Carter and Rob Cross).*

Single-molecule mechanical and optical techniques are now being applied to many biochemical processes mediated by other molecular machines; these include ATP synthesis [27], DNA transcription [28], DNA replication [29]. The folding of individual proteins [30] and RNA [31] can also be followed. The techniques can even be used to record from molecules on the surfaces of intact cells [32],[33],[34] and recordings deep inside cells should soon be possible. Thus single-molecule techniques are becoming to cell biology what the patch clamp technique and single ion-channel recordings are to neurobiology [35],[36].

5 Atomic Structures

The structural and physical basis for motility has been placed on a firm foundation by the solution of the atomic structures of actin [37] and myosin [38] and of tubulin [39],[40] and kinesin [41]. By fitting the atomic structures into electron micrographs, atomic models of the actin filament [42] and the microtubule [43] have been built. There are also reasonable guesses for how the motors dock to these filaments (e.g. Fig. 6).

The atomic structures have brought many key questions into focus. For example, how do small changes associated with the hydrolysis of ATP (on the order of a few Angstroms) lead to protein conformational changes on the order of several nanometers? What determines the directionality of a motor protein? The detailed answers to these problems will require many additional atomic structures: the motors complexed with their filaments, and the motors with different nucleotides bound to them (e.g. ATP, ADP and P_i, ADP and no nucleotide).

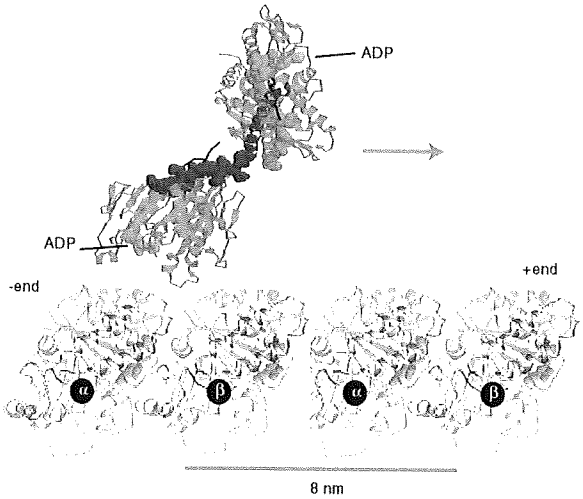


Fig. 6: *Kinesin docked to the microtubule. Kinesin has two identical heads joined by the dimerization domain (dark). Each head binds nucleotide, which was ADP in these crystallization conditions. The microtubule is composed of dimers of the closely-related α and β tubulins. The dimers associated "head-to-tail" to form a polar structure that has the β-subunit at the plus end.*

But this will be very difficult, and even when solved, these structures will provide only static pictures, with no kinetic or energetic information: photographing an internal combustion engine at top dead center and at the bottom of the down stroke does not explain how it works. In the following sections I will address some of physical and thermodynamic questions which are essential to answer in order to understand how conformational changes driven by ATP hydrolysis might generate force and produce directed motion.

6 Proteins as Machines

The structural and single-molecule results reinforce the concept of proteins as machines [44]. According to this view, a molecular motor is an assembly of mechanical parts-springs, levers, swivels and latches-that move in a coordinated fashion as ATP is hydrolyzed. How does such a molecular machine move in response to a internal and external forces? The answer depends on the Reynolds number, which determines the pattern of fluid flow around a moving object. The Reynolds number is equal to

$$Re = \frac{\rho L v}{\eta} \quad (1)$$

where ρ is the density of the liquid, L is the characteristic length of the object (in the direction of the flow), v is the speed of the movement relative to the fluid, and η is the viscosity. Note that the Reynolds number is dimensionless, and its physical meaning is that it is the ratio of the inertial and the viscous forces. For proteins in aqueous solution $\rho \approx 10^3$ kg/m^3, $L \approx$ 10nm, $v \approx$ 1m/s (corresponding to 1nm per ns, which is on the order of the fastest global conformational changes of proteins) and $\eta \approx 10^{-3}$ Pa·s. This makes $Re \approx 10^{-2}$ (and even less for slower motions). A Reynolds number much less than one means that the inertial forces can be neglected and that the motion is highly overdamped: when subject to a force, a protein will creep into a new conformation rather than undergo oscillations. The time constant of the motion is γ/κ, where γ is the drag coefficient ($\sim \eta L$ according to Stokes' law) and κ is the stiffness of the protein.

To get a feeling for how proteins move, imagine that the size of a protein were increased by a factor of 10^7, so that a 10-nm-diameter protein became a mechanical device of diameter 100 mm, fitting nicely in the palm of one's hand (Fig. 7). Now the density and rigidity (Young's modulus) of protein is similar to that of plastic or Plexiglas, so that we can consider that our device is built of plastic [4]. If the viscosity of the fluid bathing the device is increased by the same factor of 10^7 (by putting it in honey, for example), then the ratio of the inertial to the viscous forces will the same for both the protein and the macroscopic device. The Reynolds number will be unchanged (see [4] for the detailed argument) and the pattern of fluid flow will be preserved, just scaled in size. However, to deform the plastic device to the same relative extent as the protein will require a much larger force because the device has a much greater cross-sectional area: whereas a force of only 1 pN might be needed to induce a protein conformational change of 1 nm (corresponding to a strain of 10%), a force of 100 N, corresponding to a weight of 10 kg, would be required to produce the same strain in the plastic device. In response to the respective forces, the protein and the mechanical device will move at the same initial speed, but because the protein conformational change is so much smaller, the relaxation of the protein will be complete in much less time: a relaxation that took an almost imperceptible 100 ns for the protein will take a leisurely 1 s for the device. The work done scales with the volume, so the free energy corresponding to the hydrolysis of one molecule of ATP scales to ~100 J. The scaling is numerated in Table 2.

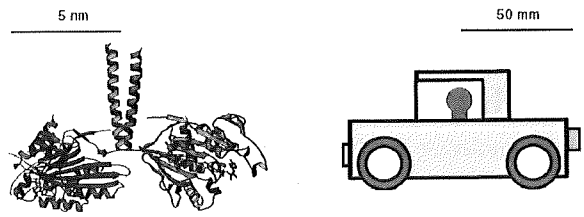


Fig. 7: *Comparison of scales between a motor protein (left) and a toy car (right).*

Table 2. Comparison of scales between a motor protein (left) and a toy car (right).

Quantity	Motor protein	Macroscopic device
Dimension	10 nm	100 mm
Material (Young's modulus)	Amino acids (2 GPa)	Plastic (2 GPa)
Solution (viscosity)	Aqueous (1 mPa·s)	Honey (10 kPa·s)
Speed	1 m/s	1 m/s
Reynolds number	0.01	0.01
Time constant	100 ns	1 s
Force	1 pN	100 N
Energy	$1\text{-}100 \times 10^{-21}$ J	1-100 J

7 Chemical Forces

In addition to mechanical forces, proteins are subject to chemical forces. By chemical forces, we mean the forces that arise from the formation or breakage of intermolecular bonds. For example, consider what happens when a protein first comes in contact with another molecule: as energetically favorable contacts are made, the protein may become stretched or distorted from its equilibrium conformation. Chemical forces could also arise from changes in bound ligands: the cleavage of ATP (Fig. 4) will relieve stresses within the protein that had been built up when the ATP bound. If a protein can adopt two different structures, the binding of a ligand or the change in a bound ligand could preferentially stabilize one of these structures and therefore change the chemical equilibrium between the structures. In this way, we imagine that a chemical change produces a local distortion that in turn pushes the protein into a new low-energy conformation.

To understand how protein machines work, it is essential to understand how proteins move in response to these chemical forces. Just as a chemical force might cause a protein to move in one direction, an external mechanical force might cause the protein to move in the opposite direction. For example, the binding of a ligand might stabilize the closure of a cleft, whereas an external tensile force might stabilize the opening of the cleft; as a result, the mechanical force is expected to oppose the binding of the ligand. Thus mechanical forces can oppose chemical reactions and conversely, chemical reactions can oppose mechanical ones. If the chemical force is strong enough, the chemical reaction will proceed even in the presence of a mechanical load: in this case we say that the reaction generates force.

8 Effect of Force on Chemical Equilibria

The influence of a mechanical force on the chemical equilibrium between two (or more) structural states of a protein can be calculated with the aid of Boltzmann's law. If the difference between two structural states is purely translational i.e. if state M_2 corresponds to a

movement through a distance Δx with respect to state M_1 as occurs when a motor moves along a filament against a constant force-then the difference in free energy is $\Delta G = - F \cdot \Delta x$, where F is the magnitude of the force *in the direction* of the translation. If the length of a molecule changes by a distance Δx as a result of a conformational change, then the difference in free energy is

$$\Delta G \cong \Delta G^0 - F\Delta x \tag{2}$$

where F is the tension across the molecule and ΔG^0 is the free energy difference in the absence of tension. The equality is exact if the molecule is composed of rigid domains that undergo relative translation, or if the two structural states have equal stiffness. Application of Boltzmann's law shows that at equilibrium

$$\frac{[M_2]}{[M_1]} = \exp\left[-\frac{\Delta G}{kT}\right] \cong \exp\left[-\frac{\Delta G^0 - F\Delta x}{kT}\right] = K_{eq}^0 \exp\left[\frac{F\Delta x}{kT}\right] \tag{3}$$

where $K_{eq}{}^0$ is the equilibrium constant in the absence of the force. *The crucial point is that an external force will couple to a structural change if the structural change is associated with a length change in the direction of the force.* If the change in length of a molecule is 4nm, then a force of 1 pN will change the free energy by 4 pN·nm $\cong kT$, the unit of thermal energy where k is Boltzmann's constant and T is absolute temperate. According to Eq. 3, this will lead to an e-fold change in the ratio of the concentrations. Because protein conformational changes are measured in nanometers, and energies range from 1 kT to 25 kT (ATP hydrolysis) (Fig. 4 legend), it is expected that relevant biological forces will be on the scale of piconewtons (10^{-12} N).

An expression analogous to Eq. 2 holds for the effect of voltage on membrane proteins [45]. If a structural change of a membrane protein such as an ion channel is associated with the movement of a charge, Δq, across the electric field of the membrane, then the energy difference between the open and closed states will include a term $V\Delta q$, and this makes the opening sensitive to the voltage, V, across the membrane (Fig. 8). The openings of the voltage-dependent Na and K channels that underlie the action potentials in neurons are strongly voltage dependent: classic experiments by Hodgkin and Huxley showed that the ratio of open probability to closed probability increased approximately e-fold per 4mV [46]. This indicates that the opening of each channel is associated with the movement of about six electronic charges across the membrane ($\Delta q = kT/V \cong 6e$, where e is the charge on the electron). The predicted movement of these electronic charges has been directly measured as a non-linear capacitance of the membrane [47]. Protein conformational changes are sensitive to many other "generalized" forces including membrane tension, osmotic pressure, hydrostatic pressure, and temperature. Sensitivity to these forces requires that conjugate structural changes occur in the protein, respectively area, solute accessible volume, water accessible volume and entropy [4].

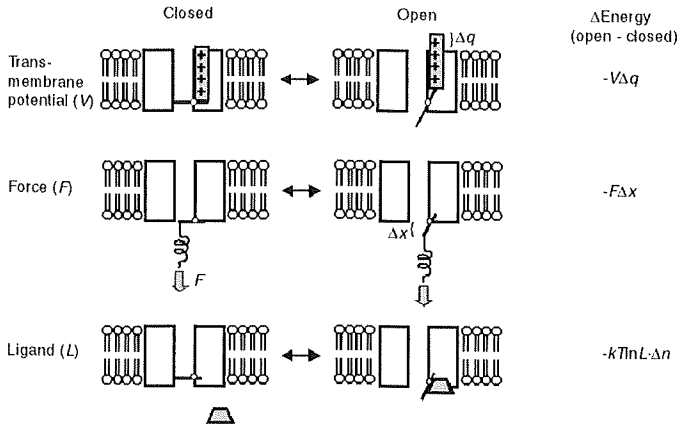


Fig. 8: *Effects of generalized forces on protein conformations. In this example an ion channel sitting in a membrane can be either open or closed depending on the position of the gate. If the gate is coupled to a movement of charge (top), then the channel's probability of being open will depend on the membrane potential. If it is coupled to a vertical movement (middle), then the probability will depend on a vertical force. And if the opening of the gate is coupled to the binding of a small molecule (bottom), then the probability will depend on that molecule's concentration (lnL) and the number of binding sites (Δn).*

The chemical analogy to force and voltage is chemical potential, a measure of the free energy change associated with a molecular reaction. For example, if a ligand at concentration L preferentially binds to one state of a protein over another, then the difference in free energy between the two states is equal to $kT\ln L \cdot \Delta n$, where $kT\ln L$ is the chemical potential of the ligand and Δn is the number of ligand-binding sites on the protein. If a protein is subject to a combination of mechanical forces, electrical forces and chemical forces, then the free energies will add, allowing one to calculate how physical and chemical forces trade off against each other under equilibrium conditions.

9 Effect of Force on the Rates of Chemical Reactions

To understand how force might influence the kinetics of protein reactions is more difficult. Proteins are very complex structures, and a full description of the transition from one structure to another would require following the trajectories of each of the amino acids. The problem is even more difficult because it expected that there are a huge number of different pathways from one structure to another, and so a full description would require the enumeration of all the different pathways, together with their probabilities. This is simply not possible at present given that we do not even understand how a protein folds into even one structure. Thus we need a simple model for the kinetics of protein reactions.

The simplest model for a chemical change between two reactants M_1 and M_2 is a first order process:

$$M_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} M_2 \qquad \frac{d[M_1]}{dt} = -k_1[M_1] + k_{-1}[M_2] \tag{4}$$

This reaction is said to obey first-order kinetics because the rate of change depends linearly on the concentration of the species. The constants of proportionality, k_1 and k_{-1} are called rate constants and they have units of s^{-1}. Many protein reactions have successfully been described by one first order processes e.g. [48], though some reactions may not be describable in this way [49].

Almost inherent within the concept of a first order reaction is that it occurs via a high-energy intermediate, called a transition or activated state. If the transition state has a free energy, G_a, then the rate constant is for the transition is equal to

$$k_1 = A\exp\left[-\frac{\Delta G_{a1}}{kT}\right] \qquad \Delta G_{a1} = G_a - G_1 \tag{5}$$

where A is a constant called the frequency factor. A similar expression holds for the reverse reaction so that the ratio,

$$\frac{k_1}{k_{-1}} = \exp\left[-\frac{\Delta G}{kT}\right] = \frac{[M_2]}{[M_1]} \tag{6}$$

accords with Boltzmann's law at equilibrium.

The activated-state concept makes specific predictions of how rate constants depend on external force. If the protein structures are very rigid and the transitions $M_1 \rightarrow M_a \rightarrow M_2$ are associated with displacements x_1, x_a, and x_2 *in the direction of the force*, F, then the energies of states will be decreased by Fx_1, Fx_a and Fx_2 respectively. This implies that

$$k_1 \equiv A\exp\left[-\frac{\Delta G_{a1} - F\Delta x_{a1}}{kT}\right] = k_1^0 \exp\left[\frac{F\Delta x_{a1}}{kT}\right] \qquad \Delta G_{a1} = G_a - G_1 \quad \Delta x_{a1} = x_a - x_1 \tag{7}$$

An analogous expression holds for k_{-1}. Note that the ratio of the forward and reverse rate constants must give the correct force dependence for the equilibrium (Eq. 3).

A useful way of thinking about the effect of force on the reaction rates is that it tilts the free energy diagram of the reaction (Fig. 9). If the displacement of the activated state is intermediate between the initial and final states ($x_1 < x_a < x_2$), then a negative external force (a load) will slow the reaction, whereas a positive external force (a push) will accelerate the reaction. However, if $x_a = x_1$ i.e. if the transition state is reactant-like then force will have has little effect on the forward rate constant. On the other hand, if $x_a = x_2$ i.e. if the transition state is product-like then the force will have little effect on the reverse rate constant. If the displacement of the activated state is not intermediate between the initial and final states, it is

even possible that a load could increase the forward rate constant (if $x_a < x_1$), though in this case the backward rate would be increased even more.

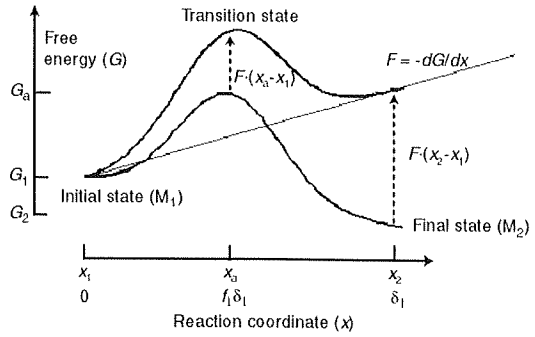


Fig. 9: *Transition -state concept. A high energy barrier limits the rate of the transitions between the initial and final states of a chemical reaction.*

10 Absolute Rate Theories

To predict the absolute rate of a biochemical reaction, a more detailed theory is needed in order to estimate the frequency factor *A*. Two such detailed theories are the Eyring rate theory and the Kramers rate theory. Both require that the reaction coordinate, the parameter that measures the progression of the reaction, be specified. If a protein changes overall length as a result of the $M_1 \rightarrow M_2$ transition, then we could make length the reaction coordinate, though, many other reaction coordinates are possible; indeed, the distance between any two atoms that move relative to one another during the reaction could be used as a reaction coordinate. If the protein is subject to a force, then the natural reaction coordinate is the length of the protein in the direction of the force.

The Eyring rate theory was originally introduced in the 1930s to describe reactions between small molecules such as the bimolecular reaction $2ClO \leftrightarrow Cl_2 + O_2$. The theory assumes that the reaction occurs via a transition state which is in equilibrium with the reactants. The transition state breaks down to the products when one of its molecular vibrations becomes a translation [50], [51], [52]. The rate is

$$k_E = \varepsilon \frac{kT}{h} \exp\left[-\frac{\Delta G_{a1}}{kT}\right] \qquad \Delta G_{a1} = G_a - G_1 \tag{8}$$

where $\varepsilon \sim 1$ is an efficiency factor (equal to the probability of making the transition when in the transition state), and the frequency factor, kT/h, is equal to ~6 x 10^{12} s^{-1}, where h is the Planck constant, and G_a is the free energy of the activated state (ignoring with the vibrational degree of freedom that breaks down). According to the Eyring Eq., the absolute rate is the frequency factor reduced by the exponential term: assuming that the efficiency is unity, a reaction with a rate constant of 10^3 s^{-1} would have an activation free energy of 22 kT. The Eyring theory proved to be a better model than collision theory to describe the rates of small-molecule reactions in the gas phase. It might also apply to covalent changes of proteins and

their ligands. But there is no reason to think that it would apply to global conformational changes of proteins in aqueous solution where there are many degrees of freedom and the motions are expected to be highly overdamped.

Kramers had a different view of the passage through the transition state [53]. He assumed that the transition corresponded to a diffusive motion of a particle out of a potential well. This might correspond to the thermal fluctuation of two protein domains held together by a flexible region. Let the (reduced) mass be m, the damping equal γ, and the spring constant of the well be κ. If the motion is underdamped, then the Kramers expression for the rate is

$$k = \frac{\omega_0}{2\pi} \exp\left(-\frac{\Delta G_{a1}}{kT}\right) \tag{9}$$

where $\omega_0 = \sqrt{\kappa / m}$ is the resonance frequency associated with the potential well. This is a generalization of the Eyring expression which holds for quantum mechanical and non-quantum cases [54]. If the motion is overdamped, the rate is

$$k = \frac{1}{2\pi\tau^*} \exp\left(-\frac{\Delta G_{a1}}{kT}\right) \tag{10}$$

where τ^* is the geometric mean of the damping time constant in the well (γ/κ) and the damping time constant at the top of the barrier (γ/κ^* where κ^* is the negative of the second derivative at the top of the energy barrier). Note that in the overdamped case, the pre-exponential factor depends on the shape of the energy barrier because the efficiency of the reaction (probability of making the transition) depends on the shape. If the peak is precipitous, if there is an absorbing boundary at the top of a harmonic well then the rate constant in the overdamped case is

$$k_K = \frac{1}{\pi\tau}\sqrt{\frac{\Delta G_{a1}}{kT}} \exp\left[-\frac{\Delta G_{a1}}{kT}\right] \tag{11}$$

Because global conformational changes of proteins are expected to be overdamped (Howard 2001), the overdamped Eq.s are more appropriate, and it is for the overdamped result that the Kramers paper is most widely appreciated [54]. According to the overdamped Kramers rate theory, the frequency factor is approximately equal to the inverse of the relaxation time, $\tau = \kappa/\gamma$ where κ is the stiffness of the protein and γ is the damping from the fluid (the other terms are close to unity). This makes intuitive sense: we can think of the protein sampling a different energy level every τ seconds, because τ is the time over which the protein's shape becomes statistically uncorrelated. The protein can react only when it attains the energy of the transition state, and the probability of this occurring is proportional to $\exp(\Delta G_{a1}/kT)$.

The Eyring and Kramers rate theories represent two extreme views of the mechanism of global conformational changes of proteins. In the Eyring model, the transition state is like the initial state. A sudden, local chemical change (such as the binding of a ligand or the chemical change in a bound ligand) creates a highly strained protein which subsequently relaxes into a new stable conformation. In the overdamped Kramers model, the transition state lies towards the final state on the reaction coordinate: after the protein has diffused into the transition state,

the reaction proceeds and locks the conformational change in. The Kramers model is a "foot-in-door" mechanism, in the which the foot plays the role of the chemical change that prevents the opening from being reversed. Obviously these are two extreme cases, and in general the transition state could be anywhere in between the initial and final states (or even outside). The important point of all this is that the various mechanisms can be distinguished: *in the Eyring model the forward rate is independent of force whereas in the Kramers model the forward rate (and perhaps the reverse rate as well) will depend on force.* Thus force dependence offers a way to tell whether a transition is chemical in the sense that it is reached by localized fluctuations, or whether the transition is physical in the sense that it is reached by global spatial fluctuations. The latter, Kramers-type mechanisms have been termed a "thermal ratchet" mechanism (see below).

11 Role of Thermal Fluctuations in Motor Reactions

From a physical viewpoint, we expect thermal fluctuations to play crucial roles in the motor reaction because thermal forces at the molecular scale are large compared to the average directed forces that motor proteins generate. For example, an unconstrained protein will diffuse through a distance equal to its own size, ~10 nm, in tens of microseconds; this is three orders of magnitude faster than the duration of the ATP hydrolysis cycle, which is typically 10 to 100 ms (motors typically hydrolyze 10-100 ATP per second while they are moving, Table 1). Indeed, it is their noisy, diffusive environment that distinguishes molecular machines from the man-made machines of our everyday world. From a chemical viewpoint, we also expect thermal fluctuations to play crucial roles: all chemical reactions require thermal energy in order for a molecule to enter the transition state (see last section), and the chemistry of the hydrolysis reaction is no exception. The challenge is to merge these two views.

An extreme model is that movement of a motor to its next binding site on the filament is purely diffusive, and the role of ATP hydrolysis is to somehow rectify diffusion so that motion in the wrong direction is blocked [55],[56],[57]. This is reminiscent of the pawl and ratchet discussed by Feynman [58]: if a pawl (the motor) and ratchet (filament) are at different temperatures then they can do work without violating the second law of thermodynamics. But motor proteins cannot literally be heat engines because the diffusion of heat is so rapid over molecular dimensions that thermal gradients will dissipate within picoseconds, much faster than the timescale of the biochemical reactions [4]. Instead of being driven by temperature differences like heat engines, motors are driven by chemical reactions (ATP hydrolysis) that are out of equilibrium.

These ratchet ideas have inspired a number of papers in the physics literature that explore the fundamental requirements of directed motility [59],[60],[61],[62],[63]. Some of the physical models are very specific, and fail to explain how real motor proteins move. For example, Astumian and Bier proposed a mechanism in which a motor alternates between times when it diffuses along a filament and times when it is subject to an asymmetric energy profile (a ratchet). However, such a purely diffusive, thermal ratchet model fails in two respects. First, the maximum force that it can generate against viscous loads is small, only $2kT/\delta$, where δ is the distance between binding sites [64]. But for kinesin, which has $\delta = 8$nm, this force is only 1pN, much less than the measured value of 4 to 5pN [64]. And second, because the motor diffuses in the right direction only half the time, it is expected that two molecules of ATP

would be hydrolyzed for each forward step. But for kinesin, there is one step per ATP hydrolyzed [65],[66]. While they may fail in some specific cases, the physical models have never-the-less been very important because they have clearly defined the two conditions that are necessary for directed motion: spatial asymmetry (a ratchet), which for motor proteins arises from their stereospecific binding to polar filaments; and temporal asymmetry which arises from an out-of-equilibrium chemical reaction (e.g. ATP hydrolysis).

A less extreme model is that only part of the motor diffuses to the next binding site, rather than the whole motor. In other words, the motor is elastic, like a spring, and picks up mechanical energy via thermal fluctuations; once it is "cocked" it can then bind to the filament and generate force. This was the picture in the original A.F. Huxley model of muscle contraction [67], and it forms the basis of other ratchet models [68]. If the cocking of the spring requires a global conformation change, then the diffusion time will be limited by the stiffness of the spring and the damping from the fluid (as well as possible internal damping within the protein). This a Kramers-type mechanism. The Huxley model was abandoned in the 1970s by [69],[70],[71] in favor of a mechanism in which a local conformational change in the nucleotide-binding pocket drives the subsequent global conformational change. This is an Eyring-type mechanism. The Kramers-type mechanism was abandoned because it was argued that it would take too long for myosin to pick up an appropriate amount of mechanical energy by diffusion; but if a more reasonable mechanical efficiency of muscle is assumed (~50%) then the diffusion time is not prohibitive [64].

Thermal energy always plays a crucial role in chemical reactions. In the local, Eyring-type model, thermal fluctuations are still needed to get the molecule into the transition state for the localized structural change. The advantage of the local mechanism is that it is faster: because a short lever is stiffer than a longer one and because the damping on a small domain is less than the damping on a large domain, high-energy states can be reached much more quickly through localized conformational changes. For example, if the lever ratio is 10, then a local conformational change occurs 1000-times faster than a global change [4]. The extent to which force-generating protein conformational changes are due to diffusive global conformational changes that are locked in by chemical changes, or to localized chemical changes that drive global changes must be determined experimentally by measuring the dependence of the rates on force.

12 A Mechanochemical Model for Kinesin

To illustrate how these concepts apply to motor proteins, I present a model for the chemomechanical cycle of kinesin. Structural, biochemical and biophysical experiments suggest that the two heads of kinesin alternate in their binding to the microtubule, first one head leading and then the other (Fig. 10, [72],[73],[74],[75]). For simplicity, we assume that the ADP and P_i concentrations are zero (as is approximately the case in the in vitro assays). This gives the following kinetic Eq.

$$\mathrm{M} \underset{k_{-1}}{\overset{k_1 \mathrm{T}}{\leftrightarrow}} \mathrm{M \cdot D \cdot P} \overset{k_2}{\rightarrow} \mathrm{M \cdot P} \overset{k_3}{\rightarrow} \mathrm{M} \tag{12}$$

where M symbolizes the motor, T the ATP, D the ADP, and P the phosphate ion.

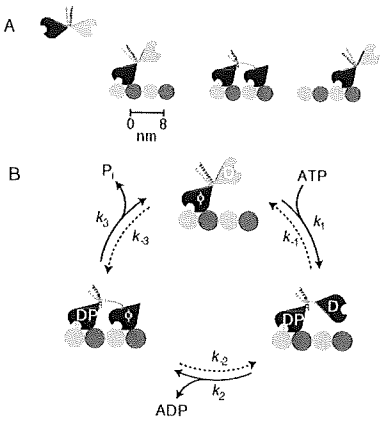


Fig. 10: *Hand-over-hand model for kinesin. T = ATP, D = ADP, Pi = phosphate, ϕ = no nucleotide.*

The solution to this Eq. is

$$v = \delta \cdot k_{\text{ATПασε}} = \delta \cdot k_{\text{cat}} \frac{[\text{T}]}{K_{\text{M}} + [\text{T}]} \qquad k_{\text{cat}} = \frac{k_2 k_3}{k_2 + k_3} \qquad K_{\text{M}} = \frac{k_3(k_2 + k_{-1})}{k_1(k_2 + k_3)} \tag{13}$$

where v is the average speed of movement, k_{ATPase} is the ATPase rate, and δ is the distance per ATP (8 nm for kinesin).

In accordance with our discussion above, we model the force-dependence by assuming that the transitions between states are associated with structural changes through distances $\delta_i = x_{i+1}\text{-}x_i$, so that the rates depend on a *load*, F, in the opposite direction as δ_i, according to

$$k_{+i} = k_{+i}^0 \exp\left[-f_i \frac{F\delta_i}{kT}\right] \tag{14}$$

where f_i is the location of the transition state as a fraction of the distance between the two states E_i and E_{i+1}, and k_{+i}^0 and k_{-i}^0 are the rate constants in the absence of load (Fig. 10). Note that $\delta_1 + \delta_2 + \delta_3 = 8$ nm, the step size. Normally we think that f_i lies between 0 and 1, but this is not an absolute requirement. For the hand-over-hand model, we make the simple assumption that all $f_i = 1$, meaning that the transition state is displaced all the way towards the final state. This puts all the force sensitivity in the forward step, like a Kramers-type mechanism. The solid curves in Fig. 11 are generated with $\delta_1 = 1$ nm, $\delta_2 = 1$ nm, $\delta_3 = 6$ nm; $k_1^0 = 100\ \mu\text{M}^{-1}\text{s}^{-1}$, $k_{-i}^0 = 3000\ \text{s}^{-1}$, $k_2^0 = 105\ \text{s}^{-1}$, and $k_3^0 = 5000\ \text{s}^{-1}$.

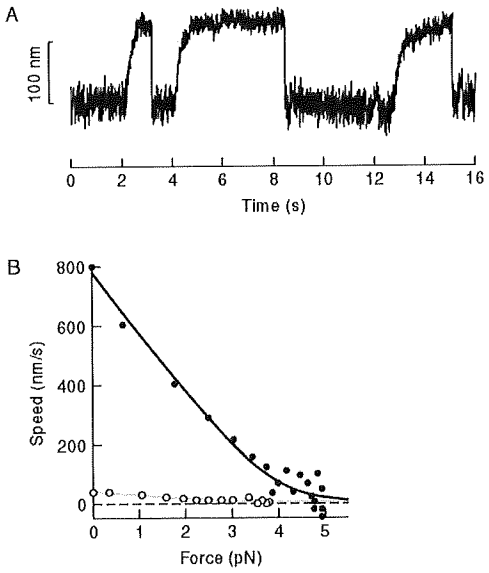


Fig. 11: *A. Three examples of a kinesin molecule walking along a microtubule at a high ATP concentration. The molecule is pulling against a flexible glass fiber and stalls when a maximum force of ~5 pN is reached. B. The speed as a function of the force at high (filled circles) and low (open circles) ATP concentration. The curves correspond to the model in the text.*

13 Conclusions and Outlook

The agreement between the model curve and the experimental data in Fig. 11 shows that the ideas presented in this review are plausible: the motor reaction can be described as a small number of first order chemical reactions in which force effect the rate constants through a Boltzmann-type exponential prefactor. However, the picture is very incomplete. What is the reaction coordinate for a molecular motor? How does diffusion and thermal agitation drive a protein's moving parts during the transition from one structural state to another? There are clearly many conceptual questions to be addressed, in addition to the practical ones. It is likely that ideas from protein folding will prove useful, and that experiments using single-molecule techniques will become increasingly useful for probing the motor reaction.

Study of the mechanics of motor proteins has given us a deeper understanding of biological force generation, and in particular how mechanical, chemical, and thermal forces act on proteins. Because conformational changes of proteins, RNA, and DNA are driven by the mutual action of forces, it is likely that the principles learned from the interactions of motors with their filaments should have wide application to other macromolecular machines.

References

[1] K. Luby-Phelps, et al., Proc Natl Acad Sci U S A, 1987. 84(14): p. 4910-3.
[2] B. Alberts, et al., Molecular biology of the cell. 4th ed. 2002, New York: Garland Pub. xliii, 1294, [67].
[3] D. Bray, Cell movements. 2nd ed. 2000, New York: Garland Pub. xviii, 406.
[4] J. Howard, Mechanics of Motor Proteins and the Cytoskeleton. 2001, Sunderland, MA: Sinauer Associates. 367.
[5] W. Kühne, et al., Untersuchungen über das Protoplasma und die Contractilität. 1864, Leipzig: W. Engelmann. vii, 158 , 8 leaves of plates (1 folded).
[6] F.B. Straub, Studies Inst. Med. Chem. Univ. Szeged, 1941-2. 2: p. 3-15.
[7] A. Szent-Gyorgyi, Stud. Inst. Med. Chem. Univ. Szeged, 1941-2. 1: p. 67-71.
[8] H.E. Huxley, J. Biophys. Biochem. Cytol. (J. Cell Biol.), 1957. 3: p. 631-648.
[9] H.E. Huxley and J. Hanson, Nature, 1954. 173: p. 973-976.
[10] A.F. Huxley and R. Niedergerke, Nature, 1954. 173: p. 971-976.
[11] I.R. Gibbons, Proc. Natl. Acad. Sci. USA, 1963. 50: p. 1002-1010.
[12] S.T. Brady, Nature, 1985. 317(6032): p. 73-5.
[13] R.D. Vale, T.S. Reese, and M.P. Sheetz, Cell, 1985. 42(1): p. 39-50.
[14] H. Miki, et al., Proc Natl Acad Sci U S A, 2001. 98(13): p. 7004-11.
[15] M.P. Sheetz and J.A. Spudich, Nature, 1983. 303(5912): p. 31-5.
[16] J.A. Spudich, S.J. Kron, and M.P. Sheetz, Nature, 1985. 315(6020): p. 584-6.
[17] A. Szent-Gyorgyi, Studies Inst. Med. Chem. of Szeged, 1941. 1: p. 17-26.
[18] J.M. Scholey, Motility assays for motor proteins. Methods in cell biology ; v. 39. 1993, San Diego, Calif.: Academic Press. xv, 304.
[19] H. Nagashima and S. Asakura, J. Mol. Biol, 1980. 136(2): p. 169-82.
[20] R.D. Allen, N.S. Allen, and J.L. Travis, Cell Motil, 1981. 1(3): p. 291-302.
[21] T. Yanagida, et al., Nature, 1984. 307(5946): p. 58-60.
[22] J. Howard, A.J. Hudspeth, and R.D. Vale, Nature, 1989. 342(6246): p. 154-8.
[23] T. Funatsu, et al., Nature, 1995. 374(6522): p. 555-9.
[24] R.D. Vale, et al., Nature, 1996. 380(6573): p. 451-3.
[25] J. Yajima, et al., Curr Biol, 2002. 12(4): p. 301-6.
[26] K. Svoboda, et al., Nature, 1993. 365(6448): p. 721-7.
[27] H. Noji, et al., Nature, 1997. 386(6622): p. 299-302.
[28] M.D. Wang, et al., Science, 1998. 282(5390): p. 902-7.
[29] G.J. Wuite, et al., Nature, 2000. 404(6773): p. 103-6.
[30] A.A. Deniz, et al., Proc Natl Acad Sci U S A, 2000. 97(10): p. 5179-84.
[31] J. Liphardt, et al., Science, 2001. 292(5517): p. 733-7.
[32] Y. Sako, S. Minoghchi, and T. Yanagida, Nat Cell Biol, 2000. 2(3): p. 168-72.
[33] M. Benoit, et al., Nat Cell Biol, 2000. 2(6): p. 313-7.
[34] G.J. Schutz, et al., Embo J, 2000. 19(5): p. 892-901.
[35] E. Neher and B. Sakmann, Nature, 1976. 260(5554): p. 799-802.
[36] B. Sakmann and E. Neher, Single-channel recording. 2nd ed. 1995, New York: Plenum Press. xxii, 700.
[37] W. Kabsch, et al., Nature, 1990. 347(6288): p. 37-44.
[38] I. Rayment, et al., Science, 1993. 261(5117): p. 50-8.
[39] J. Lowe and L.A. Amos, Nature, 1998. 391(6663): p. 203-6.
[40] E. Nogales, S.G. Wolf, and K.H. Downing, Nature, 1998. 391(6663): p. 199-203.
[41] F.J. Kull, et al., Nature, 1996. 380(6574): p. 550-5.

[42] K.C. Holmes, et al., Nature, 1990. 347(6288): p. 44-9
[43] E. Nogales, et al., Cell, 1999. 96(1): p. 79-88.
[44] B. Alberts, Cell, 1998. 92(3): p. 291-4.
[45] B. Hille, Ionic channels of excitable membranes. 2nd ed. 1992, Sunderland, Mass.: Sinauer Associates. xiii, 607.
[46] A.L. Hodgkin, The conduction of the nervous impulse. The Sherrington lectures ; 7. 1964, Liverpool: Liverpool University Press. 108.
[47] C.M. Armstrong and F. Bezanilla, J Gen Physiol, 1974. 63(5): p. 533-52.
[48] O.B. McManus, et al., Biophys J, 1988. 54(5): p. 859-70.
[49] R.H. Austin, et al., Biochemistry, 1975. 14(24): p. 5355-73.
[50] H. Eyring and E.M. Eyring, Modern chemical kinetics. 1963, New York,: Reinhold Pub. Corp. 114.
[51] W.J. Moore, Physical chemistry. 4th ed. 1972, Englewood Cliffs, N.J.: Prentice-Hall. xiii, 977.
[52] P.W. Atkins, Physical chemistry. 3rd ed. 1986, New York: W.H. Freeman. xxiv, 857.
[53] H.A. Kramers, Physica, 1940. 7: p. 284-304.
[54] P. Haenggi, T. Talkner, and B. Borkovec, Reviews of Modern Physics, 1990. 62: p. 251-341.
[55] S.M. Braxton, *Synthesis and use of a novel class of ATP carbamates and a ratchet diffusion model for directed motion in muscle.*, in *Department of Biochemistry and Biophysics*. 1988, Washington State University: Pullman.
[56] S.M. Braxton and R.G. Yount, Biophys. J., 1989. 55: p. 12a.
[57] R.D. Vale and F. Oosawa, Adv Biophys, 1990. 26: p. 97-134.
[58] R.P. Feynman, R.B. Leighton, and M.L. Sands, The Feynman lectures on physics. 1963, Reading, Mass.,: Addison-Wesley Pub. Co. 3 v.
[59] A. Ajdari and J. Prost, C.R. Acad. Sci. Paris II, 1992. 315: p. 1635.
[60] M.O. Magnasco, Physical Review Letters, 1993. 71(10): p. 1477-1481.
[61] J. Rousselet, et al., Nature, 1994. 370(6489): p. 446-8.
[62] R.D. Astumian and M. Bier, Physical Review Letters, 1994. 72(11): p. 1766-1769.
[63] H.X. Zhou and Y.D. Chen, Phys. Rev. Lett., 1996. 77: p. 194-197.
[64] A.J. Hunt, F. Gittes, and J. Howard, Biophys J, 1994. 67(2): p. 766-81.
[65] D.L. Coy, M. Wagenbach, and J. Howard, J Biol Chem, 1999. 274(6): p. 3667-71.
[66] S. Iwatani, et al., Biochemistry, 1999. 38(32): p. 10318-23.
[67] A.F. Huxley, Prog. Biophys. Biophys. Chem., 1957. 7: p. 255-318.
[68] N.J. Cordova, B. Ermentrout, and G.F. Oster, Proc Natl Acad Sci U S A, 1992. 89(1): p. 339-43.
[69] A.F. Huxley and R.M. Simmons, Nature, 1971. 233(5321): p. 533-8.
[70] E. Eisenberg and T.L. Hill, Prog Biophys Mol Biol, 1978. 33(1): p. 55-82.
[71] E. Eisenberg, T.L. Hill, and Y. Chen, Biophys J, 1980. 29(2): p. 195-227.
[72] W.R. Schief and J. Howard, Curr Opin Cell Biol, 2001. 13(1): p. 19-28.
[73] K. Kaseda, H. Higuchi, and K. Hirose, Nat Cell Biol, 2003. 5(12): p.1079-82
[74] C.L.Asbury, A.N. Fehr, and S.M. Block, Science, 2003. 302 (5653):p.2130-4
[75] W.R. Schief et.al., Proc Natl Acad Sci USA, 2004

C 9 Physics of the Cytoskeleton

A. R. Bausch
Lehrstuhl für Biophysik E22
Technische Universität München

Contents

1 Introduction – Function and Role of the Cytoskeleton 2
1.1 Actin - Treadmilling 3

2 Physics of actin networks 5

3 Rheological methods 6
3.1 Macrorheology 7
3.2 Active Microrheology 8
3.3 Passive Microrheology 9

4 Rheology of in vitro networks 10

5 Rheology of living cells 12

1 Introduction – Function and Role of the Cytoskeleton

Most cellular processes are regulated by the cytoskeleton a filamentous network inside living cells. The cytoskeleton is responsible for a manifold of tasks, one of the most obvious is maintaining the cellular shape and the cells' mechanical stability. The malfunctioning of the cytoskeleton (especially its regulation) has severe consequences for the cell and thus for the organism. In cancer the reorganisation of the cytoskeleton, concomitant with the loss of adhesion to the tissue, results in mobile cells. These mobile cancerous cells are able to cross the tissue boundaries and thus result in spreading the cancer to different organs: the metastasis of the cancer.
Here, we want to introduce the main cytoskeleton filaments, describe their major mechanical properties and introduce microrheological techniques, which allow us to study the mechanical properties of living cells, as well as of purified biopolymer networks.

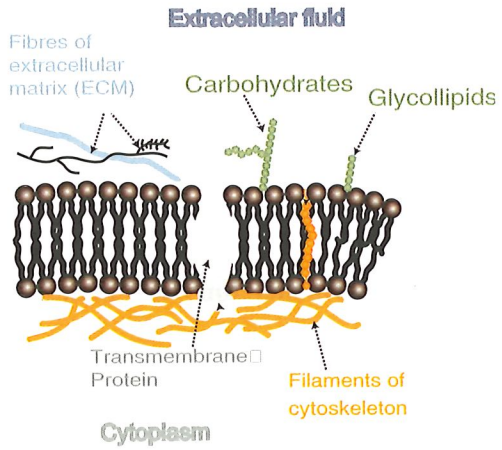


Fig. 1: *Schematic diagram of the outer shell of a cell. To the complex lipid membrane the cytoskeletal network is tightly coupled to the inside of the cell. The glycocalix and the extracellular matrix is coupled to the outer side of the membrane.*

The outer shell of cells is composed of a bilayer composed of lipid molecules (Fig.1). The mechanical properties of pure lipid shells such as vesicles are clearly not sufficient to withstand the mechanical stresses experienced by living organisms. In order to fortify the fluid membrane a rather complicated and complex polymer network, the cytoskeleton, has been developed by nature. The cytoskeleton spans the whole cell and is coupled to the membrane via specific proteins. The cytoskeleton is composed of three major types of filaments: microtubules, actin and the intermediate filaments. The latter describes a rather heterogeneous family of filaments, which has been the least studied until now. To this class belong different filaments such as keratin, desmin, vimentin and neurofilaments. It is thought, that depending on the cell type these 3 types of filaments are important for the mechanical stability. The most prominent structure is the actin cytoskeleton. The actin cytoskeleton is mostly coupled to the plasma membrane and its rapid reorganisation is important for processes like cell migration and cell differentiation. Actin is a semiflexible filament, and its function in the cell is regulated by a myriad of different helper proteins which regulate the polymerisation, cap different filament ends, crosslink the filaments and severe the filaments. In the adhesion of cells to other cells or to the extracellular matrix the coupling of the actin cytoskeleton to the transmembrane proteins, such as integrins plays an important role. Microtubules become most prominent during cell divisions, enabling the mitosis by setting up the mitotic spindle. They are also important structural elements which are commonly used for

transport processes. These filaments have a persistence length much longer than their contour length and can thus be described as rigid rods [1, 2].

1.1 Actin - Treadmilling

All mentioned cytoskeletal filaments are build by polymerisation of monomers. For actin the monomers are called G-actin (globular actin), once polymerized the resulting filamentous form is called F-actin (filamentous actin). The monomers have a binding pocket where either ATP or ADP is tightly bound. Normally the free subunits carry one ATP molecule tightly bound to a specific binding pocket within the protein. As the monomers are unsymmetrical, the resulting filaments are also not symmetric: both ends differ in their structures and thus in their affinity for monomers. The polymerisation of the actin can be described by rate constants for the association and dissociation of monomers to the filament. In order to initiate the polymerisation a nucleus composed of three monomers is needed. This nucleation explains the observed lag phase in the polymerisation kinetics as shown in Fig.2a. After the nucleation time the polymerisation of the filaments progresses rapidly, reaching a plateau, once most monomers are used up. Once the monomers are added to the filament the bound ATP is hydrolysed into ADP. The equilibrium concentration is determined by the rate constants of adding a monomer to the ends of the polymer. As can be seen in Fig.2b, the rate constants depend not only on which side of the filament the monomers are added but also if ATP or ADP is bound to the actin monomer. The end with the higher affinity for monomers is growing faster and is thus called the plus end (or pointed end). The other end, the slower growing one, is termed the minus end (barbed end). The mechanism of polymerisation can be understood with a simple model, where for both ends the growth rates are treated separately (after nucleation occurred):

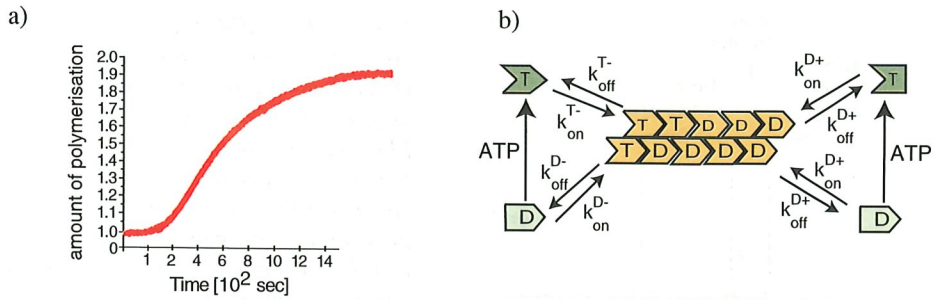


Fig. 2: *a) The polymerisation kinetics of actin shows a lag time after initiation of the polymerisation, which is due to nucleation of actin monomers. b) As the monomers and thus the filament are polar, there are 8 different rates needed to describe the polymerisation process for ATP and ADP actin. The difference in the polymerisation rates at the plus end and the minus end enables the unique treadmilling behaviour of actin.*

$$\text{For the plus end: } \frac{dn^+}{dt} = +k_{on}{}^+[M] - k_{off}{}^+ \qquad (1)$$

Where n is the number of monomers in the filament and k_{on}^- and k_{on}^+ are the *on* rates at the minus or the plus end, whereas the k_{off}^- and k_{off}^+ the respective *off* rates. The rate of change of the number of monomers n (dn/dt) is just the difference of the capture rate and the release rate. The growth or shrinkage of the filaments depends linearly on the monomer concentration

[M] in solution. The fact that the growth and shrinkage occurs only at the ends is an important result of the cooperative effect, which results from the special two stranded structure of the filaments. Neither growth nor shrinking occurs up to a critical monomer concentration M_c, which is defined as the concentration where dn/dt vanishes:

$$\text{For the plus end: } [M]_c^{+} = + \frac{k_{off}^{+}}{k_{on}^{+}}$$
$$\text{For the minus end: } [M]_c^{-} = + \frac{k_{off}^{-}}{k_{on}^{-}} \qquad (2)$$

In cases where the critical concentrations at the plus end M_c^+ is equal with the critical concentration at the minus end M_c^- a filament would grow and shrink at both ends simultaneously. However, for actin the polymerisation rates and thus the critical concentrations are different between the plus and the minus ends of the filament and thus a peculiar behaviour can be found: at a certain monomer concentration the plus ends grows while simultaneously the minus end is shrinking. This can be seen in Fig.3. There is one concentration of monomers where the rate of growth at the plus end equals the rate of shrinking at the minus end. At this steady state monomer concentrations (M_{ss}) the net filament length is constant but the filament is moving forward into the direction of the plus end. Monomers are constantly moving through the filament. This effect is termed treadmilling.

$$[M]_{ss} = \frac{k_{off}^{+} + k_{off}^{-}}{k_{on}^{+} + k_{on}^{-}} \qquad (3)$$

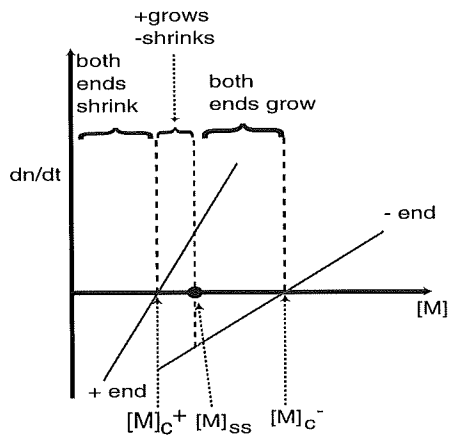


Fig. 3: *The difference in the rate constants and thus the critical concentrations of both ends results in different growth rates at both ends of the filaments in dependence of the free monomer concentration. There is one concentration at which the + end is growing at the same speed as the minus end is shrinking. At this steady state concentration the filament remains at a constant length and moves into the plus end direction simultaneously.*

For ATP actin solutions the rate constants predict treadmilling to be present at steady state concentrations of about 0.2 µM (0,0084 mg/ml) falling in-between the critical concentrations of the plus and minus end. This dynamic behaviour (and its regulation) of the filaments is crucial for the proper function of cells. The tight control of polymerisation and depolymerisation is for example used in the leading edge of motile cells. The properties of actin filaments are unique - there are no synthetic polymer with comparable dynamics [2].

For the local organization of the cytoskeleton inside living cells a myriad of helper proteins is

necessary. These so-called Actin binding proteins (ABPs) can be sorted roughly by their functions like sequestering, capping, severing or cross linking of filaments. Profilin and thymosin are prominent examples of monomer binding proteins. The binding of profilin to actin monomers inhibits the nucleation and thus the production of new filaments. It binds to the monomers in such a way that the polymerisation at the minus end is inhibited. At the same time it accelerates the ADP-ATP exchange and enables the polymerisation at the plus end. In contrast thymosin inhibits the polymerisation at the plus end and the ADP ATP exchange. It is the competition between thymosin and profilin which regulates the polymerisation of existing filaments at the plus end. The activation of profilin accelerates the polymerisation manifold. The importance of the monomer binding proteins becomes clear by looking at the typical actin concentrations found in living cells: Typically, total concentrations of actin in living cells are 2-8mg/ml (50-200μM), roughly 50% of it is in monomeric form.

Capping proteins bind specifically to one end of a filament and inhibit its further extension. Gelsolin for example binds to the plus end of a filament and as a result only the minus end is able to continue extension, this results in much slower growth rate of the filament and a lowering of the steady state length of the filament. Severing is achieved either with simultaneous capping of the new end (fragmin, gelsolin) or without capping (ADF/cofilin).

For the control of the structure of the network several cross linking proteins are present, which are categorized mainly by their structural similarities and differences. This class of ABPs is highly conserved and modular. Typically, two actin binding domains are separated by non binding repeat units. Interestingly, the actin binding domain is strongly preserved structure, similar for most cross linking proteins. Some of the cross linking proteins have a regulatory domain (e.g. the C-terminal Calmodulin-homology domain (CaM) in α-actinin). Depending on the class of crosslinkers the modular distance holders are triple stranded α-helical domains (spectrin units) or β-sheet domains. The length of the rod domain is determined by the number of modular protein domains, which varies over a wide range. Prominent examples of cross linkers are α-actinin, dystrophin, fimbrin, spectrin and utrophin, all having different functions in different cell types. Dystrophin is not only cross linking the filaments underneath the membrane but also coupling them to the fluid lipid membrane. A genetic defect in this protein results in muscle dystrophy, which is most probably due to the destabilized membrane cytoskeletal coupling.

2 Physics of actin networks

Actin is a semiflexible polymer characterized by its persistence length l_p of about 16μm. As the persistence length is of the range of the normal in vitro filament lengths of 10 to 22 μm pure actin solution have unique properties which are fundamentally different from flexible polymer networks. Due to the persistence length the relaxations along a filament can directly be observed by optical microscopy techniques. By labelling single filaments with a fluorophore the shape fluctuations can be made visible. The persistence length is defined as the contour length $l_p=s$, after which the local orientations of the filament becomes uncorrelated. The local orientation is best described by the tangent angle θ(s) at each position s (Fig.4). The curvature is the derivate of the tangent angle with respect to the arc length s. The correlation of the orientation of the local tangent decays exponentially:

$$\langle \Theta(0)\Theta(s) \rangle = e^{-\frac{s}{l_p}} \qquad (4)$$

The persistence length of a filament can be related to the bending elastic modulus B by:

$$B = l_p k_B T \tag{5}$$

which is defined by the bending energy: $\Delta G = \frac{1}{2} B \int \left(\frac{\partial \Theta}{\partial s} \right)^2 ds$. The high persistence length of actin filaments already results in a highly entangled network at very low volume fractions compared to flexible polymer networks. Entangled networks can be characterised by an average entanglement length $l_e \sim \xi$ which determines an elastic plateau, that should be constant over an extended range of frequencies. However, longitudinal fluctuations modes which relax over the persistence length of the polymer give rise to additional elasticity. This additional contribution to the elasticity is unique for semiflexible polymer networks [3].

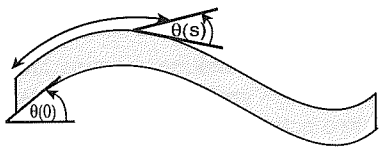


Fig. 4: *Definition of the tangent angle θ(s) at each position s of a filament. The persistence length is defined as the length over which the correlation between the tangent angles decays.*

3 Rheological methods

Despite its importance, little is known about the mechanical properties of living cells and the constituents of the cytoskeleton. Only, in the recent years several new techniques emerged which allow the quantitative determination of the mechanical properties of cells and and cytoskeletal constituents on the local micrometer scale. Depending on the time scale materials can respond to deformations dominantly viscous or dominantly elastic. This time dependent behaviour is described by the viscoelastic parameters and can be determined either in the frequency or the time domain. The most common measurement techniques apply oscillatory strain at different frequencies and measure as response function the resulting stresses. The applied shear strain is of the form $\gamma(t) = \gamma_0 \sin(\omega t)$, where γ_0 is the amplitude and ω is the frequency of the oscillation. In the linear regime the time dependent stress is linearly proportional to the strain, and is given by:

$$\sigma(t) = \gamma_0 [G'(\omega) \sin(\omega t) + G''(\omega) \cos(\omega t)] \tag{6}$$

where G'(ω) is the response which is in phase with the applied strain and is called the elastic or storage modulus, while G''(ω) is the response out of phase with the applied strain and is called the viscous or loss modulus. As the loss modulus describes the viscous dissipation it is in phase with the strain rate. The complex shear modulus is defined as $G^*=G'+iG''$. Alternatively, one can apply a defined stress and measure the strain as response, yielding, in the linear regime, the same insight into the materials properties. This is approach is used in a microrheological method with magnetic tweezers. From the response curves the moduli can be extracted by determining the strain amplitude σ_0 and the phase shift δ and using:

$$G' = \frac{\sigma_0}{\gamma_0}\cos\delta \quad \text{and} \quad G'' = \frac{\sigma_0}{\gamma_0}\sin\delta \tag{7}$$

The need for microrheological approaches come from the fact that for biological samples standard rheological experimental approaches can not be applied, not only because of the dimensions of the typical samples (microns) but also because of the softness of the materials. For example, in order to study the viscoelastic properties of pure actin networks in the linear regime only stresses as low as 20pN can be applied. This and the fact that only small sample volumes can be purified, makes the construction of special setups necessary.

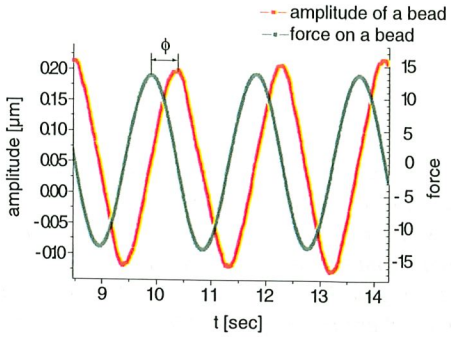


Fig. 5: *Typical experimental curves of an oscillatory magnetic tweezers experiment. The response of the bead (red) is frequency shifted from the applied force curve (green). The moduli can be extracted from the phase shift Φ and the strain and stress amplitudes.*

3.1 Macrorheology

In order to study the mechanical properties of cytoskeletal components a special macrorheometer has to be used. In Fig. 6 the principal of the magnetic force driven torsional rheometer is shown. Here, the sample is contained between a cylindrical cell and the top is closed by a silanized cover slip on which a small permanent magnet is glued. An oscillatory force is applied by electromagnetic coils which are controlled by a computer. 2 of the 4 coils are needed in order to define a resting position. An oscillatory magnetic force is applied by the other 2 coils and the strain is measured by deflected laser light. The deflection of the laser light is detected by a 1dim-photosensitive detection sensor and processed by a computer. The moment of a torque can be measured down to 1 nNm. Typical stresses that can be applied by this technique are 0,001 – 100 Pa and sample volumes are as small as 400µl can be handled with. An important detail in the setup is the prevention of denaturation of the sample by covering the water air interface with a monolayer of DMPC-lipids. The silanized cover slip is thus coupling via the lipid monolayer to the bulk network.

A typical curve of the viscoelastic response of a pure actin solution of 0.4mg/ml can be seen in Fig.8. At 1Hz the crossover frequency can be seen at which G'' is crossing G'. At higher frequencies the single filament dynamics are dominant, at lower frequency a kind of plateau is observable. The lower crossover frequency, after which the viscoelastic behaviour would be described by a reptation model, is difficult to observe with the current setup, due to mechanical instabilities of the network on time scales of hours. These so called, macro-rheometric measurements on actin networks, are ideal to test theoretical predictions on the viscoelastic nature of semiflexible polymer networks. However, in order to get a microscopic understanding of the underlying physics it is mandatory, to develop a technique, which allows the measurement of the moduli on a micron scale. This has been achieved by developing different setups of magnetic tweezers, where forces are applied on micron scaled magnetic

beads. This approach is now commonly referred to by active microrheology. Another recently introduced approach is the observation of the passive motion of micron beads embedded in the network: the passive microrheology.

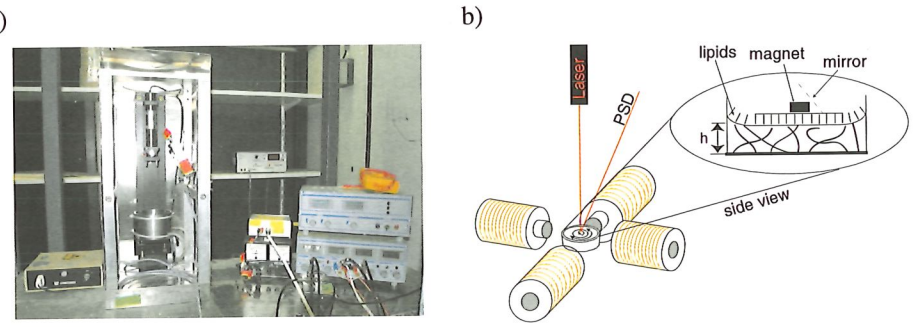


Fig. 6: *For the study of the rheological properties of actin networks a special macrorheometer is necessary. The sample (3 mm high) is placed into a sample cuvette, which is closed by a silanized cover glass, on which a small magnet is placed. On top of the magnet a mirror is glued. The deflection of the laser light from the mirror is monitored by a position sensitive detector (PSD), from which the strain can be determined. Calibration of the instrument is done with a defined viscous medium, where it is important to determine the mass momentum of the glass plate, which is limiting the frequency to frequencies lower than 4Hz.*

3.2 Active Microrheology

The principal magnetic tweezers setup for oscillatory experiments is shown in Fig.7. In order to exert a force on a super paramagnetic bead an inhomogeneous magnetic field has to be applied. The magnetic force f(t) acting on the bead is

$$f(t) = \vec{m}\frac{\partial \vec{B}}{\partial x} \tag{8}$$

Where m is the induced magnetic moment of the particle and B(t) the imposed magnetic field and x the lateral coordiante. In Fig.7 a typical magnetic field of a low force apparatus is shown, the field strength is about 30 mT and the gradient is 6 mT/mm which results in a force of up to 10 pN on particles with a diameter of 4.5 µm. In most experiments super paramagnetic beads, with 20% iron oxide content are used, but there is a variety of beads with different sizes and iron content available. The core inside of the magnetic coils is soft iron, because of the low remanescent field. The sample holder is placed exactly into the middle of both coils, which is simultaneously in the field of view of the microscope objective. By turning the sample holder one can confirm that the bead under observation is in the middle of the inhomogeneous magnetic field. From the magnetic field shown in Fig.7, it is clear that by placing the sample holder in the middle of both coils the gradient of the field can not be maximized. This is the important trade off of this setup and limits the forces that can be exerted on the beads. Higher forces can be achieved by omitting one coil and bringing the super paramagnetic bead as close as possible to the soft iron core. This setup is termed high force setup and in this setup forces up to 10nN can be applied to paramagnetic beads with a diameter of 4.5 µm [4]. These forces are high enough to determine the viscoelastic properties of composite membranes as well as of the cytoplasm of living cells. It is important to use

monodisperse beads, in order to be able to calibrate the forces reliably in calibration liquids.

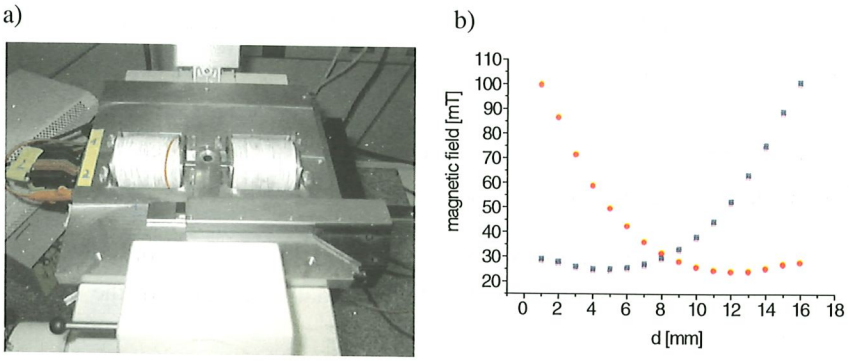


Fig. 7: *a) Typical setup of a magnetic tweezers device. The magnetic coils are driven by a self build amplifier. The control of the amplifier and the online imaging processing is done by a computer. b) The magnetic field of the low force magnetic tweezers setup, when only one coil is switched on. The resulting B field shows that in the middle of the setup the gradient can not be maximized. However, in order to apply symmetric oscillatory forces, it has to be assured that the observed microbeads are in the middle of the setup.*

An extension of the method is to embed in addition to the magnetic beads unmagnetic beads into the medium studied. This allows the determination of the deformation field. In elastic isotropic media, it is known, that a point force results in deformation field which decays with the distance R with 1/R, A point at a distance r from the point force is displaced by

$$\vec{u}(\vec{r}) = \frac{1+\mu}{8\cdot\pi\cdot E(1-\mu)}\cdot\frac{(3-4\cdot\mu)\cdot\vec{F}+\vec{n}\cdot(\vec{n}\cdot\vec{F})}{r}+O\left(\frac{1}{r^3}\right) \qquad (9)$$

Where E is the young modulus of the medium, F the applied force and μ the Poisson ratio. The same functional form of the deformation field is found in viscous media, this time the deformation is proportional to the velocity of a particle. The velocity of a point at a distance r from a moving particle is:

$$\vec{v}(\vec{r}) = \frac{3\cdot R}{4}\cdot\frac{\vec{v}_0+\vec{n}\cdot(\vec{n}\cdot\vec{v}_0)}{r}+O\left(\frac{1}{r^3}\right) \qquad (10)$$

Where R is the particle radius, and v_0 its' velocity. It was possible to show that in pure actin solutions the deformation field mapping can be applied and allows the determination of the Poisson ratio of the medium [5, 6].

3.3 Passive Microrheology

Particles embedded in a medium undergo Brownian motion. The observation of this motion can be used to extract the viscoelastic parameters of the surrounding medium [7, 8]. In the viscous limit one can think of the standard free motion of a particle in a liquid. The mean square displacement is linear in time and by the Stokes Einstein relation the Diffusion

constant D can be related to the viscosity: $\langle x^2 \rangle = 2Dt = 2\frac{k_B T}{\xi} t$, where ξ is the friction coefficient. In the elastic limit one can think of a bead connected to a spring: the spring is extended as far as the thermal energy allows. $\frac{1}{2} G' \langle x^2 \rangle = \frac{k_B T}{a}$, where a is the bead radius and G' is the elastic modulus. Thus for a viscoelastic medium the mean square displacement shows a time dependent sub diffusive behaviour and can be related to the frequency dependent moduli by generalizing the Stokes Einstein relation $D = \frac{k_B T}{\xi}$. By taking a Laplace transform (generalized for a complex frequency s=iω) $\mathfrak{J}_u$ of the mean square displacement one can relate the complex bulk modulus G*(iω) to the mean square displacement.

$$G^*(\omega) = \frac{k_B T}{\pi a i \omega \mathfrak{J}_u \langle \Delta r^2(t) \rangle} \tag{11}$$

The moduli measured are determined by the size of the used particles. By observing single particles local microenvironments at the length scale of the particle are observed.

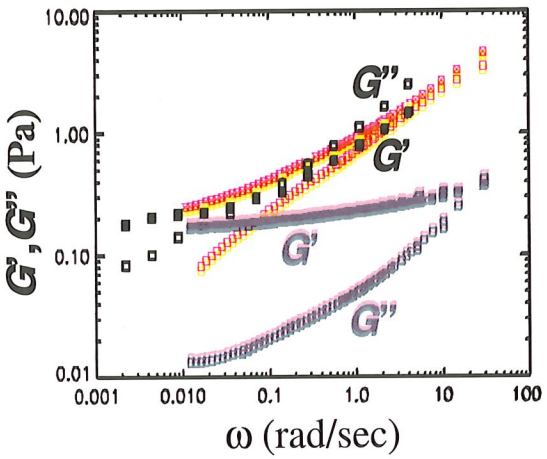


Fig. 8: *Comparison of the different methods to determine the moduli of pure actin networks. The black line are the G' (closed symbols) and G'' (open symbols) determined by macrorheometry. The blue line shows the results obtained by 1 point passive microrheology. Note the discrepancy and the frequency independent storage modulus of the curve. By calculating the correlated motion of embedded particles (2-point microrheology) the absolute values of the macrorheometrical moduli can be extracted (red curve) [9].*

4 Rheology of in vitro networks

The rheology of actin can be determined by both active and passive microrheological methods. Especially, the passive microrheology is ideal to study these soft networks and the underlying dynamics. In Fig.8 the frequency dependent moduli of a pure actin solution measured by passive microrheology and macrorheometry are shown. The agreement between both methods is rather bad: the absolute values of the moduli as well as the form of the curves differ significantly. This is due to the fact that the probe particles can only detect contributions on their dynamics which are resulting from relaxations of their own size. However, at low frequency the elastic modulus is recovered correctly for all concentrations measured.

When the correlated motion between the different probe particles is studied the moduli are in excellent agreement with the macrorheometric data. Analysing the correlated motion between

the particles has been recently introduced and is commonly referred to by 2 point microrheology. In this method the distance dependence of the correlation of the passive motion of embedded micron sized beads is observed [10, 11]. By this method it is possible to measure long wave modes in the material, much longer than the probe particle size. Typically, the positions of 100 beads over 5-20 minutes are observed by a video camera at frame rates of 30Hz and with digital imaging processing the time evolution of the position of each particle is determined. The correlated displacements of pairs of particles, i and j, separated by a distance $10\mu m<R^{ij}<100\mu m$ is determined by the calculating the averaged tensor product of the tracer displacements named the distinct mean square displacement:

$$D_{\alpha\beta}(r,\tau)=\left\langle \Delta r_\alpha^i(t,\tau)\Delta r_\beta^i(t,\tau)\delta\left[r-R^{ij}(t)\right]\right\rangle_{i\neq j,t} \tag{12}$$

Where $\Delta r_\alpha^i(t,\tau)$ is the displacement of the i-th particle in the α component of the direction at the time t with the lag time τ. The average is taken over the time and over all distinct pairs of particles ($i\neq j$). A decay of the correlated motion of 1/r is observed as expected for bulk viscoelastic materials described by the formulars (9) and (10). By extrapolating this two particle displacement correlation tensor to the bead size a, we obtain the 2 particle mean squared displacement, which reflects the one particle motion expected from long-wavelength modes in the material:

$$\left\langle \Delta r^2(\tau)\right\rangle_D = 2r/aD_{rr}(r,\tau) \tag{13}$$

The discrepancy of the 1-point microrheological data and the macrorheological (2-point) data can now be understood by considering that the measurement is dependent on the length scale with which the material is probed [9]. The bulk rheological response of semi flexible polymer networks in the intermediate frequency range between 0.1 and 30 rad/s is dominated by longitudinal density fluctuations of the filaments. These modes relax by diffusing along the filaments. The probe particle diameter sets the length scale of relaxation modes which can be measured. Only relaxations inside the medium which correspond to this length scale are contributing to the measurement. For single particles with a diameter of $a=1\mu m$ only relaxations which are faster than density fluctuations of the filaments on this length scale, can be measured. These modes relax by diffusing along the filament. The relaxations on the scale of the particle diameter will relax on time scales of $\tau_l' \sim \tau_e\left(\frac{a}{\sqrt{l_p l_e}}\right)^2 \sim 0.1\text{sec}$ where τ_e is the entanglement time which is set by the largest possible length scale, the entanglement length l_e, on which transverse fluctuations are possible. The entanglement time τ_e is given by $\tau_e \sim \frac{\zeta l_e}{l_p k_B T} \sim 0.1\text{sec}$ where l_p is the persistence length and ζ the effective friction coefficient of the filament in solution.

By calculating the correlation of movements all relaxation modes on the filaments can be probed as the length scale of the measurement is increased. The lowest frequency of the these excitations along the filaments that affects the bulk rheology is determined by the time taken for the density fluctuation to diffuse a persistence length given by $\tau_l \sim \tau_e \frac{l_p}{l_e} \sim 10\text{sec}$. Thus 1-point microrheology will not probe the elastic response of the network due to these longitudinal fluctuations along the filaments at frequencies which are lower than

$1/\tau_l' = 10$rad/sec. At higher frequencies only the contribution of the relaxations on the length scale of the particle diameter to the modulus will be measured by this technique. On the other hand by determining the correlated motion of the particles relaxations of much longer wavelengths can be measured and thus the 2 point data reflects the bulk modulus.

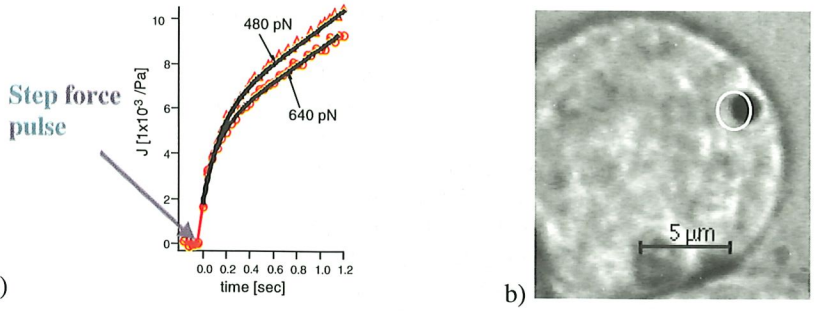
These longitudinal relaxations explain the monotonically increase of the $G'(\omega)$ above 0.1 rad/sec observed in bulk and 2 point rheology. At longer times (smaller frequencies) a frequency independent plateau is reached. This plateau modulus is time independent and just determined by the length scale of the entanglement length l_e $G_0 \sim \frac{\rho k_B T}{l_e} \sim \frac{c_a^{7/5}}{l_p^{1/5}}$. At time scales longer than the longest time needed for the modes relaxing along the filament τ_l this modulus is measured. As this modulus depends on the entanglement length l_e which is comparable to the particle diameter, it can be measured by the probe particle. This plateau is thus measured by the 1-point microrheological experiments at all frequencies smaller than 10rad/sec.

5 Rheology of living cells

The active microrheology is excellently suited to determine the viscoelastic properties of living cells. In fact the method of magnetic tweezers has been introduced 1922 by Heilbronn to study the ectoplasm of plasmodia [12]. The injection of particles into the protoplasm resulted in the observation of the viscoelastic nature of the cytoplasm. The most comprehensive study was done by Crick and Hughes who also introduced the principal of magnetic twisting: non symmetric particles were magnetised by a magnetic field and turned into the perpendicular direction by another field. The experiments were done with chicken fibroblasts [13]. There is a series of experiments using magnetic tweezers to determine the mechanical properties of the cytoplasm of different cells [14-16]. However, progress was hampered by the limited imaging processing facilities and the irregular shape of the used particles. The magnetic twisting method was refined by Valberg et al. [17] and later by Trepat et al. [18]. It was applied to different cell types and most recently the combination with imaging processing allowed the determination of the frequency dependent moduli [19].

To study the local viscoelastic properties of cells the magnetic tweezers technique is best suited. As the viscoelastic moduli of cells are rather high, it is necessary to use only a single coil, the so called high force apparatus. For the incorporation of magnetic beads into the cytoplasm macrophages are best model systems, as they are able to phagocytose micron sized particles [20]. These particles are than incorporated into the cytoplasm as normal phagosomes. It is possible to determine the modulus of the cytoplasm using ferromagnetic beads with diameters of 1 micrometer. In Fig.9 a typical response function to a step force pulse can be seen. This response function can than be analysed using equivalent circuits composed of a Voigt body and a dashpot in series. For the terminal viscosity values of about 200 Pasec and for the shear modulus values between 20-800 Pa have been obtained. The passive motion of the beads show in certain instances rather active movements, which are attributed to the active reorganization of the cytoplasm or even to molecular motors attached to the magnetic beads. The technique of magnetic tweezers is also suited to study these active forces, as the moduli can be determined locally. By introducing unmagnetic beads into the cytoplasm it is possible to study the deformation field induced by the force probes. The force is applied only to the magnetic bead, and as described above, in an isotropic medium a decay of the deformation field of 1/R is expected. This should be observable by the unmagnetic

beads. However, no motion of the unmagnetic beads was observed, even if the unmagnetic and magnetic beads were almost in contact and the forces applied are rather high. This demonstrates the highly heterogeneous nature of the mechanical properties of the cytoplasm of the macrophages. The uncoupled motion of the beads can be either explained by the presence of loosely connected microclusters or by the presence of rather hard walls decoupling the motion of the beads. To this state, there is still a lot of work to be done in order to get a detailed physical insight into the observed behaviour.



a) b)

Fig. 9: *a) Viscoelastic Response of the cytoplasm of J774 macrophages. 1 micron magnetic particles are phagocytosed by the macrophages and force pulses are applied at t=0. The response curve shows an instantaneous elastic response, a relaxation time and a long time viscous behaviour. By applying standard mechanical equivalent circuit for viscoelastic fluids the viscoelastic parameters can be determined. b) a typical microsopic image of a macrophage with an embedded magnetic particle. The white circle marks a typical displacement of the particle after a force pulse of two seconds.*

References

[1]B. Alberts, et al., Molecular biology of the cell. (Garland Science, New York, USA, 2002).
[2]J. Howard, Mechanics of Motor Proteins and the Cytoskeleton. (Sinauer, Sunderland, 2001).
[3]E. Sackmann, A.R. Bausch, and L. Vonna, Physics of composite cell membrane and actin based cytoskeleton. Physics of bio-molecules and cells. ed. H. Flyvbjerg, et al. (Springer, New York, 2002).
[4]A.R. Bausch, et al., Biophysical Journal, 75, 2038 (1998).
[5]F. Ziemann, J. Rädler, and E. Sackmann, Biophys J, 66, 2210 (1994).
[6]F.G. Schmidt, F. Ziemann, and E. Sackmann, Eur Biophys J, 24, 348 (1996).
[7]T.G. Mason and D.A. Weitz, Phys. Rev. Lett., 74, 1250 (1995).
[8]B. Schnurr, et al., Macromolecules, 30, 7781 (1997).
[9]M.L. Gardel, et al., Phys. Rev. Let., 91, 158302 (2003).
[10]A.J. Levine and T.C. Lubensky, Phys. Rev. E, 65, 011501 (2002).
[11]J.C. Crocker, et al., Phys. Rev. Let., 85, 888 (2000).
[12]A. Heilbronn, Jahrb. Wiss. Bot., 61, 284 (1922).
[13]F.H.C. Crick and A.F.W. Hughes, Exp. Cell Res., 1, 37 (1950).
[14]K. Yagi, Exp. Biochem. Physiol., 3, 73 (1961).
[15]Y. Hiramoto, Biorheology, 6, 201 (1970).
[16]M. Sato, et al., Cell Motil., 4, 7 (1984).
[17]P.A. Valberg and H.A. Feldman, Biophys J, 52, 551 (1987).
[18]X. Trepat, et al., Rev. Sci. Inst., 74, 4012 (2003).
[19]B. Fabry, et al., Phys. Rev. Let., 8714 (2001).
[20]A.R. Bausch, W. Möller, and E. Sackmann, Biophys. J., 76, 573 (1999).

C 10 Retinal proteins

G. Büldt, J. Heberle and V. Gordeliy
Institut für Biologische Strukturforschung (IBI-2)
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 The photocycle and proton translocation across bR 2

3 The ground state structure of bR 4

3.1 Electron and neutron diffraction on 2D lattices of bR 4

3.2 The crystal structure of bR 4

4 Intermediate states of the working cycle of bR 6

4.1 Infrared spectroscopy 6

4.2 Part of the mechanism evolves 6

4.3 Changes in the tertiary structure of bR 8

4.4 The crystal structure of the M state 9

5 More details of the mechanism 9

6 Sensory Rhodopsin: transmembrane signalling 11

6.1 The two-component system 11

6.2 The structure of the complex of sensory rhodopsin II and its transducer 12

6.3 Hypothesis on the mechanism of sensory transduction 14

1 Introduction

Retinal proteins, addressed in this lecture, belong to a family of seven-helix membrane proteins with retinal as the light absorbing antenna. They are distributed throughout the Bacteria, Archaea and Eukaryota. The most prominent member of this family is the visual rhodopsin in the rod-outer-segments of the retina of the eye [1]. In general, retinal proteins have two distinct functions: light-dependent ion transport and sensory function. Two archaebacterial rhodopsins will be introduced: bacteriorhodopsin (bR) from *Halobacterium salinarum*, a light driven proton pump [2], and sensory rhodopsin II (SRII) from *Natronobacterium pharaonis*. This sensor activates a signal transduction chain similar to that of the two-component system of eubacterial chemotaxis [3]. These proteins are chosen as examples to demonstrating how information is obtained from different biophysical methods contribute to an understandind of the function of these membrane proteins on the basis of their structure and dynamics.

2 The photocycle and proton translocation across bR

Almost 30 years ago, bacteriorhodopsin (bR) was found in *Haloarchea*. It represents a simple photosynthetic machinery that uses light energy to translocate protons out of the cell interior [2]. Bacteriorhodopsin is densely packed in hexagonal two-dimensional lattices, the so-called purple membranes (PM) in the plasma membrane of the *Halobacterium salinarum*. The characteriziation of the aggregation state of bR in the PM was given by X-ray diffraction [4] on these natural two-dimensional lattices (Fig.1). It was found by Oesterhelt and Stoeckenius [2] that the chromophore, retinal, covalently bound via a Schiff base to Lys 216, enables this protein to pump protons across the cell membrane.

The photocycle of bR (Fig.2) is linked to the active transport of protons by delivering part of the energy of the absorbed photon to this process. In the ground state (BR_{568}) of the photocycle the retinal is in an all-*trans* conformation, the Schiff base being protonated (pK ~ 13) and positively charged. Upon absorption of a photon, the retinal goes via an excited state to a 13-*cis* retinal conformation, which results in a dramatic pK down-shift to 3. Hence in the transition from the L to the M state the proton is released from the Schiff base and after 50 µs a proton appears at the extracellular surface. The first part of the photocycle is quite fast compared to the last part, which occurs in the ms time domain. In this part a proton has to be delivered from the inside of the cell to the Schiff base and the retinal has to reisomerize back to the all-*trans* conformation.

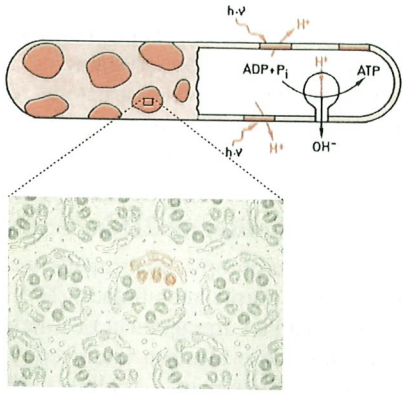


Fig. 1: *Schematic view of the Halobacterium salinarum. bR is aggregated in highly ordered two-dimensional lattices (space group p3). Three bR molecules per elementary cell are surrounded by about 9 lipid molecules. One monomer in a trimer is marked in red. The proton gradient produced by bR drives an ATP-synthase, which synthesizes ATP from ADP and P_i. The energy-rich compound ATP drives numerous cellular processes.*

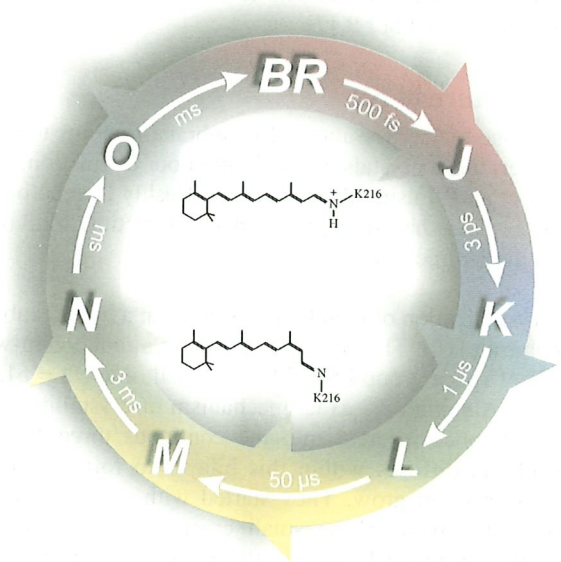


Fig. 2: *The photocycle of bR showing the light-adapted ground state BR_{568} and the intermediates J_{620}, K_{590}, L_{550}, M_{412}, N_{530}, O_{640} (indices: absorption maxima in nm)with their lifetimes. The conformation of all-trans and 13-cis retinals are also depicted. Time-resolved spectroscopy data revealed a splitting of the M intermediate into two substates , an early M_1 and a late M_2 state.*

3 The ground state structure of bR

To elucidate the mechanism of a nanomachine it is desirable to resolve its atomic structure. For water soluble proteins this is achieved by crystallizing the protein and solving the structure using the X-ray diffraction pattern. When structural research on bacteriorhodopsin was started no crystals of an integral membrane protein were ever made. The difficulty arose from the hydrophobic surfaces sticking in the lipid membrane.

3.1 Electron and neutron diffraction on 2D lattices of bR

Therefore electron diffraction and microscopy experiments were performed on the highly ordered two-dimensional lattices, the purple membranes. The basic picture of the tertiary structure, describing the three-dimensional arrangement of α-helices spanning the membrane, was given in several papers on low dose electron microscopy [5]. The knowledge of the amino acid sequence allowed the prediction of a folding model by determining seven α-helical stretches in the polypeptide chain, which were assigned to the seven α-helices seen in the electron microscope structur. In the following years, neutron diffraction was the method of choice to improve this model using bR molecules containing certain perdeuterated amino acids and partially deuterated α-helices [6].
In 1990, cryo-electron microscopy provided an additional breakthrough providing a structure of bR at 3.5 Å resolution in the *X-Y* plane and 7 Å in the *Z*-direction. This picture allowed for the first time, including all the information from other methods mentioned above, the construction of a molecular model of bR [7]

3.2 The crystal structure of bR

Bacteriorhodopsin was one of the first integral membrane proteins to be crystallized in three dimensions [8]. However, these crystals showed only poor diffraction. Considerable progress was made by Schertler et al. [9] using the surfaces of freshly formed benzymidine crystals as nucleation sites. These plate-shaped crystals gave diffraction to 3.6Å in the *a*- and *b*-directions and 4.2Å in the *c*-direction.

A novel concept for crystallization of membrane proteins in a lipidic cubic phase was reported by Landau and Rosenbusch [10]. They crystallized bR in a monoolein-water cubic phase and obtained microcrystals of space group $P6_3$ with a unit cell of a = b = 61.76Å; c = 104.16Å; α = β = 90° and γ = 120°. Although the exact mechanism of crystallization is still unknown, it seems probable that bR, once inserted into the continous three-dimensional curved lipid bilayer of the cubic phase, diffuses within this bilayer to form a nucleation particle from which well-ordered crystals can grow. They started with microcrystals of up to 30 μm in diameter and 5 μm in thickness. These small crystals already gave diffraction to 2.5Å resolution on a specially designed beamline, the microfocus beamline ID13 at the ESRF Grenoble [11]. Using a highly focused beam with a diameter of about the same size as the crystal, the signal-to-noise ratio increased significantly. The cell constant in the *a*- and *b*-directions of 61.76Å together with the hexagonal space group gave an indication for an arrangement of bR molecules similar to the PM.

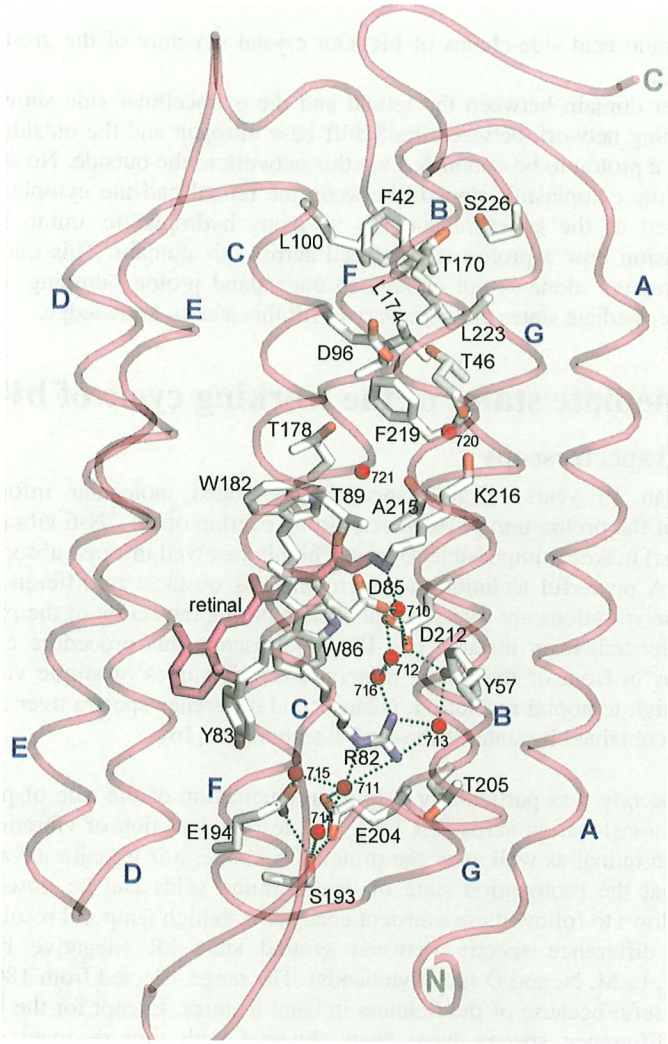


Fig. 3: *Ground-state structure of bR with the cytoplasmic side up and the extracellular side down. The protein backbone is represented by thin ribbons. Several functionally important residues and water molecules are shown. All-trans retinal is in purple colour. Dotted green lines indicate hydrogen bonding distances ranging from 2.3 to 3.4Å. Structural coordinates have been deposited on the protein data bank (entry: 1cwq).*

The crystal can be described as a stack of PM-like two-dimensional lattices with the stack axis parallel to the screw axis, which is along the c-direction of the crystal. The crystallization conditions in the lipidic cubic phase were continuously improved, so that crystals larger than 400 µm in diameter became available in our group. The crystals showed a merohedral twinning which had to be taken into account [12]. Meanwhile, additional crystal structures of bR were published using different crystallization strategies [13] [14]. The structure obtained by Essen et al [13] nicely illustrates how original lipids from the PM in the protein boundary

interact with amino acid side-chains of bR. Our crystal structure of the ground state [15] is depicted in Fig.3.
The extracellular domain between the retinal and the extracellular side shows a continuous hydrogene bonding network between the Schiff base nitrogen and the outside surface. Thus, we can imagine a proton to be conducted via this network to the outside. No such proton wire can be seen in the cytoplasmic domain between the retinal and the cytoplasmic membrane surface. This part of the structure contains so many hydrophobic amino acids, that it is difficult to envision how a proton is conduced across this domain. This makes clear that a ground state structure alone is not enough to understand proton pumping. High resolution structures of intermediate states are necessary to fill this lack of knowledge.

4 Intermediate states of the working cycle of bR

4.1 Infrared spectroscopy

Since more than 15 years infrared spectroscopy added molecular information to the understanding of the proton pump. However, strong overlap of the 3N-6 vibrations (with N = number of atoms) makes it impossible to record highly resolved infrared absorption spectra of large proteins. A powerful technique to overcome this obstacle is difference spectroscopy where only those vibrations are selected that change during the action of the protein. With the advent of Fourier-transform infrared (FT-IR) spectroscopy this procedure allows to record single vibrations in front of the whole protein. Even dynamics of single vibrations can be followed with high temporal resolution. Nanosecond difference spectra over a broad spectral range have been obtained by state-of-the-art FT-techniques [16].

Infrared spectroscopy was particularly helpful in elucidation of the role of particular amino acids in proton translocation across bR [17]. Besides the detection of vibrational changes of the chromophore retinal as well as of the protein backbone, it is a major advantage of FT-IR spectroscopy that the protonation state of acidic amino acids can be observed. Step-scan spectroscopy allows to follow these transient changes with high temporal resolution [18].
Fig.4 displays difference spectra between ground state bR (negative bands) and the intermediates K, L, M, N, and O (positive bands). The range selected from 1800 to 1000 cm^{-1} is particularly useful because of the richness in band features. Except for the K-bR spectrum, the displaced difference spectra have been obtained with time-resolved attenuated total reflection (ATR) FT-IR spectroscopy. This evanescent wave technique, where the sample is immersed in excess water, allows for the precise control of parameters like pH, temperature and ionic strength [18]. Conditions have been chosen that maximise the transient concentration of the respective intermediate state [19].

4.2 Part of the mechanism evolves

The low resolution electron microscopy structure, site-directed mutagenesis and time-resolved FT-IR spectra provided important details of the mechanism. After absorption of a photon an isomerization reaction from all-*trans* to 13-*cis* occurs. Dramatic changes in pK's are observed for the Schiff base (pK 13 → pK 3) and Asp85, (pK 4 → pK 11), the primary proton accepter group, so that in the transition from L → M the Schiff base proton is moving to Asp85. Since Asp85 stays protonated until the O-intermediate, another proton is rejected to the extracellular

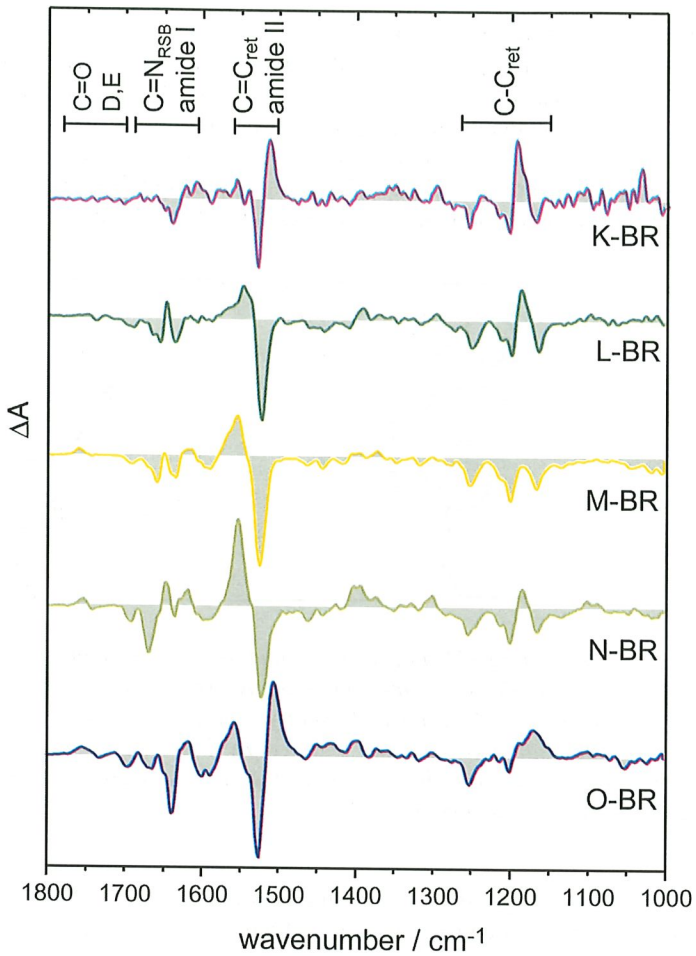


Fig. 4: *Infrared difference spectra of the photocycle intermediates of bR. All but the K-BR spectrum have been extracted from time-resolved ATR/FT-IR experiments under the following conditions: L-BR (10 μs, pH 6.6, 20C), M-BR (300-400 μs, pH 8.4, 20C), N-BR (80-100 ms, pH 8.4, 20 C), and O-BR (5-10 ms, pH 4.0, 40 C). The K-BR difference spectrum was obtained under steady-state conditions (continuous illumination of a bR microcrystal with λ = 514 nm from an argon ion laser, sample is kept at 100 K). Spectra are scaled to yield identical negative absorbance at 1252 cm^{-1}. Regions of marker bands are indicated by horizontal bars (D = aspartic acid; E = glutamic acid; RSB = retinal Schiff base; ret = Retinal)*

surface from an unknown group. In the transition from M → N the proton donor group Asp96 in the cytopolasmic domain is delivering its proton to the Schiff base. Between N and O Asp96 is reprotonated. From O to the ground state the retinal reisomerises to the all-*trans* configuration. During this step Asp85 donates ist proton to Asp212 [20]. The initial proton release group is then reprotonated by Asp212.

4.3 Changes in the tertiary structure of bR

From the photocycle Fig.2 it becomes evident that the structure of the M-intermediate would answer many questions concerning the structure-function relation of bR. Therefore already before crystals of bR were available, low resolution structural information was obtained from neutron and X-ray experiments on oriented PM films [21]. Fig.5 shows two differences Fourier maps in projection to the membrane plane of the M state minus the ground state. The PM films in these experiments were made of the mutant bR Asp96 Asn.

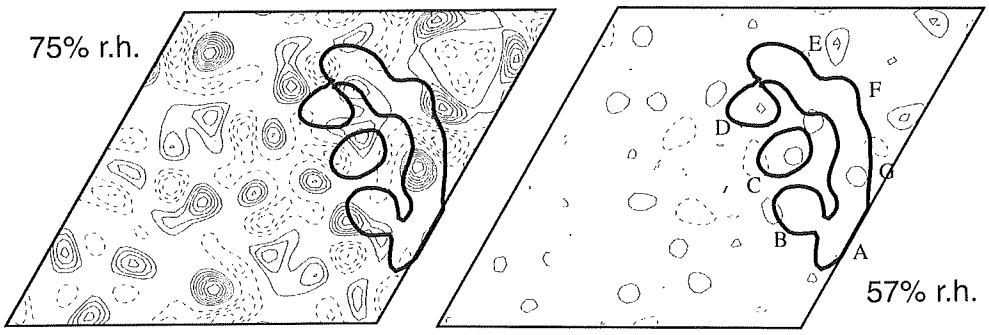


Fig. 5: *Difference density maps (M minus bR ground state) of PM in projection obtained from X-ray diffraction on films of the bR mutant Asp96Asn at different hydration corresponding to 75 % and 57 % relative humidity. The bold contour outlines one of the three bR monomers; individual helices are marked by letters A to G. Continuous lines correspond to five positive, dashed lines to five negative equidistant density levels. The contour levels in both maps are scaled to each other.*

By replacing the protonatable group Asp96 by the neutral amino acid Asn, the reprotonation of the Schiff base becomes difficult resulting in a dramatic increase in the decay time of the M-state by a factor of up to 1000. Thus, it was easy to trap the M state by continuous illumination at room temperature. Unexpectedly large structural changes were observed, especially at helix F and G. The changes at helix F were interpreted as an outward bending in comparison to the ground state. Interestingly these structural changes are completely vanished at hydration levels of the PM-films corresponding to relative humidities below 60 % (Fig.5). At both hydrations the films are completely yellow indicating up to 90 % M state. Thus, two M states existed with and without structural changes.

If one assumes that both M states would occur in the photocycle also at high hydration but in a sequential order, these changes in the tertiary structure would develop within the M intermediate between two M states M_1 and M_2 and would relax after the N state.

The concept of two M states arose from the evaluation of time-resolved spectroscopy data in the visible wavelength region [22]. An irreversible step was assumed at this position in the photocycle between M_1 and M_2, acting as a switch by changing the proton accessibility of the Schiff base from the extracellular to the cytoplasmic side and thus creating the vectoriality of the proton pump.

4.4 The crystal structure of the M state

When bR crystals diffracting to high resolution became available, immediate efforts to obtain intermediate structures were undertaken. Edman et al. [23] succeeded in solving the crystal structure of the early K intermediate and small structural changes with respect to the ground state bR could be resolved. Luecke et al. [24] trapped the M state in crystals of the mutant Asp96Asn (PDB entries: 1c8r and 1c8s) which has a prolonged decay of the M state. The advantage of this mutant is that a high amount of M can be trapped at room temperature. The disadvantage is that, as indicated by the long decay time of M, the cytoplasmic proton translocation pathway is disturbed. Therefore, Sass et al. [15] trapped the M state in wild-type bR, accumulated approximately 35 % of this intermediate and solved the crystal structure (PDB ID: 1cwq). Both M state structures show quite similar changes on the proton release side of bR and more pronounced changes on the cytoplasmic side. Fig. 6 shows the wild-type M_2 structure in comparison to the ground state.

5 More details of the mechanism

In the following we shall concentrate on several features of the M state crystal structure which contributes to our understanding of proton transfer steps. The 13-*cis* retinal orients its proton binding site to the cytoplasmic domain. The carboxyl group of Asp85 has changed from negative to neutral and in this way its electrostatic interaction with Arg82 is diminished. Consequently, this group orients downward interacting with negative charges on Glu194 or Glu204. Thus, the affinity of a proton either of these groups or in a cluster of water molecules [25] is decreased and a proton appears on the extracellular surface. In wild-type bR these reactions take place in about 70 µs.

In contrast to this, protonation steps in the cytoplasmic domain occur within several ms, indicating that these reactions are more complicated. The changes in the tertiary structure observed already at low resolution between M_1 and M_2 become now clearly visible in the M_2 crystal structure. Helix F is bending to the outside and helix G is concentrating some density by narrowing in certain parts and changing its loop conformation. These changes give place to at least four water molecules highly important for the proton transfer across these hydrophobic regions. First of all, water molecules 740 and 724 help to decrease the pK of Asp96. To reprotonate the Schiff base, a proton can follow a hydrogen-bonding network until water molecule 722, which is still 5 Å apart from the Schiff base nitrogen. We assume that this water molecule, only weakly fixed by two hydrogen bonds, can fluctuate and deliver a proton to the Schiff base nitrogen in the transition from M to N. How is Asp 96 reprotonated? According to our M_2-state structure, we can only speculate that water molecule 740 captured in a hydrophobic cavity can fluctuate to Asp96 and deliver a proton to this group. The OH-part of the molecule is escaping to the cytoplasmic surface. We believe that the ms time domain in the last part of the photocycle is caused by the high amplitude of thermal equilibrium fluctuations involved in these protonation reactions. These are coupled to the relaxation of the M_2 conformation, which originated from the altered charge distribution resulting from the protonation reactions between L and M_1.

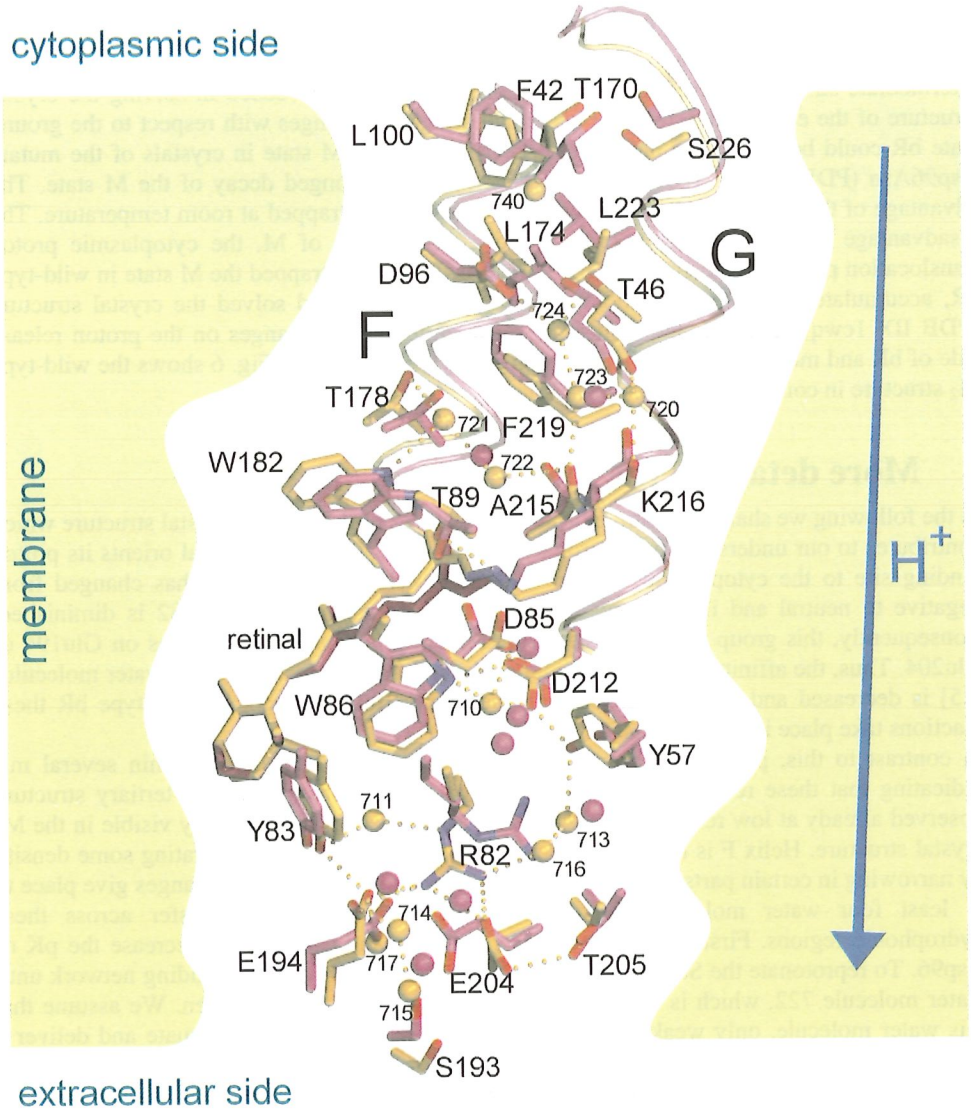


Fig. 6: *Image of the proton pathway of the bR ground state model (purple) and the M_2-state model (yellow). Shown are several functionally important amino acids, the retinal, water molecules and the cytoplasmic parts of helices F and G, showing the most important alterations between these photocycle states. The yellow dotted lines mark distances for possible hydrogen bonds in the M state. The large blue arrow gives the direction of proton pumping.*

6 Sensory Rhodopsin: transmembrane signalling

6.1 The two-component system

Archaebacteria have survived in evolution under most unfavourable conditions. One reason for their successful strategy is their ability to respond towards environmental changes by a locomotive answer. The information transfer from the external input via a membrane receptor to a change in the swimming behaviour is accomplished by several molecules in the so-called two-component system. In the case of *N. pharaonis* a photon is absorbed by the all-*trans* retinal in SRII. Similar to bR the retinal isomerises to 13-*cis* and induces a pK change at the Schiff base linkage. The proton is transferred to Asp75 which serves as the proton acceptor group equivalent to Asp85 in bR. The change in the retinal conformation together with the alteration of the charge distribution induces a conformational change in SRII or directly in the transducer molecule HtrII via electrostatic interaction. Fig.7 depicts a complex of two SRII and two HtrII molecules in the plasma membrane of pharaonis. The two transducer molecules form an elongated coiled-coil structure extending about 240 Å into the cytoplasm. The high resolution structure of this extended rod is not known for HtrII. However, the crystal structure of the serine chemotaxis receptor of *Escherichia coli* was solved [26]. Like in *E. coli* an adapter molecule (Che W) is attached to the cytoplasmic tip of the transducer [3]. The photonic signal from the receptor is transferred to the transducer and travels in an unknown way along the coiled-coil structure to the tip to activate the histidine kinase Che A. Che A phosphorylates Che Y, the switch factor of the flagellar motor, thereby reversing the rotational direction of the motor.

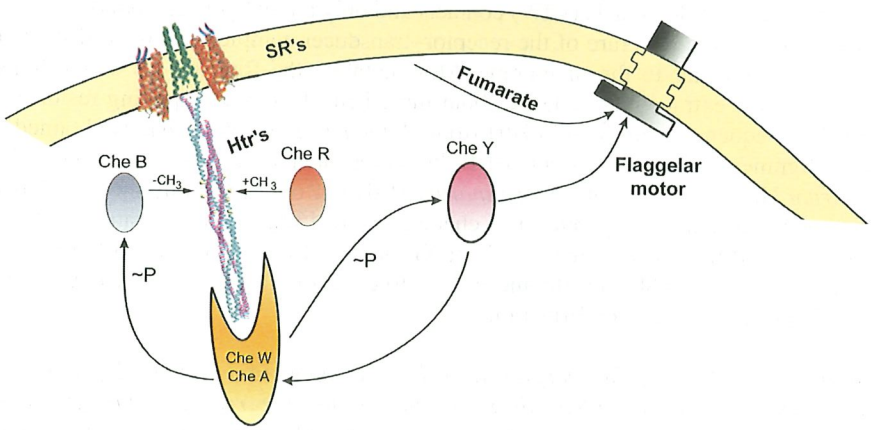


Fig. 7: *Two-component signalling cascade. The activation of the transducer HtrII by its receptor SRII leads to a conformational change of TM2 that propagates to the tip of the coiled-coil cytoplasmic domain (structure taken from ref. 26). The next steps in the signalling cascade involve—in analogy to the bacterial sensory system [3]—the homodimeric histidine kinase CheA, the coupling protein CheW, and the response regulators/aspartate kinases CheY and CheB. Phosphorylated (P) CheY functions as a switch factor of the flagellar motor. CheB (a methylesterase) together with CheR (a methyltransferase) are involved in the adaptation processes of the bacteria.*

Low concentrations of Che Y ~ P prolong the swimming period between reversals. Che A can also activate Che B, a methylesterase of methylated Asp or Glu residues. Che R, a methyltransferase re-methylates these amino acids. The de- and re-methylation reactions are involved in adaptation of the bacteria to constant stimuli.

6.2 The structure of the complex of sensory rhodopsin II and its transducer

Crystallization of the receptor–transducer complex has been achieved using a shortened transducer (residues 1–114; *N.pharaonis* HtrII114) comprising the two transmembrane helices (TM1 and TM2) and an additional small cytoplasmic fragment [27]. The crystals grown in the lipidic cubic phase of the single chain lipid monovaccenin diffracted to 1.8Å resolution. The structure was solved by molecular replacement using the structure of the receptor alone [28, 29]. The expected dimer of the complex is formed by a crystallographic two-fold rotation axis, which is located in the middle of four transmembrane helices: TM1, TM2, TM1' , TM2' (where a prime indicates the right-hand complex; Fig.8a). The transmembrane helices F and G of the receptor are in contact with the helices of the transducer. The overall X-ray structure is in good agreement with a model of the receptor–transducer complex deduced from electron paramagnetic resonance (EPR) measurements [30].

The structure of SRII complexed with its transducer is markedly similar to that of the receptor alone including the retinal conformation [28, 29]. Obviously, the binding of the transducer to helices F and G hardly interferes with the side-chain arrangement of the receptor. A notable exception is found for Tyr 199. The aromatic plane of Tyr 199 has turned in the complex by about 70 degrees and is now pointing into the direction of TM2, where its phenolic group forms a hydrogen bond to Asn 74 (Fig.9). The interface between receptor and transducer is formed mainly by van der Waals (vdW) contacts and only a few hydrogen bonds.

Further insight into the structure of the receptor–transducer complex is gained if the structure is viewed perpendicular to the membrane. Most notably, the four helices of the transducer extend beyond the extracellular side by about three helical turns, comprising residues 44–59 (Fig.8b). This sequence is missing in HtrI from *Halobacterium salinarum*, as obtained from a sequence alignment with *N. pharaonis* HtrII, but is repeated at the N-terminal end of TM2 in *H. salinarum* HtrII [31]. Notably, *H. salinarum* HtrII not only transmits the signal from the photoreceptor SRII but also operates as a chemoreceptor [32]. The temperature factors along the transducer and the receptor helices F and G increase from the centre to the cytoplasmic side of the proteins (Fig.8b). Furthermore, the structure of TM2 ends with Leu 82, leaving residues 83–114 in multiple conformations.

Fig.8 (next page): *Fold of the receptor–transducer complex (**a**), Ribbon diagram of the top view from the cytoplasmic side.Helices are in red for the receptor and green for the transducer; beta-strands are in blue and coils in grey. The labels of the symmetry related complex are marked by a prime. (**b**) Side view of the complex. The complex is coloured according to B-factor mobility: light red/green (less mobile), dark red/green (mobile). ES, extracellular side; CS, cytoplasmic side. The dotted white lines confine the major hydrophobic core of the proteins. The arrows indicate the shortened stalk in HtrI (white) and the site where the helices a1 and a4 of the chemoreceptor domain of H. salinarum HtrII are attached to the transmembrane helices TM1 and TM2 (blue).*

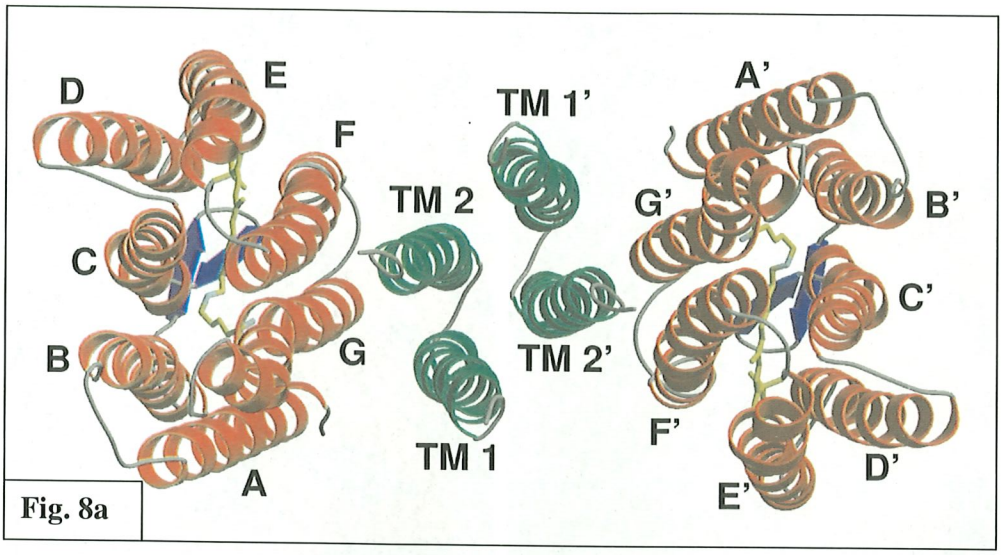


Fig. 8a

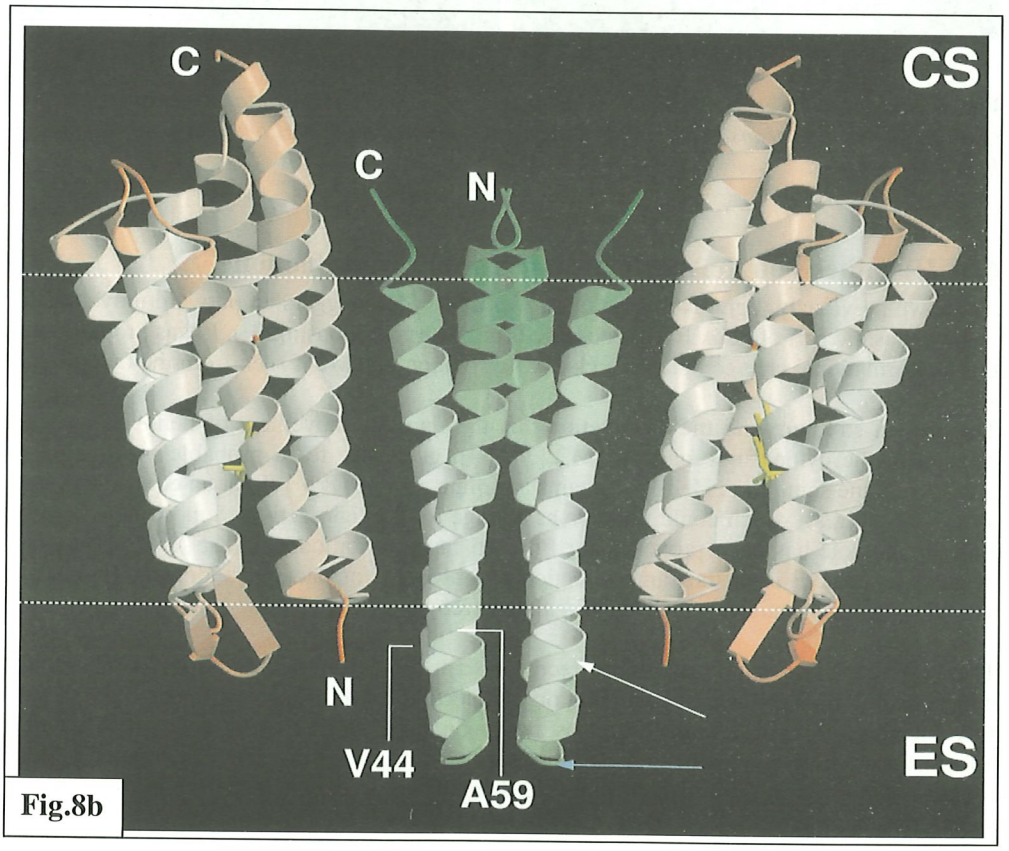


Fig.8b

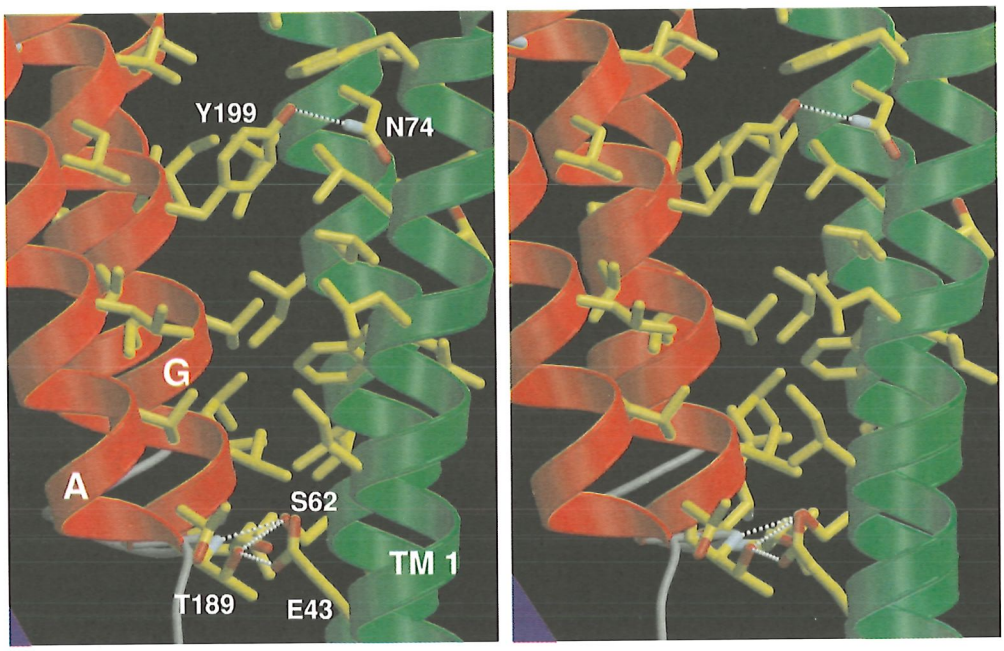


Fig. 9: *Stereo view of the hydrogen bonds and van der Waals contacts between receptor (helices in red) and transducer (helices in green). The residues that are involved inhydrogen bonds are labelled*

6.3 Hypothesis on the mechanism of sensory transduction

As shown previously, light excitation of *N. pharaonis* SRII triggers an outward movement of the cytoplasmic part of helix F (first observed for bacteriorhodopsin[33]). It was proposed that this conformational change induces a clockwise rotation of TM2 if seen from the cytoplasmic side [30].

The present structure and the observation that the flap-like movement of helix F induces a rotation of the cytoplasmic end of TM2 [30, 34], gives rise for a probable mechanism of transmembrane signalling. Helix F contains a proline residue (Pro 175) at an equivalent depth as Tyr 199 (Fig.8), which could function as a hinge for the light-activated movement of helix F (a similar role has been proposed for Pro 185 in bacteriorhodopsin [35]). If the outward bending of helix F is in the same direction as observed for bacteriorhodopsin (Fig.5,6), it should collide tangentially with TM2 (Fig.8a), thereby inducing its rotary motion.

The high-resolution structure of the receptor–transducer complex provides the basis for understanding transmembrane signalling on a molecular level. It has now become possible to solve the structure of intermediates, as has been done for bacteriorhodopsin. This new information will be of significance not only for bacterial phototaxis and chemotaxis but also for other dimeric receptors, which might lead finally to a general model of transmembrane signal transduction.

References

[1] K.Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto, M. Miyano, Science 289, 739 (2000)
[2] D. Oesterhelt, W. Stoeckenius, Proc. Natl. Acad. Sci USA 70, 2853 (1973)
[3] T.S. Shimizu, N. Le Novere, M.D. Levin, A.J. Beavil, B.J. Sutton, D. Bray, Nature Cell Biol. 2, 792 (2000)
[4] R. Henderson, Journal of Molecular Biology 93, 123 (1975)
[5] R. Henderson, P.N.T. Unwin, Nature 257, 28 (1975)
[6] J.L. Popot, D.M. Engelman, O. Gurel, G. Zaccai, Journal of Molecular Biology 210, 829 (1989)
[7] R. Henderson, J.M. Baldwin, Ceska, T.A., F. Zemlin, E. Beckmann, K.H. Downing, Journal of Molecular Biology 213, 899 (1990)
[8] H. Michel, D. Oesterhelt, Proceedings of the National Academy of Sciences USA 77, 1283 (1980)
[9] G.F.X. Schertler, H.D. Bartunik, H. Michel, D. Oesterhelt, Journal of Molecular Biology 234, 156 (1993)
[10] E.M. Landau, J.P. Rosenbusch, Proceedings of the National Academy of Sciences USA 93, 14532 (1996)
[11] E. Pebay-Peyroula, G. Rummel, J.P. Rosenbusch, E.M. Landau, Science 277, 1676 (1997)
[12] H. Luecke, H. Th. Richter, J.K. Lanyi, Science 280, 1934 (1998)
[13] L. Essen, R. Siegert, W.D. Lehmann, D. Oesterhelt, Proceedings of the National Academy of Sciences USA 95, 11673 (1998)
[14] K. Takeda, H. Sato, T. Hino, M. Kono, K. Fukuda, I. Sakurai, T. Okada, T. Kouyama, Journal of Molecular Biology 283, 463 (1998)
[15] H.J. Sass, G. Büldt, R. Gessenich, D. Hehn, D. Neff, R. Schlesinger, J. Berendzen, P. Ormos, Nature 406, 649 (2000)
[16] W. Uhmann, A. Becker, C. Taran, F. Siebert, App. Spectroscopy 45, 390 (1991)
[17] F. Siebert, Methods Enzymol. 246, 501 (1995)
[18] J. Heberle, C. Zscherp, Appl. Spectroscopy 50, 588 (1996)
[19] C. Zscherp, J. Heberle, J. Phys. Chem. B 101, 10542 (1997)
[20] C. Zscherp, R. Schlesinger, J. Heberle, Biochem. Biophys. Res. Comm. 283, 57 (2001)
[21] M.H.J. Koch, N.A. Dencher, D. Oesterhelt, H.-J. Plöhn, G. Rapp, G. Büldt, EMBO Journal 10, 521 (1991)
[22] G. Varo, J. Lanyi, Biochemstry 29, 2241 (1990)
[23] K. Edmann, P. Nollert, A. Royant, H. Belrhali, E. Pebay-Peyroula, J. Hajdu, R. Neutze, E.M. Landau, Nature 401, 822 (1999)
[24] H. Luecke, B. Schobert, H.T. Richter, J.P. Cartailler, J.K. Lanyi, Science 286, 255 (1999)
[25] R. Rammelsberg, G. Huhn, M. Lübben, K. Gerwert, Biochemistry 37, 5001 (1998)
[26] K.K. Kim, H. Yokota, S.H. Kim, Nature 400, 787 (1999)
[27] V.I. Gordeliy, J. Labahn, R. Moukhametzianov, R. Efremov, J. Granzin, R. Schlesinger, G. Büldt, T. Savopol, A.J. Scheidig, J.P. Klare, M. Engelhard, Nature 419, 484 (2002)
[28] H. Luecke, B. Schobert, J.K. Lanyi, E.N. Spudich, J.L. Spudich, Science 293, 1499 (2001)
[29] A. Royant, P. Nollert, K. Edman, R. Neutze, E.M. Landau, E. Pebay-Peyroula, J. Navarro, Proc. Natl Acad. Sci. USA 98, 10131 (2001)
[30] A.A. Wegener, I. Chizhov, M. Engelhard, H.J. Steinhoff, J.Mol. Biol. 301, 881 (2000)
[31] W.S. Zhang, A. Brooun, M.M. Mueller, M. Alam, Proc. Natl Acad. Sci. USA 93, 8230

(1996)
[32] S.B. Hou, A. Brooun, H.S. Yu, T. Freitas, M. Alam, J. Bacteriol. 180, 1600 (1998
[33] M.H.J. Koch, N.A. Dencher, D. Oesterhelt, H.J. Plöhn, G. Rapp, G. Büldt, EMBO J. 10, 521 (1991)
[34] A.A. Wegener, J.P. Klare, M. Engelhard, H.J. Steinhoff, H. J., EMBO J. 20, 5312 (2001)
[35] K. Gerwert, B. Hess, M. Engelhard, FEBS Lett. 261, 449 (1990)

C 11 Viruses as colloidal model systems

M. P. Lettinga
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Ordering in biological systems 4

3 Theory of the self-organization of rods 8
3.1 Onsager theory for the I-N transition of hard rods 8
3.2 Effect of rod charge on the I-N transition 12
3.3 Effect of flexibility of rods on the I-N transition 14
3.4 Effect of polydispersity on the I-N transition 17
3.5 Effect of attractive interactions on the I-N transition 17
3.6 Chirality and elasticity constants 18
3.7 Transitions to higher ordered states 19

4 Experimental verification 21
4.1 Isotropic-Nematic transition for nearly ideal rods 21
4.2 Higher order phase transitions for nearly ideal rods 27
4.3 Phase behavior of non-ideal rods 30

5 Mixture and external fields 33
5.1 Mixtures of rods and other particles 33
5.2 Magnetic field and shear field 34

1 Introduction

The end of the 19th century and the beginning of the 20th century was a revolutionary period in science. Not only was it the time where Quantum physics and relativity theory was developed, but also in other fields of physics and biology some important steps were made. For one the Brownian motion of colloidal particles was explained, where particles with a size in the range of nanometers to micrometers undergo a random motion when immersed in a solvent. In Chapter A.2 of this book the Brownian motion of colloids has been explained in great detail. The physicist Perrin in 1916 was the first to recogniz that the Brownian motion randomizes the position of the colloids, resembling atoms in that respect [1]. It was then reasoned that colloids could be very useful for physicists to study interactions between particles, since it is possible to engineer the interaction potential between two colloids using various chemical or physical modification schemes. In the same period it was also discovered that long molecules can form a phase where the sample can still flow, but that some crystalline structure should also be present, judging by the way light interacted with the sample (see Fig. 1). It was found that these systems have a liquid-like positional ordering, but at the same time orientational ordering of the long axis of the molecules. Zocher found then that inorganic colloidal needles dispersed in a solvent show the same features, though the concentration may be fairly low and the randomizing force is still present. Thus, in the spirit of Perrin, the shape of the colloidal particle can be a determining factor for the phase behavior. Also in biology an important development took place in that period. In 1892 the Russian scientist D. Iwanowski, found that the Mosaik illness of Tobacco plants could be inflected by juice of ill tobacco plants, which was free of bacteria. Much later it was found that the juice contained particles, which were isolated in 1936 [2]. This type of particles were since then called "virus", which makes Tobacco Mosaic Virus (TMV) the first virus discovered. Like Zocher, although not referring to him, Bawden et al. [3] showed that dispersions of TMV behave like liquid crystals, and that therefore they should be rod-like.

In 1949 Onsager made the first step in understanding the phase behavior of hard spherocylinders [4], inspired by the phase behavior of TMV, but also of more flexible chains like myosin and various linear polymers. With "hard" it is meant that the interaction energy between the particles is either zero if they do not overlap, or infinity if they do overlap. Just like billiard balls, the particles cannot interpenetrate. 'Hard' spherical particles only show a transition in the positional ordering, liquid to crystalline. The interaction between a-spherical particles, on the other hand, not only depends on the mutual position but also on the mutual orientation of the particles. Onsager predicted that a system in the *isotropic* phase, where there is no ordering in position as well as orientation, will undergo a transition to the *nematic* phase, where there is still no ordering in the position of the rods but where the rods will have mutually oriented, when the concentration of rods is increased, see Fig. 1. This orientation is characterized by a vector describing the average direction in which the rods are aligned, therefore called *director*, and an order parameter which is equal to one when the rods are perfectly aligned and zero when there is no average alinement, i.e. in the isotropic phase. The word "nematic " is derived from the Greek word for thread. The reason for this name is that under the microscope one often observes "threads" which are defect lines of the nematic crystal. These lines form borders between two small regions with a different director. Some samples will not form a nematic phase, but a cholesteric phase, which is characterized by a twisting of the director. Further increasing the concentration, a transition will occur where there is also positional order but only in one dimension. This can either be a *smectic* phase, where the rods are positioned in layers but are still liquid-like within this layer, or a *columnar* phase, where rods line-up in columns. The name

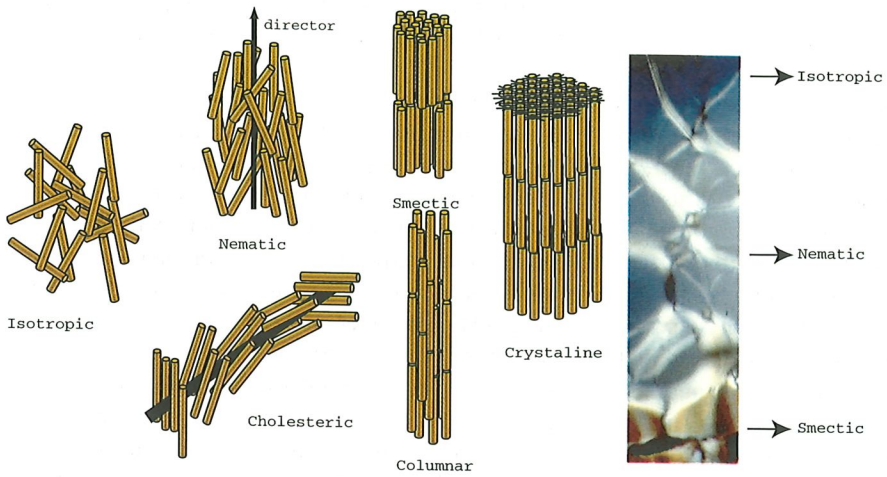


Fig. 1: *Diagram of the different phases encountered, when the concentration of rods is increased from left to right, i.e. from the isotropic phase to the nematic phase. On the right a picture of three co-existing phases of a dispersion of fd virus between crossed polarizers: the smectic phase showing different colors typical for Bragg scattering, the nematic phase showing its typical threads, and the isotropic phase which is just dark.*

"smectic" is derived from the Greek word for soap, since early investigators noticed that this phase possessed mechanical properties reminiscent for soap. Indeed the smectic phase as depicted in Fig. 1 resembles a phase of lamellar lipid bi-layers. At even higher concentrations of rods the positional order of the rods will be totally crystalline. The consequence of considering rods with a simple hard-core interaction potential is that all phase transitions which take place are entirely entropy driven. The only features one has to consider are the positional and orientational degrees of freedom of the particles in the system, or how many particles can be packed in a volume maintaining the possibility to move. For these systems temperature does not play any role, and therefore an experimental signature of entropy driven transitions is independence of temperature. This type of liquid crystals are called *lyotropic* liquid crystals.
A summary of the expected phase transitions in the system of hard particles is given in Fig. 1. The picture on the right in Fig. 1 shows a co-existing isotropic, nematic, and smectic phase between crossed polarizers. The bottom phase is smectic, since it shows different color, reminiscent for Bragg peaks, and therefore for long-ranged positional order. The middle phase is the nematic phase since it shows the typical threads and it is birefringent. This means that their refractive index depends on the polarization of the light, which can only be the case when they are to some degree of alignment. The top phase is partly black, meaning that their is no alignment, and therefore it is the isotropic phase.
The phase transitions that a system will exhibit depends on the characteristics of the rod, i.e. its flexibility, its length, its charge, and its polydispersity. In the case of spherical particles some fundamental questions have been answered due to a very good control of the system, regarding the size and interaction potential [5]. The controlled synthesis of rod-like particles, however, has proven to be a far more complicated story. The reason is that the requirements for rod-like systems are higher than for spheres: size-control is required both for its thickness and its length, moreover, ideal rods have a high aspect ratio, defined as the length (L) over the thickness (D)

of the rod, i.e. $L/D \ll 1$. The quest for suitable systems has resulted in a few promising synthesize routes [6, 7, 8, 9]. However, each of the mentioned systems lacks at least one of the requirements of an ideal system.
An alternative to the synthesis of colloidal particles is to use particles which are produced by nature. In contrast to chemical synthesis, nature uses DNA technology to produce viruses that are identical to each other, so that these systems are intrinsically monodisperse. Moreover dispersions of these viruses are colloidal stable, meaning that they do not aggregate or gel. Finally, an other important feature is that these particles can be chemically and genetically modified, changing the surface and length of the particles. It is therefore not a coincidence that TMV played an important role when the first step was made in the understanding of lyotropic liquid crystals by Onsager. More recently TMV and other viruses have been very fruitfully used to gain further understanding (see Fraden for a review [10]).
Rod-like viruses are not only used by physicists to study ordering phenomena of rods, but also as model polymers for (micro) rheology[11, 12] in isotropic suspensions. Moreover, material scientists have recognized the features mentioned above, and try to use viruses as a kind of template to construct nano-wires e.d. This will however not be treated in this chapter, but see Flynn et al. for a nice review on this work.[13]
In this chapter we first consider biological systems which show high ordering. All these systems are characterized by high concentrations of different particles. We argue that knowledge of the concentration effect on the behavior of rods is a prerequisite to under stand what is happening in such "crowded" systems. In this section we also introduce biological particles as possibly interesting model systems. Then we will give the theoretical framework needed to understand the phase behavior of rod-like particles, including the effect of charge, flexibility, polydispersity, and attraction between the rods. We will apply the theory on the phase behavior of the more ideal viruses, and the less ideal DNA and F-actin. Doing so we encounter typical liquid crystalline phenomena, of which we will discuss the phenomenological and experimental background. Finally, we will discuss what happens, or might happen, when rods are not in equilibrium, i.e. when an external field is applied.

2 Ordering in biological systems

As explained in the introduction, physicists are interested in the study of the phase behavior of particles of which the interaction can be controlled. Biological systems, on the other hand, are on average very complex, contain many components, and undergo all kinds of poorly understood interactions. Facing this problem biologists tend to investigate the specific interaction between the species of which the system consists. Here the function of each species within the biological system is of interest. On a molecular level this means for example that the interaction between a protein and, say, a membrane is investigated on the basis of the molecular structure of the protein, its charge, hydrophobicity, and enzymological activity. The physicists approach is based on more general considerations as concerning the energy state of the system and its entropy, using knowledge gained in the study of model systems such as mentioned in the introduction.
Considering, for example, a complex system like a cell. One may ask how this system deals with the transport of matter having packing fractions of up to 40 %. For hard spheres it is known that they crystallize at a volume fraction of 49 % and rods undergo a chaotic to ordered phase transition at much lower volume fractions. Therefore there must be some mechanism to

control the stability of the cell and the transport of matter. In eukaryotic cells this is done by the formation of the cytoskeleton. This is a rather rigid structure build from actin filaments and microtubules, which also form the "highways" over which particles are transported. See Chapter C 9 for further reading. In Fig. 2 a visualization of such a cell obtained by electron tomography of vitrified cells shows how crowded these cells are, and how the actin filaments are organized to form the backbone of the cell. A more highly ordered and very efficient cell is the muscle cell, which consists of interacting myosin and actin filaments, see Fig. 3a. One could even see a resemblance between this ordering and the ordering in the smectic phase. Another example of ordering in a cell is the packing of DNA in a cell nucleus or its organization in bacteria, which do not have a nucleus. The packing of DNA should be very efficient, i.e. crystalline, in order to store as much genetic material as possible. At the same time it should be liquid like, otherwise DNA can not diffuse out and be expressed. Indeed, in the electro-microscopy image of the Dinoflagellate chromosome different regions showing a cholesteric ordering of the DNA, can be distinguished, see Fig. 3b. Ordering can also be damaging for biological systems. One well known example is "Sickle cell anemia", which is a blood disease. Normal hemoglobin is stable at volume fractions of 30 to 36 %. However, spontaneous aggregation in to filaments will occur, when the hemoglobin is modified by only one amino acid, and when too many of these modified hemoglobin are contained in the cell. Sequentially, the filaments will order and deform the blood cell to a sickle shaped cell, see Fig. 3c [14]. A similar effect of filament formation also seems to be connected to the Parkinson's disease [15]. Finally, in Fig. 3d an infected Tobacco plant is depicted and a cut through a hair cell of the plant showing regions with TMV-crystals. As one may notice from these examples there are in most cases two steps of the ordering: a step where small particles (e.g. proteins or lipids) assemble to a mesoscopic particle, and a step where the mesoscopic particles mutually organize. Actin filaments (F-actin) are formed from globular actin proteins (G-actin) and are bundled as in the cytoskeleton or oriented as in a muscle cell; mutated hemoglobine aggregates to filaments which then form a liquid crystalline state; TMV is an assembly of units containing 13 different proteins which form a helical structure; DNA in in eucaryotic cells is wrapped around nucleosomes, forming stiff filaments which will sequentially form a liquid crystalline state [16].

The general question arising now is how biological systems order when they need to be ordered, and how ordering is avoided when they do not need to be ordered. Physics can contribute to this question following the same pathway as for hard rods: regard biological systems such as described above as a "crowd" of different but ideal particles, which are mutually exclusive. Then try to find an organization of this "crowd" such that the particles have as much free volume as possible, taking into account the loss of the rotational degree of freedom of the particles. In other words: try to maximize the entropy for the "crowd" of particles. The study of "crowding" is gaining more interest lately since physicists try to make the step to more complicated systems [17].

The first step in this approach is to understand the organization of rods. We will extensively treat this problem, allowing for deviations from ideal particles, i.e. including charge, flexibility and polydispersity. Having understood this problem both experimentally and theoretically, a next step can be made by disturbing this relative simple system with "external" fields. We will see for example how extra ordering is *induced* by mixing rods with spheres. We will also shortly discuss the phase behavior when a magnetic field is applied, and when is shear field is applied, both of which tend to align rods.

As stated in the introduction there is one prerequisite to be fulfilled, and that is the need for very stiff, very long, and monodisperse rods, which have a pure hard core interactions. For this

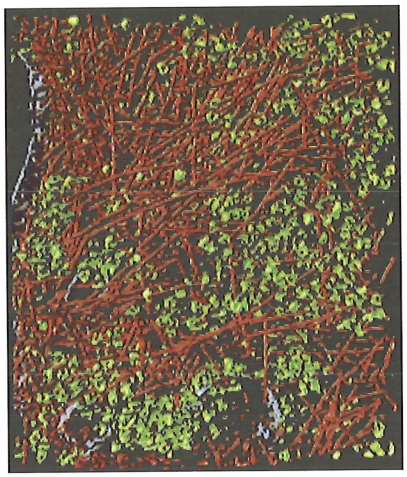


Fig. 2: *Visualization of actin network, membranes, and cytoplasmic macromolecular complexes. (a) A volume of 815 nm by 870 nm by 97 nm. Colors were subjectively attributed to linear elements to mark the actin filaments (reddish); other macromolecular complexes, mostly ribosomes (green); and membranes (blue). This figure is reprinted from Medalia et al. [18].*

biology can be very useful, since close to ideal systems can just be found in nature. In table 1 an overview is given of biological rod-like particles with varying contour length L, thickness D, and persistence length P. This latter parameter is a measure for the flexibility. The persistence length is defined as the length along the chain after which the direction of the chain becomes uncorrelated with the direction at the beginning of the chain.

In the table we show data for three different types of particles. DNA and Poly(γ-benzyl L-glutamate) (PBLG) are two polymers. When PBLG is dispersed in Dimethylsulfoxide, it is one of the stiffest of a whole class of poly-saccharides and poly-amides. These polymers play an important role in the research of lyotropic liquid crystals. F-actin and microtubules are both helical self-assembling filaments. They determine the stiffness of eucaryotic cells and play an important role in protein mobility. TMV, *fd*, M13, and Pf1 are four different viruses. They consist of DNA coated with proteins and they are intrinsically monodisperse. It is striking to observe that TMV and Microtubule are both hollow tubes and also have the stiffest structures. For engineers this is not a surprise. *fd*, M13 and Pf1 are very similar particles, which can be produced by *E-coli* bacteria. M13 and *fd* differ by a single amino acid in their major coat protein subunit, a 50-amino-acid α-helical polypeptide. As a result of this difference, M13 has one negative charge less per subunit than *fd*, corresponding to approximately a 30% reduction in surface charge density: $\sim 6\ e^-/nm$ for M13, vs. $\sim 6\ e^-/nm$. The only difference between *fd* and Pf1 is that the latter is twice as long. The charge of TMV is the highest of the four viruses: $\sim 20\ e^-/nm$.

Each of the listed systems has its advantages and disadvantages. Candidates which are very stiff do not have a fixed length (F-actin and microtubules) or have a low aspect ratio (TMV). The three last species are monodisperse and have a high aspect ratio. Their persistence length is not very long, and therefore they are semi-flexible. In the next section we will treat each of these effects separately. In the experimental section we will see that especially *fd* is a very useful system to study all the features of a phase diagram of a (almost) rigid rod.

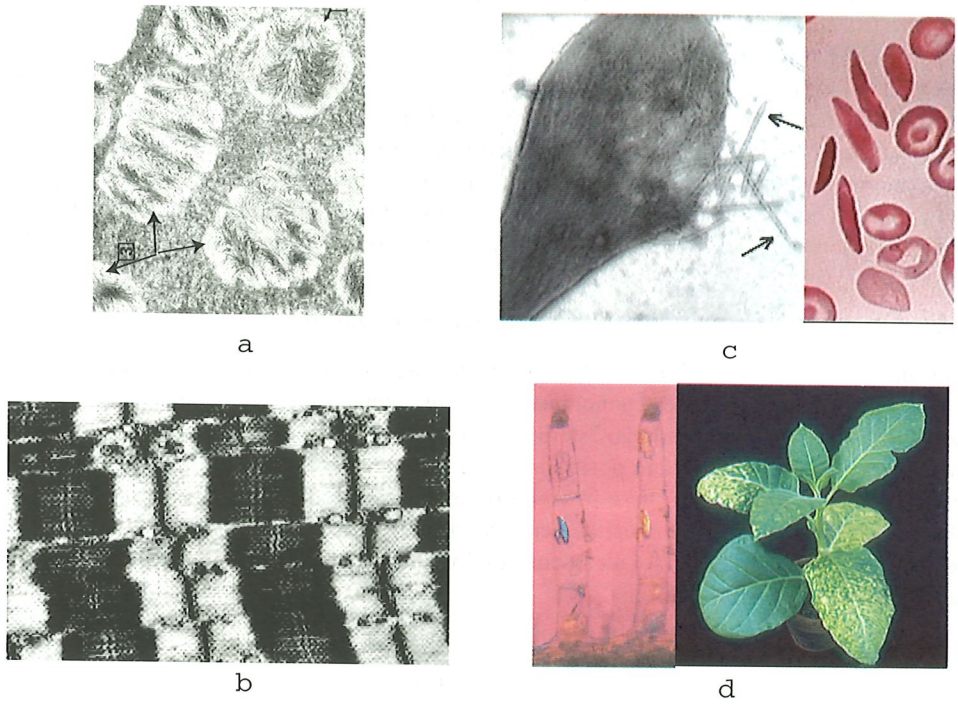


Fig. 3: *Four examples of ordering in biological systems, of which two are very use full (a and b) and two are damaging (c and d). (a) Electro-microscopy image of Dinoflagellate chromosome showing typical cholesteric structures formed by DNA [19]. The arrows indicate regions where DNA forms acholesteric phase. (b) Electro-microscopy image of a myofibril from a muscle. The Dark bands contain the myosin filaments, the light bands contain the actin filaments. (c) Ordering of fibres formed by mutated hemoglobin (left) causing the deformation of blood cells to Sickle shape cells (right). (d) TMV-crystals red and blue regions) in hair cells (left) of an ill Tobacco plant (right).*

type	L (nm)	D (nm)	P ($\times 10^3$ nm)
F-actin	> 1000	7	10-20
microtubule	> 1000	24	1000-6000
DNA	> 50	3	0.053
PBLG/DMF	16-390	1.6	0.080
TMV	300	13	~ 1000
fd	880	6.6	2.2
M13	880	6.6	2.2
Pf1	1960	6.6	2.2

Table 1: *Overview of biological rod-like particles.*

3 Theory of the self-organization of rods

In this section we will now address under which conditions a dispersion of rod-like particles will undergo a phase transition from a chaotic state to an ordered state. Two transitions will be discussed of increasing order: isotropic to nematic (I-N) and nematic to smectic (N-Sm). The theory of the I-N transition is based on the theory developed by Onsager in 1949 for long, very rigid, and monodisperse rods. We will discuss the effect of charge, flexibility, polydispersity, and attraction on the I-N transition. Theory and computer simulations of the N-Sm are more recent and more complex. We will give a short overview of the results at the end of this section.

3.1 Onsager theory for the I-N transition of hard rods

Consider a ensemble of hard rods with length L, thickness D, and persistence length P. The rods do not interact, accept that they cannot interpenetrate (interaction energy is infinitely high when they do). So the rods only undergo "sterical" interactions, which effectively means that due to the presence of a rod at a certain position $\mathbf{r}$ some volume is excluded for the presence of other rods. In order to predict the phase behavior of an assembly od such rods, one has to find out where the minima of the Helmholtz free energy are situated for varying rod concentration. This free energy has a contribution for an assembly of non-interacting rods, F_{ideal}, and a contribution due to the excluded volume F_{ex}:

$$F = F_{ideal} + F_{ex} \tag{1}$$

F_{ideal} for an assembly of N rods which can take a position $\mathbf{r}$ and orientation $\mathbf{\Omega}$ is given by

$$\frac{F_{ideal}}{k_B T} = \int_V d\mathbf{\Omega} \int d\mathbf{r} \rho(\Omega, \mathbf{r}) \ln(\rho(\mathbf{r}, \mathbf{\Omega})) = N \ln \rho + N \int d\mathbf{\Omega} f(\mathbf{\Omega}) \ln(f(\mathbf{\Omega})) \tag{2}$$

where V is the volume of the system. $\rho(\mathbf{r}, \mathbf{\Omega})$ is the density function which describes the position and the orientation of the particles. For the second step in this equation we assumed that the particles are uniformly distributed, which is reasonable before a phase transition has taken place. The density function then reduces to $\rho(\mathbf{r}, \mathbf{\Omega}) = \frac{N}{V} f(\Omega)$. $f(\Omega)$ is the Orientational Distribution Function (ODF) giving the probability that the rod is pointing at a solid angle Ω. It should be normalized as follows :

$$\int f(\Omega) d\Omega = 1. \tag{3}$$

The free energy due to the interaction between the rods can be calculated using the well known virial expansion of the pressure. Although this is a concept that is used for gases, van 't Hoff's law states that it can also be applied for solutions, replacing the pressure by the osmotic pressure:

$$\frac{\Pi}{k_B T} = \rho + \rho^2 B_2 + \rho^3 B_3 ... \tag{4}$$

B_2 and B_3 are the second and third order virial coefficients, which are determined by the two and three particle interactions, respectively. From now on, we will only consider B_2 and omit

the higher order virial coefficients. This implies that not more than two particles will interact with each other at the same time. Onsager was the first to recognize that this assumption is essentially exact at the isotropic-nematic phase transition, when the aspect ratio $L/D \rightarrow \infty$ [4]. We will see later that the I-N phase transition occurs at a volume fraction $\rho \approx 4\frac{D}{L}$. Hence, the larger the aspect ratio, the lower the volume fraction at the I-N transition and therefore expansion of the free energy in density becomes more accurate. This is confirmed by computer simulations [20].

B_2 depends on the *potential of mean force* $w(i,j)$ between two particles via the so called Mayer function $\phi(i,j) = \exp(-w(i,j)/kT) - 1$:

$$B_2 = -\frac{1}{2V}\int \mathbf{dr}_1 \int d\mathbf{\Omega}_1 \int \mathbf{dr}_2 \int d\mathbf{\Omega}_2 f(\mathbf{\Omega}_1) f(\mathbf{\Omega}_2)\phi_{12} \tag{5}$$

Hard particles are characterized by particular simple potentials and Mayer functions:

$$\begin{aligned} w = \infty \quad & \phi = -1 \quad \textit{at overlap} \\ w = 0 \quad & \phi = 0 \quad \textit{no overlap} \end{aligned}$$

This reduces the second virial coefficient to an integral over the volume excluded for one rod by the presence of a second rod. This volume is given by $v_{exclude} = 2DL^2|\sin\theta|$, where θ is the angle between the orientations of two rods. The contribution of the interaction part to the free energy can be derived from the osmotic pressure using the thermodynamic relation $\Pi = -\frac{\partial F_{ex}}{\partial V}_{N,T,\mu_0,\mu_1,\ldots}$, omitting the ideal term (the first term in eq. 4):

$$\frac{F_{ex}}{k_bT} = \rho DL^2 \int d\mathbf{\Omega}_1 \int d\mathbf{\Omega}_2 f(\mathbf{\Omega}_1) f(\mathbf{\Omega}_2)|\sin\theta| \tag{6}$$

Now we can write down the free energy for a system of hard rods in terms of the ODF, by adding up the contributions of the non-interacting (Eq.6 and interacting rods (Eq. 6):

$$\frac{F}{k_bT} = const. + N\ln\rho + N\int d\mathbf{\Omega} f(\mathbf{\Omega})\ln(f(\mathbf{\Omega})) + N\rho DL^2 \int d\mathbf{\Omega}_1 \int d\mathbf{\Omega}_2 f(\mathbf{\Omega}_1) f(\mathbf{\Omega}_2)|\sin\theta| \tag{7}$$

The first integral in Eq. 7 is the ideal part of the free energy and always attains a minimum value for the uniform probability distribution $\rho(\mathbf{r}) = constant$. Therefore this contribution to the free energy always suppresses an ordering transition. The second integral in Eq. 7 represents the second virial approximation for the interaction free energy, which under certain circumstances is lower for an ordered state. Therefore the interaction part of free energy drives the system towards ordering. The actual location of the ordering transition is determined from the competition between the ideal and the interaction contribution to the total free energy. It is important to note that in any hard-particle fluid, due to the simplicity of the interaction potential, the energy per unit volume of any allowed configuration is proportional to $\rho k_B T$. Because of this simple fact, the minimum of free energy of a hard particle fluid $F = E - ST$ is equivalent to the maximum of entropy. Furthermore, the resulting phase diagram is temperature independent (a-thermal).

In order to find the ODF $f(\Omega)$, the free energy is minimized with respect to the functional $f(\Omega)$. Doing so, we obtain the following integral equation :

$$\log[4\pi f(\theta)] = \lambda - \frac{8}{\pi}\frac{L}{D}\phi \int \sin(\theta) f(\theta) d\theta, \tag{8}$$

where $\phi = \frac{\pi}{4} L D^2 \frac{N}{V}$ is the volume fraction of rods. This integral equation can not be solved analytically but it has been solved with two different numerical procedures which yield almost identical results for the equilibrium distribution function [21, 22]. Although the numerical solution yields the most accurate results it is also possible to proceed by assuming a form of the orientational distribution function :

$$f(\alpha, \cos(\theta)) = \frac{\alpha \cosh(\alpha \cos(\theta))}{4\pi \sinh(\alpha)}. \tag{9}$$

Using this ansatz first introduced by Onsager in the second virial approximation to free energy (Eq.7) and evaluating the integrals for the case of hard rods we obtain an expression for the free energy which is a function of volume fraction ϕ and orientation parameter α:

$$\frac{F(\alpha, \phi)}{N k_B T} = \phi \ln(\phi) + \sigma(\alpha)\phi + \xi(\alpha)(\frac{L}{D}\phi)^2 \tag{10}$$

with

$$\begin{aligned} \sigma(\alpha) &= \log\left(\frac{\alpha \cosh(\alpha)}{4\pi \sinh(\alpha)}\right) - 1 + \frac{\arctan(e^{\alpha}) - \arctan(e^{-\alpha})}{\sinh(\alpha)} \\ \xi(\alpha) &= \frac{2 I_2(\alpha)}{\sinh^2(\alpha)}. \end{aligned}$$

The functions $\xi(\alpha)$ and $\sigma(\alpha)$ are the translational and orientational entropy terms. Having derived the expression for the free energy, one can use it to obtain the concentrations and the ordering at the phase transition (see next paragraph). The usual measure for ordering of the system is generally quantified by the nematic order parameter, which is calculated from the obtained ODF using the following relation:

$$S = 2\pi \int_0^{\pi} \left(\frac{3}{2}\cos(\theta)^2 - \frac{1}{2}\right) f(\theta) \sin(\theta) d\theta. \tag{11}$$

It describes the average angle between the rods and the *director*, the local main orientational axis of the system (see Introduction). This parameter is experimentally accessible for a number of techniques. One technique that is often used is optical birefringence. This technique is especially useful when their is a difference in refractive index along the long axis of the rod and perpendicular to this axis. Their will be an averaged difference in refractive index Δn, which sets the birefringence, when the rods have some overall alignment. The birefringence will cause a phase shift $\Delta\phi$ between the light with a polarization parallel to the director and perpendicular to the director, which can be detected. The order parameter S_2 is directly related to the birefringence and therefore to the phase shift by

$$S_2 = \frac{\Delta n}{\Delta n_{sat}} = \frac{\lambda \Delta\phi}{2\pi d \Delta n_{sat}}, \tag{12}$$

where d is the sample thickness, λ is the wavelength of the light and Δn_{sat} is the birefringence of perfectly aligned rods, which can be determined experimentally by aligning the rods in a magnetic field.

Calculating phase transition concentrations

By performing a linear stability analysis of the Onsager equation Eq. 7, Kayser and Ravenche found out that the isotropic (unordered) phase becomes unstable towards nematic ordering when $\frac{L}{D}\phi = 4$ [23]. This means that each fluctuation in the isotropic state of the orientation of the rods initiates spinodal phase separation, where an nematic phase of ordered rods continuously grows out of the isotropic phase. It follows that within the Onsager theory the volume fraction of the hard rods at the I-N transition scales as $\phi = 4\frac{D}{L}$. Therefore the longer the rods are, the volume fraction at the I-N transition decreased and the accuracy of the second virial approximation increases, since the probability that three rods interact at the same time reduces.

However, this (second order) transition is preempted by a first order phase transition. Starting from the isotropic phase with increasing rod concentration there is a certain concentration where the phase becomes metastable: fluctuations in orientational order whith an amplitude that is larger than a threshold value will form a nematic nucleus which will sequentially grow. Thus a discontinuity both in concentration and in the ordering of the system will occur. When reducing the concentration starting from a high concentration in the pure nematic phase, the nematic phase becomes unstable at a given concentration, and again a second phase, now the isotropic phase, will form. The concentrations at which the system becomes metastable are called the binodal points. The co-existence of two phases requires that the osmotic pressure Π and chemical potential μ of the isotropic phase, with concentration ρ_{iso}, and the nematic phase, with concentration ρ_{nem}, are equal:

$$\mu(\rho_{iso}) = \mu(\rho_{nem}) \tag{13}$$

$$\Pi(\rho_{iso}) = \Pi(\rho_{nem}) \tag{14}$$

The binodal points can thus be found as follows. First, the Onsager free energy (Eq. 7) is minimized with respect to the ODF $f(\Omega)$, by solving the integral equation Eq. 8. Next, the co-existence equations are solved by using the thermodynamic relations $\mu = \partial F/\partial N$ and $\Pi = \partial F/\partial V$). There are two approach to do this. The analytical, *but not exact*, approach is to use a trial function like the Onsager trial function (Eq. 9) and minimizing the energy , requiring $\frac{\partial F}{\partial \alpha} = 0$. Following this approach, the following co-existence concentrations and order parameter:

$$\rho_{iso} = 3.339, \quad \rho_{nem} = 4.487, \quad S_2 = 0.848. \tag{15}$$

The binodal concentrations can also be determined exactly, by numerically solving Eq. 8.[21, 22, 24]

$$\rho_{\text{iso}} = 3.289, \quad \rho_{\text{nem}} = 4.192, \quad S_2 = 0.7922, \tag{16}$$

Reliability Onsager second virial approach

The validity of the Onsager theory, where only interactions between two particles are accounted for, can be tested by including higher virial terms in the free energy. To this end a scaled particle theory for the system of hard rods was developed by Cotter and coworkers [25, 26], which accounts for all higher virial coefficients in an approximate way. Therefore this theory should

be more adequate at describing the data at higher concentration of rods or equivalently rods with lower L/D ratios. The free energy derived by Cotter is :

$$\frac{F(\delta,\phi,\alpha)}{Nk_BT} = \ln(\phi) + \ln(1-\phi) + \sigma(\phi) + \Pi_2(\delta,\alpha)\frac{\phi}{1-\phi} + \frac{1}{2}\Pi_3(\delta,\alpha)\left(\frac{\phi}{1-\phi}\right)^2 \quad (17)$$

where ϕ is the volume fraction of rods

$$\phi = \frac{N_{rods}}{V}\left(\frac{\pi}{6}D^3 + \frac{\pi}{4}D^2L\right). \quad (18)$$

The coefficients Π_2 and Π_3 are given by the following expressions:

$$\Pi_2(\delta,\alpha) = 3 + \frac{3(\delta-1)^2}{3\delta-1)}\xi(\alpha), \quad (19)$$

$$\Pi_3(\delta,\alpha) = \frac{12\delta(2\delta-1)}{(3\delta-1)^2} + \frac{12\delta(\delta-1)^2}{(3\delta-1)^2}\xi(\alpha) \quad (20)$$

and parameter δ is the overall length over diameter ratio of the spherocylinder given by

$$\delta = \frac{L+D}{D}. \quad (21)$$

The functions $\xi(\alpha)$ and $\sigma(\alpha)$ are the same terms as used in the Onsager theory (Eq. 10). We note that the expression for the free energy (Eq. 17) reduces to the Onsager second virial approximation for very long rods ($\delta \rightarrow \infty$). On the other hand for spherical particles ($\delta = 1$) it reduces to the Percus-Yevick or scaled-particle-theory free energy for hard spheres.
Comparing the results of Eq. 17 with the solution obtained through the second virial approximation allows us to test for which range of L/D ratios the second virial approximation is quantitatively valid. The results are shown in Fig. 4. We conclude that for rods with $L/D > 75$ the second virial approximation quantitatively describes the I-N transitions of hard-rods. Currently available computer simulationresults [27] agree very well with the scaled particle theory [28]. Another important conclusion we can draw from Fig. 4 is that the scaled particle theory predicts strong widening of the co-existence region with decreasing L/D of the rods.
In the following the effect of charge, flexibility, and polydispersity are treated. Doing so one has to choose a way to deal with the theory of hard rods before "perturbing" it with one of these effects. This comes down to two choices. First, one either uses a second virial approach or the approach including the higher order terms. For systems with high L/D the first approach is preferred, because it is easier to handle. Second, one either obtains an exact numerical solution for Eq. 10 or one uses an ansatz for the shape of the ODF. Onsager used Eq. 9 as an ansatz, but Gaussian or other distributions are also often used.

3.2 Effect of rod charge on the I-N transition

The Onsager theory outlined in the previous section can be extended to the case where the colloidal rods have a surface charge, which is the case for most systems. The first treatment of the isotropic phase transition of charged rods can be found in the original paper by Onsager[4]. The extension of it to the anisotropic phase can be found in the paper by Stoobants et al. [30].

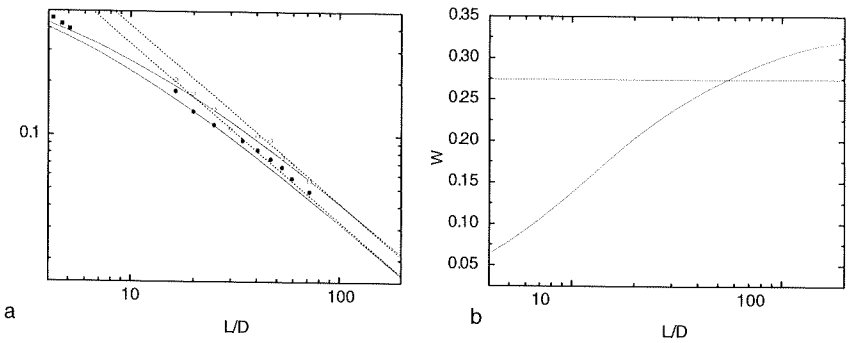


Fig. 4: *Full lines in figure A show the I-N coexistence concentrations as the function of the aspect ratio (L/D) as predicted by the scaled particle theory for rigid rods. The dashed line represents the I-N co-existence as predicted by the theory that only includes second virial coefficient. The circles are the results of the computer simulations [27]. The filled squares at low L/D represents the results from the same work but the coexistence width was too narrow to be measured. The coexistence is plotted in terms of real volume fraction* $\phi = \pi D^3/6 + LD^2\pi/4$ *while the total aspect ratio including the hemispheres is L/D+1. Figure b shows the predicted width of the coexistence regions* $W = C_n/C_i - 1$ *as predicted by the scaled particle theory (soled line) and the second virial coefficient (dashed line) for different L/D ratio. Reprinted from the thesis of Dogic [29].*

Besides the hard core repulsion, charged rods have a long range repulsive interaction of the following form:

$$\frac{U_{el}(x)}{kT} = \frac{A' e^{-\kappa(x-D)}}{\sin(\gamma)} \tag{22}$$

where x is the closest distance between two charged rods, A' is the proportionality constant obtained by solving the Poisson-Boltzman equation, κ^{-1} is the Debye screening length and γ is the angle between two rods. In the case of charged rods there are contributions to the second virial coefficient from both the hard core excluded volume interaction and the long range electrostatic repulsion interaction:

$$B_2^{\text{iso}} = \frac{1}{4}\pi L^2 D_{\text{eff}} = \frac{1}{4}\pi D L^2 + \frac{1}{4}\pi\kappa^{-1}L^2(\ln A' + C_E + \ln 2 - \frac{1}{2}) \tag{23}$$

Therefore the thermodynamics of charged rods in the isotropic suspension will be equivalent to the thermodynamics of thicker hard rods with the effective diameter (D_{eff}). On the other hand if the cluster function $\beta(\gamma)$ is integrated over an anisotropic distribution function then the relationship given by Eq. 23 is no longer exact. The reason for this is that the electrostatic energy is lower for perpendicular rods then for parallel rods. Therefore the charge effectively destabilizes the nematic phase by shifting the I-N transition to higher concentrations and reducing the order parameter of the nematic phase coexisting with the isotropic phase. This is called the "twist" effect. However, most biopolymers (including *fd* virus) are highly charged in which case it turns out that the electrostatic "twisting" effect is insignificant compared to the excluded volume interactions, when the alignment of the rods is low. For the study of the *I-N* transition the D_{eff} of the *isotropic phase* is used.

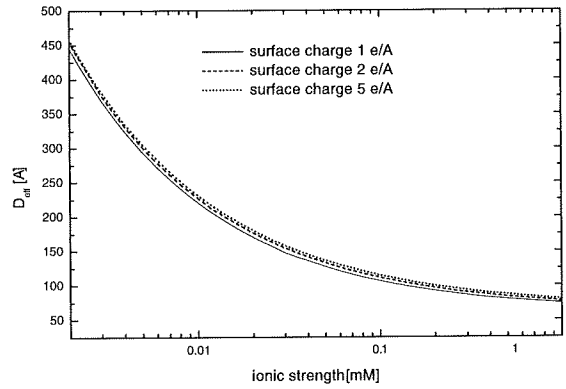


Fig. 5: *The effective diameter for a charged rod calculated from Eq. 23 for a range of ionic strengths. The bare diameter $D_{bare} = 66Å$ is that of the fd virus. The value of D_{eff} hardly changes as the surface charge is changed from $1e^-/Å$ to $10e^-/Å$. Experiments indicate that the surface charge is about $2e^-/Å$ [31]. Reprinted from the thesis of Dogic [29].*

For the calculation of higher order transitions one needs to accounted for the twisting effect to derive D_{eff}^{nem}. Using the second virial approach of Onsager it is found that:

$$D_{eff}^{nem} = D_{eff}^{iso}[1 + h\eta(f)\rho(f)] \tag{24}$$

with h the twist parameter as defined by Vroege and Lekkerkerker,[32] and both $\eta(f)$ and $\rho(f)$ functions of the angular distribution in the nematic phase, which are tabulated by Stroobants et al.[30]. As for the I-N transition, also the D_{eff}^{nem} has a limit concentration range where it is valid. The upper concentration limit is set by the point where the average angle between particles in the nematic phase is still less than D_{eff}^{nem}/L. This is the case for a typical order parameter in the nematic phase of $S_2 = 0.9$, which is close to the transition to the smectic phase (N-Sm). Again, as for the I-N, the concept of D_{eff}^{nem} becomes invalid at concentrations above the N-Sm coexistence, but can be used to derive the N-Sm transition.
The important notion is, however, that the effective diameter can be varied by changing the ionic strength. This provides the experimentalist with a useful tool to check the binodal points as derived by Onsager, which depends on the diameter D.

3.3 Effect of flexibility of rods on the I-N transition

Up to now we considered only stiff rods, which were not allowed to bend. Most systems, however, have some flexibility, which has to be accounted for. As explained in section 2, the flexibility is characterized by the persistence length P. For the limit that the actual length of the rod (the contour length L) is much smaller than the persistence length $P/L \rightarrow \infty$ we have a stiff rod. In the limit where $L \gg P$ the chain has an overall conformation of the flexible coil which in turn is composed of many persistence length units. If these units themselves satisfy the Onsager criterion ($P > D$), then the calculation of the phase diagram can be done by rescaling the exclude volume interaction. When the considered system is neither completely stiff nor flexible one has to adjust theory in such a way that the internal configurations of the semi-flexible chain

are also taken into account. There are various of ways to treat flexibility. For a clear overview we refer to Vroege and Lekkerkerker, and to Sato and Teramoto,[32, 33] although since then some progress has been made [24]. The fundament for the theory of semi-flexibility was formulated by Khokhlov and Semenov [34]. They reasoned that for semi-flexible rods besides orientational and translational entropy it is also necessary to take into account the internal configurations of the semi-flexible chain, which modifies the entropy term in the Eq. 7. Khokhlov and Semenov then solved the equation for the free energy in the limit of almost rigid rods (P≫L) and very flexible rods (L≪P) [35, 34]. After this work the challenge was to find the correct interpolation between the two limiting solutions and obtain an approximation for the configurational entropy for rods with arbitrary persistence length as was done by Hentscke, Odijk, and Dupre and Yang [36, 37, 38]. The most complete study of the influence of the semi-flexibility on the isotropic-nematic phase transition was presented by Chen [24] who numerically minimized the free energy for flexible rods as derived by Khokhlov and Seminov and subsequently calculated the co-existence concentrations. His results are shown in Fig. 6 for rods with different degrees of semi-flexibility, together with results obtained by using the Onsager approximation for the orientational distribution function (Eq. 9) and an expression for the entropy contribution of a semi-flexible rod as derived by Dupre and Yang.[38]

From Fig.6a we conclude that increasing flexibility destabilizes the nematic phase. That is, the I-N occurs at a higher volume fraction. Increasing the flexibility also drastically reduces the width of the concentration difference between the co-existing isotropic and nematic phases and the order parameter of the nematic phase as is shown in Fig. 6b and Fig. 6c respectively. The Onsager approximation (Eq. 9) for the ODF qualitatively agrees with "exact" numerical results. We observe that the difference in ρ_{iso} between the "exact" numerical result and the results obtained with approximate ODF is never larger then 5%. On the other hand, if we compare the nematic order parameter (S_2) at the I-N transition and the co-existence width (W) we observe a large quantitative disagreement (up to 30%) between two Onsager approach and "exact" approach. This demonstrates that comparing I-N co-existence concentrations between a theory and an experiment does not constitute a sensitive test of the theory. For a stringent test of the theory it is also necessary to measure the order parameter (S_2) of the nematic phase.

When the contour length is longerthan the persistence length $(L > P)$, the relevant criterion for ordering transitions to occur is that $P \gg D$ and not anymore that $L \gg D$. The coexistence concentrations and order parameter at the I-N transition rapidly decrease with increasing flexibility L/P, to reach asymptotical values for $L/P \rightarrow \infty$ of

$$(\rho_p)_{iso} = 5.409 \quad (\rho_p)_{nem} = 6.197 \quad S_2 = 0.610. \tag{25}$$

when using the Onsager ODF to obtain the concentration and order parameter at phase transition, and

$$(\rho_p)_{iso} = 5.124 \quad (\rho_p)_{nem} = 5.509 \quad S_2 = 0.46. \tag{26}$$

when using the numerical solution. Here again we note that the values for the coexistence concentrations are in much better agreement with each other than the values for the nematic order parameter. Two examples showing the influence of flexibility for $L > P$ on the concentration at phase transition are depicted in Fig. 19 for DNA and F-actin. It can be concluded that flexibility stabilizes the I-N transition, as long as the persistence length is much longer than the thickness of the polymer.

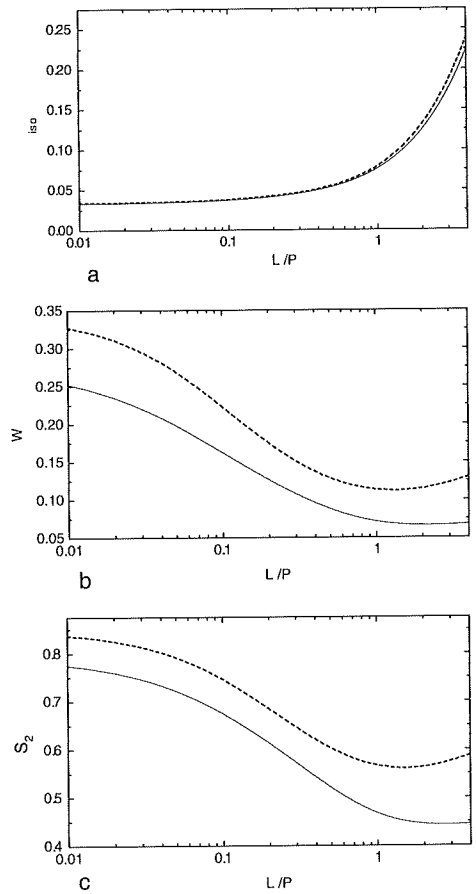


Fig. 6: *Concentration ($\rho_{iso} = (4/\pi)L^2D(N/V)$), number density difference between isotropic and nematic phase ($W = C_n/C_i - 1$) and order parameter (S_2) of the nematic phase co-existing with isotropic phase as a function of the flexibility of the particle L/P. The full line are the exact numerical results within the second virial approximation while, the dashed line are the results obtained by using the Onsager approximation for the orientational distribution function (Eq. 9). In all the figures the aspect ratio of the rod is fixed at 100 and the persistence length P varies from infinity to 25. Reprinted from the thesis of Dogic [29].*

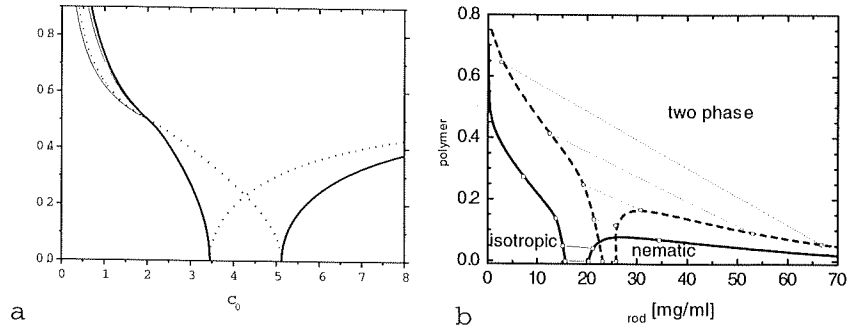


Fig. 7: *Two effects cause a widening of the co-existence region: (a) Polydispersity, where σ is a measure of the polydispersity of a system with having a Schulz distribution of length. Reprinted from Wensink and Vroege [40]. (b) Attraction, where the attraction is proportional to the polymer concentration $\phi_{polymer}$. The phase diagram is calculated for rigid rods (thick solid lines) and fsemi-flexible rods with $L/P = 0.4$ (thick dashed lines). In thin dashed lines in the graph indicate the coexisting phases. Reprinted from the thesis of Dogic [29].*

3.4 Effect of polydispersity on the I-N transition

Since most systems are not monodisperse, it is natural to study the phase behavior of a system of polydisperse rods. Although viruses like TMV and *fd* are intrinsically monodisperse, still it was the head to tail aggregation of TMV that showed the typical phenomena of a polydisperse system of rods. In this study the effect of a bidisperse mixture of rods with two different lengths already indicated some typical features of polydisperse systems: fractionation of the longer rods in the nematic phase and the shorter rods in the isotropic phase and a widening of the phase coexistence. Polydispersity has been treated theoretically, assuming various length distributions varying from bidisperse systems [22] to gaussian distributions[39]. In most studies the length distribution was assumed to be either discrete or vary narrow. Recently Wensink and Vroege presented a study where they calculated the effect of more practical distributions, using the Onsager second virial approach and a gaussian angular distribution as an ansatz for the ODF. The widening of the coexistence region upon a increasing polydispersity is nicely illustrated in Fig. 7a. Here Wensink and Vroege used a Schultz distribution of length, typical for colloidal systems [40].

3.5 Effect of attractive interactions on the I-N transition

In Chapter C 4 is was explained how an effective attraction between colloids can be induced by adding polymer to the dispersion. This is called the depletion effect. As can be seen in Fig. 16 of that chapter at the bottom left, the effect of the addition of rods (or the osmotic pressure of polymer) is that the co-existence region widens. The same effect has been found for rods by Lekkerkerker and Stroobants [41]. They showed that the approximate free energy of the colloid-polymer mixture is given by the following expression [41]:

$$F_{C+P}(\phi) = F_C(\phi) - \Pi_p < V_{free}(\phi) > \qquad (27)$$

where $F_C(\phi)$ is the free energy of the colloid suspension at volume fraction ϕ. The system

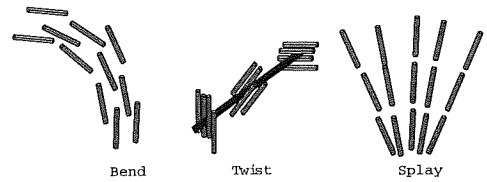


Fig. 8: *The three fundamentally different deformations of a nematic liquid crystal.*

is assumed to be in equilibrium with a polymer reservoir which is separated from the colloid-polymer mixture by a membrane permeable to polymers only. The osmotic pressure of the polymers in the reservoir is Π_p. V_{free} is the free volume available to a polymer in a solution of pure hard particle colloids. It is assumed that V_{free} in a polymer/colloid mixture is equal to the V_{free} in the pure colloid suspension. In this sense Eq. 27 is a thermodynamic perturbation theory. In order to derive the phase behavior of a rod-polymer mixture, one has to find an expression for the free energy of the colloid suspension and for the free volume. Deriving the free energy one should include higher virial terms, since attractive rods are in their lowest energy state when they are parallel to each other. These are exactly the configurations that need to be avoided if the second virial term on which the Onsager theory is based. We showed above how higher virial terms can be included using scale particle theory. In addition flexibility and charge can be incorporated in the free energy, as shown above. In Fig. 7b, two phase diagrams are plotted, showing the co-existence concentrations of the isotropic and nematic phase for rigid and flexible rods. As for hard spheres, the co-existence region widens when adding polymer to the system. Moreover, also the order parameter of the nematic phase increases upon addition of polymer. This underlines the earlier statement that the order parameter and the width of the co-existence region are two sensitive parameters to control theory. In Fig. 18a an experimental diagram is shown for a rod-polymer mixture, nicely showing the widening of the co-existing region.

3.6 Chirality and elasticity constants

As mentioned in the introduction, the nematic state is characterized by a vector describing the average orientation of the rods called the director. However, this direction is not pertained over long distances and can be deformed. There are three fundamental forms of deformation: bend, twist, and splay (see Fig. 8). Each of the deformations cost elastic energy, which are characterized by the so called Franck elasticity constants K_{ii} [42].
The fact that the bending of the director costs energy also influences the I-N phase transition. When the nematic phase nucleates out of the isotropic phase, theory would predict the formation of droplets, thus minimizing the surface energy. However, at the surface of the droplet the bending is very high, which causes a high energy penalty. Thus, finding a trade off between minimal surface energy and minimal bending, the most favorable shape of the nucleus is ellipsoidal. Such shapes are indeed found in experiments (Fig. 11), and are called "tactoids".
When all droplets are merged due to sedimentation, a full nematic *or* cholesteric phase will form. The behavior of the director in the two phases is also largely influenced by the Franck elastic constants. In the nematic state a real system is not homogeneous in the sense that the director is not constant throughout the sample. Instead nematic domains emerge, each of which having a different director. These domains give rise to the "thread"-like structures typical for the nematic phase. The formation of a domain costs elastic energy due to the strong changes of

the director, but increases the entropy. Thus, the number and size of the domains is determined by a competition between the surface energy of the domains and the entropy.

For some category of particles a cholesteric phase is formed, which is a state where the nematic director follows a helical path in space (see Fig. 1). These particles are chiral, which means that they have a helical morphology or charge distribution. The "twist" of the director is caused by the torque which a chiral molecule exerts on its neighbor, and is characterized by a Chiral twist constant K_t. Creation of an elastically distorted state characterized by the usual twist constant K_{22} raises the free energy. Therefore the ratio of the two constants sets the "Pitch" which is defined as the length over which the director has made a full turn. In most cases the elastic constant for the twist K_{22} is so low that the free energy difference between the nematic (N) and cholesteric phases is small. Thus the theory described above applies equally well to the isotropic-Cholesteric as to the I-N transitions.

3.7 Transitions to higher ordered states

Although the Onsager theory was successful in predicting the isotropic-nematic phase transition, no further transition at higher concentrations was predicted. From experiments it was known, however, that a transition to a state where there is also positional ordering occurs, since typical Bragg scattering peaks were observed. A second transition can be reasoned in a similar way as done for the I-N transition. There we argued that although orientational entropy is lost at the transition, configurational entropy is gained due to the much smaller volume that is excluded for aligned rods. When the density is further increased it also becomes favorable for the rods to positionally order. A nice intuitive picture was given by Wen and Meyer [43]. In the nematic phase, neighboring rods overlap one another by random amounts along their axial direction. This tends to create voids at the end of the rods, which are available only to that one rod and not to another. In the Smectic-A phase, with rods distributed in layers, the random overlapping of rods along their length is avoided, so the voids disappear. Hence, though positional entropy is lost at the transition to the Smectic-A phase, freely available volume is gained, and therefore also configurational entropy. Another way to eliminate the voids is to align the rods end-to-end; these strings then tend to pack together in a two-dimensional hexagonal array known as the columnar phase.

Even considering perfectly aligned rods one can argue that the smectic phase at some point becomes more stable. A smectic layer may be considered a 2D-liquid. If however, one rod would stick out of one layer into the layer on top, simultaneously voids will be created in the top and bottom layer as a result of the partially overlap it has with the neighboring layer. In this way configurational entropy is lost at high volume fractions and the rod is pushed back in its original layer. Thus, at high volume fractions the smectic phase is stable.

The approach to find the N-Sm phase transition is similar as used for the I-N transition. First one needs to find an expression for the free energy in the different phases, making the proper assumptions about the interaction terms of the particles, i.e. for the orientation dependence and the free volume. Then the osmotic pressure and chemical potential are derived, which should be simultaneously equal for the two phases. However, the difference with the I-N transition is that now one is not looking for an instability in the order parameter, but for instabilities in the density in one specific direction. Now the "trial" function includes the spacial distribution of the rods and assumes a periodic modulation of the average density ρ along the alignment axis:

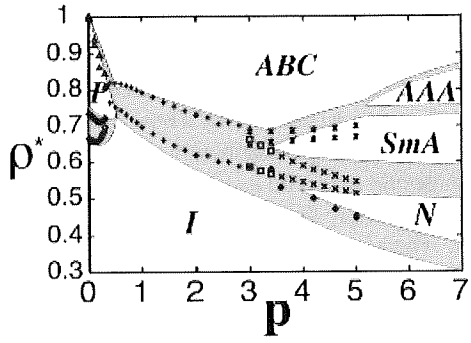


Fig. 9: *Phase diagram for hard spherocylinders in the $\rho* - P$ plane, where $\rho* = \rho/\rho_{closepacking}$ and in this case $P = L/D$, the aspect ratio. The shaded area corresponds to the coexistence region calculated within the cell theory and the dots are the simulation data from Ref. [27]. AAA, ABC, and P indicate different kinds of solid phases. SmA is the smectic-A phase, N the nematic phase, and I an isotropic fluid. The symbols for the simulation data are +, I-ABC transition; ◇, I-P transition; □, I-SmA transition; ♦, I-N transition; ×, N-SmA transition; *, SmA-ABC transition; ▲, P-ABC transition.*

$$\rho(\mathrm{r}) = \rho\psi(\frac{2\pi}{\lambda}\mathrm{z}) = \rho(1 + \sum_{\mathrm{l}=1}^{\infty} \hat{\psi}(\mathrm{l})\cos(\mathrm{l}\frac{2\pi}{\lambda}\mathrm{z})), \tag{28}$$

using the amplitudes $\hat{\psi}(l)$ and the wavelength λ to minimize the free energy [44]. The wavelength λ characterizes the spacing between the layers. Following this approach Mulder found a value of $\frac{\lambda}{L} = 1.34$ for the spacing, where L is the length of the rod. For the coexistence concentration he found a value of $\phi = 0.369$, with $\phi = c_s \frac{\pi}{4} LD^2$ the dimensionless volume fraction and c_s the number density [44]. There is still debate about the order of the transition. Most theories and simulations predict a second order, continuous, transition. However, Polson and Frenkel found evidence of a first order transition, when $L/D \longrightarrow \infty$ [45].

How the different characteristics of the smectic phase depend on the aspect ratio L/D was studied for hard rods by simulations [27] and theory [46, 47]. It was shown that at low L/D there is a direct transition from the isotropic to the smectic phase. At higher L/D the volume fraction at the nematic-smectic transition decreases initially with increasing L/D, but becomes constant above $L/D > 10$ [46]. More recently, Graf and Löwen used a scaled-particle theory to calculate the full phase diagram for hard spherocylinders. Beside the I-N and N-Sm transitions, also transitions to crystalline phases are predicted. Their phase diagram is depicted in Fig. [47], which also includes the simulation results of Bolhuis and Frenkel, showing a good agreement with theory.

The influence of charge can be treated using the expression for the thickness of the rods in the nematic phase, D_{eff}^{nem}, as derived in Eq. 24 in the above mentioned theory. A more rigorous way to treat the repulsive interaction was worked out by Graf and Löwen. They described the long-tail characteristics of the repulsive interaction on a mean-field level, and used the results for the short-ranged part of the interaction potential (i.e. pure hard sphero-cylinders) as a reference system.[47] In this way they obtained the free energies in the various liquid crystalline phases and the overall phase diagram including all significant phases, which they specifically applied

to TMV.

Again, like for the I-N, perturbations to the theory of ideal rods can be discussed. The influence of the flexibility was studied independently by van der Schoot [48] and Tkachenko[49]. Two important conclusions were drawn. First, it was found that the flexibility suppresses the N-Sm transition to higher concentrations. An intuitive picture can be given that the voids at the end of the rods, which are available only to that one rod and not to another, can be partly "filled" when the rods have some flexibility, thus increasing the concentration where the excluded volume effect is really taking over. Second it was found that the spacing $\frac{\lambda}{L}$ is decreasing with increasing flexibility. Tkachenko also found that the transition from nematic to smectic was continuous (second order) for rigid rods and first order for semi-flexible rods.

Polydispersity is also expected to suppress the N-Sm transition. A simple explanation for this is that rods of different lengths do not pack into layers as efficiently as rods of the same length. Therefore, when the polydispersity is very high, a more efficient way to pack will be that rods align in columns, which is another way to deal with voids (see above). Sluckin was the first to treat polydispersity using Mulder's approach of the N-Sm transition and calculating the perturbation due to the polydispersity [39]. Bates and Frenkel performed Monte Carlo simulations on a polydisperse hard rod system in the limit of infinite aspect ratio. The results of their simulations are plotted in Fig. 10, together with the theoretical prediction by Sluckin. They found three different regimes of phase behavior: corresponding to small (standard deviation $S_2 < 0.08$), intermediate ($0.08 < S_2 < 0.18$), and large ($S_2 < 0.18$) polydispersity. In the first regime, the monodisperse limit, the behavior is unaltered. This is not surprising since we saw already before that the spacing between the layers exceeds the length of the rod. Small deviations in the heights of the particles will not destroy the smectic phase, since the different lengths can be accommodated in the layer structure. In the intermediate regime the maximum density at which the smectic phase is stable decreases, while the maximum density where the nematic phase is stable is increased. In other words, the concentration range where the smectic phase is stable decreases. For even higher concentrations the smectic phase is not found and there is a direct transition from the nematic phase the the columnar phase.

4 Experimental verification

4.1 Isotropic-Nematic transition for nearly ideal rods

The Onsager theory for hard rod packings predicts a discontinuous transition between an unordered, isotropic, state to an ordered, anisotropic, state. In other words, at a fixed concentration the system phase separates in a part with a higher concentration and order, and a part with a lower concentration and no order. A phase separated system consists of droplets of nematic phase in a surrounding isotropic phase when the isotropic phase is dominating, and vise versa when the nematic phase is dominating. Such droplets are shown in Fig. 11a as first observed by Brawden et al.[51] 1936 in a suspension of TMV, using a polarization microscope. Also shown are the biphasic structures of *fd* (b) and DNA (c) suspensions. Clearly, the suspensions of the rigid rods show typical ellipsoidal form, tactoids, as was discussed in subsection 3.6. The "crosses" that can be seen in the droplets are typically due to birefringence, and are a signature for the ordering of the rods in the droplet. DNA in the biphasic region forms spherical birefringent droplets. The nematic elastic constants for this system is not high enough to compete with the surface energy. These droplets will merge to one phase when they sediment. For some

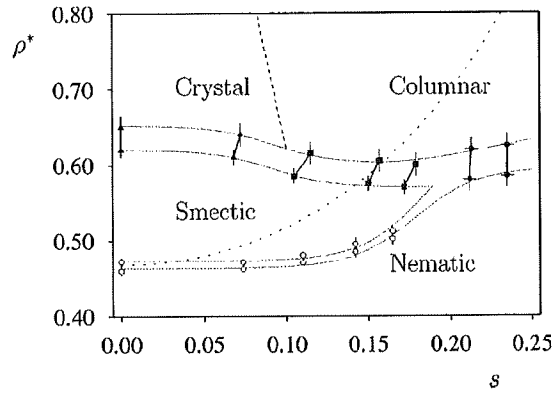


Fig. 10: *The phase behavior of polydisperse spherocylinders in the limit of infinite aspect ratio as obtained by semigrand ensemble Monte Carlo simulations by Bates and Frenkel [50], where S is a measure for the polydispersity. The symbols indicate the NSm (○), SmCrystal (▲), SmCol (■) , and NCol (●) transitions and the coexisting phases are joined by heavy solid lines. The dotted lines indicate the approximate phase boundaries and are intended as a guide to the eye only. The dashed line is an intuitive estimate of the CrystalColumnar transition and the dotted line gives the position of the NSm transition from the bifurcation analysis polydisperse system of Sluckin [39].*

systems the final phase separated phase is cholesteric and for others it is nematic, depending on the morphology of the rods and the charge distribution. Fig. 11d shows a cholesteric phase of *fd* which has indeed a chiral structure. Interestingly, Pf1 (Fig. 11e) and TMV do not show a cholesteric phase. Especially for Pf1 this is striking, since it has a very similar structure as *fd*. This is still a subject of ongoing research [52, 53].

In order to test Onsager theory on a quantitative level one needs to verify the relation between the aspect ratio and the concentration at which the transition takes place, and the order parameter of the nematic phase. For that purpose one would ideally need rods which are monodisperse, very long ($L/D > 100$), very stiff and only interacting via excluded volume. None of the systems mentioned in table 1 satisfy all these conditions. The viruses, which are intrinsically monodisperse, are the most promising candidates to test the theory. However also the viruses are not ideal. Although the TMV virus is extremely stiff it has an aspect ratio of only L/D=17. Also *fd* is not ideal, but it has the great advantage that it is extremely long and reasonably stiff. As all the other systems it is charged so that the interaction is long-ranged due to the electro-static repulsion. When testing the Onsager theory one has to consider these effects. As mentioned in subsection 3.2 one way of dealing with the electro-static repulsion is by assigning an effective diameter to the rod, corresponding to the extra excluded volume due to the electro-static repulsion between the rods. Though this might seem a disadvantage, it provides the experimentalist with a tool to study the relation given by Onsager between aspect ratio and the concentration at which the isotropic-nematic transition is found. After all, by adding salt to the dispersion the effective diameter is reduced. This has been done for both *fd* in Fig. 12 a [55] and TMV in Fig. 12 b [56]. Grelet and Fraden also made use of the possibility to modify *fd* by irreversibly binding neutral polyethylene glycol (PEG) of different molecular weights (5,000 or 20,000 g/Mol) onto it. In this way the bare diameter of the virus increases with two times the

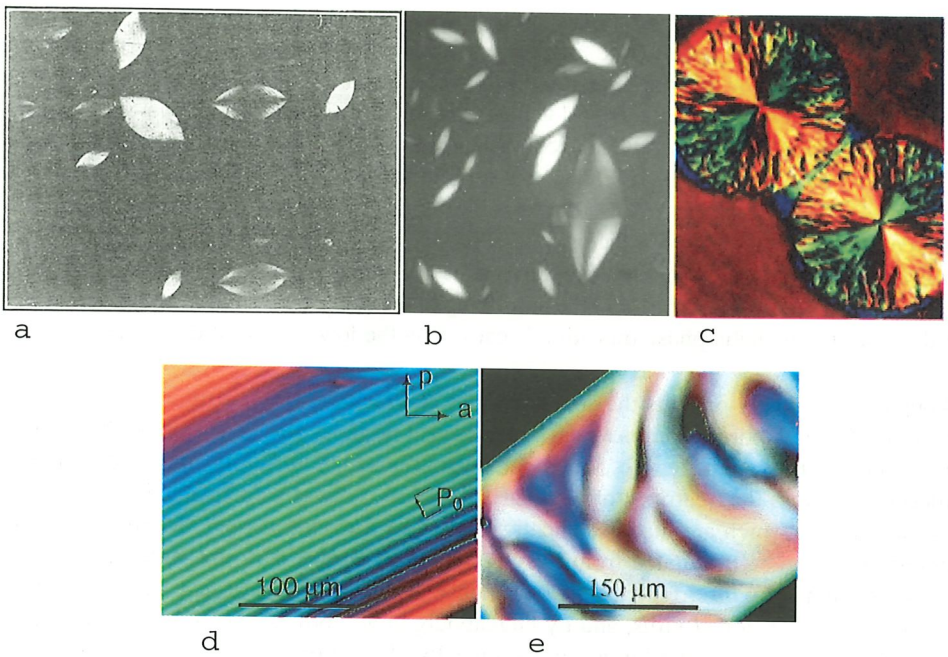


Fig. 11: *Polarization micrographs showing nematic tactoids on an isotropic background for TMV (a) and fd(b, photo by K.Kang/IFF, Weiche Materie). Fig. a is a historical picture published in 1936 by Brawden et al.[51] showing for the first time tactoidal structures for lyotropic liquid crystals. (c) Nematic droplets of DNA, as reprinted from Zakharova et al. [54]. d) Typical cholesteric phase of fd as observed with polarization microscopy. The length of the cholesteric pitch (P_0) spans two bright and two dark lines. The positions of the polarizer (p) and analyzer (a) are indicated in the upper right corner. (e) Texture observed with a polarization microscope of Pf1 in the nematic phase, not showing any cholesteric structure, despite of the similarity with fd.*

radius of gyration of the polymer (respectively 3 and 7 nm). When varying the ionic strength, *fd* with grafted PEG shows identical behavior to non-grafted, up to an ionic strength corresponding to a D_{eff} that matches the size of the stabilizing polymer layer. For a higher amount of added salt, the phase diagram becomes independent of ionic strength, and the fd-PEG system behaves as a sterically stabilized suspension. This is schematically represented by the cartoon of a virus with polymers grafted on its surface where $D_{eff}^{electrostatic} < D_{eff}^{polymer}$.

Comparing the results with theory it is for TMV important to incorporate the higher virial coefficients since $L/D = 22$, even when the charge is not taken into account. In Fig. 12 this is done by using theory by Lee[58], similar to the Cotter theory explained in the theoretical section. For the *fd* it is important to incorporate the flexibility and to a lesser extent the higher virial coefficients. For this the numerical results of Chen are used [24], which treats excluded volume interactions at the second virial while the orientational distribution function is calculated exactly. Figure 12 shows that the concentrations at the I-N transition can be predicted well by theory, provided that charge, flexibility and a limited aspect ratio is accounted for. Although *fd* is fairly rigid with $P/L \approx 2.5$, the flexibility is important enough to narrow the width of the coexistence region, w, by a factor of 3 from the rigid rod limit and suppress the isotropic-cholesteric transition to higher concentrations, as was predicted in Fig. 6 a and b. The same effect is observed for TMV, but here the narrowing of the coexistence region and the suppression of the isotropic-nematic phase transition is caused by the low aspect ratio, as predicted in Fig. 4.

Though theory nicely predicts the concentrations at which the I-N transition occurs, we know from Fig. 6 that the order parameters is a more sensitive check for the theory, at the I-N transition and in the fully nematic sate. Moreover, one of the key points of any theory describing the ordering of rods is the treatment of the ODF. The ODF is directly accessible by X-ray diffraction on samples which are macroscopically aligned by a magnetic field. Moreover, the order parameter S_2 can be calculated from the ODF (11) and compared to the order order parameter as determined by birefringence measurements. This rigorous treatment has been followed by Purdy et al.[59] for *fd* virus, and by Oldenbourg et al. [60] for TMV. In Fig. 13 the scattered intensity is shown of *fd* in a macroscopically aligned nematic phase for three different concentrations at wide and small angles, i.e. due to intraparticle scattering and to interparticle scattering, respectively. It was shown that the ODF can be determined independently from both scattering patterns. The intraparticle interference, or form factor, contains information about the structure of the individual particles. The interparticle interference, or structure factor term, contains information about the positional and orientational correlations between particles. In Fig. 13 c angular intensity scans at a fixed q-value are shown, with the resulting ODF fit. However, the Onsager and Gaussian trial angular distribution functions fit the angular distribution of both the intraparticle and interparticle diffraction peaks equally well and returned similar values for the nematic order parameter, so no decisive conclusions can be drawn about the shape of the ODF.

The concentration dependance of the order parameter was measured from both the interparticle and intraparticle peaks. The resulting values are graphed in Fig. 14. At high ionic strength, or large L/D_{eff}, the nematic order agrees with that predicted by a scaled-particle-theory of charged semi-flexible rods. At low ionic strength, theoretical predictions only qualitatively reproduce the concentration dependance of the order parameter. With decreasing ionic strength however, a weak systematic increase in the nematic coexistence order parameter was found. This is consistent with both a decrease in the twist parameter and a deviation of L/D_{eff} from the long rod limit. The TMV data are compared with calculations of hard rods, where higher virial

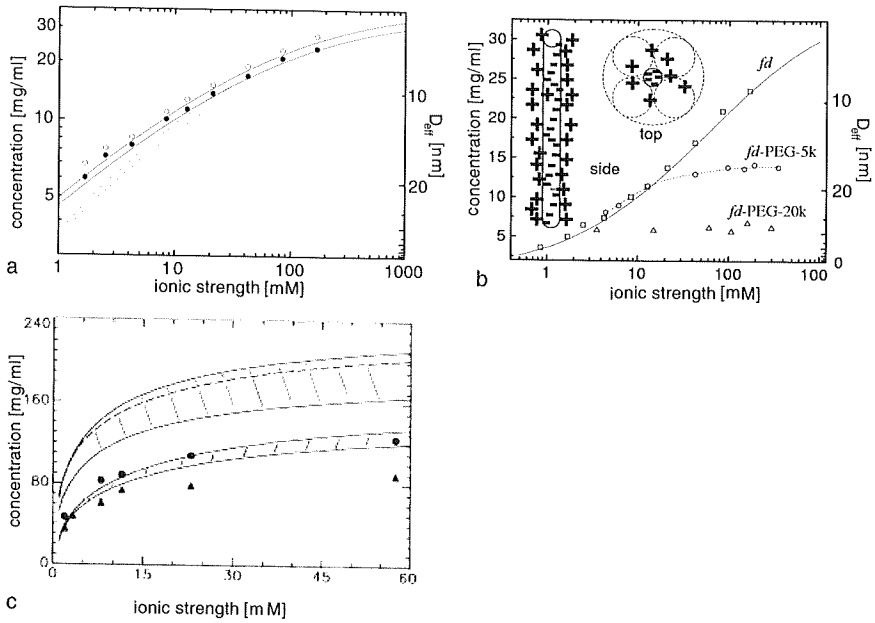


Fig. 12: *The concentrations of the co-existing isotropic (lower points) and nematic (lower points) for aqueous suspension of fd virus (a, data from Tang and Fraden[57]), fd coated with neutral polymer (fd-PEG-5k and fd-PEG-20k) (b, reprinted from Grelet and Fraden)[53], and TMV (c, reprinted from Fraden et al. [56]). In Fig. a the full lines indicate the coexisting boundaries as calculated from the numerical solutions of Chen.[24] The dashed lines indicate the co-existing boundaries as calculated from the scaled particle theory, in which all virial coefficients and the orientational distribution function are included in an approximate way. The scale on the right side indicates the effective diameter of a rod at at I-N transition for a given ionic strength. In Fig. b the lines represent the highest concentration for which the isotropic phase is stable (below line), whereas the range of the cholesteric phase is above each line. Squares and circles indicate the I-N transition in fd and in fd coated with PEG-5k respectively, while triangles refer to the fd virus coated with PEG-20k. The cartoon shows the situation where where* $D_{eff}^{electrostatic} < D_{eff}^{polymer}$. *In Fig c, the cubes and the triangles indicate the concentration of the co-existing isotropic and nematic phase, respectively. The coexistence region is calculated using Onsager theory (negative-sloping hatching), and the scaling theory of Lee (positive-hatching) accounting for the higher viral coefficients.*

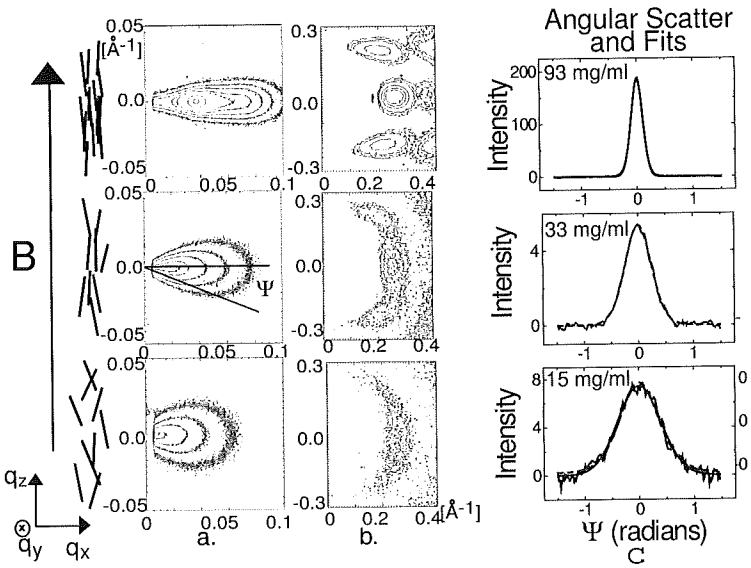


Fig. 13: *(a) Contour plots of scattering from nematic fd due to interparticle scatter.(b) Contour plots of scattering from nematic fd due to intraparticle interference. The interparticle scatter shown in (a) is hidden behind the beamstop in (b) which is located on the left of the images and can not be seen in the contour plot. From top to bottom the concentration of the samples are 93 mg/ml, 33 mg/ml, and 15.5 mg/ml. Samples shown are at an ionic strength of 10 mM (20mM Tris buffer), pH 8.2. The magnetic field and virus orientation are perpendicular to the scatter as shown in the schematic. Ψ is the angle from the equator on the detector film. Axes of contour plot are labeled in $Å^{-1}$, note the scales are different in (a) and (b). (c) Angular intensity scan at q_r=0.07 ± 0.001$Å^{-1}$ from the three data plotted in (a) with best-fit curves calculated from the Gaussian and Onsager ODF's (solid line), and Oldenbourg's ODF (dotted line). Reprinted from Purdy et al. [59].*

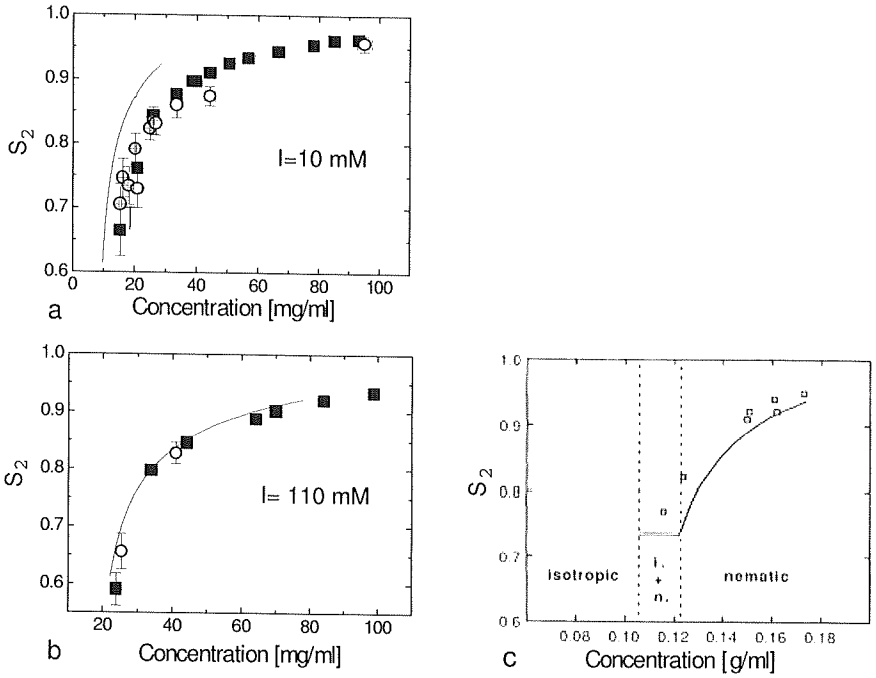


Fig. 14: *Concentration dependence of the nematic order parameter for fd (a and b) [59] and TMV (c) [60]. (a) fd in a buffer with an ionic strength of 10 mM, and (b) 110 mM ionic strength at pH 8.2. Squares (■) in (a) and (b) are from the interparticle interference peak, and open circles (○) are results from the intraparticle peak. The solid lines shown are for a scaled-particle-theory for charged semi-flexible rods.*

coefficients and the limited L/D of TMV are taken into account, showing that for this system no flexibility needs to be included.

The values for the order parameter calculated from the x-ray diffraction data were also compared to birefringence measurements for the whole range of concentrations and ionic strengths. The results are plotted in Fig. 15. The saturation birefringence per unit concentration was measured as $\Delta n_{\mathrm{sat}}/c = 3.8 \times 10^{-5} \pm 0.3 \times 10^{-5}$ ml/mg. Having established this strict relationship between x-ray measurements and birefringence, the order parameter can also be obtained simply by measuring the birefringence of a sample of nematic *fd* and rescaling it by the saturation birefringence. This is useful notion for future research.

4.2 Higher order phase transitions for nearly ideal rods

Viruses like *fd* and TMV are ideal systems to study the smectic phase, due to their intrinsic monodispersity. In fact, the smectic phase was first observed in a dispersion of TMV, which had a soapy structure at high concentrations. Studying these samples with white light, different colors emerged depending on the angle which the light source made with the sample. These observations are typical for crystal structures. Indeed electron microscopy pictures showed a very regular layered structures and scattering measurements showed intense Bragg peaks.

Fd at high concentrations also shows a smectic phase, which can directly be visualized under

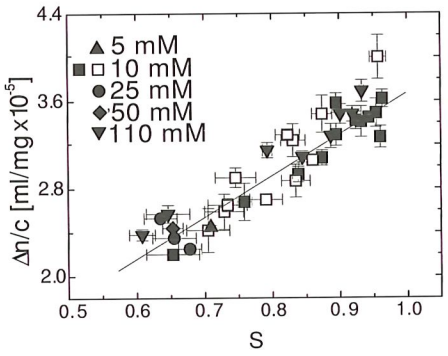


Fig. 15: *Comparison of measured birefringence $\Delta n/c$ to the deduced x-ray order parameter S_2. Open shapes are from intraparticle interference peak measurements. Closed shapes are from interparticle interference measurements. Reprinted from Purdy et al. [59].*

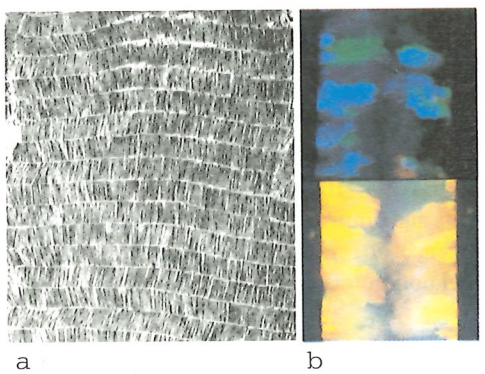


Fig. 16: *(a) A TEM micrograph of a ultra thin layer of the smectic phase of TMV. (b) Bragg scattering from the same sample using two different angles. Both pictures are reproduced from Wetter.[61]*

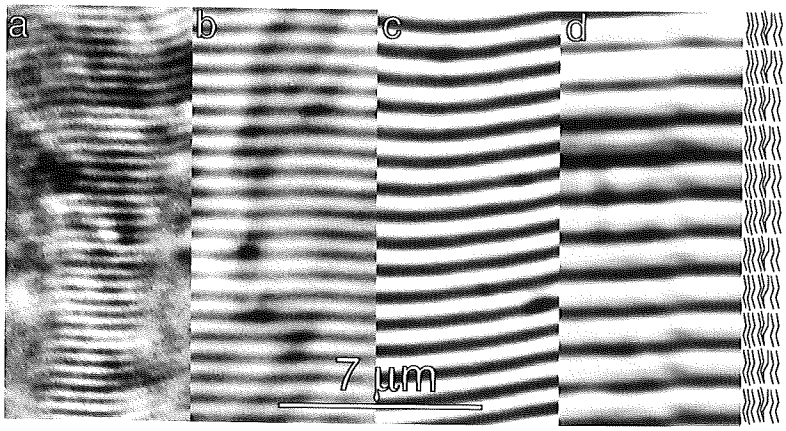


Fig. 17: *Optical micrographs of smectic phases of three different M13 constructs and fd (c). The periodic pattern is due to smectic layers that are composed of two dimensional liquids of essentially parallel rods, as indicated in the cartoon on the right. From left to right, the contour length of the rod-like viruses forming the smectic phase are 0.39 μm, 0.64 μm, 0.88 μm, and 1.2 μm. The smectic spacing measured from optical micrographs is 0.40 μm, 0.64 μm, 0.9 μm and 1.22 μm from image (a) to (d) respectively. Reprinted from Dogic [62]*

the microscope, since it is more than two times longer as TMV [62]. In subsection 3.7 the theoretical background of this transition was exposed. The smectic phase is not only characterized by the order parameter of the rods and the concentration at the N-Sm transition, but also by the spacing between the layers. For *fd* it was found that the volume fraction at the cholesteric-smectic transition was constant, as predicted, even though L/D_{eff} increases from about 30 to 70 nm as the ionic strength varies from 4 to 65 mM. The value of $\phi_{eff} = 0.75 \pm 0.05$ at the N-Sm phase transition is substantially higher than observed in simulations or theory of the N-Sm transition for rigid and hard particles, where $\phi = 0.5$ is found. In subsection 3.7 we saw that both flexibility and polydispersity may lead to higher volume fractions of the transition. Since *fd* is intrinsically monodisperse we contribute the high volume fraction to the flexibility of *fd*. The smectic layer spacing *fd* has a width of $\lambda = 920 \pm 10\ nm$, so $\lambda/L = 1.03$. These values did not vary with ionic strength, and were lower than predicted for hard rods. Again it can be argued that this is due to the flexibility of the rods, but theory is not accurate enough yet to use for a direct comparison.

In order to establish a link between flexibility and the layer spacing, the smectic structure of clones of M13 (which is almost identical to *fd*) of different length were studied.[52] Here Dogic and Fraden used the very useful tool from biology that allows to genetically modify the bacteria that produce M13 in such a way that viruses of different lengths can be produced. The lengths of the viruses varied between 0.4 and $1.4\mu m$, while it was assumed that the persistence length ($P \sim 2.2\mu m$) of all the clones is the same. Thus it was expected that the shorter rods would be relatively stiffer than the longer ones and consequently that the layer spacing would increase for shorter rods. It was found from Fig. 17, however, that the measured spacing of the smectic phase (λ) is almost identical to the contour length (L) for all the mutants studied.

Another issue is the order of the transition to the smectic state. Most theories predict a continuous transition from the nematic to the smectic state, but there is also prove from simulations

that there is a jump in concentration, suggesting that it is a first order transition. Using optical microscopy coexisting regions of the nematic and smectic phases were observed with abrupt boundaries between the phases, which is evidence of a first-order transition. For TMV, which is effectively a rigid rod, a second order transition was observed. All viruses are charged and it seems that the only significant difference between the viruses is their flexibility. It can be concluded that flexibility drives the second order N-Sm transition first order, in agreement with a recent calculation [49]. Also the spacing between the smectic layers and the concentration at the transition were observed for TMV were closer to the predictions of simulations for hard rods: $\lambda/L = 1.2$ (1.3 from simulations), $\phi_{N-Sm} \sim 0.5$ for both, simulations and experiment.
It has been said above that the main drawback of polymeric systems is, their intrinsic polydispersity. This problem has been circumvented by, again, using a trick from biology. The same *E-coli* bacteria which are used to produce *fd* virus can be modified in such a way that they produce very mono-disperse poly-saccharides [63]. It was observed that monodisperse PBLG of $11.4nm$ and $14.1nm$ form a smectic phase. Interestingly the spacing of the smectic layers were observed to be $\lambda/L = 1.0$ for both. This is lower than expected, because for this length they are very stiff.
More complex phase behavior and structures are found when the rods are forced to very high concentrations by the addition of coiled polymers to the dispersion. In section 3.5 we explained how the co-existence region should widen by the addition of polymers. Fig. 18a depicts a phase diagram for a mixture of *fd* and dextran, taken from Dogic [64]. Indeed a strong widening of the co-existing region is observed, as escribed in subsection 3.5. Thus, for a critical polymer concentrations, the concentration of the nematic phase gets so high that a transition to the smectic phase is expected. However, this is a local effect when the system is still in the co-existing region and therefore contains droplets of one phase embedded in the other. As a result the tactoids which are normally observed (Fig. 11 will not be in the equilibrium state, and higher ordered structures will grow out of them. In Fig. 18 four regions are identified for which increasingly more ordered and complex structures are detected. The kinetics of the formation of the structures and the morphology itself is still poorly understood.

4.3 Phase behavior of non-ideal rods

Having established Onsager theory for the virus model systems, we proceed and see what happens when using less ideal, but biologically very relevant, systems like DNA and F-actin. DNA is the most flexible polymer we tabulated in table 1, but at sufficiently high concentrations it will form liquid crystalline phases, as was predicted for polymers with a large L/P *and* P/D (see subsection 3.3). As mentioned in the section 2, closely packed DNA in primitive cells show nice ordering. Strzelecka et al.[65] were the first to systematically study the phase behavior DNA using both optical microscopy and NMR. A sequence of transitions from isotropic to cholesteric to smectic was suggested. Later, Livolant et al [66, 67] showed that the highly concentrated phase is a columnar hexagonal state, using electron microscopy in combination with the freeze fracture technique. This complies with the argumentation we gave in subsection 3.4 that for polydisperse systems the columnar phase is more stable than the smectic phase, since DNA samples always have some polydispersity. For a review on the phase behavior of DNA see Livolant and Leforestier [68], and Bloomfield [69].
In order to gain more insight on the influence of the flexibility on phase behavior, one can fractionated DNA into portions of varying length L and limited polydispersity. The stiffness of DNA is supposed to be unaffected by the fractionation, one can test the influence of the

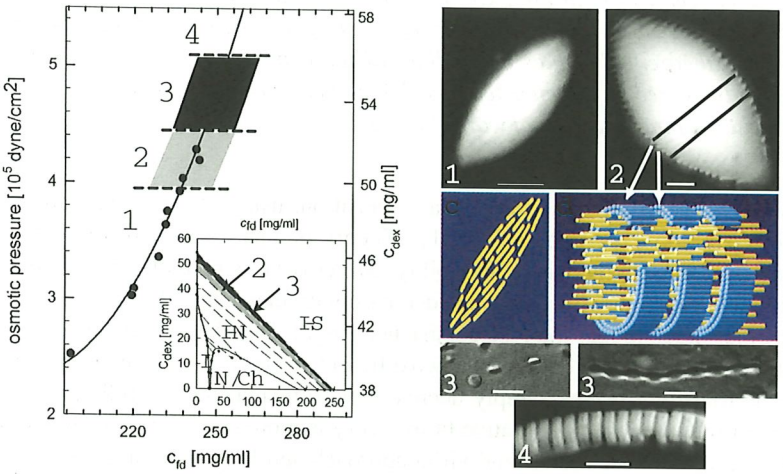


Fig. 18: *(a) The coexistence concentrations of an immiscible fd-Dextran mixture. The y axis shows the concentration and the osmotic pressure of Dextran in the isotropic phase that coexists with the rod-rich liquid crystalline phase (droplets). The rod concentration is shown on the x axis. Numbers 1 through 4 indicate regions where different phase behaviors are observed. Images of the structures observed in these regions are shown in (b). A stable surface smectic phase wets the isotropic-nematic interface in region 2. Colloidal membranes are stable in region 3. Inset: The complete phase diagram of a fd-Dextran mixture. Tie lines along which the phase separation proceeds are indicated by dashed lines. Regions of the isotropic-nematic (I-N) and isotropic-smectic (I-S) coexistence are indicated. Reprinted from Dogic [64].*

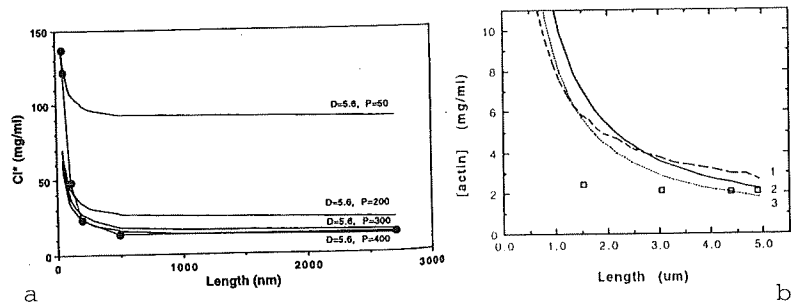


Fig. 19: *The concentration at the I-Ch transition as studied for DNA (a, reprinted from Merchant and Rill[70]) and F-actin (b, reprinted from Furukawa et al.[72]). (a) Comparison of actual (•) and theoretical values (solid line) of the concentration Ci^*, as a function of the length for the indicated persistence lengths (P) diameter (D). (b) Comparison with theoretical models for charged and uncharged worm-like chains.*

flexibility on the isotropic-cholesteric phase transition, using NMR, polarization microscopy and phase separation experiments [70, 71]. Theory predicts (Fig. 6b) that the width of the co-existence region with increasing flexibility. Moreover it is predicted for a fixed persistence length that both transition concentrations decrease with increasing flexibility, until the contour length becomes much longer than the persistence length. Qualitatively this latter prediction has been verified by experiments on DNA by two different groups [70, 71]. Indeed the concentration at which the transition occurs steeply decreases with increasing flexibility, until a plateau is reached (see Fig. 19). The quantitative fit to theory is more convincing using the treatment of Kassapidou et al, who use the second virial approach and Dupre's treatment of the flexibility, but there range of L/P is more limited. There is some dispute on the width of the biphasic region. The problem is that one has to account for some polydispersity, which also widens the coexistence region (see Fig. 7). Kassapidou et al. also produced a complete phase diagram varying the aspect ratio dependence [71]. Doing so they used the same tool as had been applied for *fd* and TMV: using a rather monodisperse sample of DNA with length $L = 42\ nm$, they changed the ionic strength and thus the effective thickness. They obtained a phase diagram that can be compared with that of *fd* (see Fig. 20). Note that now the phase at high concentrations is a columnar phase and not a smectic phase.

F-actin is a potentially very interesting model system, since it is very stiff. The length of the filaments can be controlled using gelsolin, which is a nucleating and filament severing protein. The average length of the filaments is thus set by the ratio of the actin monomers (Globular or G-actin) over gelsolin. Furukawa et al.[72] used this tool to get a relationship between the length of the filaments and the concentration at which the phase transition to the aligned phase occurs. As can be seen in Fig. 19 the correspondence is poor for each of the three theories tried. Only a good correspondence is found at the limiting value for very long filaments is reached. One possible explanation is that the persistence length might be different than assumed. The disadvantage of F-actin as a model system is, however, that it is a self-assembling polymer and it is difficult to control the polydispersity. This is probably the reason why for the short filaments the transition takes already place at low concentrations, induced by the few long filaments still present in the distribution of filaments. In Fig. 7 it was indeed shown that polydispersity of the length distribution makes that anisotropy occurs at lower concentrations than expected. Most

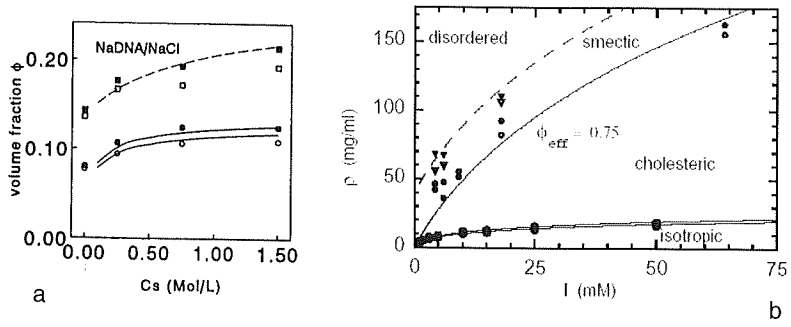


Fig. 20: *Complete phase diagrams for monodisperse DNA of* 50 μm *length (a, reprinted from Kassapidou et al.[71]) and fd virus (b, reprinted from Fraden[10]): the concentration at the I-N (bullets) and N-Columnar (DNA, squares), respectively, N-Sm (fd) as a function of ionic strength, i.e. rod diameter.*

probably it is also the polydispersity that causes the order parameter in the nematic phase to be much lower than expected for such stiff rods. The order parameter was measured using the polarized fluorescence technique by Coppen and Leavis.[73] The maximum value found by these authors was $S_2 \sim 0.2$, which is a lot lower than observed for e.g. *fd* (see Fig. 14. A future challenge is to use the measured length distributions to predict the phase behavior on the basis of such a theory as developed by Wensink and Vroege.[40] As mentioned before the I-N transition is expected to be of first order transition, which is characterized by a phase coexistence between the nematic and the isotropic phase. For F-acing, varying in length between 3 and 12 μm, this was not observed [74], although F-actin meets all the requirements regarding aspect ratio and stiffness. Another particle that meets the requirements is Pf1 but also for this particle no phase coexistence has been observed, although this might be due to its flexibility. Already for *fd* we have seen that the coexistence region is narrowed by a factor two, and Pf1 has the same persistence length and is twice as, long, i. e. twice as flexible. There is some debate on the reason why long rods undergo a continuous transition. One possibility is that the rotational diffusion of these long rods is so slow that the transition is just very slow. After all, in order to form an ordered state the rods have to reorient, and the rotational diffusion is strongly dependent on the length of the rod.

5 Mixture and external fields

5.1 Mixtures of rods and other particles

One of the main motivations to study so thoroughly the effect of packing hard rods, was to understand how this system can undergo different phase transitions, purely driven by geometrical effects like the shape of the rod, its flexibility, and the number of rods present. In real live, systems are scarcely pure and different species are included in the same system. Again it can be studied, purely on the basis of geometrical arguments, how a multi-component system organizes. This effect is called "crowding". For example we saw that F-actin forms ordered phases at concentrations much lower than the *in vivo* concentration in the cell. However, in the cell their is no sign that the filaments order. Instead they are forced into a different organization

by the surrounding particles. For DNA self-assembly with nucleosomes or other 'crowding' mechanisms are most regularly used to get the most efficient packing [75, 16]. Herzfeld gave a nice insight in this problem and shows how entropy can drive complicated phase transitions. For a more recent overview see also Rivas et al. [17].
An important step to systematically study the effect of crowding was undertaken by Adams et al.[76] introducing colloidal spheres in dispersions of rods which were in the isotropic or nematic phase. The particles used were *fd* viruses and synthesized monodisperse polystyrene spheres of various size. For *fd* we have shown above that it is a close to ideal hard rod. That the used spheres behave like model hard spheres has already been shown by Pusey et al. [5] and others. Depending on the size and volume fraction of the spheres, either demixing in to rod-rich and rod-poor phases is found, or the formation highly ordered structures, varying from "lamellar" structures where layers of rods and spheres are formed, to even two-dimensional columnar structures of rods, see 21. Since all interactions are purely hard core interactions entropy, must be the driving force behind the observed phase transitions. As for the ordering phenomena we have discussed for the pure virus systems, it is the maximization of chaos/entropy that induces order. Though this might be counterintuitive, one should realize that due to higher degree of organization, their can be an enormous gain of free volume. The fact that more highly ordered structures are formed when mixing two species of different shape than would have been formed for the two separate systems illustrates the possible influence of such effects in biological systems like cells, were globular proteins are mixed with filaments like F-actin and microtubules.
In the sample we discussed above a mean-field of rod-like particles was disturbed by big colloidal spheres. Visa versa, the mean-field of rods can also influence the ordering of particles emerged into the sample. Fig. 22 shows micrographs of F-actin and DNA immersed in a dispersion of *fd*. F-actin and DNA are labelled with a fluorescent dye so they can be observed individually under using fluorescence microscopy. At the left side the dispersion is in the isotropic phase, on the right side it is in the nematic phase. Clearly F-actin undergoes a coil-rod transition such that it dissolves in the nematic phase by stretching out along its director. As mentioned in subsection 3.6, deformation of the director costs elastic energy. On the other hand also the F-actin has a bending constant. Apparently the cost to stretch the F-actin is less then the cost to deform the nematic phase, resulting in the stretching of F-actin in the nematic phase, see Fig. 22 a. DNA, on the other hand, does not undergo a coil-rod transition but forms a slightly anisotropic droplet which is shown in Fig. 22b. It can be concluded that sufficiently stiff semi-flexible polymers stretch out when changing the background phase from the isotropic to nematic phase, while more flexible polymers micro-phase separate to anisotropic droplets.

5.2 Magnetic field and shear field

For birefringence measurements it is used that (most) rod-like particles have a different dielectric constant along their long axis and perpendicular to their long axis. Similarly they also often have different diamagnetic constants along and perpendicular to their long axis, resulting in a diamagnetic anisotropy. This feature can be exploited to align rods in a magnetic field. It thus can be used to align nematic domains, unwind cholesteric phases, or align smectic domains, such that one single monodomain is created. In this way one can determine the ordering of the rods in the phase of interest, as was done for the X-ray experiments in subsection 4.1 [59].
There is important information contained in the strength of the magnetic field needed to unwind the cholesteric phase, H_c, inducing the nematic phase. This critical magnetic field strength is related to the twist elastic constant, K_{22}, which can thus be determined by measuring H_c [42].

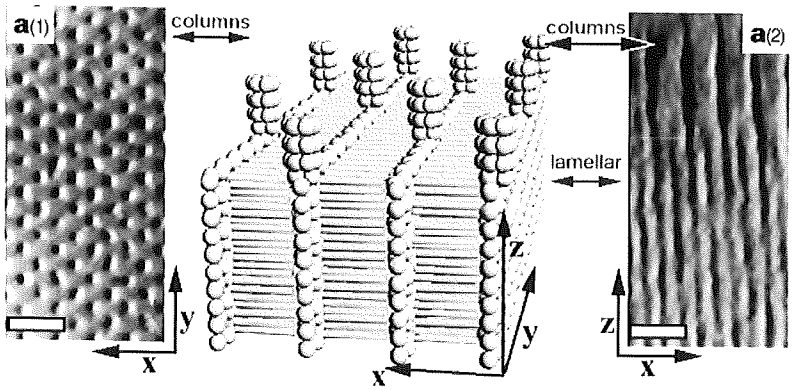


Fig. 21: *One exemplary structure formed by a mixture of fd and 100-nm diameter polystyrene spheres, using differential interference contrast (DIC) microscopy. The micrograph cannot clearly resolve single particles, so the drawings give only an approximate representation of the structure. The actual volume fraction of the spheres within the columns and lamellae is 10 to 20%. Length bars are 3mm. a(1), Section in the xy plane perpendicular to the columns, corresponding to the upper part of the adjacent drawingeach column is about 3 spheres in diameter; a(2), section in the xz plane corresponding to the front of the drawing. The columns (top) have twice the spacing of the lamellae (bottom). Figure is reprinted from Adams et al.[76]*

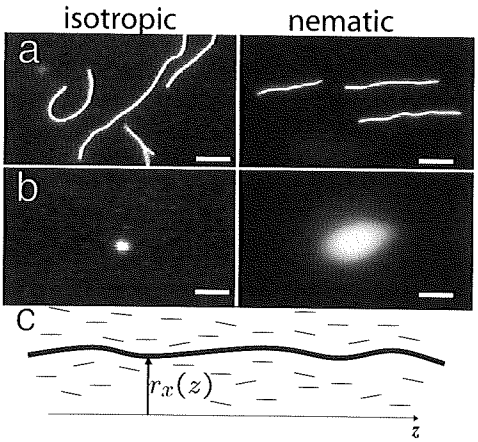


Fig. 22: *Images of fluorescently labelled F-actin (a) and DNA (b) in the isotropic (left) and nematic (right) background phase of fd virus. (c) Schematic of a biopolymer in the background nematic field. Photo by Dogic/Rowland institute at Harvard.*

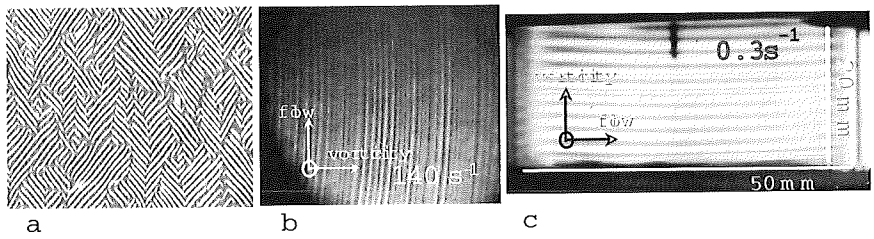


Fig. 23: *Structure formation by applying external fields as observed between crossed polarizers. (a) Micrograph of the nematic phase of TMV in a magnetic field, after switching the sample* 90^o. *Reprinted from Hurd et al. [77]. (b) Micrograph of the nematic phase of fd in a shear field, with a shear rate of* $140\ s^{-1}$. *Photo by Lettinga/IFF, Weiche Materie. (c) Image of shear cell containing a fd-dextran mixture in the bi-phasic region, showing the formation of shear bands. Reprinted from Dhont et al. [78].*

Using the relation between the twist elastic constant and the Chiral twist constant, K_t, the latter can also be determined, see subsection 3.6. When applying a magnetic field perpendicular to the director of a *nematic* field, then patterns will be induced, instead of destroyed, as is the case for the cholesteric phase (see Fig. 23a). These patterns have a typical field strength dependent wavelength and angle relative to the initial director, which depend not only on the twist elastic constant but also on the bend and splay elastic constants. The ratio of the constants can be determined from this experiment, as has been done for TMV. Again, since TMV is nearly a rod, results can be nicely compared with theory [77].
It is expected that aligning external fields enhance the I-N phase transition. The most suited way to verify this is by taking sample from an isotropic phase, which co-exists with a nematic phase. Since this sample is right at the boundary of the region where it is stable, a small increase in concentration *or* orientation is expected to lead to a phase transition. Tang and Fraden[55] did this experiment for *fd* and M13 at the I-N concentration by analyzing the kinetics of the development of the birefringence. They indeed found that the isotropic sample entered the bi-phasic region on applying a magnetic field. They also found a critical magnetic field strength, above which the fully nematic state was entered.
Another way to align rods is to expose them to a shear flow. The gradient in the flow velocity gives rise to a torque on the rods, such that they will align with the flow. Similar to the effect of the magnetic field, it also expected for the flow field that it enhances the I-N transition. Indeed Lenstra et al. found that the I-N spinodal concentration of a *fd* virus dispersion decreased with increasing shear rate [79]. A full phase non-equilibrium phase diagram including the binodal points has not been lined out so far.
Also structure formation in the nematic phase of *fd* under shear has been observed, similar to that observed when applying a magnetic field (see see Fig. 23b). All information on the elasticity constants of the nematic phase is also contained in the morphology and kinetics of these structures, but it is so far not known how. Even more puzzling is the formation of structure when applying shear at very low shear rates, to a mixture of *fd*-dextran in the bi-phasic region. This "shear banding" phenomenon includes all features we have discussed in this chapter: hard rods, mixtures, ordering, and external fields. Complex systems like this, where one has full control over all the parameters, set the challenge for future research.

References

[1] J. Perrin, *Atoms* (Constable, London, 1916).

[2] Stanley, Phytopathology **26**, 305 (1936).

[3] F. C. Bawden, N. W. Pirie, J. D. Bernal, and I. Fankuchen, Nature 1051 (1936).

[4] L. Onsager, Annals of the New York academy of science **51**, 62 (1949).

[5] P. N. Pusey and W. van Megen, Nature **320**, 340 (1986).

[6] A. Maeda and Y. Maeda, Phys. Rev. Lett. **90**, 018303 (2003).

[7] R. Piazza, S. Iacopini, M. Pierno, and E. Vignati, J. Phys.: Condens. Matter **14**, 75637580 (2002).

[8] M. P. B. van Bruggen, Langmuir **14**, 2245 (1998).

[9] J. Buitenhuis *et al.*, J. Colloid Int. Sci. **175**, 46 (1995).

[10] S. Fraden, in *Observation, Prediction, and Simulation of Phase Transitions in Complex Fluids*, Vol. 460 of *NATO-ASI - Series C*, m. baus and l. f. rull and j. p. ryckaert, ed. (Kluwer Academic Publishers, Dordrecht, 1995,), pp. 113–164.

[11] C. Graf *et al.*, J. Chem. Phys. **98**, 4921 (1993).

[12] A. W. C. Lau, K.-H. Lin, and A. G. Yodh, Phys. Rev. E **66**, 020401(R) (2002).

[13] C. E. Flynn, S.-W. Lee, B. R. Peelle, and A. M. Belcher, Acta Materialia **51**, 58675880 (2003).

[14] J. Hezfeld, Acc. Chem. Res. **29**, 31 (1996).

[15] M. D. Shtilerman, T. T. Ding, and J. P. T. Lansbury, Biochemistry **41**, 3855 (2002).

[16] S. Mangenot, A. Leforestier, D. Durand, and F. Livolant, J. Mol. Biol. **333**, 907916 (2003).

[17] G. Rivas, F. Ferrone, and J. Herzfeld, EMBO Reports **5**, 23 (2004).

[18] O. Medalia *et al.*, Science **298**, 1209 (2002).

[19] F. Livolant, Physics A **176**, 117 (1991).

[20] D. Frenkel, J. Chem. Phys. **91**, 4912 (1987).

[21] J. Hezfeld, A. E. Berger, and J. W. Wingate, Macromolecules **17**, 1718 (1984).

[22] H. N. W. Lekkerkerker, P. Coulon, R. van der Haegen, and R. Deblieck, J. Chem. Phys. **80**, 3427 (1984).

[23] R. F. K. Jr. and H. J. Raveche, Phys. Rev. A **17**, 2067 (1978).

[24] Z. Y. Chen, Macromolecules **26**, 3419 (1993).

[25] M. A. Cotter and D. C. Wacker, Phys. Rev. A **18**, 2669 (1978).

[26] M. A. Cotter, *Hard particle theory of nematics* (Academic Press, London, 1979).

[27] P. Bolhuis and D. Frenkel, J. Chem. Phys. **106**, 666 (1997).

[28] E. M. Kramer and J. Herzfeld, Phys. Rev. E **58**, 5934 (1998).

[29] Z. Dogic, Ph.D. thesis, Brandeis Universtity, 2000.

[30] A. Stroobants, H. N. W. Lekkerkerker, and T. Odijk, Macromolecules **19**, 2232 (1986).

[31] K. Zimmermann *et al.*, J. Bio. Chem. **261**, 1653 (1986).

[32] G. J. Vroege and H. N. W. Lekkerkerker, Rep. Prog. Phys. **55**, 1241 (1992).

[33] T. Sato and A. Teramoto, Adv. Pol. Science **126**, 85 (1996).

[34] A. R. Khokhlov and A. N. Semenov, Physica A **112**, 605 (1982).

[35] A. R. Khokhlov and A. N. Semenov, Physica A **108**, 546 (1981).

[36] R. Hentschke, Macromolecules **23**, 1192 (1990).

[37] T. Odijk, Macromolecules **19**, 2313 (1986).

[38] D. B. DuPré and S. Yang, J. Chem. Phys. **94**, 7466 (1991).

[39] T. J. Sluckin, Liquid Crystals **6**, 111 (1989).

[40] H. H. Wensink and G. J. Vroege, J. Chem. Phys. **119**, 6868 (2003).

[41] H. N. W. Lekkerkerker *et al.*, Europhys. Lett. **20**, 559 (1992).

[42] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals,*, 2nd ed. (Clarendon Press, Oxford, 1992).

[43] X. Wen and R. B. Meyer, Phys. Rev. Lett. **59**, 1325 (1987).

[44] B. Mulder, Phys. Rev. A **35**, 3095 (1987).

[45] J. M. Polson and D. Frenkel, Phys. Rev. E **56**, R6260 (1997).

[46] A. Stroobants, H. N. W. Lekkerkerker, and D. Frenkel, Phys. Rev. A. **36**, 2929 (1987).

[47] H. Graf and H. Löwen, PRE **59**, 1932 (1999).

[48] P. van der Schoot, J. Phys. II France **6**, 1557 (1996).

[49] A. V. Tkachenko, Phys. Rev. E **58**, 5997 (1998).

[50] M. A. Bates and D. Frenkel, J. Chem. Phys **109**, 6193 (1998).

[51] F. C. Brawden, N. W. Pirie, J. D. Bernal, and I. Fankuchen, Nature **138**, 1051 (1936).

[52] Z. Dogic and S. Fraden, Langmuir **16**, 7820 (2000).

[53] E. Grelet and S. Fraden, Phys. Rev. Lett. **90**, 198302 (2003).

[54] S. S. Zakharova, W. Jesse, C. Backendorf, and J. R. C. van der Maarel, Biophys. J. **83**, 11191129 (2002).

[55] J. Tang and S. Fraden, Phys. Rev. Lett. **71**, 3509 (1993).

[56] S. Fraden, G. Maret, D. L. D. Caspar, and R. B. Meyer, Phys. Rev. Lett. **63**, 2068 (1989).

[57] J. Tang and S. Fraden, Liquid Crystals **19**, 459 (1995).

[58] S.-D. Lee and R. B. Meyer, J. Chem. Phys. **84**, 3443 (1986).

[59] K. R. Purdy *et al.*, Phys. Rev. E. **67**, 031708 (2003).

[60] R. Oldenbourg, X. Wen, R. B. Meyer, and D. L. D. Caspar, Phys. Rev. Lett. **61**, 1851 (1988).

[61] C. Wetter, Biologie in unserer Zeit **15**, 81 (1985).

[62] Z. Dogic and S. Fraden, Phys. Rev. Lett. **78**, 2417 (1997).

[63] S. M. Yu *et al.*, Nature **389**, 167 (1997).

[64] Z. Dogic, Phys. Rev. Lett. **91**, 165701 (2003).

[65] T. E. Strzelecka, M. W. Davidson, and R. L. Rill, Nature **331**, 457 (1988).

[66] F. Livolant, A. M. Levelut, J. Doucet, and J. P. Benoit, Nature **339**, 724 (1989).

[67] F. Livolant, J. Mol. Biol. **218**, 165 (1991).

[68] F. Livolant and A. Leforestier, Prog. Polym. Sci. **21**, 1115 (1996).

[69] V. A. Bloomfield, Curr. Opin. Sruc. Biol. **6**, 334 (1996).

[70] K. Merchant and R. L. Rill, Biophys. J. **73**, 3154 (1997).

[71] K. Kassapidou, W. Jesse, J. A. P. P. van Dijk, and J. R. C. van der Maare, Biopolymers **46**, 31 (1998).

[72] R. Furukawa, R. Kundra, and M. Fechheimer, Biochemistry **32**, 12346 (1993).

[73] C. M. Coppin and P. C. Leavis, Biophys. J. **63**, 794 (1992).

[74] J. Viamontes and J. X. Tang, Phys. Rev. E **67**, 040701 (2003).

[75] A. Leforestier and F. Livolant, Biophys. J. **73**, 1771 (1997).

[76] M. Adams, Z. Dogic, S. L. Keller, and S. Fraden, Entropically driven microphase transitions in mixtures of colloidal rods and spheres, 1998.

[77] A. J. Hurd, S. Fraden, F. Lonberg, and R. B. Meyer, J. de Physique **46**, 905 (1985).

[78] J. K. G. Dhont *et al.*, Faraday Discuss. **123**, 157 (2003).

[79] T. A. J. Lenstra, Z. Dogic, and J. K. G. Dhont, J. Chem. Phys. **114**, 10151 (2001).

MEMBRANES

D

D 1 Biomembranes

B. Hoffmann
Institut für Schichten und Grenzflächen
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 The lipid bilayer 2

2.1 The lipid composition of biomembranes 2
2.2 Membrane proteins 4
2.3 Membrane structures and fluidity 5

3 The compartimentalization of cells 8

4 Principles of biomembrane flow 11

4.1 Vesicle transport between organells 12
4.2 Vesicle recognition 14

5 A short comment 15

1 Introduction

The boundary between a living cell and its surroundings is the incredibly thin (7 – 10 nm) plasma membrane. This vital partition, which controls the flow of materials into and out of a cell and which senses and controls the response for cells to hormones and other external signals, consists largely of phospholipids together with embedded proteins. Membranes serve many purposes. The most obvious is to give cells their individuality by separating them from their environment. Membranes are highly selective permeability barriers rather than impervious walls because they contain a tremendous number of specific channels and pumps. These transport systems regulate the molecular and ionic composition of the intracellular medium. Eukaryotic cells also contain internal membranes that form the boundaries of organelles such as endoplasmic reticulum, Golgi apparatus, mitochondria, and lysosomes. Functional specialization in the course of evolution has been closely linked to the formation of such compartments.
Membranes control the flow of information within a cell and between cells and their environment. They contain specific receptors necessary for cell adhesion and cell movement and approximately 20% of all enzymatic reactions within a eukaryotic cell happen in an membrane attached state. Additionally, membranes with their incorporated proteins are able to create membrane potentials essential for many metabolic processes as the formation of ATP in mitochondria but also for the transport of signals in neurons. Since biomembranes are an essential vital factor for living cells, the whole knowledge about them is too complex to summarize them in one chapter. Here, the focus is given on the general basics of biomembranes. More specific details about membrane receptors, membrane channels, and membrane vesicles are given in other chapters.

2 The lipid bilayer

2.1 The lipid composition of biomembranes

Biomembranes are composed of lipids, proteins and partially carbohydrates. The most abundant lipid components in most biomembranes are **phospholipids**, which are, as all lipids, amphipathic molecules, meaning that they contain both a hydrophobic and a hydrophilic end. They are derived from either glycerol, a three-carbon alcohol, or sphingosine, a more complex alcohol. The ones derived from glycerol are **phosphoglycerides** and consist additionally of two fatty acid chains (the hydrophobic part) and an esterified phosphate (Fig. 1A). The fatty acid chains in biolipids usually contain an even number of carbon atoms, typically between 14 and 24. The 16- and 18- carbon fatty acids are the most common ones. Fatty acids may be saturated or unsaturated, meaning a double bond between two carbon atoms. In phosphoglycerides, the hydroxyl groups at C-1 and C-2 of glycerol are esterified to the carboxyl groups of two fatty acid chains. The C-3 hydroxyl group of the glycerol backbone is esterified in the most simple case just to phosphate. The phosphate can then additionlly esterify to choline, alcohols such as ethanolamine or sugar derivatives such as inositol forming the tremendous number of different phospholipids present in biomembranes (Fig. 1B).
Sphingomyelin, a phospholipid that lacks a glycerol backbone, is found mainly in plasma membranes. Instead of a glycerol backbone, it contains sphingosine, an amino alcohol derived from serine with a long unsaturated hydrocarbon chain.

In sphingomyelin, the terminal hydroxyl group of sphingosine is esterified to phosphocholine forming the hydrophilic head (Fig. 1b).

Cholesterol and its derivatives constitute another important class of membrane lipids, the steroids. The basic structure of steroids is the four-ring hydrocarbon shown in Fig. 1C. Although cholesterol is almost entirely hydrocarbon in composition, it is amphipathic because its single hydroxyl group can interact with water. Cholesterol is especially abundant in the plasma membrane of mammalian cells but is absent from most prokaryotic cells. As much as 30 to 50% of the lipids in plant plasma membranes consists of cholesterol and certain steroids unique to plants.

Carbohydrates are found in many membranes, covalently bound either to membrane proteins or to lipids as constituents of **glycolipids** (Fig. 1D). Glycolipids show the most extreme asymmetry in membranes with an exclusive noncytosolic localization. Bound carbohydrates increase the hydrophilic character of lipids and proteins and help to stabilize the conformations of many membrane proteins.

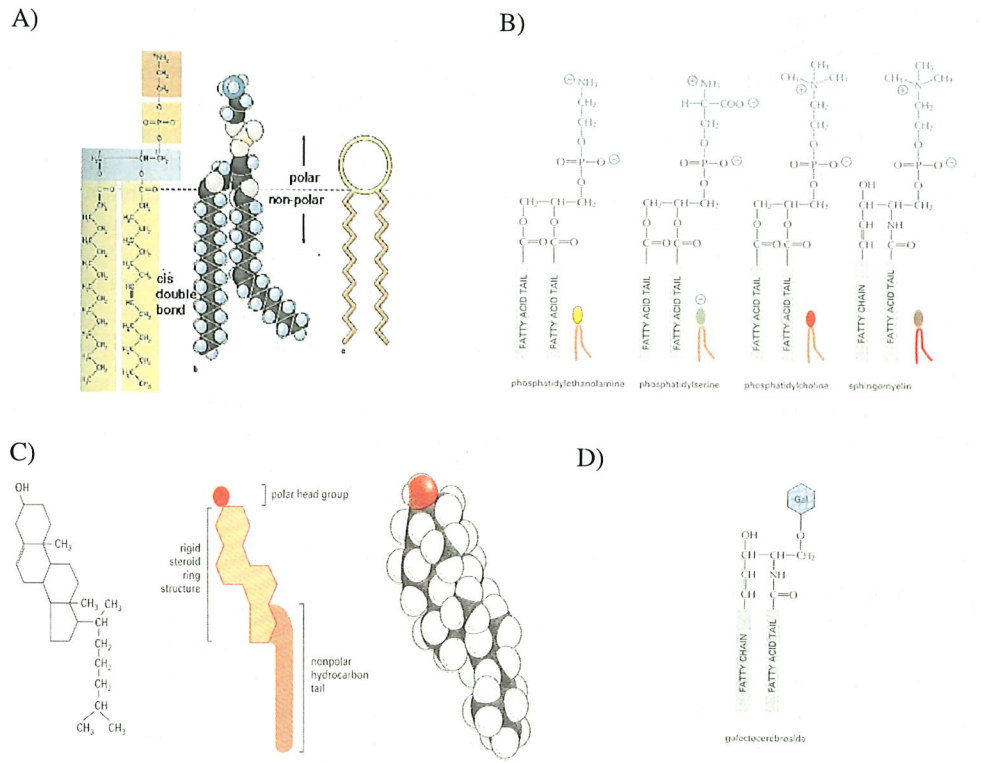


Fig. 1: *Biomembrane lipids. A) Molecular parts of phosphatidylethanolamine as example for phospholipids. The separate parts are from the top: ethanolamine (red), phosphate (orange), glycerol backbone (blue) and the two fatty acyl chains (green). (B) Most common phospholipids in mammalian membranes. First three lipids are derived from glycerol and differ in their hydrophilic headgroup esterified to phosphate. The fourth lipid, phingomyelin, is derived from the amino acid serin. C) Structure of cholesterol given in different schemes. D) Glycolipid. For one of the most simple glycolipids, galactocerebroside, one single galactose molecule is esterified to the C3 atom of the glycerol backbone.*

2.2 Membrane proteins

Although the basic structure of biological membranes is provided by the lipid bilayer, membrane proteins perform most of the specific functions of membranes. Membrane proteins can be classified into two broad categories – integral (intrinsic) and peripheral (extrinsic) – proteins based on the nature of the membrane-protein interactions. Most biomembranes contain both types of membrane proteins.

Integral membrane proteins have one or more segments that are embedded in the phospholipid bilayer. Typically, integral proteins contain residues with hydrophobic side chains that interact with fatty acyl groups of the membrane phospholipids, thus anchoring the protein to the membrane. Most integral proteins span the entire phospholipid bilayer. These transmembrane proteins contain one or more membrane-spanning domains as well as domains, from four to several hundred residues long, extending into the aqueous medium on each side of the bilayer. In all the transmembrane proteins examined to date, the membrane-spanning domains are α-helices or multiple β-strands. Typically the α-helix is composed entirely of hydrophobic and partially uncharged amino acids with a length of around 3.75 nm (>21 amino acids), just sufficient to span the hydrocarbon core of the phospholipid bilayer (Fig. 2A). The hydrophobic side chains form van der Waals interactions with the fatty acyl chains and shield the polar carbonyl (C=O) and imino (NH) groups of the peptide bond. The hydrophobic helices are normally prevented from slipping across the membrane by a flanking set of positively charged amino acids (lysine and arginine) that are thought to interact with negatively charged phospholipid head groups. Many proteins do not only contain just one membrane spanning α-helix but two or more. Membrane-spanning α-helices of the same protein typically stabilize themselves by specific interactions between the amino acid side chains at their interfaces. The most prominent group of integral proteins containing multiple transmembrane α-helices is the one with seven membrane spanning domains (Fig. 2B). Many channel proteins, the bacteriorhodopsin or the opsins (eye proteins that absorb light) and odorous receptors belong to this group.

In contrast to proteins with hydrophobic amino acids embedded into the membrane, some integral proteins are anchored to one of the membrane monolayers by covalently bound fatty acids. In these proteins, the bound fatty acid is embedded in the membrane, but the polypeptide chain does not enter the phospholipid bilayer.

Peripheral membrane proteins generally do not interact with the hydrophobic core of the phospholipid bilayer. Instead they are bound to the membrane indirectly by interactions with integral membrane proteins or directly by interactions with lipid polar head groups. For the plasma membrane most peripheral proteins are localized to its cytosolic face but a few proteins including certain proteins of the extracellular matrix, are localized to the outer surface of the plasma membrane. An important group of peripheral membrane proteins are water-soluble enzymes that associate with the polar head groups of membrane phospholipids. One well-understood group of such enzymes are the phospholipases, which hydrolyze various bonds in the head groups of phospholipids.

These enzymes have an important role in degradation of lipids during aging processes or in modifications of the lipid composition of certain biomembranes as well as in forming short-living intracellular mediators by specific phospholipid cleavage. The mechanism of action of phospholipase A2 illustrates how such water-soluble enzymes can reversibly interact with membranes and catalyze reactions at the interface of an aqueous solution and lipid surface. When this enzyme is in aqueous solution, its Ca^{2+}-containing active site is buried in a channel lined with hydrophobic amino acids. Binding of the enzyme to a phospholipid bilayer induces

a small conformational change that fixes the protein to the phospholipid heads and opens the hydrophobic cleft. As a phospholipid molecule moves from the bilayer into that cleft, the enzyme-bound Ca^{2+} binds to the phosphate in the head group and positions the ester bond to be cleaved next to the catalytic site.

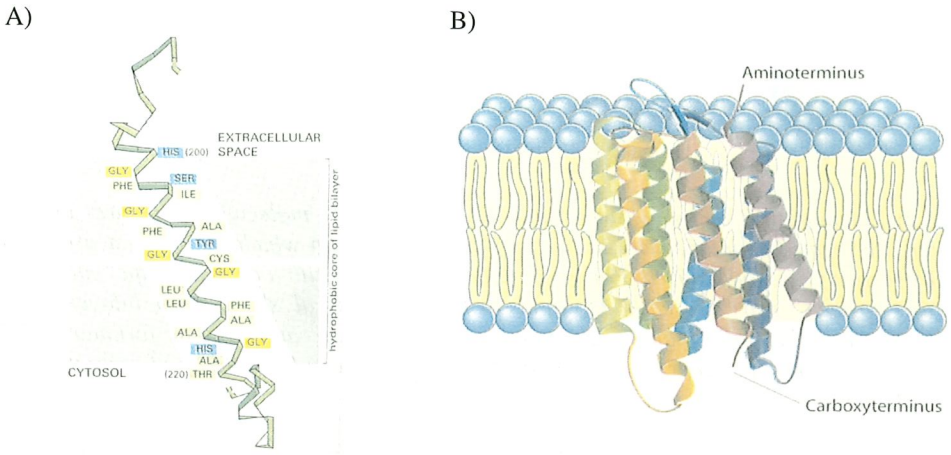


Fig. 2: *Transmembrane domains. A) At least 21 consecutive amino acids with hydrophobic (green and yellow) or uncharged (blue) character are necessary to span the complete hydrophobic core of a lipid bilayer. B) An important group of integral proteins is composed of seven transmembrane α-helices (each helice is given in a different colour). α-helices stabilize each other by side chain interactions forming a stable tertiary structure. Amino acids at the N- and C-terminus are not shown.*

2.3 Membrane structures and fluidity

As described above, lipids are amphipathic molecules. In contact with water they form spontaneously a lipid monolayer with their polar heads arranged to the water surface. Within an aqueous medium one way of fulfilling its amphipathic character is to form micelles. In such globular lipid structure the polar head groups are surrounded by water and the hydrocarbon tails are sequestered inside, facing one another. In contrast, the favored structure for most phospholipids and glycolipids in aqueous media is a bimolecular sheet, the so called lipid bilayer, rather than a micelle (Fig. 3). The reason is that their two fatty acyl chains are too bulky to fit into the interior of a micelle. In contrast, salts of fatty acids, as constituents of soap, which contain only one chain, readily form micelles. The rapid and spontaneous formation of bilayers instead of micelles by phospholipids is of critical biological importance, since only bilayers can have macroscopic dimensions. Bilayers can form either lateral structures or round up to form vesicles.

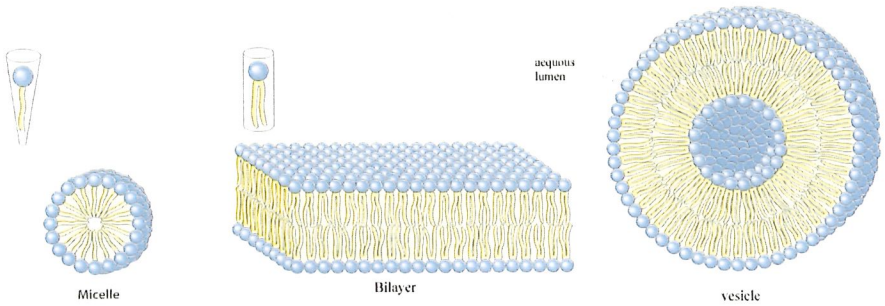


Fig. 3: *Membrane structures. Since lipids are amphipathic molecules, structures of higher order are formed spontaneously in the presence of water in which the contact area of the hydrocarbon tails is minimized. For amphipathic molecules with a conical shape, micelles are getting formed easily. In contrast, lipids with a cylindrical shape form bilayers. These bilayers can either have planar macroscopic dimensions or can round up forming vesicles. Artificial vesicles can either be unilamellar (as shown) or can be build up of several bilayers (multilamellar).*

Biological membranes are not rigid structures. On the contrary, lipids and many membrane proteins are constantly in lateral motion. Biomembranes are a two-dimensional fluid (Fig. 3A). The rapid lateral movement of membrane proteins has been visualized by means of fluorescence microscopy and fluorescence photobleaching recovery techiques. The experiments showed that some mammalian proteins have a diffusion coefficient of 0.4 $\mu m^2/s^{-1}$ at 37°C (rhodopsin) while phospholipids diffuse an average distance of 2 µm/second. In contrast, the movement of a lipid molecule in z-direction, meaning the transition from one site of a bilayer to the other site is a very rare phenomenon. In fact, this so called flip-flop mechanism takes place for lipid molecules in a time range of hours to days. A flip-flop of integral membrane proteins has never been observed until now. The temperature is very important for membrane behavior since the fatty acyl chains in bilayer membranes can exist in an ordered, rigid state or in a relatively disordered, fluid state. In the ordered state, all the C-C bonds have a trans conformation (no steric hindrance between side groups and more stable), whereas in the disordered state, some are in the gauche conformation (steric hindrance). The phase transition from the rigid to the fluid state occurs rather abruptly as the temperature is raised above Tm, the melting temperature. This transition temperature is specific for each lipid and depends on the length of the fatty acyl chains and on their degree of unsaturation. The rigid state is favored by the presence of saturated fatty acyl residues because their straight hydrocarbon chains interact very favorably with each other, and so Tm is increased. The length of the fatty acyl chain also affects the transition temperature. Long hydrocarbon chains interact more strongly than do short ones. Specifically, each additional $-CH_2-$ group makes a favourable contribution of about – 0.5 kcal/mol to the free energy of interaction of two adjacent hydrocarbon chains. Both, the modification of the ratio of saturated to unsaturated fatty acid chains as well as of the chain length is most efficiently used by prokaryotes able to life at a large temperature range. At all these temperatures the organisms have to make sure for proper function that the plasma membrane remains fluid. In eukaryotes, cholesterol additionally is a key regulator of membrane fluidity. Cholesterol inserts with its bulky steroid hydrocarbon tail described before into bilayers with its long axis perpendicular to the plane of the membrane. The hydroxyl group of cholesterol hydrogen-

bonds to a carbonyl oxygen atom of a phospholipid head group, whereas the hydrocarbon tail of cholesterol is located in the nonpolar core of the bilayer. Cholesterol prevents the crystallization of fatty acyl chains by fitting between them. In fact, high concentrations of cholesterol abolish phase transitions of bilayers. An opposite effect of cholesterol is to sterically block large motions of fatty acyl chains, which makes membranes less fluid. Thus, cholesterol moderates the fluidity of membranes.

The many noncovalent interactions between lipid molecules within a bilayer as well as its fluidity at temperatures suitable for growth and metabolism leads to three significant biological consequences:

1. lipid bilayers have an inherent tendency to be extensive
2. lipid bilayers will tend to close on themselves so that there are no edges with exposed hydrocarbon chains. Such tendency results in the formation of a compartment
3. lipid bilayers are self-sealing because a hole in a bilayer is energetically unfavourable.

The lipid bilayer as core structure as well as the tremendous number of proteins embedded in the bilayer or attached to it, were brought together by Singer and Nicolson in 1972 in the fluid mosaic model (Fig. 4B).

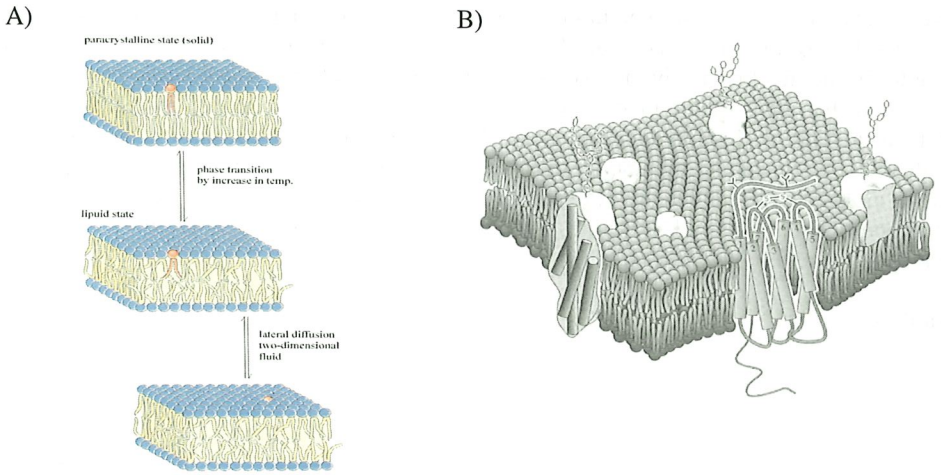


Fig. 4: *A) At temperatures below Tm, membranes are in a paracristalline state. Lateral movement is abolished and all fatty acyl C-C bonds are in trans conformation leading to ordered fatty acyl chains. By increasing the temperature above Tm a rapid phase transition occurs, allowing lateral movement of biomembrane components. A higher percentage of gauche conformation for C-C bonds additionally causes a reduced order in the fatty acyl chains. B) The knowledge about lipid bilayers and membrane proteins was brought together by Singer and Nicolson in creating the fluid mosaic model. Proteins are shown either as solid blocks or as transmembrane α-helices. Structures on top of some proteins symbolize glycosylations.*

3 The compartimentalization of cells

Although all biomembranes have the same basic bilayer structure and certain common functions, each type of cellular membrane has certain distinctive activities determined largely by the unique set of proteins associated with that membrane and their specific lipid composition. Approximately 1500 different lipids have been identified in the myelin of the central nervous system of humans. About 30 of these are present in substantial amounts. The distribution of the different lipids varies strongly between membranes from different sources (Table 1) making generalization difficult. However, phospholipids are apparently always present and, except in chloroplasts, make up from 40% to over 90% of the total lipid. Additionally the lipid compositions of the two monolayers of the bilayer in many membranes are strikingly different (Table 2). In the human red blood cell membrane, for example, almost all of the lipid molecules that have choline in their head group are in the outer monolayer, whereas almost all of the phospholipid molecules that contain a terminal primary amino group are in the inner monolayer. Lipid asymmetry is functionally important since many cytosolic proteins bind to specific lipid head groups found in one or the other monolayer of the lipid bilayer. Therefore the lipid asymmetry causes simultaneously a protein asymmetry and these proteins give each type of membrane in the cell its characteristic functional properties. Accordingly, not only the types but also the amounts of proteins in a membrane are highly variable. In the myelin membrane, which serves mainly as electrical insulation for nerve cell axons, less than 25% of the membrane dry mass is protein. By contrast, in the membranes involved in ATP production approximately 75% is protein. A typical plasma membrane has about 50% protein of the membrane mass. Because lipid molecules are small compared to protein molecules, there are always many more lipid molecules than protein molecules in membranes. About 50 lipid molecules for each protein molecules are found in membranes (corresponds to 50% protein by mass).

Table 1

	Percentage of total dry weight of membrane[a]					
Compound	Myelin (bovine)	Retinal rod	Plasma membrane (human erythrocyte)	Mitochondrial membranes	E. coli[b] (inner and outer membranes)	Chloroplasts[c]
Protein	22	39	60	76	75	18
Total lipid	78	41	40	24	25	82[d]
Phosphatidylcholine	7.5	13	6.9	8.8		
Phosphatidylethanolamine	11.7	6.5	6.5	8.4	18	
Phosphatidylserine	7.1	2.5	3.1			
Phosphatidylinositol	0.6	0.4	0.3	0.75		
Phosphatidylglycerol					4	
Cardiolipin[e]		0.4		4.3	3	
Sphingomyelin	6.4	0.5	6.5			
Glycolipid	22.0	9.5	Trace	Trace		23
Cholesterol	17.0	2.0	9.2	0.24		
Total phospholipid	33	27	24	22.5	25	47
Phospholipid as a percentage of total lipid	42	66	60	94	>90%	9

Table 2

Phospholipids in plasma membrane	Percentage of phospholipids in plasma membrane	Orientation in the membrane: inner monolayer 100 — 0 — outer monolayer 100
Phosphatidylethanolamine	30	
Phosphatidylcholine	27	
Sphingomyelin	23	
Phosphatidylserine	15	
Phosphatidylinositol	5	
Phosphatidylinositol-4-phosphate		
Phosphatidylinositol-4,5-biphosphate		
Phosphatidyl acid		

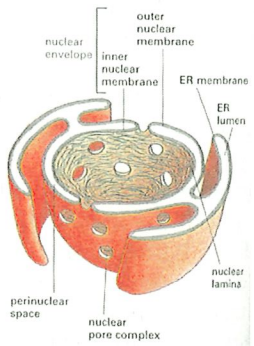
Table 1: *Percentage of protein and lipids in different membranes. Amounts for the most prominent biomembrane lipids are given separately* .

Table 2: *Lipid composition of the inner and outer monolayer of an erythrocyte plasma membrane.*

To give a short overview, some of the most important organells within a eukaryotic cell, their functions and membranous structures are listed below.

Nucleus: The nuclear envelope encloses the DNA and defines the nucleus. It consists of two concentric membranes that are penetrated by nuclear pore complexes. Although the inner and outer nuclear membranes are continuous, they maintain distinct protein compositions. The inner nuclear membrane contains specific proteins that act as binding sites for chromatin and for the protein meshwork of the nuclear lamina that provides structural support for this membrane. The inner membrane is surrounded by the outer nuclear membrane, which is also continuous with the membrane of the endoplasmic reticulum (ER). Transport processes in and out of the nucleus occur through nuclear pore complexes with an inner diameter of 9 nm estimated from diffusion measurements. Such small diameter makes sure that nuclear inflow and outflow are tightly regulated and performed by active transport allowing the movement of particles of up to 26 nm in diameter through the pore. It also means that particle transport via vesicles does not take place for the nucleus which does not means that there is no biomembrane flow for the nucleus since nuclear membranes and ER membrane are a continuous construct (Fig. 5A).

A)



B)

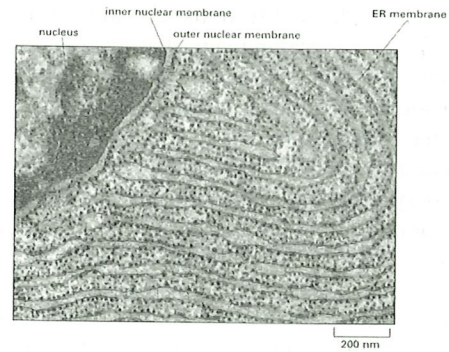


Fig. 5: *Membrane interconnections between nucleus and ER. The cell nucleus is surrounded by an inner and an outer membrane. These membranes are a continuous construct with the nucleus surrounding ER (A and B). Black dots covering the ER define ribosomes on the electron micrograph (B).*

Endoplasmic reticulum (ER): The endoplasmic reticulum is part of every eukayotic cell. For an average cell half of the total membrane dry mass is part of the ER. The ER membrane is forming a fine interconnected meshwork of tubes and flat sacks. The inner volume of the ER takes generally more than 10% of the total cell volume. The ER is the central place for protein biosynthesis. The synthesis of all transmembrane proteins and most lipids not only for the ER itself but also for almost all other organells, the Golgi apparatus, lysosomes, endosomes, secretory vesicles, peroxisomes, mitochondria and the plasma membrane takes place at the ER membrane. The protein synthesis starts on ribosomes which are also bound to the ER membrane (rough ER, Fig. 5B). After translation transmembrane proteins are getting incorporated into the ER membrane independent of their final destination. Soluble proteins often pass the ER membrane and remain for certain times in the inner ER lumen. Since many of the translated proteins and formed lipids are not a functional part of the ER but of other

organells, they have to be transported. Transport occurs almost exclusively via membrane vesicles. Lipids and transmembrane proteins are incorporated into the vesicle bilayer whereas soluble proteins are present in the inner vesicle lumen. Formation of ER vesicles takes place on ER membranes not coated with ribosomes (smooth ER). Formation and guidance of vesicles to the specific target membranes is described below.

Golgi apparatus: The typical way of an ER vesicle goes to the Golgi apparatus. The Golgi apparatus consists of a collection of flattened, membrane-enclosed cisternae, somewhat resembling a stack of pancakes. Each of these Golgi stacks usually consists of four to six cisternae. Most eukaryotic cells have hundreds of individual Golgi stacks. Usually, the Golgi apparatus is located close to the ER and the nucleus. Each Golgi stack has two distinct sites: the cis Golgi network (entry site) and the trans Golgi network (exit site). Proteins and lipids enter the cis Golgi network in transport vesicles arriving from the ER and exit from the trans Golgi network. During their passage through the Golgi apparatus, transported proteins and lipids undergo an ordered series of covalent modifications, especially glycosylations. For the transport of molecules from stack to stack vesicle formation takes place again. Even the exit of lipid and protein molecules from the Golgi occurs in transport vesicles. These vesicles serve then either as secretory vesicles or are further guided mostly to lysosomes or the plasma membrane. Transport of lipid and protein molecules occurs not only in one direction from the ER to the Golgi and from the cis- to the trans Golgi network. Vesicle transport takes place in both directions making sure that glycosylated molecules reenter the ER and that ER marker proteins essential for vesicle formation can be reused again.

Lysosomes: Lysosomes are membrane-enclosed sacks, filled with hydrolytic enzymes. These enzymes are used for a well controlled digestion of macromolecules. They contain at least 40 different types of hydrolysing enzymes able to digest everything like proteins, lipids, DNA and polysaccharids. All of these enzymes have their pH optimum at about 5.0. This safety mechanism prevents damage for the rest of the cell with a pH of 7.2 in case of a lysosome leakage. The low pH within lysosomes is made possible by a lysosome-specific H^+ pump located in the lysosome membrane and hydrolysing ATP as energy source for H^+ transport. In addition, no other organell membrane has as many integral transport proteins incorporated as the one of lysosomes. These proteins transport the endproducts of the hydrolytic digestion, as nucleotides, monosaccharids and amino acids back into the cytosol. An additional characteristic for lysosomal membranes is the intense glycosylation of the inner monolayer which helps to protect them from the lysosomal proteases in the lumen.

Endosomes: Endosomes describe the only vesicles coming from the plasma membrane and entering the cell lumen. They are formed by endocytosis. Formed vesicles can be small (pinocytosis; around 100 nm) or big (phagocytosis; more than 250 nm in size) Specific receptors in the plasma membrane as well as proteins coating the endosomes on their outside concentrate molecules of interest and take them up by bud and finally vesicle formation (described below). These vesicles are called early endosomes. The early endosome acts as the main sorting station in the endocytic pathway. Most of the macromolecules in the endosome lumen will end up in lysosomes while most lipids and proteins involved in macromolecule binding and vesicle formation are retransported to the plasma membrane.

Mitochondria and Chloroplasts: Although functionally both organells have nothing in common they are described here in the same paragraph. Mitochondria form the ATP essential for live. Without mitochondria, animals would be dependent on anaerobic glycolysis for all of their ATP with an output of 15 times less than produced in mitochondria. In contrast, chloroplasts are specific for specialized plant cells.

They capture energy from sunlight and use it to fix carbon by forming carbohydrates. All reactions occurring in a chloroplast can be grouped into two main categories: the photosynthetic electron-transfer (light reaction) with the formation of ATP, $NADPH_2$ and O_2 and the carbon-fixation reaction in which CO_2 is converted to carbohydrate (dark reaction). The reason why they are described here in parallel is that for both organells their energy interconversion by chemiosmotic mechanisms is much the same. Also their organization follows the same principles. They both have a doubled bilayered membrane. The outer bilayer is highly permeable. The inner membrane is much less permeable. Here, most of the membrane transport proteins are embedded. The lipid composition is strongly different between the inner and outer membranes of both organells. Several lipids of the inner membrane are known to be present only here. In between, both organells have a narrow intermembrane space. Both have their own genome and genetic system with a special set of ribosomes. In addition, both organells show an immense amount of membranes of different shape and function in the lumen. While the inner membrane of mitochondria forms infoldings, known as cristae (Fig. 6B), in chloroplasts a third type of membrane, the thylakoid membrane, forms a set of flattened disclike sacs, known as grana (Fig. 6A). In both organells almost all reactions take place on these inner membranes and even their different reactions are separated in place by having many membranous bound proteins localized only in specific areas. Although both organells have their own genome, they depend on biomembrane flow coming from other organells. Since both organells are too complex just to summarize the basics, I must refer for further details to excellent text books in cell biology and biochemistry.

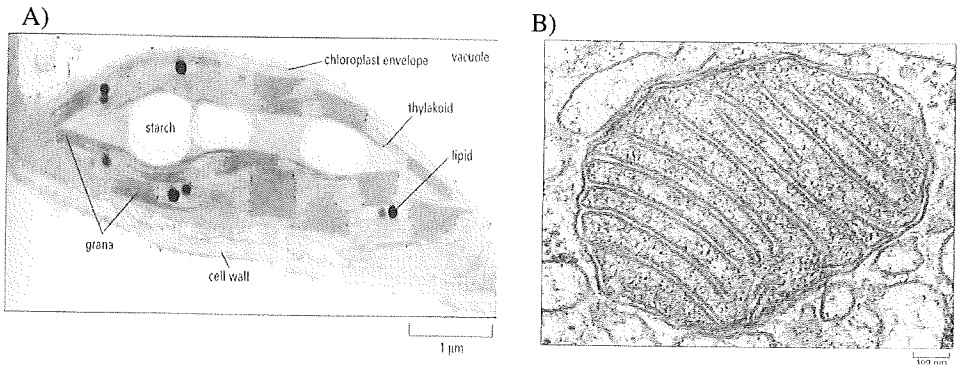


Fig. 6: *Organization of chloroplasts and mitochondria. Electron micrographs of a chloroplast (A) and a mitochondria (B) show the extensive formation of membranous structures in the lumen of these organells.*

4 Principles of biomembrane flow

Transport processes are essential for endocytic and biosynthetic-secretory pathways. They mediate a continual exchange of lipid and protein components between the chemically and functionally distinct compartments of a cell. Since every compartment is enclosed by its own membrane, biomembane flow is only possible by budding forming bilayered vesicles.

These vesicles are then moved to a target compartment and fuse to it. Biomembrane flow occurs in both directions, to the plasma membrane (**secretion**) as well as from the plasma membrane (**endocytosis**) (Fig. 7).

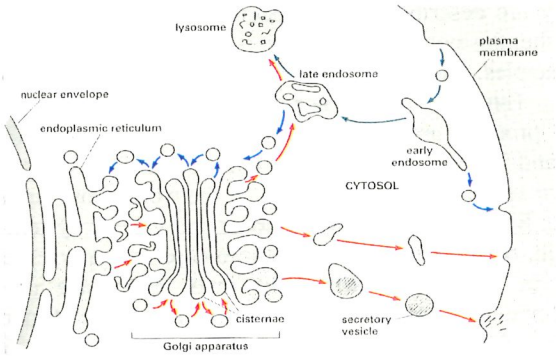


Fig. 7: *Biosynthetic-secretory and endocytic pathways. Each compartment of a cell is surrounded by a lipid bilayer. Transport of lipids and proteins from one compartment to an other is largely performed by vesicle formation causing in membrane flow from the ER to the plasma membrane (secretion) and from the plasma membrane to inner compartments (endocytosis).*

4.1 Vesicle transport between organells

Most transport vesicles form from specialized, coated regions of membranes which are part of the ER, Golgi apparatus, endosomes, lysosomes and plasma membrane. They bud off as coated vesicles that have a distinctive cage of proteins covering their cytosolic surface. Before the vesicle fuses with a target membrane, the coat is discarded, as is required to allow the two cytosolic membrane surfaces to interact directly and fuse. The coat is thought to perform two principal functions. First, it concentrates specific membrane proteins in a specialized membrane patch that then gives rise to the vesicle membrane. It thus helps to select the appropriate molecules for transport. Second, the assembly of the coat proteins into curved, basketlike lattices deforms the membrane patch and thereby molds the forming vesicles.
There are at least three types of coated vesicles, which differ in their coat proteins: clathrin-coated, COPI-coated and COPII-coated vesicles. Each type is used for different transport steps in the cell. Highly simplified one can say that Clathrin-coated vesicles mediate transport from the Golgi apparatus and from the plasma membrane, whereas COPI- and COPII-coated vesicles most commonly mediate transport from the ER and the Golgi cisternae (Fig. 8A). Since vesicles are essential for biomembrane flow within a cell, it is important to understand the basic character of vesicle coating. As example clathrin vesicles are taken. The major protein component of clathrin-coated vesicles is clathrin itself. Each clathrin subunit consists of three large and three small polypeptide chains that together form a three-legged structure called a triskelion (Fig. 8B). Clathrin triskelions assemble into a basketlike framework of hexagons and pentagons to form coated pits on the cytosolic surface of membranes.

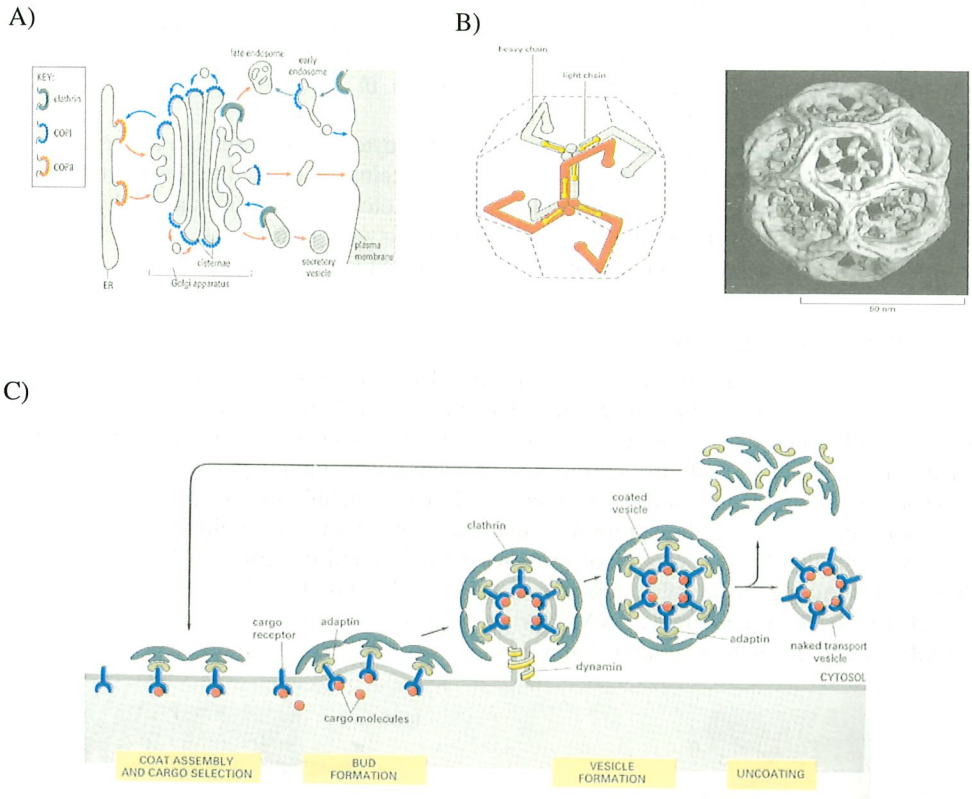


Fig. 8: *(A) Clathrin and COP proteins coat distinct pools of vesicles. Pathways of clathrin-, COPI- and COPII-coated vesicles are shown in a simplified scheme to clarify how different coating proteins mediate movement of a distinct class of vesicles. (B) A schematic drawing of the arrangement of light and heavy clathrin chains forming triskelions (left). On the right a cryo electron image of a clathrin coat is shown. The triskelions can be arranged as pentagons or as hexagons. Together they form the coat enclosing a membrane vesicle. The architecture resembles the one of a soccer ball. (C) Coat assembly and disassembly. Binding of coating proteins to specific membrane receptors cause bud and finally vesicle formation under involvement of additional proteins. The coat of vesicles remains on the vesicle only for a short period of time since it would block fusion with the target membrane. Coat proteins can be reused for a new round of vesicle assembly.*

A second major coat protein in clathrin-coated vesicles is a multisubunit complex called adaptin. It is required both to bind the clathrin coat to the membrane and to trap various transmembrane proteins, including transmembrane receptors that capture soluble cargo molecules inside the vesicle (Fig. 8C). Several types of adaptins are known each specific for a different set of cargo receptors. Clathrin-coated vesicles budding from different membranes use different adaptins and thus package different receptors and cargo molecules. The formation of a clathrin-coated bud is driven by forces generated by the successive assembly of adaptins and the clathrin coat on the surface of the membrane. As a clathrin-coated vesicle

grows, dynamin assembles as a ring around the neck of each bud. Dynamin is a GTPase able to bring the two noncytosolic leaflets of the membrane into close proximity causing fusion and sealing off the formed vesicle. The membrane bending is directly caused by local distortion of the bilayer structure and by a change in lipid composition, performed by dynamin recruited proteins.
An additional component in vesicular transport is the cytoskeleton. Since a eukaryotic cell is too large for vesicular diffusion processes, several proteins, associated to the outer membrane of vesicles, are able to bind to subunits of molecular motors such as kinesins or dynein. These molecular motors are able to bind to the microtubule network to transport in an active manner bound vesicles from one organell to another along the cytoskeletal track.

4.2 Vesicle recognition

To ensure that membrane traffic proceeds in an orderly way, transport vesicles must be highly selective in recognizing the correct target membrane with which to fuse. Because of the diversity of membrane systems, a vesicle is likely to encounter many potential target membranes before it finds the correct one. Specificity in targeting is ensured because all transport vesicles display surface markers while target membranes display complementary receptors that recognize the appropriate markers. This crucial recognition step seems to be mainly controlled by two classes of proteins: SNAREs and targeting GTPases called Rabs. Rabs work together with other proteins to regulate the initial docking and tethering of the vesicle to the target membrane. SNARE proteins seem to have a central role both in providing specificity and in catalyzing the fusion of vesicles with the target membrane.

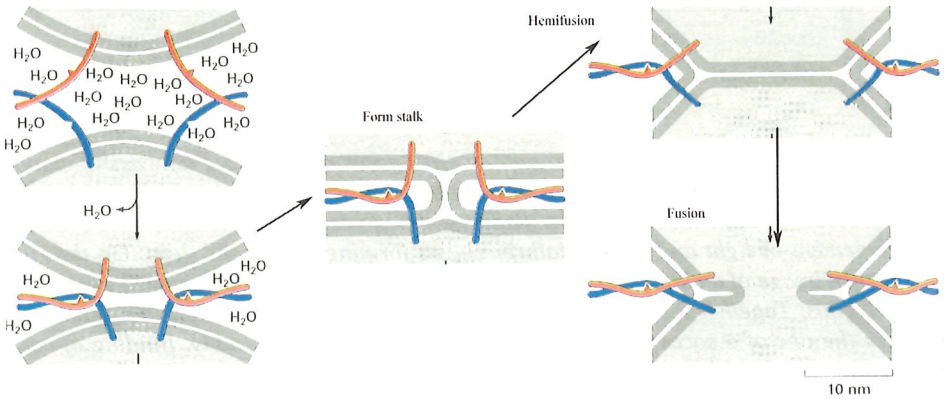


Fig. 9: *Model of how SNARE proteins may mediate membrane fusion. v-SNARE proteins from the vesicle (red) interact with t-SNARE proteins from the target membrane (blue). By pairing the SNARE proteins both membranes are getting brought in close contact. After forming first a new bilayer (stalk formation and hemifusion) this bilayer opens and completes the vesicle fusion. All Fig.s shown in this article have been taken from the text books "Molecular Biology of the Cell"[1], "Lehninger Biochemie"[2], "Molecular Cell Biology"[3] and "Biochemistry"[4].*

There are many different SNARE proteins in animal cells, each associated with a particular membrane-enclosed organelle involved in the biosynthetic-secretory or endocytic pathway. These transmembrane proteins exist as complementary sets – vesicle membrane SNAREs, called v-SNAREs and target membrane SNAREs, called t-SNAREs. v-SNAREs and t-SNAREs have characteristic helical domains. When a v-SNARE interacts with a t-SNARE, the helical domains of one wrap around the helical domains of the other to form stable trans-SNARE complexes, which lock the two membranes together (Figure 9). After fusion, the v- and t-SNARE proteins remain associated till a catalytic dissociation takes place. After retransport to the original membrane the t-SNARE protein can accept a new vesicle while the v-SNARE is getting recycled for participation in a new round of vesicular transport.

5 A short comment

Biomembranes are a vital factor for live. This fact is true for prokaryotes as well as for eukaryotes. The only difference is that membranes become even more important the higher and more complex a cell is evolved. An average eukaryotic cell contains intracellular membranes that enclose nearly half the cell's total volume in separate intracellular organell compartments. About 3000 to 4000 different lipids are part of various biomembranes. Their composition as well as the composition of bound proteins varies drastically from cell type to cell type, from organell to organell and even within organells itself. How biomembranes work, how they are able to keep their specific composition despite the immense biomembrane flow and many other fascinating questions are still far from complete understanding. An increasing number of scientists is analyzing these questions *in vivo* as well as *in vitro* using specific, pure lipids and single proteins. It will be a major challenge to use these highly simplified model systems to answer the right questions for clarifying step by step the function of presumably the most heterogeneous cell component of all.

References

[1]B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, Molecular Biology of the Cell, Fourth Edition (Garland Science, New York, 2002)
[2]D. Nelson and M. Cox, Lehninger, Principles of Biochemistry, Third Edition (Springer, 1999)
[3]H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell, Molecular Cell Biology, Fourth Edition (Freeman and Company, New York, 2000)
[4]D.E. Metzler, Biochemistry, Second Edition, Volume I (Harcourt/Academic Press, 2001)

D 2 Mechanics and Statistical Mechanics of Membranes

G. Gompper
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Shapes and Fluctuations of Fluid Membranes 4
2.1 Short Introduction to Differential Geometry of Surfaces 5
2.2 Curvature Energy 7
2.3 Gauss-Bonnet Theorem and Euler's Formula 9
2.4 Almost Planar Fluid Membranes 9

3 Crystalline Membranes 14
3.1 Elasticity Theory of Solids 14
3.2 Continuum Theory of Flat Membranes 17
3.3 Continuum Theory of Buckled Membranes 22

4 Monte Carlo Simulations of Dynamically Triangulated Surfaces 29

5 Biological Systems 31
5.1 Virus Shapes 31
5.2 Clathrin-Coated Pits 33

6 Summary and Conclusions 36

A Appendices 39
A.1 Curvature Tensor of Almost Planar Membranes 39
A.2 Stress Tensor in Polar Coordinates 39
A.3 Stress Function of a Planar Disclination 40
A.4 Bending Energy of Buckled Disclinations 40

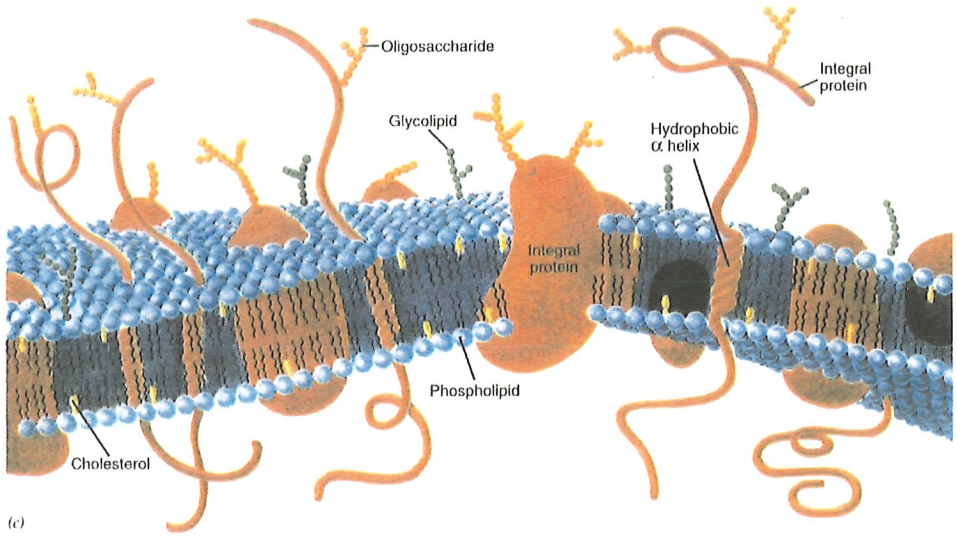


Fig. 1: *The basic structural element of the plasma membrane of eukaryotic cells is a lipid bilayer. This bilayer contains embedded and attached membrane proteins. From Ref. [1].*

1 Introduction

The most important examples of membranes are bilayers composed of amphiphilic molecules such as surfactants or lipids. Amphiphilic molecules self-assemble in aqueous solution into a large variety of phases. The driving force for this aggregation process is the hydrophobic effect, which strongly disfavors contact between water and non-polar solutes such as hydrocarbons. Amphiphilic molecules therefore aggregate in such a way that the hydrocarbon tails are shielded from contact with water by a layer of polar head groups. This leads to the formation of spherical and cylindrical micelles, and of bilayer membranes. Micelles typically form when the size of the head groups is larger than that of the tails, while bilayer membranes are stable when head and tail are of roughly equal size. A more detailed discussion of amphiphilic self-assembly can be found in Chap. A.3.

Lipid bilayers are the basic building blocks of the plasma membrane as well as the membranes of all cell organelles in living organisms. Cell membranes are typically composed of several species of lipids, have different lipid composition in the inner and outer leaflet of the bilayer, and contain many embedded, adsorbed, or attached proteins — as shown schematically in Fig. 1, and explained in more detail in Chaps. A.4 and D.1.

Membranes can be classified according to their internal properties. Membranes can be

- Polymerized – with fixed connectivity:
  neighboring atoms or molecules are chemically connected.
  Examples: graphite sheets, fullerene molecules, polymer network attached to plasma membrane of red blood cells.

- Crystalline – with a finite density of thermally generated lattice defects:

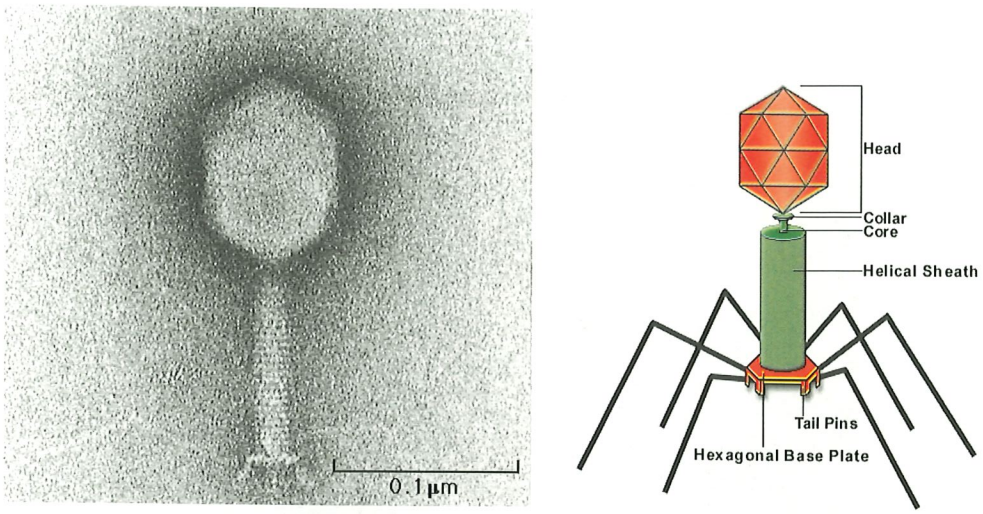


Fig. 2: *Bacteriophage T4, a virus which infects E. coli.* **Left:** *Electron microscopy picture. Note the faceted shape of the viral capsid. From Ref. [2].* **Right:** *Schematic view. The DNA is contained in the head of the virus, and is injected into the bacterium through the cylindrical tail. From Ref. [3].*

interactions between neighboring molecules induce long-range orientational order, but defects can appear spontaneously according to their Boltzmann weight.
Examples: gel phase of lipid bilayers, membrane proteins crystallized in a fluid lipid bilayer, the protein shell of viruses (the "capsid", compare Fig. 2).

- Fluid, with fixed topology:
  fluidity implies free diffusion of molecules within the membrane; fluid membranes cannot support any shear stress.
  Examples: lipid-bilayer vesicles (closed bilayers of spherical topology), plasma membrane of cells (like the red blood cells shown in Fig. 3).

- Fluid, with fluctuating topology:
  different parts of the membranes can fuse, or connected parts of a membrane can separate.
  Examples: microemulsion and sponge phases (mesoscopically structured phases in mixtures of water, oil and amphiphilic molecules), endo- and exocytosis of transport vesicles.

In general, membranes are two-dimensional sheets of atoms or molecules which are different from the three-dimensional medium in which they are embedded. Because their thickness is on the order of nanometers, but their lateral extension can be as large as microns, membranes are very flexible and can be easily deformed. From a theoretical point of view, this separation of length scales implies that membranes can, in many cases, be modeled as idealized surfaces of vanishing thickness. In this approach, the shape and fluctuation spectrum of a membrane is determined by an elastic energy; the molecular properties, architecture, and interactions of the membrane constituents enter only through the functional form of the elastic energy and the values of the elastic moduli.

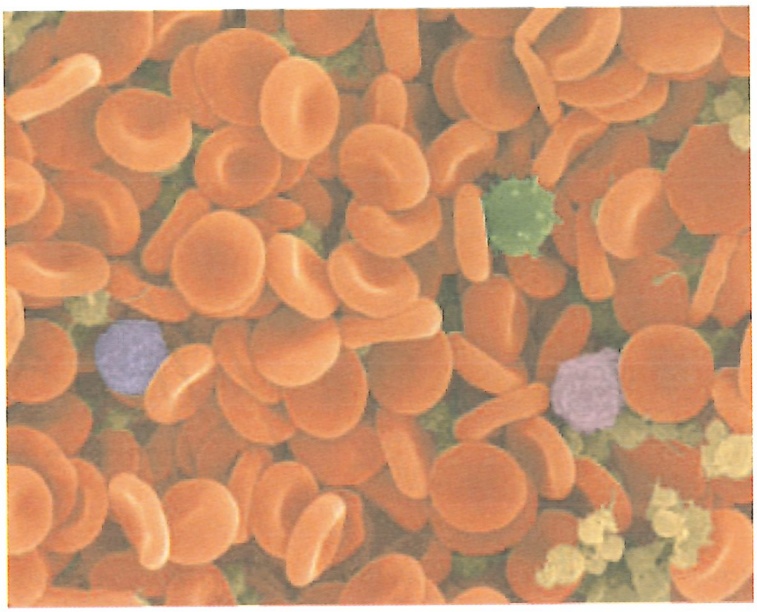


Fig. 3: *Red blood cells have a characteristic biconcave shape, which is called a discocyte. This smooth shape minimizes the bending energy at a given enclosed volume. From Ref. [4].*

It is interesting to note that the shapes of fluid membranes, like the red blood cells shown in Fig. 3, are generally very smooth, while the shape of crystalline membranes, like the viral capsid shown in Fig. 2 is more faceted. Another interesting observation is that not all viruses look the same. Small viruses, like the polyoma virus, are more rounded, while larger viruses, like the adenovirus, are more faceted, see Fig. 4. It is the aims of this lecture to understand and explain the physical origin of these and other mechanical and thermal properties of membranes.

The statistical mechanics of membranes and related subjects have been discussed in recent years in a number of books and reviews by Nelson et al. (1989) [8], Gompper and Schick (1994) [9], Gelbart et al. (1994) [10], Safran (1994) [11], Lipowsky and Sackmann (1995) [12], David et al. (1996) [13], Seifert (1997) [14], Gompper and Kroll (1997) [15], Bowick and Travesset (2001) [16], Dhont et al. (2002) [17], Boal (2002) [18], and Gompper and Kroll (2004) [19].

2 Shapes and Fluctuations of Fluid Membranes

Fluid membranes are characterized by free diffusion of the molecules within the membrane. Therefore, a fluid membrane cannot sustain any shear stress. The area of the membrane is fixed by the number of constituent molecules. However, it costs energy to deform a membrane from its preferred curvature state. Therefore, the shapes and fluctuations of fluid membranes are controlled by their curvature elasticity.

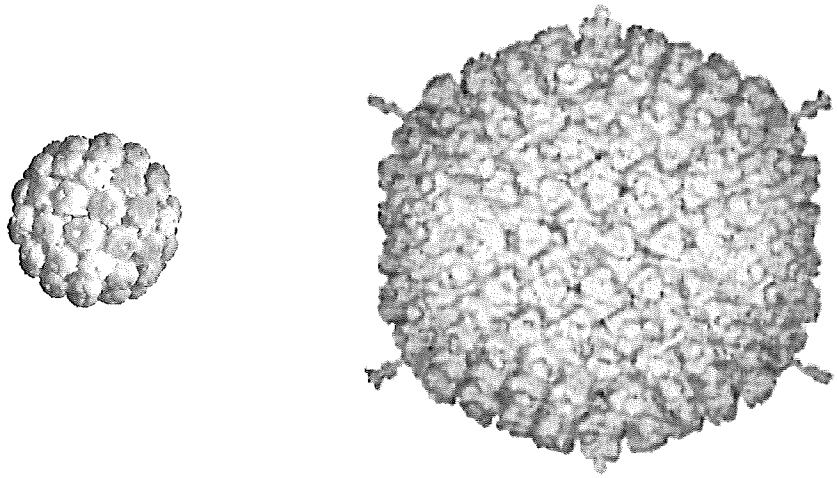


Fig. 4: *The polyoma virus (left) with a diameter of about 440 Å is approximately spherical, while the larger adenovirus (right) with a diameter of about 1450 Å is more faceted. From Refs. [5], [6] and [7].*

2.1 Short Introduction to Differential Geometry of Surfaces

A surface is defined mathematically by a vector field $\mathbf{R}(\underline{\sigma}) \in \mathbb{R}^3$, which maps a two-dimensional coordinate system $\underline{\sigma}$ onto a smooth two-dimensional object in the three-dimensional embedding space, see Fig. 5. A basis for the tangent plane to the surface at some point P is given by the two *tangent vectors*

$$\mathbf{t}_i = \partial_i \mathbf{R}(\underline{\sigma}) \equiv \frac{\partial \mathbf{R}}{\partial \sigma^i}, \quad i = 1, 2 \tag{1}$$

With the tangent vectors, the *metric tensor*

$$g_{ij}(\underline{\sigma}) = \partial_i \mathbf{R} \cdot \partial_j \mathbf{R} = \mathbf{t}_i \cdot \mathbf{t}_j \tag{2}$$

can be defined. The metric tensor determines the infinitesimal Euclidean distance ds between two points on the surface, which is given by [20]

$$\begin{aligned} ds^2 &\equiv [\mathbf{R}(\underline{\sigma} + d\underline{\sigma}) - \mathbf{R}(\underline{\sigma})]^2 = \partial_i \mathbf{R}(\underline{\sigma}) \cdot \partial_j \mathbf{R}(\underline{\sigma})\, d\sigma^i d\sigma^j \\ &= g_{ij}(\underline{\sigma})\, d\sigma^i d\sigma^j \end{aligned} \tag{3}$$

Here, the Einstein convention of summation over repeatedly appearing indices is used. Similarly, the infinitesimal area element is found to be

$$\begin{aligned} dS &\equiv |\mathbf{t}_1 d\sigma^1 \times \mathbf{t}_2 d\sigma^2| = \sqrt{(\mathbf{t}_1)^2 (\mathbf{t}_2)^2 - (\mathbf{t}_1 \cdot \mathbf{t}_2)^2}\, d\sigma^1 d\sigma^2 \\ &= \sqrt{\det g(\underline{\sigma})}\, d\sigma^1 d\sigma^2 \end{aligned} \tag{4}$$

The unit *normal vector* $\mathbf{n}(\underline{\sigma})$ at some point P on the surface is perpendicular to the tangent plane, so that it can be written as

$$\mathbf{n} = \frac{\mathbf{t}_1 \times \mathbf{t}_2}{|\mathbf{t}_1 \times \mathbf{t}_2|} \tag{5}$$

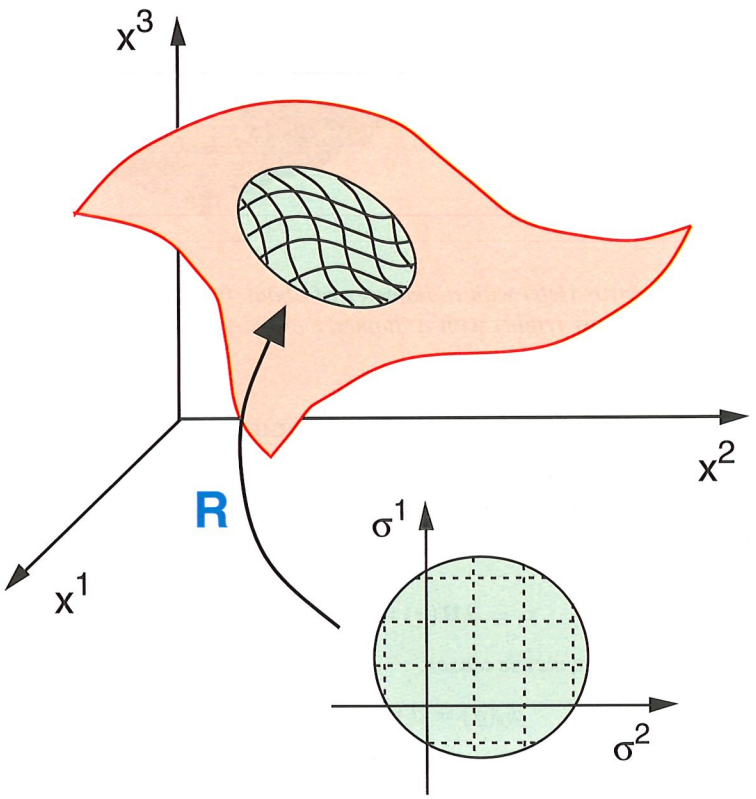


Fig. 5: *A surface is described mathematically by a map* $\mathbf{R}(\underline{\sigma})$ *of a two-dimensional coordinate system* $\underline{\sigma}$ *onto a smooth, two-dimensional object in the three-dimensional embedding space.*

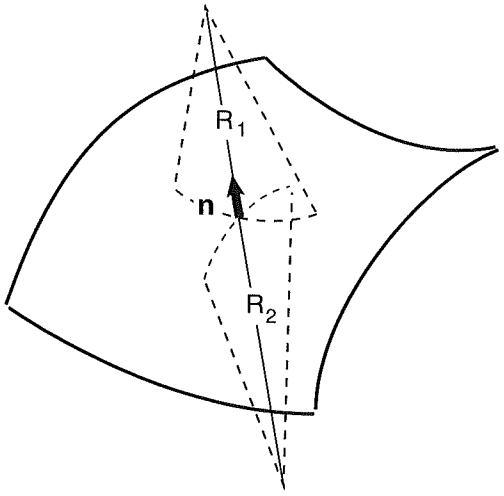


Fig. 6: *The two principal radii of curvature, R_1 and R_2, of a point on the surface.*

The change of the normal vector for an infinitesimal displacement along the surface finally defines the *curvature tensor* K_{ij}, via [20]

$$\partial_i \mathbf{n} = -K_{ij}\mathbf{t}^j \qquad [= -K_{ij}g^{jl}\mathbf{t}_l = -K_i^j \mathbf{t}_j] \tag{6}$$

where $g^{jl} = (g^{-1})_{jl}$.
The curvature tensor K_i^j can be diagonalized. The eigenvalues are the two *principal curvatures*, $c_1 = 1/R_1$ and $c_2 = 1/R_2$. Since the curvature tensor is symmetric, the two eigenvectors are perpendicular to each other. These results can be interpreted as follows [20]. Consider a plane, which intersects the surface at the point P and which is perpendicular to the tangent plane at P. A circle can be drawn in this plane, which optimally approximates the shape of the curve of intersection of the plane with the surface in the vicinity of P. As this plane is rotated around the normal vector, the radius of curvature of the "optimal" circle varies. The minimum and maximum of these radii are the principal radii of curvature, R_1 and R_2; their corresponding planes are perpendicular to each other. The principal radii of curvature are illustrated in Fig. 6. The local geometry of a surface can now be classified by the signs of c_1 and c_2. The corresponding geometries are shown in Fig. 7.

2.2 Curvature Energy

The energy functional of a *fluid membrane* must satisfy the following conditions

- Motion invariance:
  The energy must be invariant under translations and rotations of the whole membrane.
- Reparametrization invariance:
  Since the molecules can diffuse freely within the membrane, no preferred coordinate system can exist — which would correspond physically to a labelling of distinct constitutive elements of the membrane. Therefore, the energy must be invariant under a change of the coordinate system.

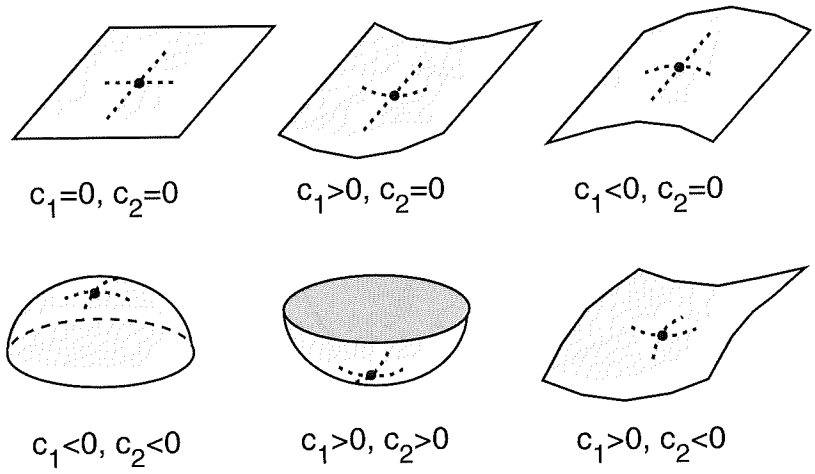


Fig. 7: *The six possible cases of local curvatures of a surface: Plane, concave cylinder, convex cylinder, convex spherical cap, concave spherical cap, and saddle.*

These conditions imply that in an expansion in powers of the inverse radii of curvature, the curvature energy to leading order is given by [21, 22]

$$\mathcal{H}_{curv} = \int d^2\sigma \sqrt{\det g}\,[\gamma + 2\kappa(H - C_0)^2 + \bar{\kappa}K + ...] \tag{7}$$

where

$$2H = Tr(K_i^j) \equiv g^{ij}K_{ij} = K_i^i = c_1 + c_2 \tag{8}$$

$$K = \det(K_i^j) \equiv \left[K_1^1 K_2^2 - K_1^2 K_2^1\right] = c_1 c_2 \tag{9}$$

are the mean and Gaussian curvatures, respectively. The parameters γ, κ, $\bar{\kappa}$ and C_0 are the elastic constants of the membrane, which depend on the structure and interactions of the constitutive molecules.
It can easily be seen that all the terms in Eq. (7) satisfy the conditions of motion invariance and of reparametrization invariance since

- only derivatives of the position vector $\mathbf{R}$ appear in the form of scalars
- it is well known that the trace and the determinant of a tensor are invariant under similarity transformations.

The different terms in the elastic energy (7) have simple physical interpretations. The first term is proportional to the area of the membrane. Its elastic constant γ is therefore just the surface tension. The second term measures the deviation of the mean curvature H from a preferred or "spontaneous" curvature C_0; the elastic constant κ is the bending rigidity. The last term is a bit more difficult to interpret. It will be shown in Sec. 2.3 below that the surface integral over the Gaussian curvature K is related to the topology of the membrane; the elastic constant $\bar{\kappa}$ is called the saddle-splay modulus.

For amphiphilic membranes, the surface tension γ is very small or vanishes completely. The reason is that the membranes are self-assembling, with the aggregation of the amphiphiles driven by the hydrophobic effect, which tries to minimize the contact of a polar solvent with the hydrocarbon chains of the amphiphilic molecules, and similarly of a apolar solvent with the polar headgroups. For a more detailed discussion see Chap. A.3. Therefore, it costs free energy for an amphiphilic mono- or bilayer to increase its area per molecule. On the other hand, when the amphiphilic molecules are tightly packed, squeezing them together even more also costs energy. The result of these competing interactions is an optimal area per molecule, which corresponds to a state of zero surface tension.

The curvature energy provides the basis for the calculation of the shape of fluid vesicles, i.e. closed membranes of usually spherical topology. The dependence of vesicle shapes on the spontaneous curvature and on the enclosed volume (for fixed membrane area) is discussed in detail in Chap. D.3.

2.3 Gauss-Bonnet Theorem and Euler's Formula

The Gauss-Bonnet theorem states that for a *closed* membrane, the surface integral over the Gaussian curvature K is a topological invariant,

$$\int dS\ K = 2\pi\chi_E \tag{10}$$

which for fixed topology has always the same value, independent of its shape. The topological constant χ_E is the Euler characteristic. It can be obtained very easily from the expression

$$\chi_E = 2 \times (\text{No. of objects} - \text{No. of handles}) \tag{11}$$

For a surface, which is composed of polygons, there is also the famous Euler's formula, which states that

$$\chi_E = N_{plaquettes} - N_{edges} + N_{vertices} \tag{12}$$

where $N_{plaquettes}$, N_{edges} and $N_{vertices}$ denote the numbers of polygons, edges and vertices, respectively. For a closed, triangulated surface of spherical topology, with $\chi_E = 2$ and $N_{edges} = \frac{3}{2}N_{triangles}$, Eq. (12) implies $N_{vertices} = 2 + \frac{1}{2}N_{triangles}$.
Some examples are shown in Fig. 8. For a sphere, Eqs. (10) and (11) give $\chi_E = 2$, while for a topologically equivalent cube, Eq. (12) gives — consistently — $\chi_E = 6 - 12 + 8 = 2$. For a torus, which has one handle, Eq. (11) gives $\chi_E = 0$, while for a torus formed by 8 connected cubes, Eq. (12) gives — again consistently — $\chi_E = 32 - 64 + 32 = 0$. For a plane, Eq. (10) gives $\chi_E = 0$. Two planes with a connecting passage, as shown in the lower right of Fig. 8, according to Eq. (11) then have $\chi_E = -2$.
With the Gauss-Bonnet theorem and Eq. (11), the interpretation of the last term in the curvature energy (7) is now straightforward. A comparison of surfaces with different topologies indicates that $-2\kappa < \bar{\kappa} < 0$ is required for membranes with curvature radii on mesoscopic scales to be stable [9].

2.4 Almost Planar Fluid Membranes

Despite of its deceptive simplicity, the general form (7) of the curvature energy cannot be used directly for analytical calculations. Therefore, we consider now a membrane, which deviates

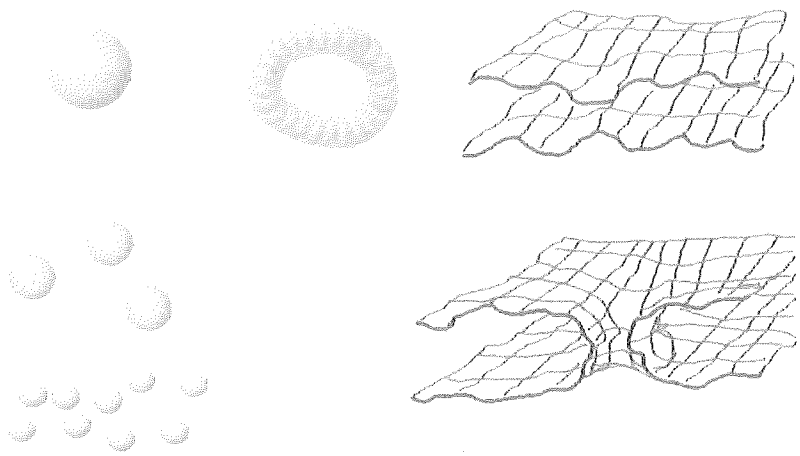


Fig. 8: *The Euler characteristic χ_E of a membrane ensemble is twice the number of objects minus twice the number of handles.* **Left:** *For the collections of spheres, $\chi_E = 2$, 6 and 18, from top to bottom.* **Middle:** *A torus has $\chi_E = 0$.* **Right:** *Two parallel membranes have $\chi_E = 0$ as long as they are disconnected, but $\chi_E = -2$ if they are connected by a passage.*

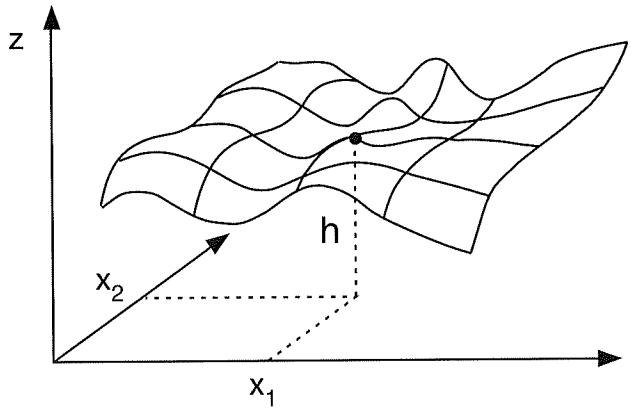


Fig. 9: *In the Monge representation, the surface is described by a single-valued function $z = h(x_1, x_2)$.*

not too strongly from a planar reference state. As long as there are no overhangs, the membrane configuration can then be described by a single-valued function $z = h(x_1, x_2)$, see Fig. 9. This is called the *Monge representation*. The position vector $\mathbf{R}(\underline{\sigma})$ of Sec. 2.1 then takes the form

$$\mathbf{R}(\underline{\sigma}) = (\sigma^1, \sigma^2, h(\sigma^1, \sigma^2)) \tag{13}$$

The calculation of the tangent vectors, and of the metric and curvature tensors is now straightforward. The tangent vectors are

$$\mathbf{t}_1(\underline{\sigma}) = \partial_1 \mathbf{R} = (1, 0, \partial_1 h(\sigma^1, \sigma^2)) \tag{14}$$

$$\mathbf{t}_2(\underline{\sigma}) = \partial_2 \mathbf{R} = (0, 1, \partial_2 h(\sigma^1, \sigma^2)) \tag{15}$$

which gives the normal vector

$$\mathbf{n} = \frac{\mathbf{t}_1 \times \mathbf{t}_2}{|\mathbf{t}_1 \times \mathbf{t}_2|} = \frac{(-\partial_1 h, -\partial_2 h, 1)}{\sqrt{1 + (\nabla h)^2}} \tag{16}$$

The expressions for the tangent vectors show that the mean curvature can be obtained very easily from the relation

$$\partial_1 n_x + \partial_2 n_y = K_1^1 + K_2^2 = 2H \tag{17}$$

Therefore, the mean curvature is found to be

$$2H = \nabla \cdot \frac{\nabla h}{\sqrt{1 + (\nabla h)^2}} \tag{18}$$

A more detailed derivation is given in Appendix A.1, where also the result

$$K = \frac{\partial_1^2 h \partial_2^2 h - (\partial_1 \partial_2 h)^2}{[1 + (\nabla h)^2]^2} \tag{19}$$

for the Gaussian curvature is obtained.

Thermal Membrane Fluctuations — With these results, an explicit form of the curvature energy (7) is now available. It is important to notice that in the derivation of Eqs. (18) and (19) no approximations have been made. Thus, these expressions are still exact in the Monge representation. In order to proceed with the calculation of membrane *fluctuations*, we now assume that the deviations from the planar reference state are small enough that an expansion in powers of $|\nabla h|$ is justified. To leading order in this expansion, the curvature energy for $\gamma = 0$ then becomes

$$E_b \simeq \frac{\kappa}{2} \int d^2x \, \left(\nabla^2 h(\mathbf{x})\right)^2 \tag{20}$$

Note that the Gaussian curvature term does not appear in this equation. This can be seen in general from the Gauss-Bonnet theorem and $\chi_E = 0$ for planar membranes, or explicitly in this case from Eq. (19), which for $|\nabla h| \ll 1$ can be written as a perfect derivative

$$2K \simeq -\epsilon_{im}\epsilon_{jn}\partial_m \partial_n [(\partial_i h)(\partial_j h)] \tag{21}$$

where ϵ_{ij} is the totally anti-symmetric matrix

$$\underline{\epsilon} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{22}$$

With the help of Stokes integral theorem, the surface integral over K can then be written as a boundary integral, which can be neglected in the thermodynamic limit.
In Fourier space, with

$$h(\mathbf{x}) = \int \frac{d^2q}{(2\pi)^2} h(\mathbf{q}) e^{i\mathbf{q}\cdot\mathbf{x}} \tag{23}$$

the expression (20) simplifies to

$$\begin{aligned} &\int d^2x \, \left(\nabla^2 h(\mathbf{x})\right)^2 = \\ &= \int \frac{d^2q}{(2\pi)^2} \int \frac{d^2q'}{(2\pi)^2} (-q^2)(-q'^2) h(\mathbf{q}) h(\mathbf{q}') \underbrace{\int d^2x \, e^{i(\mathbf{q}+\mathbf{q}')\cdot\mathbf{x}}}_{(2\pi)^2\delta(\mathbf{q}+\mathbf{q}')} \\ &= \int \frac{d^2q}{(2\pi)^2} \, q^4 |h(\mathbf{q})|^2 \end{aligned} \tag{24}$$

where

$$h(-\mathbf{q}) = h^*(\mathbf{q}) \tag{25}$$

since $h(\mathbf{r}) \in \mathbb{R}$. Therefore we finally obtain

$$E_b = \frac{\kappa}{2} \int \frac{d^2q}{(2\pi)^2} \, q^4 |h(\mathbf{q})|^2 \tag{26}$$

This is an important result, since it shows that the fluctuation modes, which are called *undulation modes* for membranes, *decouple* in Fourier space. The calculation of the fluctuation spectra then becomes very simple, since the *equipartition theorem* — which states that the energy of each independent mode is $k_BT/2$ on average — then implies

$$\langle h(\mathbf{q}) h(\mathbf{q}') \rangle = \frac{k_BT}{\kappa q^4} \delta(\mathbf{q} + \mathbf{q}') \tag{27}$$

This behavior has indeed been verified by studying fluctuation spectra of quasi-spherical [23, 24] and non-spherical vesicles [25].

Flickering and Steric Repulsion — When a membrane is brought into the vicinity of a hard wall, it collides more and more frequently with the wall the smaller the average membrane-wall separation d. Since less space is available for membrane undulations, the fluctuation entropy decreases and the free energy increases. This leads to an effective repulsive force, which is called *steric membrane interaction*. It has first been shown by Helfrich (1978) [26] that the suppression of undulation modes implies an effective interaction potential V_{steric} (per unit area) of the form

$$V_{steric} \sim \frac{(k_BT)^2}{\kappa} \frac{1}{d^2} \tag{28}$$

In biological systems, the steric interaction is believed to be responsible, for example, for preventing red blood cells from sticking to the walls of the blood vessels. The steric repulsion induced by the thermal membrane fluctuations — which are called *flickering* in the case of red blood cells — keeps the cells apart from each other and from walls, because it dominates over the attractive van der Waals interaction.

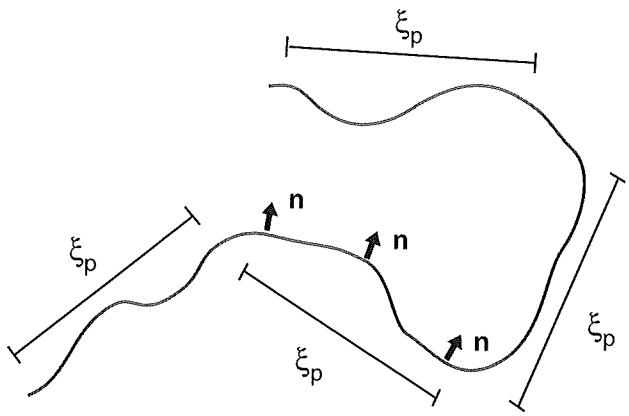


Fig. 10: *On length scales short compared to the persistence length ξ_p, normal vectors point roughly into the same direction, while on larger scales, normal vectors are uncorrelated.*

Persistence Length — We now want to calculate the fluctuations of the angle θ of the normal vector $\mathbf{n}$ with the z-axis,

$$n_z \equiv \mathbf{n} \cdot \mathbf{e}_z = \cos\theta \simeq 1 - \frac{1}{2}\theta^2 \tag{29}$$

for a membrane of *linear size* L with periodic boundary conditions. In the Monge representation, we have found (see Eq. (16))

$$n_z = 1/\sqrt{1 + (\nabla h)^2} \simeq 1 - \frac{1}{2}(\nabla h)^2 \tag{30}$$

so that with the equipartition theorem (27)

$$\langle \theta^2 \rangle = \langle (\nabla h)^2 \rangle = \int \frac{d^2q}{(2\pi)^2}\, q^2\, \frac{k_B T}{\kappa q^4} = \frac{k_B T}{2\pi\kappa} \ln(L/a) \tag{31}$$

where a is a molecular length. The fluctuations of the angle θ thus increase *logarithmically* with increasing membrane size L. Clearly, this is not possible without violating the condition $|\nabla h| \ll 1$, which was used in the derivation of the expression (20) of the curvature energy. The length where this happens, the *persistence length* ξ_p, can be defined — somewhat arbitrarily — by $\langle \theta^2 \rangle = 1$ (de Gennes & Taupin 1982) [27]. The inversion of Eq. (31) gives immediately

$$\xi_p = a \exp(2\pi\kappa/k_B T) \tag{32}$$

with molecular length a. The interpretation of the persistence length is explained in Fig. 10. For a membrane patch, which is much smaller than the persistence length, the fluctuations of the normal vector around the average orientation are small — the membrane is almost planar on this scale. On the other hand, membrane patches, which are much larger than the persistence length, have a wildly fluctuating normal vector, which means that no preferred orientation exists. This is the situation sketched in Fig. 10.
It is instructive to consider two examples:

- Lipids: $a = 1$ nm, $\kappa = 10 k_B T \rightarrow \xi_p = 10^{24}$ km. The persistence length is *astronomical*, certainly much larger than the size of a typical cell.
- Surfactants: $a = 1$ nm, $\kappa = 0.5 k_B T \rightarrow \xi_p = 25$ nm. The persistence length is *mesoscopic*.

3 Crystalline Membranes

3.1 Elasticity Theory of Solids

The theory of elasticity has been developed in the middle of the last century to explain the deformation of three-dimensional solids — such as bars, plates or shells — under an external load [28]. The same theory can also describe the deformation of crystalline solids due to internal defects such as interstitials, voids, and grain boundaries. Elasticity theory is therefore sometimes considered as "old-fashioned". I want to demonstrate in this section that this is not the case, and that in contrast elasticity theory forms the basis of the theoretical description of many new, complex soft materials, which have been discovered and investigated in recent years.
A particularly interesting case is melting and freezing in two dimensions. Kosterlitz, Thouless, Halperin, Nelson and Young (KTHNY) [29, 30, 31] have discovered in the seventies of the last century that melting of two-dimensional crystals is driven by a proliferation of two types of topological defects. These defects are called "dislocations" and "disclinations", and are shown in Figs. 11 and 12, respectively. A *dislocation* is the end-point of an additional line of lattice sites, which is inserted into an otherwise perfect two-dimensional crystal lattice. In a generic lattice of six-fold coordinated vertices, a *disclination* is a lattice site with five-fold or seven-fold coordination. It can be generated by removing from the lattice or inserting into the lattice a sector with an opening angle of 60^o. A close inspection of Fig. 11 shows that a dislocation consists of a pair of five- and seven-fold coordinated vertices, and can therefore be considered as a bound pair of disclinations (see also Fig. 14 below). The main purpose the this section is to calculate the elastic energy of these two types of topological defects.

The Strain Tensor — In any D-dimensional crystalline solid, a preferred coordinate system exists, which is tied to the relaxed, unperturbed crystalline lattice with long-range positional order. The positions of the molecules in this lattice are given by a set of lattice vectors $\mathbf{r}_n$. Deformations of the lattice are then represented by a *displacement vector field* $\mathbf{u}(\mathbf{r})$, which maps the point $\mathbf{r} = (x_1, ..., x_D)$ to $\mathbf{r} + \mathbf{u}$, compare Fig. 13.
When a body is deformed, the distances between its points change. Consider two points which are separated by a short distance vector $d\mathbf{r} = (dx_1, ..., dx_D)$ before the deformation. In the deformed state, the distance vector changes to $d\mathbf{r}' = d\mathbf{r} + d\mathbf{u}$. The squared distance between the two points then is $ds^2 = dx_i^2$ before and $ds'^2 = dx_i'^2 = (dx_i + du_i)^2$ after the deformation. Substituting $du_i = (\partial u_i/\partial x_k) dx_k$ for $i = 1, ..., D$, we obtain

$$ds'^2 = ds^2 + 2\frac{\partial u_i}{\partial x_k} dx_i dx_k + \frac{\partial u_i}{\partial x_k}\frac{\partial u_i}{\partial x_l} dx_k dx_l \tag{33}$$

where the Einstein convention of summation over repeatedly appearing indices is again employed. Since the summation in the second term is over both indices i and k, it can be written in the explicitly symmetrical form $(\partial u_i/\partial x_k + \partial u_k/\partial x_i) dx_i dx_k$. Then, ds'^2 takes the final form

$$ds'^2 = ds^2 + 2u_{ik} dx_i dx_k \tag{34}$$

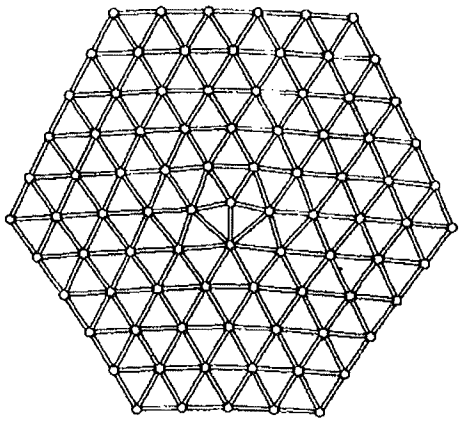


Fig. 11: *Flat dislocation in a triangular lattice. It is the end of an additional line of vertices, which extends from the upper right corner of the network to the center. From Ref. [32].*

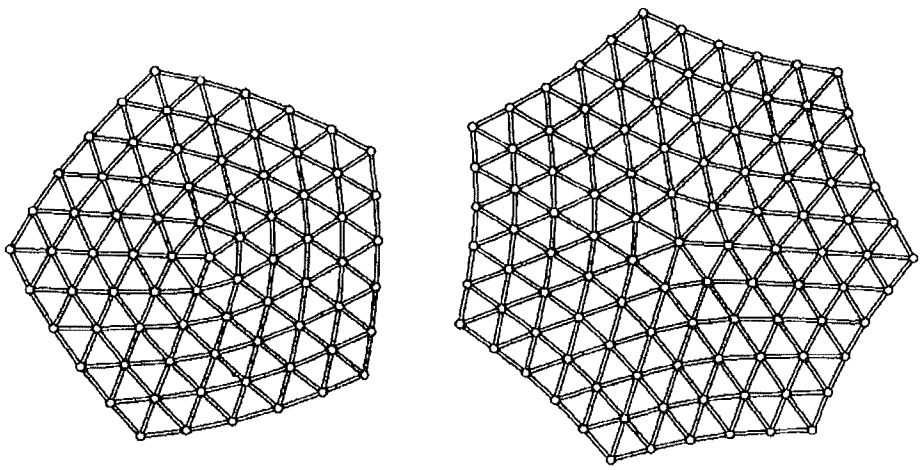


Fig. 12: *Flat five-fold and flat seven-fold disclination in a triangular lattice. From Ref. [32].*

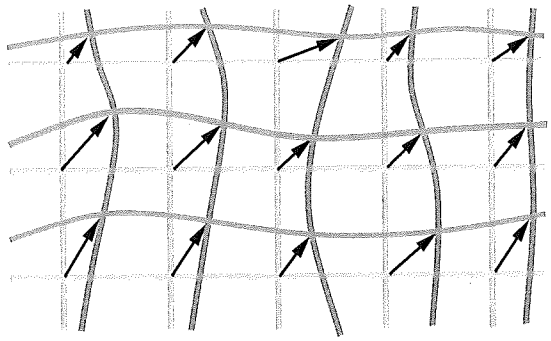


Fig. 13: *The displacement field* $\mathbf{u}(\mathbf{r})$ *determines the positions of the lattice sites in a deformed solid relative to the relaxed state.*

where the *strain tensor* u_{ik} is defined as

$$u_{ik} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} + \frac{\partial u_l}{\partial x_i}\frac{\partial u_l}{\partial x_k}\right) \tag{35}$$

For sufficiently small displacement gradients, which is the case in most applications, the quadratic terms in Eq. (35) can be omitted, so that

$$u_{ik} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right) \tag{36}$$

As any symmetrical tensor, the strain tensor can be diagonalized locally, with the eigenvalues $u^{(1)}, ..., u^{(D)}$. This is useful, for example, to calculate the local relative volume change. In the local coordinate system in which the strain tensor is diagonal, the lengths dx_i change into $dx_i' = (1 + u^{(i)})dx_i$. Therefore, the volume element dV changes into $dV' = (1 + u^{(1)})...(1 + u^{(D)})dV \simeq (1 + \sum_i u^{(i)})dV$ for small strains, i.e.

$$dV' = dV(1 + u_{ii}) \tag{37}$$

Thus, the trace u_{ii} of the strain tensor determines the relative volume change after a deformation.

Elastic Stretching Energy and the Stress Tensor — For small deformations, the *stretching energy* is quadratic in the strain. For an isotropic material, it is given by [28]

$$E_s = \frac{1}{2}\int d^D r\ \left(2\mu u_{ij}^2 + \lambda u_{kk}^2\right) \tag{38}$$

where λ and μ are the Lamé coefficients, and u_{ij}^2 is the trace of the square of the strain tensor, while u_{kk}^2 is the square of the trace of the strain tensor. It should be noticed that this does not contradict the statement in Sec. 2.2 that in *two* dimensions the squared trace and the determinant of a tensor are the two quadratic invariants, because the determinant can be expressed as $\det(u_{ij}) = \frac{1}{2}(u_{kk}^2 - u_{ij}^2)$. The meaning of the Lamé coefficients becomes clear by considering two simple deformations of a hyper-cubic solid. For a homogeneous compression, $\mathbf{r}' = \gamma\mathbf{r}$, so

that $\mathbf{u} = (\gamma - 1)\mathbf{r}$. This implies $u_{ij} = (\gamma - 1)\mathbf{1}$, so that $u_{kk} = D(\gamma - 1)$ and $u_{ij}^2 = D(\gamma - 1)^2$. The stretching energy is then found to be $E_s = (1/2)K_b u_{ll}^2 V$, with the *bulk modulus*

$$K_b = \frac{2}{D}\mu + \lambda \tag{39}$$

Similarly, a pure shear deformation is given by $u_1 = \alpha x_2$, $u_2 = ... = u_D = 0$. In this case $u_{12} = u_{21} = \alpha/2$, $u_{kk} = 0$ (i.e. no volume change), and $u_{ij}^2 = \alpha^2/2$. The stretching energy in this case is given by $E_s = (1/2)\mu u_{ij}^2 V$, and the coefficient μ is therefore called the *shear modulus* .

Let us now consider an elastic body, which is deformed by an external force density $\mathbf{f}(\mathbf{r})$, so that the total energy is given by

$$E = \frac{1}{2}\int d^D r \left(2\mu u_{ij}^2 + \lambda u_{kk}^2\right) + \int d^D r \int_0^{\mathbf{u}(\mathbf{r})} d\mathbf{u}' \cdot \mathbf{f}(\mathbf{r} + \mathbf{u}') \tag{40}$$

In equilibrium, the strain is determined by minimizing E respect to variations in $\mathbf{u}$, i.e. $\delta E/\delta \mathbf{u}(\mathbf{r}) = 0$. This implies

$$\frac{\partial}{\partial x_i}\sigma_{ij} + f_j = 0 \tag{41}$$

where

$$\sigma_{ij} = 2\mu u_{ij} + \lambda u_{kk}\delta_{ij} \tag{42}$$

is the *stress tensor* . Since u_{ij} is symmetric, Eq. (42) implies immediately that the stress tensor σ_{ij} is also symmetric.

The physical meaning of the stress tensor can be seen by integrating Eq. (41) over the total volume of a solid body. With the use of Gauss' integral theorem, this gives

$$\int dV\, f_j = \int dV \frac{\partial}{\partial x_i}\sigma_{ij} = \oint \sigma_{ik} dA_k \tag{43}$$

where the last integral extends over the body surface. This shows that $\sigma_{ik}dA_k$ is the ith component of the force on the surface element $\mathrm{d}\mathbf{A}$. By taking elements of area in the planes (x_1, x_2), (x_2, x_3) and (x_3, x_1) in three dimensions, we find that the component σ_{ik} of the stress tensor is the ith component of the force in the unit area perpendicular to the x_k-axis. For example, the force on an unit area element perpendicular to the x_1-axis, normal to the surface element (i.e. along the x_1-axis), is σ_{11}, while the tangential forces (i.e. along the x_2 and x_3 axes) are σ_{21} and σ_{31}.

3.2 Continuum Theory of Flat Membranes

In the absence of external body forces, the condition of mechanical equilibrium, Eq. (41), reads

$$\frac{\partial}{\partial x_i}\sigma_{ij} = 0 \tag{44}$$

Because σ_{ij} is symmetric and has zero divergence, it can be expressed in two spatial dimensions in terms of a scalar potential, the *Airy stress function* χ, as [28, 32]

$$\sigma_{ij} = \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l\chi \tag{45}$$

where ϵ is the totally anti-symmetric tensor of rank 2 (see Eq. (22)), or, equivalently,

$$\sigma_{11} = \frac{\partial^2 \chi}{\partial x_2^2} \ , \quad \sigma_{22} = \frac{\partial^2 \chi}{\partial x_1^2} \ , \quad \sigma_{12} = -\frac{\partial^2 \chi}{\partial x_1 \partial x_2} \tag{46}$$

Although any choice for χ yields a stress tensor that satisfies Eq. (44), the choice cannot be arbitrary. For any physically realizable stress distribution, there must be a corresponding displacement field related to χ via

$$\begin{aligned} \frac{1}{2}(\partial_i u_j + \partial_j u_i) &= u_{ij} \\ &= \frac{1+\sigma_P}{K_0}\sigma_{ij} - \frac{\sigma_P}{K_0}\sigma_{ll}\delta_{ij} \\ &= \frac{1+\sigma_P}{K_0}\epsilon_{ik}\epsilon_{jl}\partial_k\partial_l\chi - \frac{\sigma_P}{K_0}\nabla^2\chi\delta_{ij} \end{aligned} \tag{47}$$

The parameters K_0 and σ_P are the two-dimensional Young modulus and Poisson ratio, respectively, and can be expressed in terms of the two-dimensional Lamé coefficients as

$$K_0 = \frac{4\mu(\mu+\lambda)}{2\mu+\lambda} \ , \quad \sigma_P = \frac{\lambda}{2\mu+\lambda} \tag{48}$$

The Poisson ratio has a direct physical interpretation. For a uniaxial extension in the x_1-direction with *free* boundary conditions in the orthogonal direction(s), the stress σ_{11} is non-zero, which implies that also the strains u_{11} and u_{22} are non-zero. The Poisson ratio is then defined as the relative contraction in the x_2-direction, i.e. $\sigma_P = -u_{22}/u_{11}$. Since the boundary condition in the x_2-direction is free, $\sigma_{22} = 0$. Eq. (42) then immediately yields the relation (48). Note that thermodynamic stability requires $\mu > 0$ and $K_b > 0$. For a two-dimensional solid, this implies $-1 < \sigma_P < +1$.

The equality of the first and second line of Eq. (47) can be verified easily with the use of Eq. (42). The equality of the second and third line follows immediately with the relation $\epsilon_{ik}\epsilon_{jl} = \delta_{ij}\delta_{kl} - \delta_{il}\delta_{jk}$. We need a constraint on χ which guarantees that Eq. (47) can be solved for the displacements. This condition is found be applying $\epsilon_{ik}\epsilon_{jl}\partial_k\partial_l$ to both sides of Eq. (47), which yields

$$\frac{1}{K_0}\nabla^2\nabla^2\chi = \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l u_{ij} \tag{49}$$

In order to analyze further the right-hand side of Eq. (49), we have to take a closer look at the topological defects — dislocations and disclinations — introduced in Sec. 3.1. For a weakly deformed regular lattice, $\oint du_i = 0$ for any closed contour. However, this is no longer correct in the presence of topological defects. Now, for any closed contour, which encloses the dislocation core, the displacement vector $\mathbf{u}$ receives a finite increment $\mathbf{b}$, which is equal to one of the lattice vectors in magnitude and direction, compare Figs. 14 and 11. The vector $\mathbf{b}$ is called the *Burgers vector* of the dislocation. Mathematically, this property can be expressed as [28, 32]

$$\oint du_i = \oint \partial_k u_i dx_k = b_i \tag{50}$$

Therefore, in the presence of dislocations, the displacement field $\mathbf{u}$ becomes a multi-valued function. Since the Burgers vector is always equal to a lattice vector, this multi-valuedness

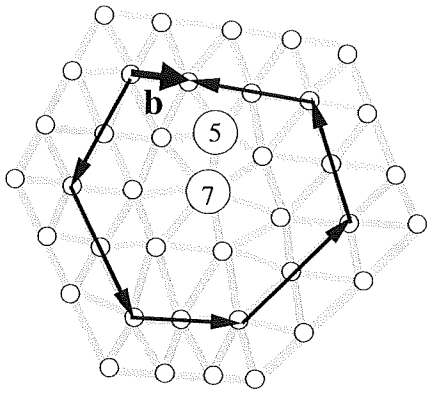


Fig. 14: *Definition of the Burgers vector* b *of a dislocation. A path on the lattice with equal number of steps in each of the six lattice directions, which encloses a dislocation, does not lead back to the starting point — as it would in a defect-free crystal. The distance between the starting and the end point of this path is the Burgers vector.*

represents no physical ambiguity. The condition (50) may also be written in differential form by transforming the line integral into a surface integral with the help of Stokes integral theorem,

$$\oint \partial_j u_i dx_j = \int \epsilon_{lj} \partial_l \partial_j u_i dS \tag{51}$$

Since $\partial_l \partial_j u_i$ is symmetric in the indices l and j and ϵ_{lj} is antisymmetric, the integrand vanishes everywhere except in the dislocation core, where the derivatives of u_i are singular. Therefore, Eqs. (50) and (51) together imply

$$\epsilon_{lj} \partial_l \partial_j u_i = b_i \delta(\mathbf{r} - \mathbf{r}_0) \tag{52}$$

where $\mathbf{r}_0$ is the position of the dislocation core.
Disclinations are defined in terms of the bond-angle field θ, which measures the orientation of bonds around each vertex [32], as illustrated in Fig. 15. For any loop, which encloses the disclination core, the angle θ receives an increment by the *disclinity* s,

$$\oint d\theta = \oint \partial_i \theta dx_i = s \tag{53}$$

For a triangular lattice with six-fold symmetry, s must be a multiple of $2\pi/6$ to ensure that the multi-valuedness of θ represents no physical ambiguity. The five- and seven-fold disclinations mentioned above correspond to disclinities $s = +2\pi/6$ and $s = -2\pi/6$, respectively. The same arguments leading to Eq. (52) give immediately

$$\epsilon_{lj} \partial_l \partial_j \theta = s \delta(\mathbf{r} - \mathbf{r}_0) \tag{54}$$

where $\mathbf{r}_0$ is now the position of the disclination core. Finally, since the bond-angle field describes the local rotation of the bond-orientations, as explained in Fig. 15, it can be written for small strain as

$$\theta = \frac{1}{2} \epsilon_{ij} \partial_i u_j \tag{55}$$

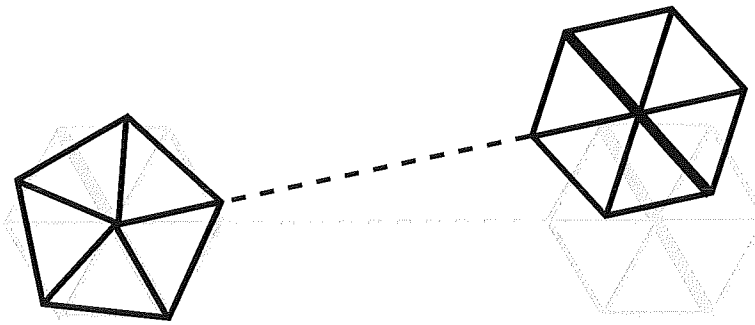


Fig. 15: *The bond-angle field θ defines the orientation of the local crystalline order at a lattice site. In a defect-free hexagonal lattice (gray lines), the local bond orientation, indicated by the thick gray lines, is independent of the location. For a five-fold disclination (black lines), the lattice points are shifted by the displacement field $\mathbf{u}(\mathbf{r})$. In the defect core (left), this introduces a five-fold coordinated vertex. The local hexagonal order outside the dislocation core is retained (compare Fig. 12), but each elementary hexagon is both shifted and rotated by the displacement field. This implies that the local bond orientation (indicated by the thick black line) now has a spatial dependence.*

since $\epsilon_{ij}\partial_i u_j \equiv (\nabla \times \mathbf{u})$ just corresponds to the curl of the vector field $\mathbf{u}$.
With these results, we can now go back to Eq. (49) to obtain

$$\begin{aligned} \frac{1}{K_0}\nabla^2\nabla^2\chi &= \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l\frac{1}{2}(\partial_i u_j - \partial_j u_i) + \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l\partial_j u_i \\ &= \epsilon_{kl}\partial_k\partial_l\theta + \epsilon_{kl}\partial_k(\epsilon_{jl}\partial_l\partial_j u_i) \end{aligned} \tag{56}$$

so that finally [32]

$$\frac{1}{K_0}\nabla^2\nabla^2\chi = \sum_\alpha s_\alpha\delta(\mathbf{r}-\mathbf{r}_\alpha) + \sum_\beta b_i^\beta\epsilon_{ik}\partial_k\delta(\mathbf{r}-\mathbf{r}_\beta) \tag{57}$$

where s_α denotes the disclination charge of a disclination at $\mathbf{r}_\alpha$ and $\mathbf{b}_\beta$ denotes the Burgers vector of a disclination at $\mathbf{r}_\beta$. This is a remarkable result. It shows that two-dimensional elasticity theory consists of just one biharmonic equation. Furthermore, the solutions do not depend in the two Lamé coefficients λ and μ independently, just on the two-dimensional Young modulus K_0 (compare Eq. (48)).
Eq. (57) has to be supplemented by boundary conditions to obtain a unique solution. We will focus on the case of free boundary conditions, where the edges of the membrane are unconstrained. This implies that at the boundary, the components of the stress tensor parallel to the boundary normal vector must vanish.

Stretching Energy of Dislocations and Disclinations — For an isolated dislocation in the center of a circular membrane, the solution of Eq. (57) is given by [32, 33]

$$\chi = \frac{K_0}{4\pi}b_i\epsilon_{ij}r_j\ln(r) \tag{58}$$

as can be checked easily by calculating the derivatives. To satisfy the boundary conditions on a finite region, extra terms have to be added to Eq. (58). However, these terms vanish in the limit of large membrane size, and may therefore be neglected.
It is instructive to calculate from Eq. (58) the stress field (compare Eqs. (45)). The stress tensor in polar coordinates is derived in Appendix A.2. For a Burgers vector pointing in the x_1-direction (without loss of generality), the stress tensor is then found to be

$$\sigma_{rr} = \frac{1}{r}\frac{\partial\chi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\chi}{\partial\phi^2} = \frac{K_0\, b}{4\pi\, r}\sin(\phi) \tag{59}$$

$$\sigma_{r\phi} = -\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial\chi}{\partial\phi}\right) = -\frac{K_0\, b}{4\pi\, r}\cos(\phi) \tag{60}$$

$$\sigma_{\phi\phi} = \frac{\partial^2\chi}{\partial r^2} = \frac{K_0\, b}{4\pi\, r}\sin(\phi) \tag{61}$$

Therefore the stress and also the strain (compare Eq. (47)) decays as $1/r$ with the distance r from the dislocation core.
The energy of a dislocation is most conveniently evaluated by expressing Eq. (38) in terms of the Airy stress function χ, so that

$$E_s = \frac{1}{2K_0}\int d^2r(\nabla^2\chi)^2 - \frac{1+\sigma_P}{2K_0}\int d^2r\epsilon_{ik}\epsilon_{jl}\partial_k\partial_l(\partial_i\chi\partial_j\chi) \tag{62}$$

The substitution of the solution (58) for χ shows that the second term in Eq. (62) vanishes. The first term is given by Eqs. (59) and (61), since $\nabla^2\chi = \sigma_{rr} + \sigma_{\phi\phi}$ (compare Appendix A.2). Thus, we finally obtain the stretching energy

$$E_s = \frac{K_0 b^2}{8\pi}\ln\left(\frac{R}{a}\right) \tag{63}$$

contained in the region $a < r < R$ of an infinite membrane. This expression is usually used to approximate the dislocation energy of a membrane of radius R. The length a in Eq. (63) is a microscopic cutoff, which describes the size of the dislocation core, in which the description by continuum elasticity breaks down and the molecular structure of the crystal becomes important.
The calculation of the stretching energy of a disclination is a bit more subtle, and is therefore discussed in Appendix A.3. In this case, the energy is found to be [32]

$$E_s = \frac{K_0 s^2}{32\pi}R^2 \tag{64}$$

This shows that the stretching energy of disclinations is much larger (for large R) than of dislocations — in agreement with the intuitive expectation.

Melting in Two Dimensions – The defect energies can now be used to predict the melting behavior of crystals in two dimensions [34]. The free energy F of dislocations and disclinations is given by $F = E_s - TS$. The energy E_s can be approximated by Eq. (63) and (64) for dislocations and disclinations, respectively. The entropy is determined by the number of possible locations of the defect core, i.e. by the number of lattice sites, so that $S = 2k_B\ln(R/a)$. Therefore, the free energy of a dislocation is given by

$$F_s = \left[\frac{K_0 b^2}{8\pi} - 2k_B T\right]\ln(R/a) \tag{65}$$

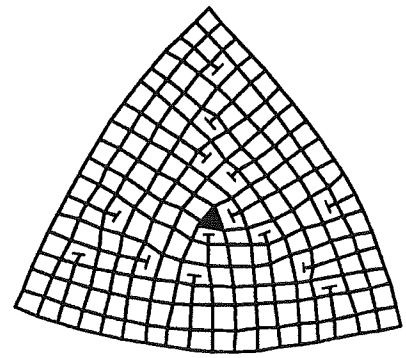


Fig. 16: *Schematic of an unbound three-fold disclination (▲) in an approximately four-fold coordinated medium surrounded by its screening cloud of dislocations (⊥). The dislocations relieve the elastic stress of the planar disclination, and allow all bond lengths to be approximately equal. From Ref. [35].*

for a membrane of radius R. For small temperatures, the free energy (65) is positive, i.e. it costs free energy to generate dislocations, and the crystalline phase is stable. However, at a temperature

$$k_B T_m = \frac{K_0 b^2}{16\pi} \tag{66}$$

the free energy vanishes. Therefore, dislocations start to proliferate at $T = T_m$, and destroy the long-range translational order — the crystal melts. However, note that the free energy of disclinations for large R is still positive at T_m. Therefore, the bond-orientational order, which is *not* destroyed by dislocations, persists for $T > T_m$. This phase, with long-range orientational but short-range translational order, is called a *hexatic phase* [34]. In the hexatic phase, the free energy of disclinations is reduced by "screening" due to free dislocations, as shown schematically in Fig. 16. A detailed calculation [34] shows that it now increases logarithmically in R. A similar argument as for Eq. (65) then indicates that there is a second transition from the hexatic to the fluid phase, where disclinations proliferate and destroy the long-range orientational order [34].

3.3 Continuum Theory of Buckled Membranes

Biological membranes are not restricted to be planar — except when strongly bound to a planar substrate — but are free to buckle out of plane. Then any deformed state is described by the displacement field $\mathbf{u}(\mathbf{r})$ and the deflection $h(\mathbf{r})$, which maps the point $(x_1, x_2, 0)$ of the reference state to $(x_1 + u_1, x_2 + u_2, h)$. The exact form of the strain tensor is now $u_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i + \partial_i u_k \partial_j u_k + \partial_i h \partial_j h)$. For small displacement gradients, the terms quadratic in u_k can again be omitted, so that

$$u_{ij} = \frac{1}{2}(\partial_i u_j + \partial_j u_i + \partial_i h \partial_j h) \tag{67}$$

It is important to notice that the term quadratic in h must be retained, because it presents the only coupling term between in-plane and out-of-plane deformations.
The total energy is the sum of the stretching energy (38) with strain tensor (67), and the curvature energy (7). The stress tensor is related to the non-linear strain as in Eq. (42). Employing again the Airy stress function as in Eq. (45) and following the same analysis as in Sec. 3.2, we arrive at [32]

$$\kappa\nabla^2\nabla^2 h + \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l(\partial_i\chi\partial_j f) = 0 \tag{68}$$

$$\frac{1}{K_0}\nabla^2\nabla^2\chi = s(\mathbf{r}) - K(\mathbf{r}) \tag{69}$$

where $s(\mathbf{r})$ and $K(\mathbf{r})$ are the disclination density and the Gaussian curvature, respectively. Dislocations are contained in Eq. (69) since they can be considered as disclination dipole pairs. It is important to note that Eq. (69) is almost the same as Eq. (57), except that the Gaussian curvature now can screen out the stress induced by disclinations!
For a defect-free membrane, Eqs. (68) and (69) are the von Kármán equations for large deflections of thin plates. These coupled non-linear partial differential equations are "very complicated, and cannot be solved exactly, even in very simple cases" [28].

For a membrane with defects, the simplest case is an isolated five-fold disclination in the center of a circular membrane. In this case, an exact solution can be found, but only in the *inextensional limit* $K_0 \to \infty$. The inextensional limit is exemplified by a piece of paper, which is free to bend but essentially impossible to stretch. In this case, the left-hand side of Eq. (69) vanishes and leaves simply

$$K(\mathbf{r}) = s(\mathbf{r}) \tag{70}$$

i.e. the Gaussian curvature must exactly equal the disclination density. This equation can only be satisfied by a cone (which has zero Gaussian curvature everywhere except at the tip), with an opening angle which is determined by the disclinity,

$$h = \pm\left(\frac{s}{\pi}\right)^{1/2} r \tag{71}$$

The stretching energy vanishes in the inextensional limit. The bending energy of the cone is easily calculated to be

$$E_b = s\kappa\ln(R/a) + s\bar{\kappa} \tag{72}$$

compare Appendix A.4. Note that again a small-distance cutoff is necessary because the curvature energy is no longer applicable on the molecular scale. Eq. (72) is an important result. It shows that a disclination can reduce its energy considerably by buckling — from the strong R^2 dependence for the flat case to the very weak logarithmic R dependence in the buckled case.
The case of a seven-fold disclination can be treated in a similar way. In this case, the deflection field is given by

$$h = \left(\frac{2|s|}{3\pi}\right)^{1/2} r\,\sin(2\phi) \tag{73}$$

with bending energy

$$E_b = 3|s|\kappa\ln(R/a) - |s|\bar{\kappa} \tag{74}$$

compare Appendix A.4.
Numerically calculated shapes (compare Sec. 4 below) of positive and negative disclinations are shown in Fig. 17.

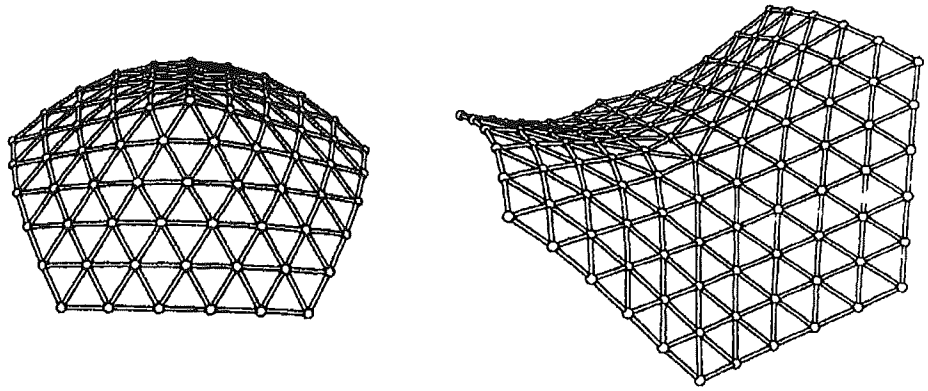


Fig. 17: *Buckled five-fold and seven-fold disclinations. From Ref. [32].*

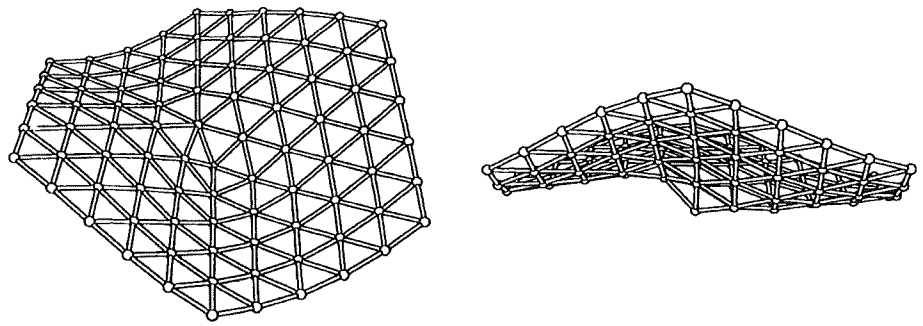


Fig. 18: *Two different views of a buckled dislocation. From Ref. [32].*

The case of dislocations is very difficult. Because the equations are non-linear, a simple superposition of the solutions of positive and negative disclinations fails. Even in the inextensional limit, no analytical solutions is known so far. Thus, numerical solutions have to be used. The numerically calculated shape of a dislocation is shown in Fig. 18. The energy is found to be a slowly increasing function of R, which seems to saturate for large R [32]. This is consistent with the results for disclinations discussed above, which also showed a much smaller increase of the energy with R in the buckled state.

Phase Behavior of Flexible Membranes — This result has dramatic consequences for the phase behavior of flexible membranes. As in Sec. 3.2, we can again write the free energy of a dislocation as $F = E_b - TS$. Since the energy approaches a constant, but the entropy still increases logarithmically with R, the free energy is always negative for sufficiently large membrane sizes. Thus, the crystalline phase is destabilized for any finite temperature T (in the limit $R \to \infty$). The low temperature phase of a flexible membrane is therefore a hexatic phase. The phase transition from the hexatic to the fluid phase is determined by the free energy of disclinations. Since the bending energy of disclinations has now a logarithmic R-dependence,

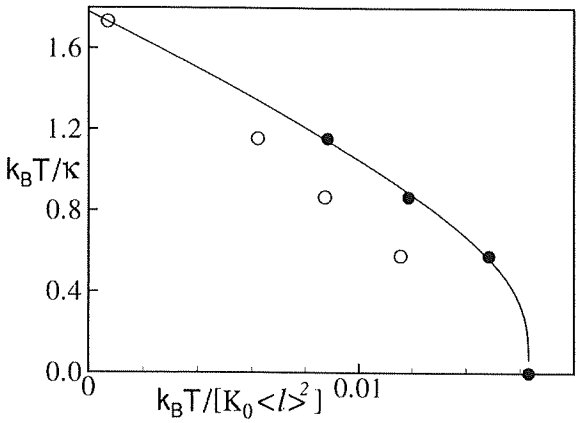


Fig. 19: *Phase diagram of flexible membranes, as calculated by Monte Carlo simulations. For small k_BT/κ and small $k_BT/(K_0\langle\ell\rangle^2)$ (where $\langle\ell\rangle$ is the average nearest-neighbor distance), the hexatic phase is stable. At high temperatures, the fluid phase is stable. From Ref. [36].*

see Eqs. (72) and (74), we can use the same arguments as in Sec. 3.2 to estimate

$$F_b = [2|s|\kappa - 2k_BT]\ln(R/a) \tag{75}$$

[since five- and seven-fold disclinations appear in equal numbers, the defect energy is taken to be the algebraic average of Eqs. (72) and (74)] and thus the transition temperature

$$k_BT = \frac{\pi}{3}\kappa \tag{76}$$

The phase diagram [36] obtained from Monte Carlo simulations of dynamically triangulated surfaces (see Sec. 4 below) is shown in Fig. 19. Here, the in-plane elasticity is measured in terms of the two-dimensional Young modulus K_0 of a "tethered network", i.e. a network with fixed neighbor relations in which no defects can occur, with the same average nearest-neighbor distance. The hexatic phase is found to be stable for high bending rigidity and large in-plane density (large Young modulus K_0). A transition to the fluid phase is found *both* with decreasing density and with decreasing in-plane density, in agreement with the arguments given above:

- The melting for high bending rigidity is driven by the reduction of the stretching energy of disclinations on length scales smaller than the buckling radius.
- The melting for high Young modulus (large in-plane density) is driven by the reduction of the bending energy of cone- and saddle-shaped five- and seven-fold disclinations, respectively.

The Monte Carlo results agree well with analytical predictions [37], which are derived in the limit of small k_BT/κ.

Buckling Radius and Shapes of Crystalline Vesicles — The elastic energy of a planar disclination increases as R^2, while the elastic energy of a buckled disclination only increases as

$\ln(R/a)$, compare Eqs. (64) and (72), respectively. Therefore, the buckled state is favorable for large membrane radius R. However, for small membrane sizes, the energy of the planar disclination is smaller. There is a critical radius, the *buckling radius* R_b, which the membrane has to exceed before buckling is favorable. It can immediately be seen from Eqs. (64) and (72) that the buckling radius of disclinations scales as $R_b \sim \sqrt{\kappa/(K_0|s|)}$. A more detailed calculation [32] gives (for $\bar{\kappa} = -\kappa$)

$$R_b \simeq 10 \left(\frac{\kappa}{K_0|s|} \right)^{1/2} \tag{77}$$

for both positive and negative disclinations (in fact, the numerical prefactors are slightly different). Similarly, the buckling radius of a dislocations is found to be [32] (again for $\bar{\kappa} = -\kappa$)

$$R_b \simeq 120 \left(\frac{\kappa}{K_0 b} \right) \tag{78}$$

The buckling radius can be seen directly in the shape of crystalline vesicles, i.e. of closed membranes of spherical topology, which for topological reasons (Euler's formula, compare Sec. 2.3) must have exactly 12 five-fold disclinations. Let us consider the shape of such a vesicle with fixed bending rigidity κ and fixed Young modulus K_0 as a function of the vesicle radius R. For $R < R_b$, the contribution of the stretching energy is small, the bending energy dominates, and the vesicles shape is spherical. For $R > R_b$, however, the total energy can be reduced by buckling, which gives the disclinations a local cone shape. Therefore, the vesicle now looks like a very smoothed-out icosahedron, see Fig 20a. As R increases further, an icosahedral shape slowly emerges. In this regime, the energy increases logarithmically with R, compare Eq. (72). Finally, for $R \gtrsim 100 R_b$, another scaling regime emerges, see Fig 20b. In this case, the faces of the icosahedron are almost planar, and all the stress condenses into the ridges which connect the corners of the icosahedron [38, 39]. In this regime, the energy scales as

$$E \sim \kappa \left(\frac{K_0 R^2}{\kappa} \right)^{1/6} \tag{79}$$

A very nice derivation of this result, based on a detailed analysis of the von Kármán equations (68) and (69) with $s(\mathbf{r}) = 0$, can be found in Ref. [40].

Crystalline Order on a Sphere — Another interesting question in this context is whether an ideal lattice with exactly 12 five-fold disclinations is actually the minimal-energy configuration on a sphere, or whether a configuration with more defects might have lower energy. This problem was already recognized by J.J. Thomson nearly 100 years ago, who attempted unsuccessfully to explain the periodic table in terms of rigid electron shells [41].
We can get immediately an intuitive idea about what might happen from elasticity theory in the inextensional limit, see Eq. (70). In contrast to the discussion of the shape of isolated topological defects — in which the membrane shape had to adjust in such a way as to make the Gaussian curvature $K(\mathbf{r})$ equal to a given defect distribution — we now have the inverse problem of finding a defect distribution which mimics a given Gaussian curvature $K(\mathbf{r}) = 1/R^2$ as closely as possible. For a shell with a finite number of lattice sites, it is obviously impossible to satisfy the equation $s(\mathbf{r}) = 1/R^2$ exactly. Therefore, some local stretching is always required in addition to an optimal defect distribution.
Considerable progress has been made recently in solving the Thomson problem. Both numerical simulations [42] and analytical theory [43, 44] shows that the energy can be reduced by grain

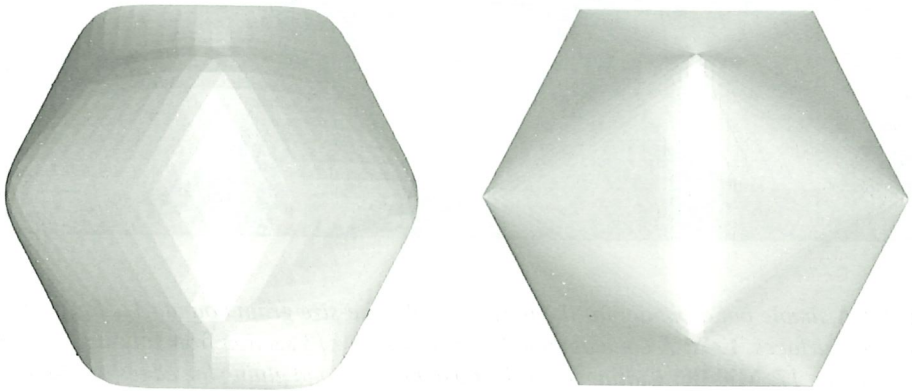


Fig. 20: *Two calculated shapes of crystalline vesicles for Foppl-von Kármán numbers* $\gamma \equiv K_0 R^2/\kappa$. **Left:** $\gamma = 690$, *in the scaling regimes of "cones".* **Right:** $\gamma = 8 \times 10^6$, *in the scaling regime of "stretching ridges". From Ref. [7].*

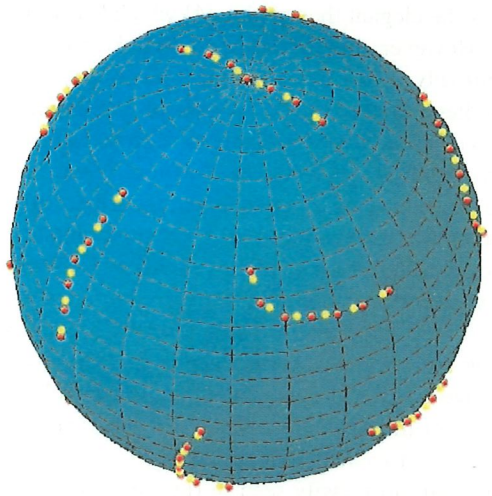


Fig. 21: *Finite size grain boundaries, which are strings of dislocations, reduce the elastic energy of a crystal on a spherical surface. Five-fold disclinations are shown by red, seven-fold disclinations by yellow points. Note that each grain boundary contains one excess five-fold disclination. From Ref. [45].*

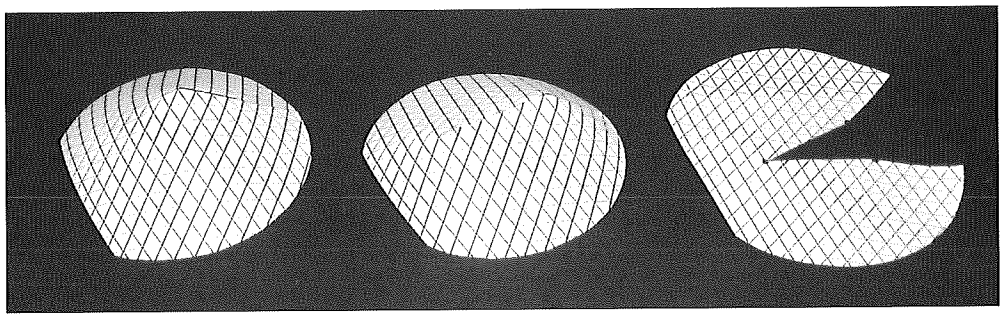


Fig. 22: *A simple paper model for the formation of finite-size grain boundaries for crystals on spherical surfaces.* **Left:** *A cone obtained by removing a $\pi/3$ section from a hexagonal lattice, compare Fig. 17.* **Middle:** *Cutting away the tip of the cone and gluing the edges together brings the shape closer to that of a spherical cap. This generates a grain boundary.* **Right:** *The planar pattern, which generates the middle figure when the edges are glued together correctly. The end points of the grain boundary are marked (•).*

boundaries of finite length, which are located at each of the 12 corners of an icosahedron, as shown in Fig. 21. These predictions have already been verified experimentally [45].

A detailed description of the elegant theory [43, 44] which leads to this result goes beyond the purpose of this lecture. However, this formation of grain boundaries can be demonstrated and thereby understood intuitively with a sheet of paper. As already explained just before Eq. (70) above, paper is an excellent example for a crystalline membrane in the inextensional limit, since it is free to bend but essentially impossible to stretch. Let us consider one corner of the icosahedron, which corresponds to a five-fold disclination. When a $\pi/3$ section is removed from a planar sheets, and the free edges are glued together, the paper forms a cone, see the left-hand part of Fig. 22. The cone is obviously not a very good approximation of a spherical cap. Since we are interested in a crystal lattice on a sphere, we have to deform the sheet by pushing the tip onto the surface of an underlying sphere. For paper, this is essentially impossible, since the stress which is generated is so large that it would crumple or rupture the sheet. A better way of making our cone look more like a spherical cap is to take a pair of scissors, cut away the tip, and again glue the edges together. This generates a shape which is shown in the center part of Fig. 22. It is much closer to the desired shape of a spherical surface. The crucial point is now that while the lattice fits together perfectly along the cuts of the cone, it does not when the tip is removed. This can most easily seen in right-hand part of Fig. 22, which shows the two-dimensional pattern of the truncated cone. The line between the marked points is the grain boundary.

This construction shows also that for larger spheres, with more lattice sites, more complex patterns of grain boundaries become favorable. For example, by truncating the remaining two tips in the center part of Fig. 22, the resulting sheet can be made even more spherical, so that even smaller stresses are required if such a crystal is confined to the surface of a sphere.

4 Monte Carlo Simulations of Dynamically Triangulated Surfaces

As in all entropically dominated, strongly fluctuating systems, simulation methods provide a very important tool to obtain information about the system properties, which are not easily accessible by purely analytical approaches. For sufficiently small systems, thermal averages can be calculated by simulation techniques with very high precision. The extrapolation to large system sizes is sometimes difficult, but often possible with good reliability. A general introduction to Monte Carlo simulation techniques can be found in Chap. B.5.

Simulation method — The simulation of membranes and vesicles with characteristic sizes on the order of 10 to 1000 nm is impossible on the basis of a molecular model, since it would require an enormous number of lipid and solvent molecules. Therefore, a model in which the membrane is described by an elastic surface with an energy as discussed in the previous sections, is most appropriate on these mesoscopic length scales. In order to make this model suitable for simulations, the continuous surface has to be approximated by a network of vertices and bonds, see Fig. 23. The simplest potential for the interaction of vertices which are connected by bonds is a *tethering potential*,

$$V(r) = \begin{cases} 0 & \text{if } r < \ell_0 \\ \infty & \text{otherwise} \end{cases} \tag{80}$$

which causes the particles to behave as tethered by a string. A triangular network is usually used because it provides the most homogeneous and isotropic discretization of the surface.

Tethered Membranes — A Monte Carlo step consists of a random displacement of a randomly selected vertex. This step is accepted with the probability determined by the Boltzmann weight, as long as the vertices remain within the maximum bond lengths with their neighbors. The energy, which appears in the Boltzmann weight is the curvature energy, which can be discretized in the form [46, 47]

$$E_b = \lambda_b \sum_{<ij>} (1 - \mathbf{n}_i \cdot \mathbf{n}_j) \tag{81}$$

where $\mathbf{n}_i$ and $\mathbf{n}_j$ are the normal vectors of neighboring triangles, and the sum runs over all pairs of neighboring triangles. The coupling constant λ_b in Eq. (81) is related to the bending rigidity and saddle-splay modulus by $\kappa = \sqrt{3}\lambda_b/2$ and $\bar{\kappa} = -\kappa$ [32]. The discretization (81) of the curvature energy is not without problems; better discretizations are discussed in Ref. [48].
Such a model has first been suggested and simulated for *polymerized membranes*, i.e. for membranes of fixed connectivity, by Kantor and Nelson (1987) [46, 47]. When hard spheres of diameter σ_0 are placed on the vertices, and the bond lengths ℓ_0 are restricted to be $\ell_0 \leq \sqrt{3}\sigma_0$, the surface is *self-avoiding*, since an arbitrary sphere does *not* fit through the holes of the network, so that no interpenetration of different parts of the network is possible.
The model described so far applies to regular network without defects, but can of course also be used to study the properties of "static" defects, which are inserted "by hand" into a regular lattice. The model can not only be studied by Monte Carlo simulations. With a harmonic bond interactions, it can also be used to minimize the elastic energy. With such a energy minimization, the shapes of membranes with static defects shown in Figs. 11, 12, 17, 18, 20 and 25 have been determined.

Fluid Membranes — For a study of fluid membranes, the connectivity of the membrane cannot remain fixed during the simulation, because otherwise a diffusion of vertices within the mem-

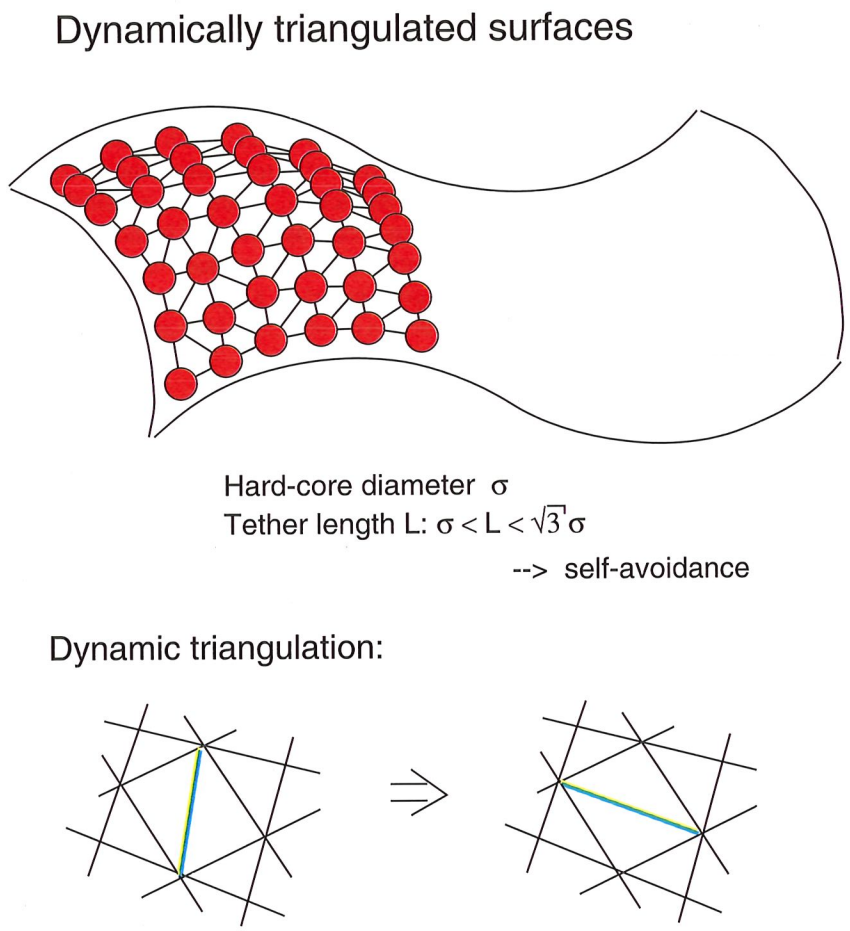


Fig. 23: *A triangulated network model of a fluctuating surface. Top: Hard spheres connected by bonds of maximum extension ℓ_0 are used to describe self-avoiding membranes. Bottom: The Monte Carlo step, which makes the triangulation dynamic, is required to model fluid membranes.*

brane is not possible. Therefore, dynamically triangulated surfaces [15, 19] have to be used in this case. The essential step of the dynamic triangulation procedure is shown at the bottom of Fig. 23. Among the four vertices of two neighboring triangles, the "diagonal" bond is switched from one of the two possible positions to the other. This bond-switching is only allowed if the vertices remain connected to at least three neighbors after the switch. Also, the distance between the newly connected vertices has to be smaller than the maximum bond length. This Monte Carlo step has the advantages that

- it is local, i.e. only the vertices of two neighboring triangles are involved, and
- it guarantees that the network retains its two-dimensional connectivity during the whole simulation run.

Crystalline Membranes — The same model used for fluid membrane can also be used to study crystalline membranes, thermally excited defects in crystalline membranes, the interaction between defects, the hexatic phase, and the phase transitions between these phases [49, 50, 36, 19]. This is possible because the in-plane density of the membrane model can be controlled by the tether length ℓ_0. In a very similar way as for hard discs in two dimensions, the system crystallizes when the in-plane density exceeds a critical value. This can be demonstrated easily with a planar box of ball bearings. When the density of hard spheres in the box is small, the box is held horizontally and is gently shaken (to mimic the effects of thermal motion), the spheres move randomly like in a gas or fluid. However, if the box is tilted sufficiently strongly, so that the density on one side increases strongly due to gravity, the spheres quickly order into a hexagonal packing. In the tether-and-bead model described above, the crystalline-to-hexatic transition is found to occur at a tether length $\ell_0 \simeq 1.48\sigma_0$, while the hexatic-to-fluid transition is located at $\ell_0 \simeq 1.52\sigma_0$ [50]. The simulations of this network model therefore confirm the melting scenario for two-dimensional systems discussed in Sec. 3.2.
The main advantage of the dynamically-triangulated network model compared to hard discs with an external pressure is that it can immediately be used for buckled membranes. This model has indeed been used to calculate the phase diagram of these systems discussed in Sec. 3.3.

5 Biological Systems

5.1 Virus Shapes

The analysis of approximately spherical viruses dates back to the pioneering work of Crick and Watson in 1956 [51], who argued that the small size of the viral genome requires identical structural units packed together with an icosahedral symmetry. Furthermore, it has been found recently that most viral coat and capsid proteins have about the same size, molecular weight, amino acid composition, and, most importantly, the same folded structure in three dimensions [52]. Moreover, the protein structure determines the mechanical properties of proteins [53].
It has therefore been suggested very recently by Lidmar, Mirny and Nelson (2003) [7] to describe the shape of viral capsids by the simple elastic model explained in Sec. 3.3. Such a description provides new insight into the physics of viral capsids:

- Can the shape of viral shells be well described by a simple elastic model?
- Is the bending and stretching elasticity of different capsids similar?

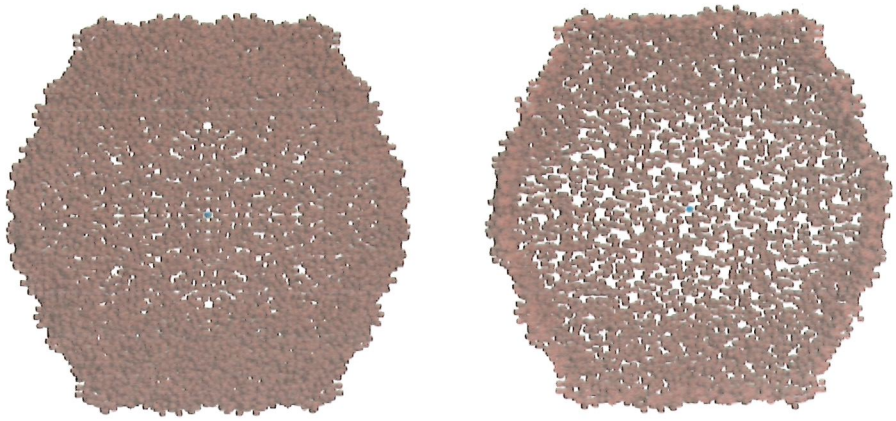


Fig. 24: *Experimental result for the shape of the virus shell of the bacteriophage HK97 (full virus and cross section), with 72 capsomers and a diameter of 60 nm. From Ref. [7].*

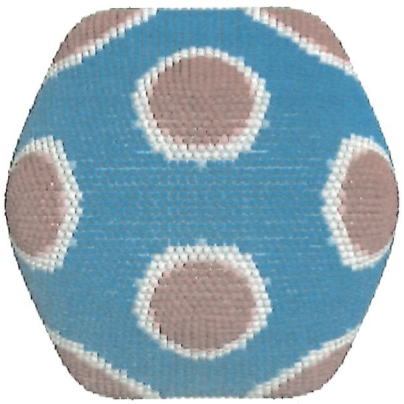


Fig. 25: *Calculated shape of a crystalline vesicle of icosahedral symmetry for the Foppl-von Kármán number $\gamma = K_0 R^2/\kappa = 1480$, which gives the best fit to the shape of the bacteriophage HK97 shown in Fig. 24. The colors indicate the distance of the shell from the center. From Ref. [7].*

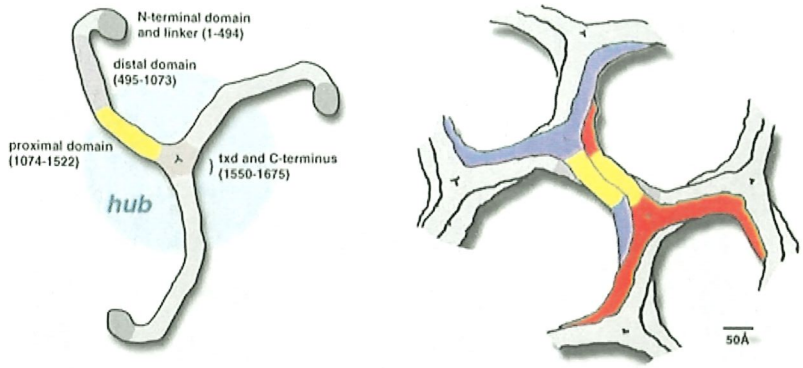


Fig. 26: Left: *Schematic picture of the clathrin triskelion, which is a three-armed protein.* **Right:** *After adsorption on the membrane, clathrin proteins assemble into a perfect honeycomb lattice in such a way that each bond is formed by the arms of three different triskelions. From Ref. [55].*

The answer to the first question is clearly positive, as demonstrated in Figs. 24 and 25, which show good agreement of the experimental and theoretical shapes for the virus shell of the bacteriophage HK97. The results of Sec. 3.3 suggest that the shape of a closed elastic membrane is determined by a single parameter, the Foppl-van Kármán number $\gamma = K_0 R^2/\kappa$, where shapes for small γ are spherical and become more and more faceted with increasing γ. For fixed elastic constants κ and K_0, the Foppl-van Kármán increases rapidly with increasing size. This is consistent with the viral shapes shown in Fig. 4.
More quantitatively, in the particular case of the bacteriophage HK97 shown in Fig. 24, the best fit is obtained for $\gamma = 1480$. With the radius $R = 30$ nm, one obtains $[K_0/\kappa]_{HK97} = 1.64\text{nm}^{-2}$. A similar fit for the yeast L-A virus (the shape is shown in Refs. [5, 7]) gives $\gamma = 547$. With a smaller radius $R = 21.5$ nm of the yeast L-A virus, one obtains $[K_0/\kappa]_{yeast} = 1.24\text{nm}^{-2}$. This result gives strong support for the idea that the shells of different viruses have very similar elastic properties.

5.2 Clathrin-Coated Pits

A famous and biologically very important example of domain formation and budding is the adsorption of clathrin molecules on the plasma membrane. Clathrin molecules are three-armed proteins, see Fig. 26, which assemble to form a regular hexagonal network on the membrane surface. By forming first a coated pit and then a complete bud, see Fig. 27, these clathrin coats control endo- and exocytosis, i.e. the formation and detachment of small transport vesicles from the cell membrane, as illustrated in Fig. 28. In the cell, the budding process involves several other proteins, like endophilin and dynamin [54], which generate a spontaneous curvature and assist in the detachment of the vesicle from the plasma membrane, respectively. In vitro, the budding process can also be driven by a change of pH, as in Fig. 27, which induces a change of the spontaneous curvature [25]. The formation of clathrin cages is therefore an example for the budding of crystalline patches embedded in a fluid lipid membrane.
In a two-component membrane with nearly immiscible components A and B, an important

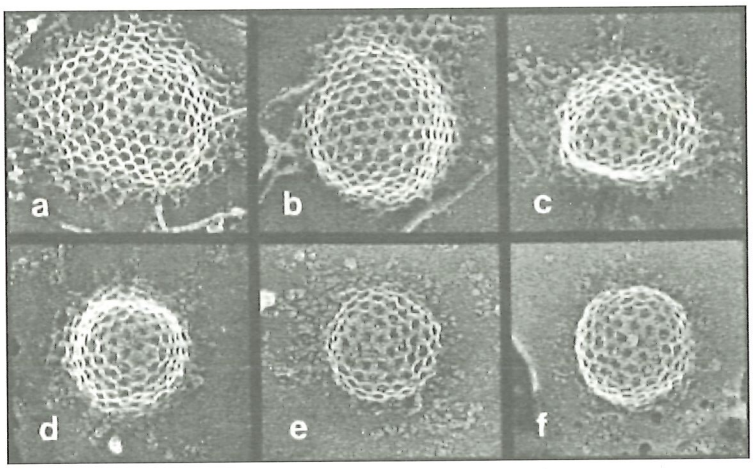


Fig. 27: *Rounded clathrin coated pits in normal chick fibroblasts (a-d) and coated pits on membrane fragments derived from cells that have been broken open and left in pH 7 buffer for 10 minutes at* $25°$ *before fixation and freeze drying (e-f). The width of field of view of the individual pictures is* $0.4\mu m$. *Reproduced from Ref. [56] by copyright permission of The Rockefeller University Press.*

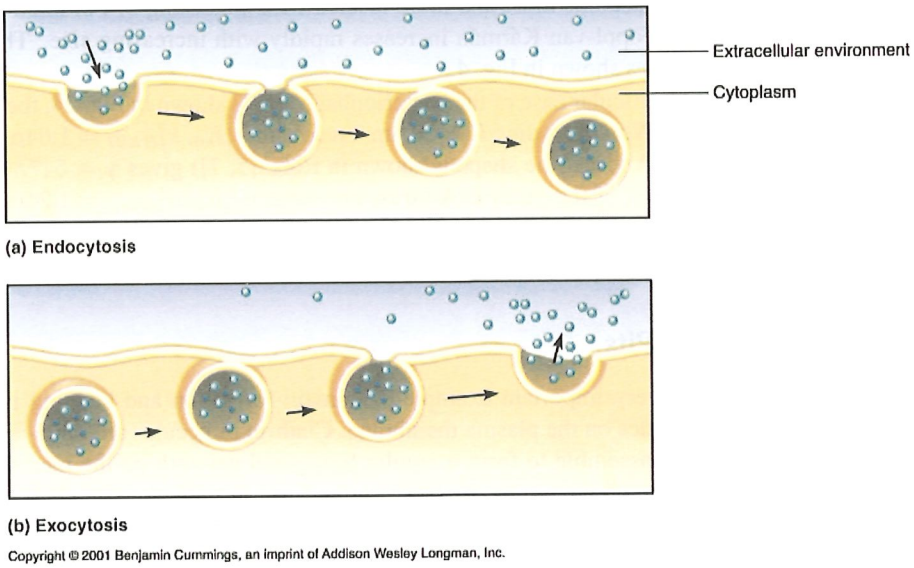


Fig. 28: *Schematic illustrations of the process of (a) endocytosis, by which small transport vesicles can be loaded with large molecules and detached from the plasma membrane, and (b) exocytosis, by which transport vesicles fuse with the plasma membrane and expel their content. From Ref. [57].*

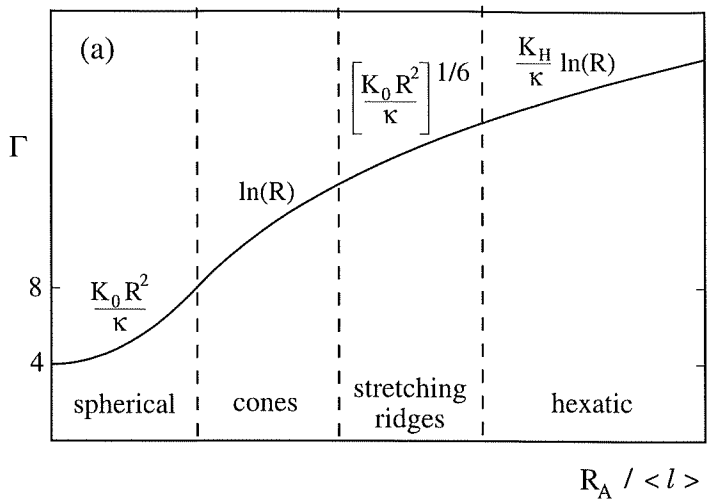


Fig. 29: *Schematic plot of the scaling regimes of* $\Gamma(R_A)$*, compare Eq. (83), which characterizes the dependence of the budding transition on the domain size* R_A *of a planar crystalline domain (in units of the average nearest-neighbor distance* $\langle \ell \rangle$*). Different scaling regimes are indicated. From Ref. [63].*

question is the dependence of the domain size R_A of an A-rich domain on the line tension λ and spontaneous curvature C_0^A at the budding transition. In two-component membranes with a *fluid* A-rich membrane domain, the location of the budding transition can be found approximately from the following simple estimate [58, 59]. For a planar membrane, the energy of an A-rich domain of radius R_A is given by $E_{planar} = 2\pi\lambda R_A + (\pi/2)\kappa(C_0^A R_A)^2$. On the other hand, a complete bud has the energy $E_{bud} = (\pi/2)\kappa(4/R_A - C_0^A)^2$. These two energies are equal at

$$\lambda R_A/\kappa + 2C_0^A R_A = 4 \tag{82}$$

Thus, a budding transition occurs with increasing line tension λ or with increasing spontaneous curvature C_0^A. More elaborate calculations [60, 61, 62] confirm this picture, and yield more precise quantitative results, as discussed in more detail in Chap. D.4.

A flat, crystalline membrane domain cannot be deformed into a spherical bud without the introduction of topological defects. For the generic case of six-fold coordinated vertices in the crystalline phase, Euler's formula (12) requires an excess of exactly 12 five-fold disclinations to form a spherical cage. This can be seen as follows. For any closed triangulated surface, $N_{vertices} = 2 + \frac{1}{2}N_{triangles}$ (compare Sec. 2.3). In addition, $\frac{3}{2}N_{triangles} = N_{bond} = 3N_6 + \frac{5}{2}N_5$ with $N_{vertices} \equiv N_6 + N_5$. These two equations can be solved easily to give $N_5 = 12$.

Therefore, the budding behavior of *crystalline* membrane domains is more complex. The results described in the Sec. 3.3 imply the following behavior. The location of the budding transition is now given by

$$\lambda R_A/\kappa + 2C_0^A R_A = \Gamma(R_A) \tag{83}$$

which for fixed domain size shows the same linear dependence of C_0^A and λ as in the fluid case, Eq. (82). However, there is now an additional dependence on the domain size, determined by

the function $\Gamma(R_A)$ which is the energy (in units of κ) of a spherical bud of radius $R = R_A/2$ — without the spontaneous-curvature contributions. This function has several scaling regimes, compare Fig. 29:

- For small R_A, the bending energy dominates, the bud is spherical, and the energy increases quadratically from the Γ-value which applies for fluid domains, compare Eq. (64).
- For larger R_A, as soon as the stretching energy becomes larger then the bending energy, the bud deforms to lower its energy. The shape becomes a smoothed-out icosahedron, with a logarithmic dependence on R_A, compare Eq. (72).
- For even larger R_A, the stress condenses into stretching ridges, the shape looks very much like an icosahedron with smooth edges, and the energy increases as a power law with the radius, compare Eq. (79).
- Finally, the membrane enters the hexatic phase, and the energy increases again logarithmically.

These results are consistent with Monte Carlo simulations of dynamically triangulated surfaces [63].

Another important question is the generation of the 12 topologically required five-fold disclinations. As explained in Sec. 3.1, disclinations are topological defects which cannot be generated locally. In order to explain bud formation, it is therefore necessary to understand how the disclinations appear inside the crystalline domain. Two principal mechanisms are possible:

- an edge-acquisition mechanism, in which disclinations form at the edge of the domain and then diffuse into the interior, and
- an interior-acquisition mechanism, in which dislocation pairs are generated in the interior; each of these dislocations subsequently dissociates, leaving behind a five-fold disclination while the seven-fold disclination diffuses to the domain boundary.

This question can be addressed very well by Monte Carlo simulations. Fig. 30 shows a sequence of snapshots of the budding process, which starts with a completely defect-free crystalline clathrin domain. It can be seen very clearly in the initial stages, Fig. 30a, that the defects are generated at the domain boundary. This strongly supports the exterior acquisition hypothesis. It is also interesting to note that in the later stages, Fig. 30c,d, lines of dislocations are visible, which correspond to the finite grain boundaries discussed in Sec. 3.3.

6 Summary and Conclusions

The investigation of membrane properties is an interdisciplinary endeavor. It brings together mathematicians (minimal surfaces), theoretical physicists (statistical physics, field theory), experimental physicists (microscopy, neutron scattering), chemists (synthesis, phase behavior), biologists (cell membranes), and material scientists (nanoscale materials).
Over the last twenty years, the joint effort of researches from these different disciplines has lead to significant progress in our understanding of membranes. In particular, the behavior of polymerized, crystalline, and fluid membranes has been found to be governed by very different

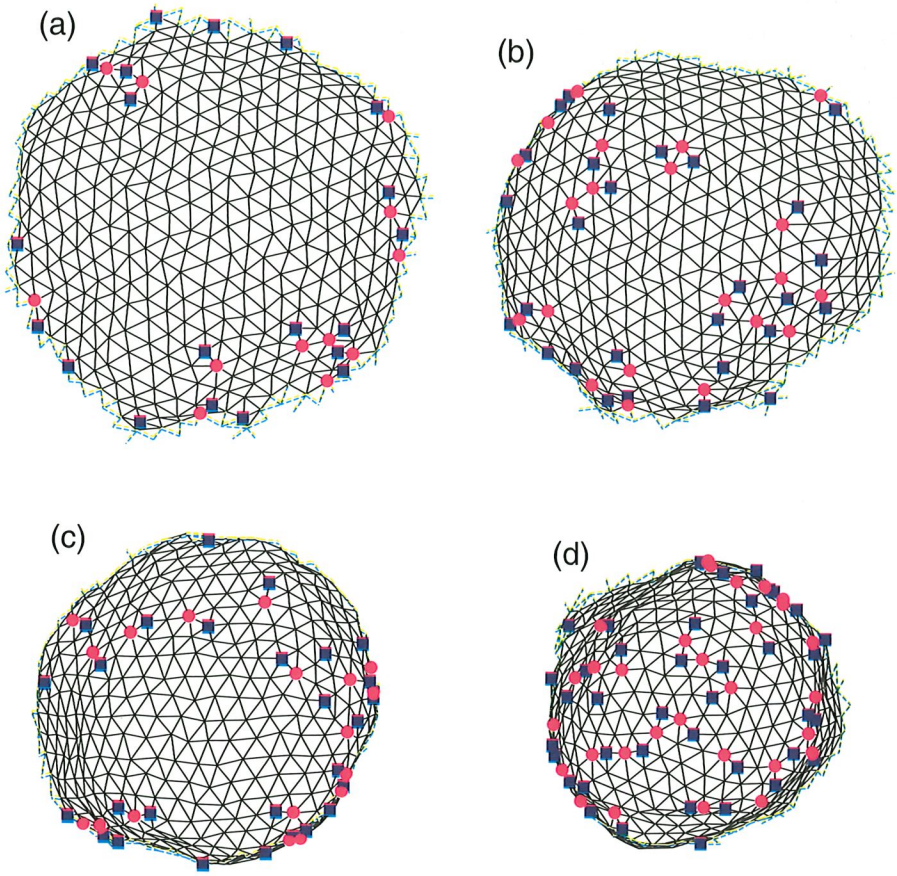


Fig. 30: *Snapshots of configurations during the budding process after a quench from a completely spherical initial state without defects in the clathrin domain. The pictures show a top view of the crystalline domain; the fluid-membrane part of the vesicle is not shown. Five- and seven-fold vertices are indicated by blue squares and purple circles, respectively. Snapshots are shown at time (a)* $t = 0.1 \times 10^6$*, (b)* $t = 0.3 \times 10^6$*, (c)* $t = 1.0 \times 10^6$*, and (d)* $t = 2.0 \times 10^6$ *Monte Carlo steps after the quench. Parameters are* $\kappa = 10 k_B T$*,* $\lambda\sigma_0 = 2.0$*,* $C_0^A \sigma_0 = 0.2$*,* $C_0^B \sigma_0 = 0$*,* $\ell_0^A = 1.50$*, and* $\ell_0^B = 1.68$*. From Ref. [63].*

physical mechanisms. However, even the "simple", single-component membranes will require more studies before a satisfactory level of understanding has been reached.
In the future, the trend will undoubtedly go towards more complex membrane systems. With the increasing number of components, the system will become more flexible, with properties which can be controlled and varied externally. A good example is the red blood cell membrane, which consists of several lipid species, has a different lipid composition in its inside and outside layer, has a polymer network attached to its inside, and contains several different kind of functional proteins. For Statistical Physics, it is an enormous challenge to predict the properties of such complex systems quantitatively.

A Appendices

A.1 Curvature Tensor of Almost Planar Membranes

For almost planar membranes, the full curvature tensor can be calculated explicitly. Eq. (6) implies

$$\mathbf{t}_l \cdot \partial_i \mathbf{n} = -K_{ij}\mathbf{t}^j \cdot \mathbf{t}_l = -K_{il} \tag{84}$$

Differentiation of the orthogonality relation $\mathbf{t}_j \cdot \mathbf{n} = 0$ gives

$$K_{ij} = -\mathbf{t}_j \cdot \partial_i \mathbf{n} = \mathbf{n} \cdot \partial_i \mathbf{t}_j = \mathbf{n} \cdot \partial_i \partial_j \mathbf{R} \tag{85}$$

This relation demonstrates that the curvature tensor is symmetric. It can now easily be evaluated with Eqs. (14), (15) and (16), with the result

$$K_{ij} = \frac{\partial_i \partial_j h}{\sqrt{1+(\nabla h)^2}} \tag{86}$$

This implies

$$2H = Tr(K_i^j) = \nabla \cdot \frac{\nabla h}{\sqrt{1+(\nabla h)^2}} \tag{87}$$

$$\begin{aligned} K &= \det(K_i^j) = \det(K_{il})\det(g^{lj}) = \det(K_{il})/\det(g_{lj}) \\ &= \frac{\partial_1^2 h \partial_2^2 h - (\partial_1 \partial_2 h)^2}{[1+(\nabla h)^2]^2} \end{aligned} \tag{88}$$

A.2 Stress Tensor in Polar Coordinates

The components of the stress tensor in polar coordinates are defined by

$$\sigma_{rr} = \mathbf{e}_r \underline{\sigma} \mathbf{e}_r \ , \quad \sigma_{r\phi} = \mathbf{e}_r \underline{\sigma} \mathbf{e}_\phi \ , \quad \sigma_{\phi\phi} = \mathbf{e}_\phi \underline{\sigma} \mathbf{e}_\phi \tag{89}$$

where $\mathbf{e}_r = (x_1/r, x_2/r)$ and $\mathbf{e}_\phi = (-x_2/r, x_1/r)$. The derivatives $\partial^2\chi/\partial x_i \partial x_j$ are expressed in polar coordinates in the form

$$\frac{\partial^2\chi}{\partial x_i \partial x_j} = \frac{\partial^2\chi}{\partial r^2}\partial_i r \partial_j r + \frac{\partial^2\chi}{\partial \phi^2}\partial_i\phi\partial_j\phi + \frac{\partial^2\chi}{\partial\phi\partial r}(\partial_i\phi\partial_j r + \partial_j\phi\partial_i r) + \frac{\partial\chi}{\partial r}\partial_i\partial_j r + \frac{\partial\chi}{\partial\phi}\partial_i\partial_j\phi \tag{90}$$

With $r^2 = x_1^2 + x_2^2$ and $\phi = \arctan(x_2/x_1)$, we obtain

$$\partial_i r = x_i/r, \ \ \partial_i\phi = -\epsilon_{ik}\frac{x_k}{r^2}, \ \ \partial_i\partial_j r = \frac{1}{r}\left(\delta_{ij} - \frac{x_i x_j}{r^2}\right), \ \ \partial_i\partial_j\phi = -\frac{1}{r^2}\epsilon_{ij} + 2\epsilon_{ik}\frac{x_j x_k}{r^4} \tag{91}$$

Now, σ_{rr} for example is given by

$$\sigma_{rr} = \frac{1}{r^2} x_i x_j \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l\chi \tag{92}$$

Since $x_i\epsilon_{ik}x_k = 0$ and $x_i x_j \epsilon_{ik}\epsilon_{jl}\epsilon_{kl} = 0$, the first, third and fifth contribution on the left-hand side of Eq. (90) do not contribute. Since $x_i\epsilon_{ik}\epsilon_{kl}x_l = -r^2$ and $x_i x_j \epsilon_{ik}\epsilon_{jl}\delta_{kl} = r^2$, we finally arrive at

$$\sigma_{rr} = \frac{1}{r}\frac{\partial\chi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\chi}{\partial\phi^2} \tag{93}$$

compare Eq. (59). Similarly, we obtain

$$\sigma_{\phi\phi} = \frac{1}{r^2}\epsilon_{im}\epsilon_{jn}x_m x_n \epsilon_{ik}\epsilon_{jl}\partial_k\partial_l\chi = \frac{\partial^2\chi}{\partial r^2} \tag{94}$$

compare Eqs. (61). The consistency of these results can be checked by calculating

$$\sigma_{xx} + \sigma_{yy} = \nabla^2\chi = \frac{1}{r}\frac{\partial\chi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\chi}{\partial\phi^2} + \frac{\partial^2\chi}{\partial r^2} \tag{95}$$

Since the trace of a tensor is invariant in any coordinate system, $\sigma_{xx} + \sigma_{yy} = \sigma_{rr} + \sigma_{\phi\phi}$, in agreement with Eqs. (93) and (94).

A.3 Stress Function of a Planar Disclination

The case of a disclination is more subtle than of a dislocation, because the boundary conditions cannot be so cavalierly neglected [32]. For an isolated disclination, the solution of Eq. (57) is given by

$$\chi = \frac{K_0 s}{8\pi}(Ar^2 + r^2\ln(r)) \tag{96}$$

Without the term Ar^2, which is a solution of the homogeneous equation $\nabla^2\nabla^2\chi = 0$, Eq. (96) would imply stresses and strains at the boundary that diverge as $\ln(R)$. This is obviously unacceptable, since Eq. (38) is a harmonic approximation, which is only valid for small strains and small stresses. The requirement of σ_{rr}, the rr-component of the stress tensor, to vanish at the boundary $r = R$ determines $A = -0.5 - \ln(R)$. The Airy stress function in then

$$\chi = \frac{K_0 s}{8\pi}r^2\left(\ln(r/R) - 1/2\right) \tag{97}$$

A.4 Bending Energy of Buckled Disclinations

With the ansatz $h(\mathbf{r}) = r[A + B\sin(2\phi)]$, the bending energy (20) of buckled disclinations can be calculated straightforwardly. In polar coordinates,

$$\nabla^2 h(\mathbf{r}) = \frac{1}{r}\frac{\partial h}{\partial r} + \frac{1}{r^2}\frac{\partial^2 h}{\partial\phi^2} + \frac{\partial^2 h}{\partial r^2} = \frac{1}{r}[A - 3B\sin(2\phi)] \tag{98}$$

This result is inserted into Eq. (20) for the curvature energy,

$$E_b = \frac{\kappa}{2}\int dS(\nabla^2 h)^2 = \frac{\kappa}{2}\int_0^{2\pi} d\phi \int_a^R dr\, r(\nabla^2 h)^2 = \frac{\pi\kappa}{2}\left(2A^2 + 9B^2\right)\ln(R/a) \tag{99}$$

Thus, for a positive disclination with $A = \sqrt{s/\pi}$ and $B = 0$,

$$E_b = s\kappa\ln(R/a) \tag{100}$$

while for a negative disclination with $A = 0$ and $B = \sqrt{2|s|/(3\pi)}$,

$$E_b = 3|s|\kappa\ln(R/a) \tag{101}$$

compare Eqs. (72) and (74), respectively.

References

[1] D. Voet and J. G. Voet, *Biochemistry*, 2nd ed. (Wiley, New York, 1995).

[2] C. Starr, *Biology: Concepts and Applications*, 2nd ed. (Wadsworth, Belmont, CA, 1994).

[3] Southwest Biotechnology and Informatics Center, New Mexico State University.

[4] Source unknown.

[5] V. Reddy *et al.*, J. Virol. **75**, 11943 (2001).

[6] P. L. Stewart, R. M. Burnett, and S. D. Fuller, Cell **67**, 145 (1991).

[7] J. Lidmar, L. Mirny, and D. R. Nelson, Phys. Rev. E **68**, 051910 (2003).

[8] *Statistical Mechanics of Membranes and Surfaces*, edited by D. Nelson, T. Piran, and S. Weinberg (World Scientific, Singapore, 1989).

[9] G. Gompper and M. Schick, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and J. Lebowitz (Academic Press, London, 1994), Vol. 16, pp. 1–176.

[10] *Micelles, Membranes, Microemulsions, and Monolayers*, edited by W. M. Gelbart, A. Ben-Shaul, and D. Roux (Springer-Verlag, Berlin, 1995).

[11] S. A. Safran, *Statistical Thermodynamics of Surfaces, Interfaces, and Membranes* (Addison-Wesley, Reading, MA, 1994).

[12] *Structure and dynamics of membranes - from cells to vesicles*, Vol. 1 of *Handbook of biological physics*, edited by R. Lipowsky and E. Sackmann (Elsevier, Amsterdam, 1995).

[13] *Fluctuating geometries in statistical mechanics and field theory*, edited by F. David, P. Ginsparg, and J. Zinn-Justin (Elsevier, Amsterdam, 1996).

[14] U. Seifert, Adv. Phys. **46**, 13 (1997).

[15] G. Gompper and D. M. Kroll, J. Phys. Condens. Matter **9**, 8795 (1997).

[16] M. J. Bowick and A. Travesset, Phys. Rep. **344**, 255 (2001).

[17] *Soft Matter — Complex Materials on Mesoscopic Scales*, Vol. 10 of *Matter and Materials*, edited by J. K. G. Dhont, G. Gompper, and D. Richter (Forschungszentrum Jülich, 2002).

[18] D. H. Boal, *Mechanics of the Cell* (Cambridge University Press, Cambridge, 2002).

[19] G. Gompper and D. M. Kroll, in *Statistical Mechanics of Membranes and Surfaces*, 2nd ed., edited by D. R. Nelson, T. Piran, and S. Weinberg (World Scientific, Singapore, 2004).

[20] E. Kreyszig, *Differential Geometry* (Dover Publications, New York, 1991).

[21] P. B. Canham, J. Theor. Biol. **26**, 61 (1970).

[22] W. Helfrich, Z. Naturforsch. **28c**, 693 (1973).

[23] H. P. Duwe, J. Käs, and E. Sackmann, J. Phys. France **51**, 945 (1990).

[24] P. Méléard *et al.*, Biophys. J. **72**, 2616 (1997).

[25] H.-G. Döbereiner *et al.*, Phys. Rev. Lett. **91**, 048301 (2003).

[26] W. Helfrich, Z. Naturforsch. **33a**, 305 (1978).

[27] P.-G. de Gennes and C. Taupin, J. Phys. Chem. **86**, 2294 (1982).

[28] L. D. Landau and E. M. Lifshitz, *Theory of Elasticity*, 3rd ed. (Pergamon Press, Oxford, 1986).

[29] J. M. Kosterlitz and D. J. Thouless, J. Phys. C **6**, 1181 (1973).

[30] B. I. Halperin and D. R. Nelson, Phys. Rev. Lett. **41**, 121 (1978).

[31] A. P. Young, Phys. Rev. B **19**, 1855 (1979).

[32] H. S. Seung and D. R. Nelson, Phys. Rev. A **38**, 1005 (1988).

[33] F. R. N. Nabarro, *Theory of Crystal Dislocations* (Clarendon, Oxford, 1967).

[34] D. R. Nelson, in *Phase Transitions and Critical Phenomena*, edited by C. Domb and J. Lebowitz (Academic Press, London, 1983), Vol. 7, pp. 1–99.

[35] D. R. Nelson and L. Radzihovsky, Phys. Rev. A **46**, 7474 (1992).

[36] G. Gompper and D. M. Kroll, Europhys. Lett. **58**, 60 (2002).

[37] J.-M. Park and T. C. Lubensky, Phys. Rev. E **53**, 2648 (1996).

[38] T. A. Witten and H. Li, Europhys. Lett. **23**, 51 (1993).

[39] A. Lobkovsky *et al.*, Science **270**, 1482 (1995).

[40] A. Lobkovsky, Phys. Rev. E **53**, 3750 (1996).

[41] J. J. Thomson, Phil. Mag. **7**, 237 (1904).

[42] A. Perez-Garrido and M. A. Moore, Phys. Rev. B **60**, 15628 (1999).

[43] M. J. Bowick, D. R. Nelson, and A. Travesset, Phys. Rev. B **62**, 8738 (2001).

[44] M. J. Bowick, A. Cacciuto, D. R. Nelson, and A. Travesset, Phys. Rev. Lett. **89**, 185502 (2002).

[45] A. Bausch *et al.*, Science **299**, 1716 (2003).

[46] Y. Kantor and D. R. Nelson, Phys. Rev. Lett. **58**, 2774 (1987).

[47] Y. Kantor and D. R. Nelson, Phys. Rev. A **36**, 4020 (1987).

[48] G. Gompper and D. M. Kroll, J. Phys. I France **6**, 1305 (1996).

[49] G. Gompper and D. M. Kroll, J. Phys. I France **7**, 1369 (1997).

[50] G. Gompper and D. M. Kroll, Eur. Phys. J. E **1**, 153 (2000).

[51] F. Crick and J. D. Watson, Nature **177**, 473 (1956).

[52] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).

[53] M. S. Kellermayer, C. Bustamante, and H. L. Granzier, Biochim. Biophys. Acta **1604**, 105 (2003).

[54] L. Brodin, P. Löw, and O. Shupliakov, Curr. Opin. Neurobiol. **10**, 312 (2000).

[55] J. A. Ybe *et al.*, Nature **399**, 371 (1999).

[56] J. Heuser, J. Cell Biol. **108**, 401 (1989).

[57] M. D. Johnson, *Human Biology: Concepts and Current Issues* (Benjamin Cummings, New York, 2001).

[58] R. Lipowsky, J. Phys. II France **2**, 1825 (1992).

[59] R. Lipowsky, Biophys. J. **64**, 1133 (1993).

[60] F. Jülicher and R. Lipowsky, Phys. Rev. Lett. **70**, 2964 (1993).

[61] F. Jülicher and R. Lipowsky, Phys. Rev. E **53**, 2670 (1996).

[62] W. T. Góźdź and G. Gompper, Europhys. Lett. **55**, 587 (2001).

[63] T. Kohyama, D. M. Kroll, and G. Gompper, Phys. Rev. E **68**, 061905 (2003).

D 3 Fluid vesicles

U. Seifert
II. Institut für Theoretische Physik
Universität Stuttgart

Contents

1 Introduction 2

2 Shapes 3
2.1 Physics of the curvature model 3
2.2 Phase diagram of a minimal model 5
2.3 Refined curvature models 7
2.4 Thermal fluctuations 8

3 Adhesion 9
3.1 Two length scales 9
3.2 Macroscopic model 10
3.3 Mesoscopic description 12
3.4 Unbinding transition 13
3.5 Gravity 14

4 Dynamics 14
4.1 Classification and dissipative mechanisms 14
4.2 Equation of motion 15
4.3 Equilibrium dynamics 18
4.4 Vesicle in shear flow 18
4.5 Dynamically induced unbinding 20

5 Perspectives 21

A Differential geometry for vesicles 22

B Steric of fluctuation potential 23

C The Oseen-tensor 24

1 Introduction

Vesicles constitute a well-defined model system for studying basic physical aspects of the more complex biological cells. These cells consist of a structured inhomogeneous interior, where the nucleus, cell organelles such as ribosomes, mitochondria and the golgi apparatus are embedded in a highly viscous complex fluid formed by all sorts of proteins in an ionic aequous solution. Cells are bounded by the cell membrane, which itself has a complex architecture, sketched in Fig. 1. Its main element is a bilayer of various types of lipids into which proteins such as ion channels and pumps are embedded. The sugar polymers of the outer glycocalix are anchored to this membrane as well as an actin-spectrin network is attached to it on the cellular side in the case of red-blood-cells.

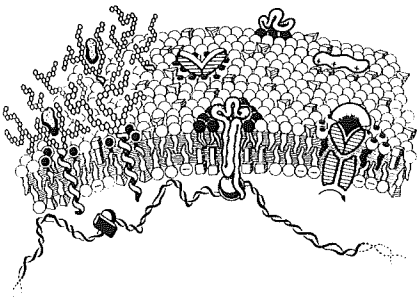


Fig. 1: *Cartoon of the cell membrane*

Vesicles, to the contrary, are formed by a closed lipid bilayer only. Lipid molecules consist of a hydrophilic dipolar or charged head group to which two hydrophobic hydrocarbon chains are attached. In a bilayer these chains are shielded from the aqueous environment. Bilayers form spontaneously in a process of self-aggregation (or self-organization) if lipids are dissolved in aequous solution due to their amphiphilic nature. Given some experience, unilamellar vesicles with sizes of several micron diameter can be obtained. Once a nice vesicle is formed, it can be studied using videomicroscopy, see Fig. 2. Such observations have revealed an amazing variety of shapes, some of which are reminiscent on the shapes of red-blood-cells, despite the fact that vesicles lack so much of the complexity of cells and, in particular, their non-equilibrium nature. A selection of these shapes of various symmetries and topologies is shown in Figs. 3-6.
This variety of shapes among which transformation have been induced by changing the temperature or the osmotic condition prompted in the late eighties an intense and finally successful search for a quantitative physical explanation of these phenomena.
In this lecture, I will outline the basic physical principles of fluid vesicles and their theoretical description focussing on three main topics. In Section 2, the curvature model will be introduced that leads to a quantitative understanding of the whole zoo of vesicle shapes and their thermal fluctuations. Section 3 treats the adhesion or interaction of a vesicle with a substrate. Section 4 deals with the dynamics of vesicles in particular in hydrodynamic flow fields. Such non-equilibrium aspects of vesicles have become increasingly important starting in the mid-nineties.
These lecture notes are mainly pedagogical with only a few selected references. A systematic and comprehensive presentation of the physics of fluid vesicles can be found in the review article [1].

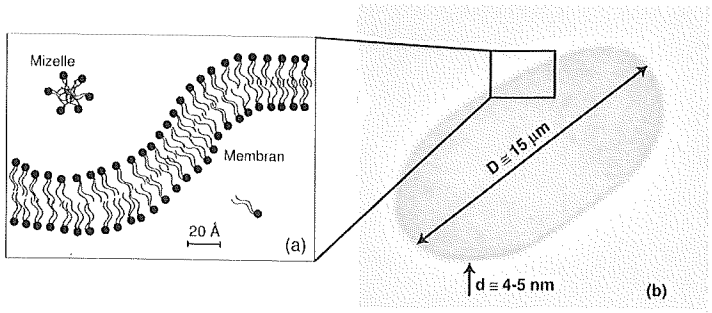


Fig. 2: *Sketch of a lipid bilayer membrane and phase contrast microscopy picture of a giant vesicle (courtesy of H.-G. Döbereiner).*

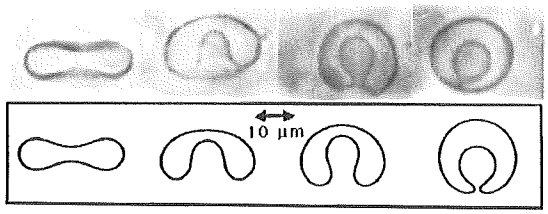


Fig. 3: *Discocyte-stomatocyte transition. With increasing temperature, the up/down symmetric discocyte (left shape) turns into a symmetry-broken stomatocyte. The contours below are calculated based on the bilayer-couple model* [2, 3].

2 Shapes

2.1 Physics of the curvature model

For a physical approach for understanding the factors which determine the variety of vesicle shapes and their transformation we search for an appropriate (free) energy whose minimization yields these shapes. For a symmetric membrane, the chemical composition and environment of both monolayers is identical. Therefore, the flat conformation is locally the state of lowest energy. For a closed configuration, which is necessarily non-flat, the selection of the correct energy has to be guided by the essential physical properties of closed bilayer membranes.

Length-scale separation and curvature model: The bilayer is about 4 nm thick, a giant vesicle is about 10^4 times larger. This separation of length-scales allows to describe the membrane as a two-dimensional surface $\mathbf{R}(s_1, s_2)$ parametrized by two internal coordinates (s_1, s_2) embedded in the three-dimensional space. Any such surface can locally be characterized by the two radii of curvature R_1 and R_2 from which one forms the mean curvature, $H \equiv (1/R_1 + 1/R_2)/2$, and the Gaussian curvature, $K \equiv 1/R_1R_2$, see Fig. 7. The detailed mathematics is recalled in Appendix A. The basic assumption of the curvature model is that bending such a membrane costs locally the energy [7, 8]

$$f \equiv (\kappa/2)(2H)^2 + \kappa_G K, \tag{1}$$

where κ and κ_G are the ordinary and the Gaussian bending rigidity, respectively. These material parameters have dimension of energy (since f is an energy/area). For phospholipid membranes,

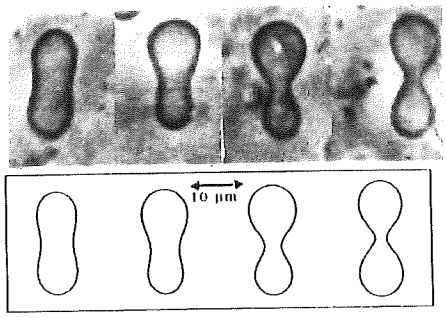


Fig. 4: *Prolate and pear-like shapes* [2, 3].

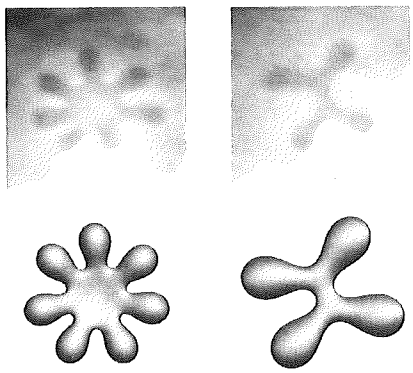


Fig. 5: *Starfish vesicles as observed experimentally (top) and modelled theoretically (bottom)* [4].

a typical value is 10^{-12} erg $= 10^{-19}$ J.

Large edge energy and fixed topology: Any open edge of a membrane patch exposed to water costs a large energy. This is the reason why free membrane patches do not exist. Likewise, the topology of a vesicle does usually not change since this would imply to form transient edges. Therefore, we have to consider only closed membrane configurations, i.e., vesicles, for which the energy

$$F_\kappa \equiv (\kappa/2) \oint dA(2H)^2 \tag{2}$$

is given by the integral of (1) over its closed surface. Due to the mathematical Gauss-Bonnet theorem the integral of the second term of (1) depends only on the topology (i.e. number of handles g) of a vesicle but not on its shape, i.e. $\oint KdA = 4\pi(1-g)$. Therefore, this second term can be discarded as long as we consider vesicles of fixed topology.

Fluidity, insolubility and incompressibility: Since the membrane is in the fluid state of matter, it cannot withstand shear in its plane. Moreover, the solubility of the double chain phospholipids is extremely low. Therefore, there is practically no exchange of material between membrane and solution. This fact together with the quite small compressibility of the membrane implies that for almost all vesicle phenomena the membrane can be considered as locally incompressible. For a closed vesicle, the total area A is thus fixed.

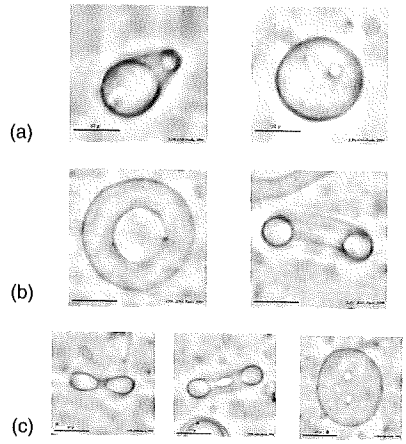


Fig. 6: *Shapes of non-spherical topology. a) a non-axisymmetric torus, b) an axi-symmetric circular torus, c) the "button" surface of genus 2* [5, 6].

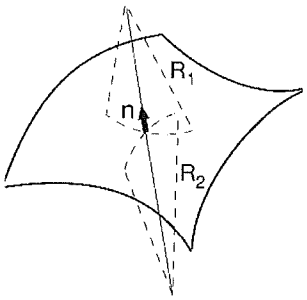


Fig. 7: *Local radii of curvature R_1 and R_2 and normal vector* $\mathbf{n}$ *of a membrane patch.*

Osmotic control of volume: Experimentally, one controls in these systems often the osmotic conditions of the external solution. Even if one works with "pure water", there are necessarily osmotically active impurities around. Since the membrane is permeable for water, the volume changes up the point where practically no more osmotic pressure difference acts. The bending moments that can be derived from (2) are just too weak to balance any non-zero osmotic pressure. Therefore, vesicle configurations must be determined at a prescribed volume V.

2.2 Phase diagram of a minimal model

We are now looking for the shapes which minimize the curvature energy (2) for given area A and volume V. These constraints are added with Lagrange multipliers Σ and P to the energy (2) which leads to

$$F \equiv F_\kappa + \Sigma A + PV. \tag{3}$$

The minimization of (3) must be performed numerically. This is done by solving the Euler-Lagrange-equations, $\delta^1 F = 0$, in a suitable parametrization of an axisymmetric shape [9]. This procedure leads to stationary shapes which are local minima and saddle points in the shape

space, among which careful inspection of the energy diagrams allows to distinguish. The stationary shapes depend only on one parameter which is the reduced volume

$$v \equiv V/[(4\pi/3)(A/4\pi)^{3/2}]. \tag{4}$$

The value of the bending rigidity κ will not affect the result at this stage, since κ is the only energy scale in the problem.

It turns out that there are three types or "branches" of local minima: prolates/dumbbells, oblates/discocytes and cup-shaped stomatocytes, see Figs. 8,9. Obviously, this minimal model captures important aspects of vesicle shapes since we obtain theoretically shapes found previously in the lab. However, something important is missing since this minimization does not lead to pear-like shapes as shown in Fig. 4 or starfish vesicles as shown in Fig. 5. The very observation of these shapes proves that the model described so far is incomplete.

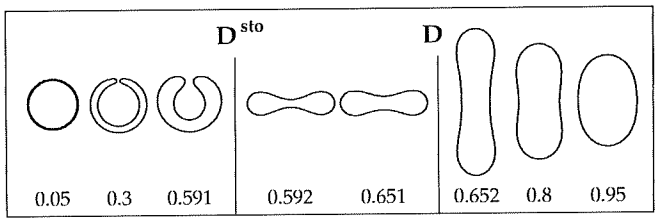


Fig. 8: *Shapes of minimal curvature energy F_κ for various values of the reduced volume v. The shapes of lowest energy are prolate for $v \gtrsim 0.65$, oblate for $0.65 \gtrsim v \gtrsim 0.59$, and stomatocyte for $v \lesssim 0.59$* [9].

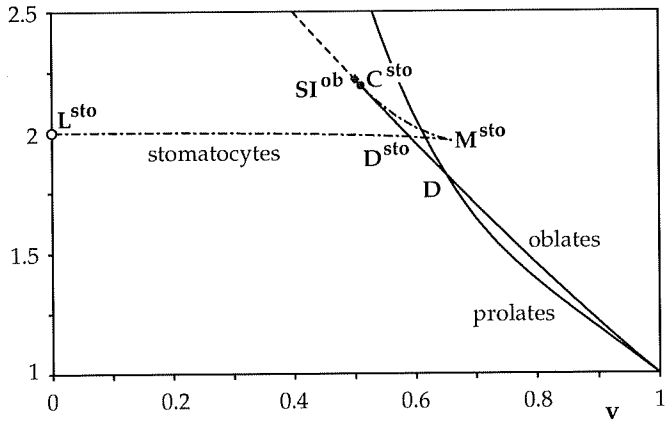


Fig. 9: *Curvature energy $F_\kappa/8\pi\kappa$ of the stationary shapes $[\delta^1 F = 0]$ as a function of the reduced volume v. D and $D^{\rm sto}$ denote the two discontinuous transitions shown in Fig. 8. The further capital letters denote distinct points described in detail in* [9].

2.3 Refined curvature models

The missing ingredient in (2) is the fact that it is not possible to describe the membrane just as a structure-less two dimensional surface. One key signature of its molecular architecture must be retained and that is the bilayer feature [10, 11]. The two monolayers are tightly coupled but the exchange of lipid molecules between them is slow compared to the experimental time-scale of these observations. Therefore, the numbers N^+ and N^- of molecules in the outer and inner monolayer, respectively, are conserved. The number difference $N^+ - N^-$ leads to a prefered area-difference

$$\Delta A_0 \equiv (N^+ - N^-)a_0 \tag{5}$$

between the two layers. Here, a_0 is the equilibrium area per lipid molecule. The actual area difference, $\Delta A = A^+ - A^-$, can be expressed as an integral over the mean curvature as

$$\Delta A = 2D \oint dA\, H \tag{6}$$

where D is the distance between the neutral surfaces of the two monolayers, i.e., roughly half the bilayer thickness. If the area of each molecule was strictly conserved, ΔA would be conserved too. This would lead to a third constraint on the total mean curvature

$$M \equiv \oint H dA. \tag{7}$$

Such a model had become popular under the name of the bilayer couple model [11, 2]. Detailed comparisons with experiments and a thorough theoretical analysis of bilayer bending finally revealed that this hard constraint on ΔA must be softened [12]. The correct form of the bilayer feature is therefore to add to F_κ an area difference elasticity,

$$F_{ADE} = \frac{\alpha\pi\kappa}{2D^2 A}(\Delta A - \Delta A_0)^2, \tag{8}$$

where α is somewhat model dependent but close to 1.

Minimizing the total energy,

$$F \equiv F_\kappa + F_{ADE}\,, \tag{9}$$

now leads indeed to pears, shapes and a multitude of starfish vesicles [12, 13, 4]. In this model, the minimal shapes depend not just on the reduced volume v but on one more parameter

$$\bar{m}_0 \equiv \Delta A_0 / 2DR, \tag{10}$$

which is basically a scaled "frozen-in" number difference of lipid molecules in the two layers. The shapes of lowest energy can be arranged in a two-dimensional phase diagram [12, 13], see Fig. 10. As one changes temperature or the osmotic conditions, the parameters v and $\bar{m}_0$ change due to thermal expansion of the area [9]. Thus, a trajectory in this phase diagram cuts across the different region of stability of the different shapes. Transitions from one shape to another, say from prolate to pear, can be abrupt (i.e., discontinuous, or first order) or smooth (i.e. continuous or second order). All experiments performed so far on the shapes of such vesicles and their transitions seem to be quantitatively compatible with this theoretical model.

The discussion so far has been confined to membranes which are symmetric with respect to both the chemical composition of the two leaflets of the bilayer and the liquid environment.

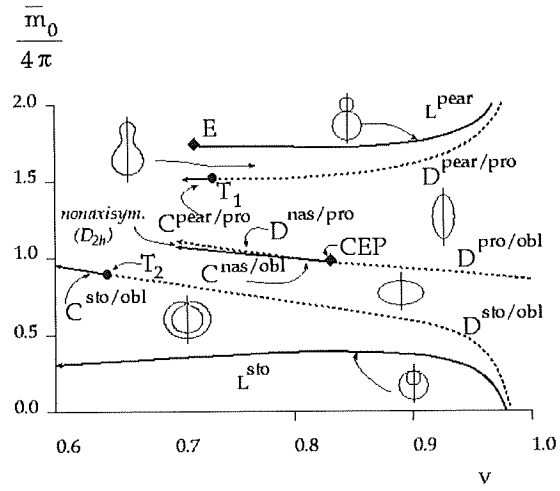


Fig. 10: *Phase diagram of the ADE model. Shown are the shapes of lowest energy as a function of the reduced volume v and the scaled prefered area-difference $\bar{m}_0$ for $\alpha = 1.4$. The capital letters denote shape transitions, limiting lines and special points. The region of stable starfish vesicles lies at smaller v* [13] .

Originally, Helfrich suggested a spontaneous curvature model in which $(2H)$ in (2) is replaced by $(2H - C_0)$ [8, 14]. In this case, C_0 is the curvature that a free membrane patch cut from a vesicle in a Gedanken-experiment would acquire spontaneously. One can show that minimizing such a modified energy (2) leads to the same multitude of shapes as minimizing (9) although with different stability properties [12, 15]. Even for a vesicle under such asymmetric conditions the second energy term in (9) must be retained if its two leaflets do not exchange molecules. In the latter case, the spontaneous curvature and the area difference term can then be combined into an "effective" spontaneous curvature.
Minimization of the curvature energy happens at fixed topology, so far spherical topology. For vesicles of higher topology as those shown in Fig. 6, similar phase diagrams have been calculated [16, 17]. For surfaces with genus $g \geq 2$ an interesting new phenomenon occurs. The fact that F_κ in (2) is invariant under conformal transformation of the three-dimensional space, leads to a continuous degeneracy of the shape of minimal energy at fixed v and $\bar{m}_0$. This phenomenon, called conformal diffusion, has been first predicted theoretically [17] and later been verified experimentally [4].

2.4 Thermal fluctuations

The vesicle shapes obtained by minimizing the appropriate curvature energy subject to the geometric constraints formally correspond to solving a zero temperature problem. Video microscopy reveals that these shapes typically exhibit visible thermal fluctuations [18]. How can

these fluctuations be described?
It is instructive to start from a planar membrane. The local position of such a membrane with respect to the mean plane can be parametrized by a height variable $h(x, y)$. The energy (2) then becomes

$$F_\kappa = \frac{\kappa}{2}\int dxdy(\nabla^2 h)^2 = \frac{\kappa}{2}\int \frac{d^2q}{(2\pi)^2} q^4 h_{\mathbf{q}} h^*_{\mathbf{q}} \tag{11}$$

where the Fourier transformation is defined as

$$h(\mathbf{x}) \equiv \int \frac{d^2q}{(2\pi)^2} h_{\mathbf{q}} e^{i\mathbf{qx}} \,. \tag{12}$$

Invoking the equipartition theorem [19] then tells us the mean amplitudes at temperature T

$$< |h_{\mathbf{q}}|^2 >= \frac{k_B T}{\kappa q^4}, \tag{13}$$

where $< ... >$ is the thermal expectation value taken with the Boltzmann weight, $\exp(-F_\kappa/k_B T)$.
In principle, one could attempt to calculate the fluctuations around any of the zero temperature ($T = 0$) shapes using a similar approach. In practice, the differential geometry involved is tricky and taking care of the geometrical constraints is not entirely trivial. The latter problem can be solved by introducing an effective tension and pressure difference whose numerical value are given by the Lagrange multiplier Σ and P, used to implement the area and volume constraints, respectively [15]. The former must be attacked numerically which is somewhat cumbersome. So far, only a few attempts have been made to calculate such fluctuation spectra around non-trivial ($T = 0$) shapes [20, 21, 22]. The notable exception are quasi-spherical vesicles from whose fluctuation spectrum the value of the bending rigidity can be deduced by comparing data to theory, for relatively recent work, see e.g. [23, 24].
In a very recent development, fluctuation around non-spherical shapes can be simulated numerically and then be compared to experimental spectra. This technique allows a precise determination of membrane parameters like bending rigidity and spontaneous curvature [25].

3 Adhesion

3.1 Two length scales

From the point of view of having a well-defined system, an isolated vesicle freely floating in solution is almost perfect. For interesting applications or a higher level in complexity, interaction of the vesicle with other objects should be considered. The conceptually simplest case corresponds to adhesion of a vesicle to a rigid substrate. For this case, a comprehensive set of experimental data can be obtained using reflection interference microscopy [26, 27]. This technique allows not only to study the gross features of a bound vesicle such as its contour radius or the size of the adhesion disc but also the fluctuation spectrum of its bound part. Such experiments show clearly that the problem of a vesicle adhering to a substrate has two complementary aspects, a macroscopic one (overall shape) and a mesoscopic one (fluctuation spectrum in the vicinity of the substrate), see Fig. 11. Recognizing in a first step this distinction, is key for a successful theoretical modeling of this problem. In a second step, the two aspects must be reconciled self-consistently.

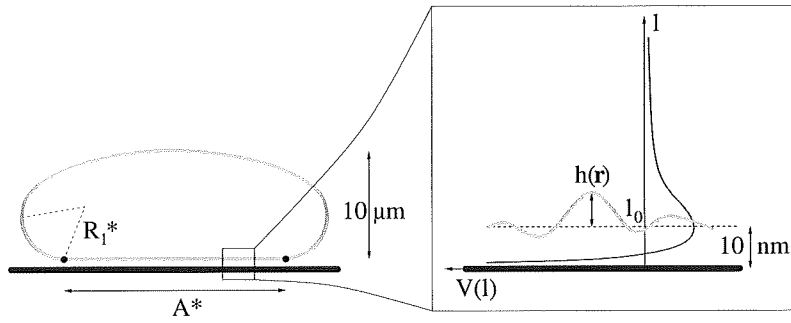


Fig. 11: *Adhesion of vesicles: Macroscopic view with contact area A^* and contact curvature $1/R_1^*$ and mesoscopic view with the bound part fluctuating in a potential $V(l)$ which has a minimum typically $\simeq 10$ nm above the substrate.*

3.2 Macroscopic model

From the macroscopic perspective, a minimal model for adhesion of a vesicle to a substrate consists in adding an adhesion energy

$$F_a \equiv -WA^* \tag{14}$$

to the bending energy (2) [28]. Here, W is the contact potential and A^* the size of the adhesion patch with which the vesicle adheres to the substrate. Such a contact potential enters the boundary condition at the point of contact. First, the contact angle is necessarily π since any sharp bend would have an infinite curvature energy. This implies that the membrane is curved only in one direction and $1/R_2^* = 0$ along the line of contact. Second, the contact curvature obeys

$$1/R_1^* = (2W/\kappa)^{1/2} , \tag{15}$$

which follows from minimizing with respect to the area of contact [28]. This boundary condition, which does not depend on the area of the vesicle, holds for all variants of the curvature model. Depending on both the size of the vesicle,

$$R \equiv (A/4\pi)^{1/2} , \tag{16}$$

and the strength of the adhesion potential W, adhesion can either be "weak" or "strong". The crucial quantity is the scaled strength of the adhesion potential

$$w \equiv WR^2/\kappa . \tag{17}$$

Typical bound shapes as those shown in Fig. 12 demonstrate that there are two regimes.

Weak adhesion (w of order 1): It turns out, somewhat surprisingly, that one needs a threshold value w_c for adhesion to occur. If the adhesion strength is too small, the possible gain in adhesion energy is smaller that the cost in deforming its shape so that the vesicle will not adhere even in the presence of the weakly attractive substrate [28]. The specific value of w_c depends on the reduced volume but is of order 1, see Fig. 13 [29]. For $w > w_c$, the vesicle is bound. Its contact area A^* increases with increasing strength of the potential W.

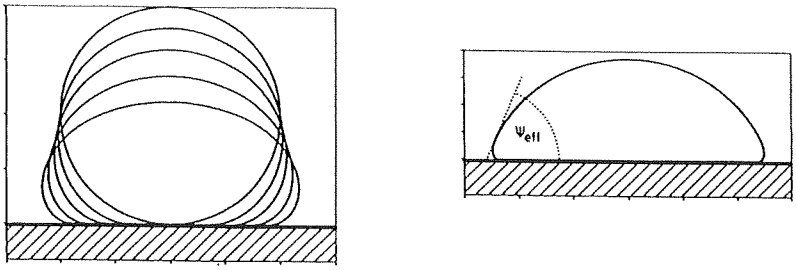


Fig. 12: *Calculated shapes of bound vesicles. (Left): regime of weak adhesion; (right): regime of strong adhesion. The latter shape is almost spherical cap-like with an effective contact angle* Ψ_{eff} [28].

Strong adhesion ($w >> 1$)*:* Physically, in this regime, the bending energy becomes irrelevant and the shape is determined solely by the desire to maximize the contact area A^* for given total area and enclosed volume. The optimal shape then is a spherical cap with an effective contact angle $\Psi_{\text{eff}}(v)$ determined by the geometrical constraints, see Fig. 12. The explicit relation $\Psi_{\text{eff}}(v)$ is given in [30]. The limiting cases are $\Psi_{\text{eff}} = \pi$ for a sphere ($v = 1$) and $\Psi_{\text{eff}} = 0$ for a flat pancake ($v = 0$). Interestingly, this effective contact angle also obeys a Young equation [28],

$$W = \Sigma(1 + \cos \Psi_{\text{eff}}). \tag{18}$$

In contrast to adhesion of liquid droplets where Σ would be the surface tension, for vesicle adhesion Σ is not an independent quantity but rather given by the (numerical) value of the Lagrange multiplier used to implement the area constraint. As (18) shows, Σ is a function of both the reduced volume (through $\Psi_{\text{eff}}(v)$) and of the adhesion strength W.

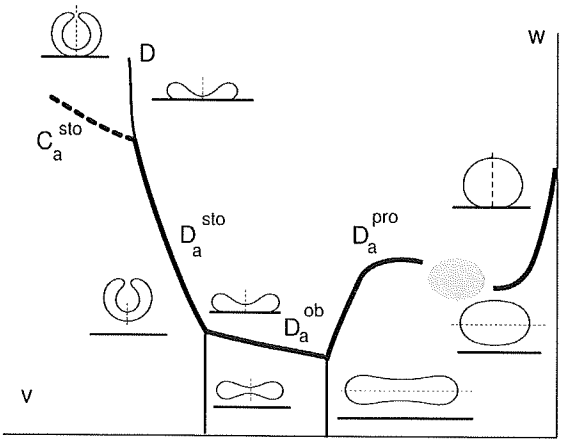


Fig. 13: *Semiquantitative phase diagram for adhering vesicles as a function of adhesion energy* w *and reduced volume* v*. The thick lines denote the curvature driven adhesion transition* $w_c(v)$*. Depending on the reduced volume* v*, bound and free shapes can be prolate, oblate or stomatocyte-like* [29].

Experimentally, the difference between strong and weak adhesion has clearly been seen. With light microscopy [31], the mutual interaction of vesicles has been studied and large rounded contact regions corresponding to weak adhesion were found. Spherical cap-like configurations have been seen in micro-pipet experiments [32] as well as in freeze-fracture electron microscopy for smaller vesicles [33].

From the macroscopic point of view, the problem of a vesicle adhering to a substrate is thus solved if we know the strength of the contact potential W which must be extracted from a more mesoscopic approach.

3.3 Mesoscopic description

The bound part of a closed vesicle is typically several nanometers away from the substrate and still shows thermal fluctuations [27]. These conditions require that we refine our modeling by paying due attention to the mesoscopic interactions between membrane and substrate. As reviewed in [34], various forces are involved in this interaction such as van der Waals interaction, electrostatic contributions in the case of charged membranes, and short range repulsive forces such as the hydration force or protrusion forces. All these direct forces can be combined into an effective potential $V_d(l)$. This quantity gives the energy of a (fluctuationless) planar membrane at a distance l from the substrate. Typically $V_d(l)$ increases dramatically for small l and vanishes for large l. In between, there are either one or two minima. If these direct forces were the full story, we could take as the effective adhesion strength of the macroscopic approach the value of the direct potential at the minimum l_0, i.e. $W = -V_d(l_0)$. However, the situation is more complex due to thermal fluctuations.

The presence of the substrate confines its fluctuations [35]. As derived in Appendix B, this effect can be described by a steric or fluctuation potential [30]

$$V_f(\Sigma; l) = \left(\frac{6b^2(k_BT)^2}{\kappa l^2}\right)\left(\frac{y^2}{\sinh^2 y}\right) \quad \text{with } y \equiv (\Sigma/bT)^{1/2} l/2 \text{ and } b \simeq 1/2\pi. \tag{19}$$

This potential expresses the increase in free energy due to a decreasing entropy for thermal fluctuations with increasing confinement. It diverges as $1/l^2$ for $l \to 0$ and vanishes for $l \to \infty$.

Within a superposition approach, the fluctuation potential $V_f(\Sigma; l)$ should be added to the direct potential $V_d(l)$ to form the total potential $V(l) \equiv V_d(l) + V_f(\Sigma; l)$. The energy of the bound part of the vesicle, which is at a local distance $l(x, y)$ from the substrate, is then given by

$$F_a = \int dA \left\{\frac{\kappa}{2}(\nabla^2 l)^2 + V(l) + \Sigma\right\} \approx \frac{1}{2}\int \frac{d^2q}{(2\pi)^2}\left(\kappa q^4 + (\Sigma - W)q^2 + V''(l_0)\right)|h_{\mathbf{q}}|^2. \tag{20}$$

In the first equality we have to include Σ from the area constraint. The second equality follows after expanding for small fluctuations,

$$h(x, y) \equiv l(x, y) - l_0, \tag{21}$$

around the minimum at l_0. The area element dA becomes

$$dA \approx dxdy(1 + (\nabla h)^2/2) \tag{22}$$

and for the potential $V(l)$ an expansion reads

$$V(l) \approx -W + V''(l_0)h^2/2. \tag{23}$$

The prefactor of the q^2 term in (20) is the effective tension

$$\bar{\Sigma} \equiv (\Sigma - W) = -W\frac{\cos\Psi_{\rm eff}}{1+\cos\Psi_{\rm eff}}, \tag{24}$$

which carries two contributions. The first arises from the area constraint, which acts like an external pulling force at the bound part of the membrane. The second one describes the fact that a fluctuating membrane exposes more area to the attractive potential minimum [36]. The final equality, which arises from using (18), shows that for strong adhesion there is a universal relation between the effective adhesion energy W and the effective tension $\bar{\Sigma}$ acting on the fluctuations of the bound part. This relation depends only on the reduced volume v which determines $\Psi_{\rm eff}$. This relation states in particular that the effective tension becomes negative for those vesicles with a relatively small reduced volume for which $\Psi_{\rm eff} < \pi/2$. Such a negative tension should show up in the measurable spectrum of fluctuations of the bound part [27], which can be obtained using equipartition on (20) as

$$< |h_{\mathbf{q}}|^2 >= \frac{k_B T}{\kappa q^4 + \bar{\Sigma} q^2 + V''(l_0)}. \tag{25}$$

The final step in a self-consistent description consists in determining not only the ratio between the tension Σ and the depth of the adhesion potential W as given by (24) but also their individual value. Since W corresponds to the depth of the total potential at its minimum l_0, one has

$$W = -[V_d(l_0) + V_f(\Sigma; l_0)]. \tag{26}$$

With this equation, the problem of vesicle adhesion taking into account steric repulsion is closed and solvable for any given form of the direct potential $V_d(l)$. For a specific example, see [30], where the distinction between Σ and $\bar{\Sigma}$ was not yet made. This refinement is irrelevant for large reduced volume or tense vesicles which were the focus in [30] but becomes important for small reduced volume where this refinement [36] predicts a negative $\bar{\Sigma}$.

3.4 Unbinding transition

Two mechanisms for an adhesion or unbinding transition for vesicles should be distinguished. Ignoring fluctuations, the competition between adhesion energy and bending energy predicts a scale dependent adhesion transition at $w_c(v)$. For given W, this relation implies that large vesicles with $R > R_c \equiv (w_c\kappa/W)^{1/2}$ are predicted to adhere while smaller ones are free [28]. Physically, the large ones can take advantage of the adhesion energy without having to pay too much in bending.

The second mechanism happens via fluctuations. Even if the direct potential V_d has a minimum at l_0 with $V_d(l_0) < 0$, thermal fluctuations can be so strong that the membrane prefers to be unbound. This unbinding transition has been studied first for free planar tensionless membranes [37, 38]. The theory described above shows that a similar effect also holds for the vesicles except for quite tense ones [30]. In the latter case the fluctuations may be restricted so much that such a vesicle is still bound even though an open membrane patch (with $\Sigma = 0$) would

unbind due to fluctuations. This regime corresponds to genuine tension-induced adhesion [39]. The crossover between the two different mechanisms can be estimated to happen at a length-scale in the 100 nm range [28].

3.5 Gravity

A particular type of adhesion arises from gravitational effects often used in the laboratory to stabilize vesicles at the bottom of a measuring chamber. If a vesicle is filled with a fluid slightly denser than the surrounding aqueous solution, it will sink to the substrate. Even without an explicit adhesion energy, this geometrical confinement leads to shape changes [40]. The phase diagram is shown in Fig. 14 as a function of the reduced volume v and the (dimensionless) gravitational interaction

$$g \equiv g_0 \Delta\rho R^4/\kappa, \tag{27}$$

where $g_0 \simeq 9.81$ m/s^2 and $\Delta\rho$ is the density difference between the aqueous solutions inside and outside the vesicle.

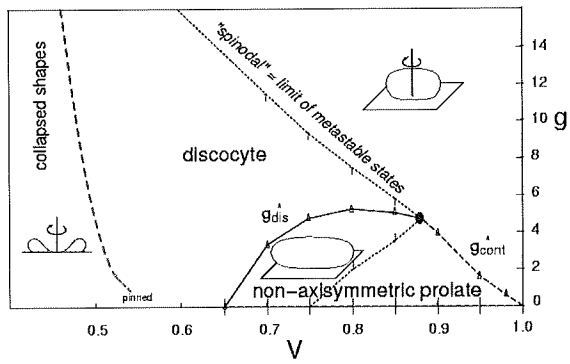


Fig. 14: *Phase diagram including gravity as a function of reduced volume v and gravitational parameter g. In the grey area, the minimal shape is a non-axisymmetric prolate, for larger g an axisymmetric oblate. The transition between the prolate and the oblate is continuous along $g^*_{\rm cont}$ and discontinuous (with associated spinodal) along $g_{\rm dis}$* [40].

4 Dynamics

4.1 Classification and dissipative mechanisms

Characteristic for giant vesicles is their slow dynamics visible in the microscope. The typical time-scale for shape changes on the micron scale lies in the range of a second which makes vesicles a perfect system to study dynamical aspects in real space directly. For a classification of vesicle dynamics, one has to distinguish dynamics in equilibrium from non-equilibrium phenomena.

Equilibrium dynamics of vesicles comprises the dynamical fluctuations around locally stable mean shapes. *Non-equilibrium dynamics* is the more challenging and less well explored regime. One should distinguish two classes. First, there is relaxation into a new equilibrium after a

parameter change. This class comprises the decay of a metastable shape such, e.g., as the budding process induced by raising the temperature [3, 41]. The driving force in this case is the curvature energy which favors a different shape at the new external conditions. Conceptually related is the spectacular pearling instability of cylindrical vesicles, which develops upon action of optical tweezers [42]. In this case, the laser generates a tension for which a homogeneous cylinder is no longer a stable configuration. The second class of nonequilibrium behavior refers to genuine non-equilibrium states of vesicles induced by external fields such as hydrodynamic flow as described below.

Dynamics in the micron world of vesicles is overdamped, i.e. inertial effects can safely be ignored [43]. This realm is often somewhat counter-intuitive to our experience of fluid phenomena shaped mostly by the so-called large Reynolds number regime where inertia matters. For example, in the micron world, motion is reversible and objects come to rest right after the driving forces cease to act. Driving forces for vesicles arise from bending energy and from external fields such as optical tweezers or hydrodynamic flow. Dissipation (or "friction") takes place both in the surrounding liquid and in the membrane, in principle. For giant vesicles of micron size, the dominant dissipation is viscous dissipation in the embedding fluid [44], which requires a full treatment of the hydrodynamics of this fluid.

Dissipation in the membrane can be classified into three phenomena: Drag between the two monolayers [45], shear viscosity within each layer and permeation through the membrane. Calculation of the relaxation spectra of bending fluctuations involving the first two mechanisms show that on scales of microns and larger, hydrodynamic dissipation is dominant [46]. In the submicron range, friction between the layers becomes relevant. On even smaller scales of several tenths of nanometers, shear viscosity within each layer should be included. Finally, permeation through the membrane seems to be irrelevant on all length-scales with the possible exception of membranes in the vicinity of a substrate [47]. Based on this hierarchy of dissipative mechanisms, for giant vesicles it is often permissible to ignore all but viscous dissipation in the surrounding liquid, as we will do in the following.

4.2 Equation of motion

We now derive the equation of motion for a fluid membrane embedded in a viscous liquid [48]. From a conceptual point of view, it is best to start with the Navier-Stokes equation for the surrounding incompressible liquid given by

$$\rho(\partial_t \mathbf{v} + \mathbf{v}\nabla v) = -\nabla p + \eta\nabla^2\mathbf{v} + \mathbf{f}(\mathbf{r}) \tag{28}$$

$$\nabla\mathbf{v} = 0 \tag{29}$$

where $\mathbf{v}(\mathbf{r})$ is the 3-d velocity field of the liquid with density ρ. $p(\mathbf{r})$ is the local pressure which is basically determined by the incompressibility condition (29). η is the shear viscosity of the liquid (with $\eta \simeq 10^{-3}$ Js/m^3 for water at room temperature), and $\mathbf{f}(\mathbf{r})$ is the force density acting on the liquid. Fortunately, we don't have to deal with the full non-linear Navier-Stokes equation because the left-hand side (lhs) of equation (28), the inertia and connective terms, is much smaller than the right-hand side (rhs) for typical membrane dynamics. This can be understood by defining the Reynolds number

$$Re \equiv \left(\frac{\rho v_{\mathrm{typ}}}{T_{\mathrm{typ}}}\right) / \left(\frac{\eta v_{\mathrm{typ}}}{L^2_{\mathrm{typ}}}\right) = \frac{\rho L^2_{\mathrm{typ}}}{\eta T_{\mathrm{typ}}} \tag{30}$$

as the ratio between the lhs and the rhs. Here, $v_{\rm typ}$, $L_{\rm typ}$ and $T_{\rm typ}$ are typical velocities, length-scales and time-scales. Choosing $\rho = 10^3$ Js^2 / m^4 , $L_{\rm typ} = 1\mu$m, $\eta = 10^{-3}$ Js/m^3 we get $Re = 10^{-6}$ for $T_{\rm typ} = 1$ s and $Re = 10^{-3}$ for $T_{\rm typ} = 10^{-3}$s. So on time-scales accessible to video-microscopy, the Reynolds number Re is small and the lhs of (28) can be neglected. This defines the regime of Stokes-flow governed by

$$\begin{aligned} \nabla p - \eta \nabla^2 \mathbf{v} &= \mathbf{f}(\mathbf{r}) \\ \nabla \mathbf{v} &= 0. \end{aligned} \tag{31}$$

The general solution of these inhomogeneous linear equations is given by the super-position of a solution of the homogeneous equation, $\mathbf{v}^{\rm ex}(\mathbf{r})$, which is typically the externally imposed flow field, and a special solution

$$\mathbf{v}^{\rm ind}(\mathbf{r}) = \int d\mathbf{r}' \mathcal{O}(\mathbf{r}, \mathbf{r}') \mathbf{f}(\mathbf{r}'), \tag{32}$$

of the inhomogeneous equation. The Green's function for (31) is called the Oseen-tensor $\mathcal{O}(\mathbf{r}, \mathbf{r}')$. As shown in Appendix C, for infinite flow, the Oseen tensor $\mathcal{O}(\mathbf{r}, \mathbf{r}')$ has Cartesian matrix elements

$$\mathcal{O}_{\alpha\beta}(\mathbf{r}, \mathbf{r}') \equiv \frac{1}{8\pi\eta|\mathbf{r} - \mathbf{r}'|} \left[\delta_{\alpha\beta} + \frac{(r_\alpha - r'_\alpha)(r_\beta - r'_\beta)}{|\mathbf{r} - \mathbf{r}'|^2} \right]. \tag{33}$$

Thus, the hydrodynamics mediates a long-range interaction ($\sim 1/|\mathbf{r} - \mathbf{r}'|$) between the force $\mathbf{f}$ acting at $\mathbf{r}'$ and the velocity $\mathbf{v}$ induced at $\mathbf{r}$. The total velocity field becomes

$$\mathbf{v}(\mathbf{r}) = \mathbf{v}^{\rm ex}(\mathbf{r}) + \mathbf{v}^{\rm ind}(\mathbf{r}). \tag{34}$$

Such a simple superposition is possible since the Stokes equation is linear due to the absence of the convective term.

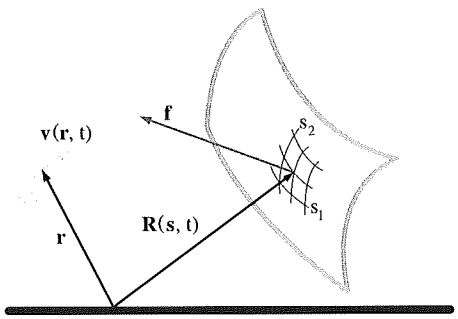


Fig. 15: *Sketch of the hydrodynamic set-up: The membrane at* $\mathbf{r}' = \mathbf{R}(s_1, s_2, t)$ *exerts a force* $f(\mathbf{r}')$ *onto the liquid which affects the velocity* $\mathbf{v}(\mathbf{r})$.

The force density $\mathbf{f}(\mathbf{r}')$ arises from both the embedded membrane and from boundary conditions at confining surfaces, see Fig. 15. If the membrane is locally in a non-equilibrium configuration, it wants to change its shape accordingly which acts as a force on the liquid. The formal

expression of this effect is

$$\mathbf{f}(\mathbf{r}') \equiv -\int ds_1 ds_2 \sqrt{g} \left(\frac{1}{\sqrt{g}} \frac{\delta F}{\delta \mathbf{R}} \right) \delta(\mathbf{r}' - \mathbf{R}(s_1, s_2)). \tag{35}$$

This 3-d force density is localized along the two-dimensional membrane as expressed by the δ-function and the integration over the membrane surface. The driving force arises from the variational derivative (virtual displacement) of the appropriate energy F. In a minimal model

$$F = F_\kappa + F_\Sigma, \tag{36}$$

where F_κ is the curvature energy (2). The second term

$$F_\Sigma \equiv \oint dA \; \Sigma(s_1, s_2) \tag{37}$$

is necessary in the dynamics to ensure local area conservation as discussed below. Explicitly, the force density then reads

$$\left(\frac{1}{\sqrt{g}} \frac{\delta F}{\delta \mathbf{R}} \right) = \left(-2\Sigma H + \kappa[(2H)(2H^2 - 2K) + 2\Delta H] \right) \mathbf{n} - g^{ij} \mathbf{R}_i \partial_j \Sigma. \tag{38}$$

Here,

$$\Delta \equiv (1/\sqrt{g}) \partial_i (g^{ij} \sqrt{g} \partial_j) = g^{ij} \partial_i \partial_j. \tag{39}$$

is the Laplace–Beltrami operator on the surface. The normal part of (38) [49] corresponds to the stationarity condition $[\delta^1 F_\kappa = 0]$ discussed in Section 2. The tangential part arises from inhomogeneities in the surface tension $\Sigma(s_1, s_2)$ which will be needed to ensure local incompressibility of the induced flow.

For given forces $\mathbf{f}(\mathbf{r}')$ we know through (32) and (33) the velocity field $\mathbf{v}(\mathbf{r})$ anywhere in the liquid. Hence, we know $\mathbf{v}(\mathbf{r})$ particularly at the location of the membrane $\mathbf{R}(s_1, s_2; t)$ (where we make the time-dependence of the configuration explicit). Since the membrane is impermeable to water flow, the normal velocity of the fluid at the membrane pushes along the membrane and leads to a shape change. Tangential motion of the fluid along the membrane induces lipid flow within the membrane. As boundary conditions for this tangential flow one uses so-called non-slip boundary conditions for which the tangential velocity of the liquid and the induced lipid flow in the membrane are equal. Therefore, the velocity field leads to an equation of motion for the membrane as

$$\partial_t \mathbf{R}(s_1, s_2, t) = \mathbf{v}(\mathbf{R}(s_1, s_2, t)) = v^{\text{ex}}(\mathbf{R}(s_1, s_2, t)) + \mathbf{v}^{\text{ind}}(\mathbf{R}(s_1, s_2, t)). \tag{40}$$

This deterministic equation of motion includes both, normal motion that signifies a shape change of the membrane, and tangential motion that corresponds to lipid flow within the membrane. The so far unknown local tension $\Sigma(s_1, s_2, t)$ is determined by requiring that the membrane flow induced by this equation obeys local incompressibility. Thus, we must demand that the surface element does not change,

$$\partial_t \sqrt{g} = 0, \tag{41}$$

which implies explicitly

$$g^{ij} \mathbf{R}_i \partial_j (\partial_t \mathbf{R}) = 0. \tag{42}$$

Upon insertion into the equation of motion (40) with (35) and (32), the condition (42) becomes a partial differential equation for the unknown tension $\Sigma(s_1, s_2, t)$.
Equations (40) and (42) yield a deterministic evolution equation for a membrane configuration under the action of bending energy and hydrodynamics [48]. For any given initial membrane configuration, the solution to these equations will run into the next dynamically accessible local minimum of bending energy. In general, these equations must be solved numerically.

4.3 Equilibrium dynamics

For small perturbations around highly symmetric membrane configuration, the above scheme can be solved analytically. This allows the determination of relaxation times for thermal fluctuations in equilibrium. The simplest case is a planar membrane for which dynamical fluctuation can be calculated as follows. The time-dependent membrane configuration is written in Monge representation as

$$\mathbf{R}(x,y,t) = \begin{pmatrix} x \\ y \\ h(x,y,t) \end{pmatrix} \tag{43}$$

If an initial displacement $h(x,y,0) = h_{\mathbf{q}} e^{i\mathbf{qr}}$ with $\mathbf{r} = (x,y)$ is put into the equation of motion (40) with ($\mathbf{v}^{\rm ex} = 0$) one finds after some calculation a relaxation behaviour

$$h(x,y,t) = h(x,y,0) e^{-\gamma_q t} \tag{44}$$

with $\gamma_q = \kappa q^3/\eta$, first derived in [44] slightly differently. According to the usual fluctuation-dissipation relation [19], the dynamical correlation functions follow as

$$< h_q(t) h_q^*(0) > = < |h_q|^2 > e^{-\gamma_q t}, \tag{45}$$

with $< |h_q|^2 > = < h_q(0) h_q^*(0) >$ given by (13). In a quasi-spherical geometry, a similar calculation using spherical harmonics yields relaxation rates $\gamma_{l,m}$ which can be compared to experiments, for some of the latest studies see [50, 51].
For non-spherical shapes, dynamical equilibrium fluctuations have been investigated for prolate shapes in the vicinity of the budding transition [52]. Another particularly interesting example of dynamical fluctuations has been observed at the prolate-oblate transition. Since the activation energy between the prolate and the oblate shapes is just a few $k_B T$, occasionally thermal fluctuations are large enough to drive the vesicle into the other minimum. This system thus constitutes one of the few examples showing a thermally induced macroscopic bistability [53].

4.4 Vesicle in shear flow

A vesicle in infinite shear flow was the first case for which the equation of motion derived in 4.2 was applied numerically [54]. An external shear flow with shear rate $\dot\gamma$ in x-direction

$$\mathbf{v}^{\rm ex}(\mathbf{r}) = \dot\gamma y \mathbf{e}_x = \dot\gamma[(y/2)\mathbf{e}_x + (x/2)\mathbf{e}_y)] - (\dot\gamma/2)\mathbf{e}_z \times \mathbf{r} \equiv \mathbf{v}^{\rm el}(\mathbf{r}) + \mathbf{v}^{\rm rot}(\mathbf{r}) \tag{46}$$

can be decomposed into an elongational component $\mathbf{v}^{\rm el}(\mathbf{r})$ which is tilted at $45°$ and a rotational component $\mathbf{v}^{\rm rot}(\mathbf{r})$, see Fig. 16.
Fig. 17 shows the result of the numerical implementation of the dynamical equations. An initial

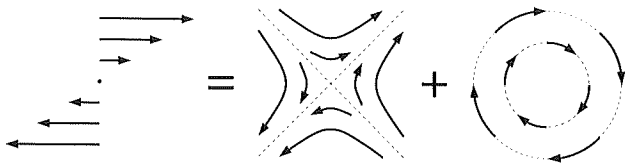


Fig. 16: *Decomposition of shear flow into an elongational and a rotational component.*

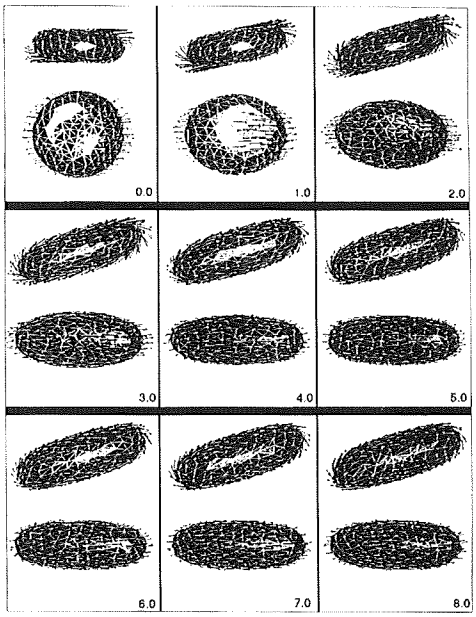


Fig. 17: *Evolution of an initially oblate vesicle under shear flow. The arrows denote the local velocity* [54].

equilibrium shape is at time $t = 0$ exposed to shear flow. The shape begins to elongate which immediately breaks the initial axi-symmetry. It tilts. After some time it acquires a stationary tilted elongated shape around which the membrane perpetually revolves in a so-called tank-treading motion. The qualitative features of this dynamically induced shape transformation can be understood on the basis of the decomposition (46). The elongational component leads to the tilt, whereas the rotational component leads to the tank-treading motion.

The quantitative analysis [54] shows that the tilt θ depends on the reduced volume (with $\theta \to \pi/4$) for $v \to 1$ as expected from the decomposition (46). The tilt is almost independent of shear rate $\dot{\gamma}$. This behaviour cannot persist to $\dot{\gamma} \to 0$ since one would expect to recover the equilibrium shape in this limit, which is however, difficult to reach numerically. The tank-treading frequency $\bar{\omega}$ (averaged over the shape) increases with reduced volume approaching $\dot{\gamma}/2$ for $v \to 1$ as expected.

Experimentally, a vesicle in shear flow has been studied in a slightly different regime [55]. The vesicle was initially spherical with thermal fluctuations. With increasing shear flow, some area stored in these fluctuations is "pulled out" and the shape elongates. With increasing shear rate

this elongation saturates since the total area is conserved. The tilt is initially 45° and becomes smaller with increasing shear rate. Since the algorithm described above does not include the effect of fluctuations (yet) a direct comparison is not possible. Analytically, however, one can include thermal fluctuation as long as one keeps the quasi-spherical approximation. Details of such a calculation using Langevin equations for the shape hydrodynamics [56] are given in [48].

4.5 Dynamically induced unbinding

For vesicles interacting with a substrate, shear flow can lead to a dynamically induced unbinding or detachment transition as observed in two recent experiments [57, 58], see Fig. 18. For small shear rate $\dot{\gamma}$, the vesicle tanktreads along the substrate. At a critical shear rate $\dot{\gamma}_c$, the vesicle starts to detach from the substrate.

A lift force occurs because of a symmetry breaking. The Stokes equation (31) are time-reversal invariant. This implies that a rigid sphere (or a tense vesicle) in shear flow at a certain height above a substrate will neither experience lift nor a force towards the substrate but rather translate at constant height. For a non-spherical shape, however, the fore-aft symmetry is broken and lift can occur.

If the vesicle is filled with a denser fluid than the environment, it hoovers at a distance h above the substrate. In such a stationary state the gravitational force will be compensated by the lift force. In the absence of a density difference between interior and exterior, the lift force leads the vesicle continuously further away from the substrate.

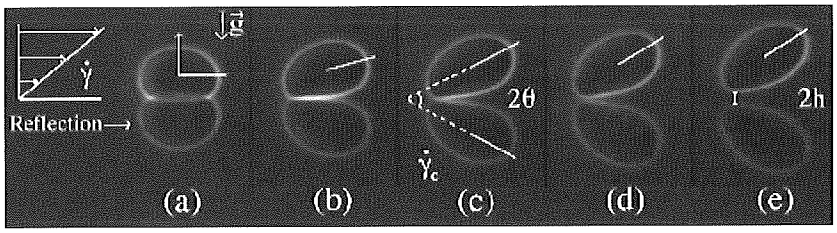


Fig. 18: *Experimental study of shear-flow induced unbinding of a vesicle from a substrate; (a): rest state; (b-d): vesicle elongates and tilts; (e): a gap develops between vesicle and substrate* [58].

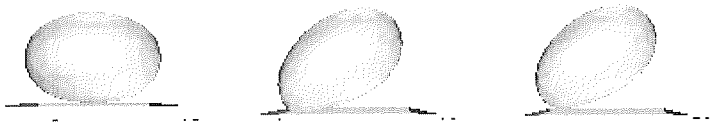


Fig. 19: *Theoretical study of shear-flow induced unbinding in the same geometry* [59].

The theoretical work on this effect can be roughly classified as follows. The simpler two-dimensional problem (a semiflexible ring polymer above a one-dimensional line substrate) has been studied in [60]. A perturbative approach is possible at large separation $h \gg R$ for ellipsoidal shapes. It leads to a lift force $\sim \dot{\gamma}/h^2$ [61]. In three dimensions, a simple scaling

approach has given a rough picture of this unbinding transition in the regime of strong adhesion [62].

We now describe the results of a numerical study of the full three-dimensional problem [59]. Such a numerical analysis is possible only, since the Oseen-tensor for a half-space geometry is known exactly [63] as reproduced in footnote 22 of [59]. To retrieve the bare contribution of hydrodynamics to the lift force, adhesion and gravity were first turned off. The vesicle is initially at a small distance from the wall and then shear flow is imposed. In Fig. 19, a typical sequence of "snapshots" are shown. The initially oblate-like shape tilts, elongates, becomes prolate and finally takes off. The steady tilt of the vesicle plays the leading role in breaking the fore–aft symmetry of the vesicle with respect to shear flow. Since the tilt is almost independent of shear rate, the excess pressure in the space between the vesicle and the wall causes a lift force which is proportional to shear rate $\dot{\gamma}$. Additional effects due to the deformation are of $O(\dot{\gamma}^2)$. When gravity is included by implementing a density difference between the fluids inside and outside the vesicle, the vesicle indeed hovers at a constant distance from the substrate . Finally, for vesicles adhering to a substrate by the influence of a potential $V(l)$, a critical shear rate was found above which the vesicle unbinds from the wall. This critical shear rate is roughly linear in the adhesion strength w [59].

5 Perspectives

In the three main sections of this lecture, I described the analysis of the simplest model systems which show the phenomenon we wanted to understand, like the different shapes, adhesion and dynamics in hydrodynamic flow. By focusing on the simplest cases, the parameter space is small enough to allow for a comprehensive study. These studies prepared the ground for investigating more complex systems. Many applications or a more detailed modelling of biological systems require that some of the simplifying assumptions must be relaxed and further ramifications be introduced. These more complex models necessarily contain more parameters which makes a comprehensive study more difficult and often not even desirable. In this concluding section, I will mention a few such perspectives with some recent references. The interested reader will easily find more literature by searching the various data bases via the internet.

Vesicle membranes need not consist of only one sort of lipid. Lipid mixtures on curved shapes allow for a coupling between local composition and curvature as it has been shown recently in a beautiful experimental study [64]. Shape transformations induced by light or by changing the chemical composition are reviewed in [65]. Fusion of vesicles has recently been visualized using microfluoresence spectroscopy [66]. A quantitative modelling of such topological changes will require a more microscopic approach.

Adhesion needs not to be on a laterally homogeneous substrate as described here. Technologically interesting substrates may be chemically or morphologically structured as the one recently been used in the context of adhesion on silicon chips [67]. Adhesion can be counteracted by pulling with localized point forces which gives rise to force-induced detachment [68].

Dynamics is a huge field where we described the Oseen-tensor approach in some detail. A complimentary theoretical technique uses a phase field for studying vesicles in shear flow [69]. Other geometries like flow of vesicles in a cylinder or in a porous network are important for certain applications. Hydrodynamic flow also introduces long range interactions between two or several vesicles. For denser systems like vesicle suspensions, macroscopic rheological parameters should be determined on the basis of such mesoscopic studies.

One of the original motivations for studying vesicles as mentioned in the introduction was the quest for understanding the shapes and the elastic behaviour of red-blood cells. For a successful modelling, the polymeric spectrin network which endowes these cells with shear elasticity must be included. Shape transformations of these cells [70] and their fluctuations [71] have recently indeed been described by the interplay of the various elasticities of this compound membrane.
For other cells like leucocytes one wants to understand their adhesion to capillary walls in blood flow. Quite generally, biological adhesion is dominated by specific receptor/ligand molecules rather than by an unspecific adhesion due to physical interactions described here. How this specific interaction competes with hydrodynamically induced forces in cell detachment is a fascinating and open problem [72].
A general characteristics of cells is that they are genuine non-equilibrium systems. While we have discussed non-equilibrium aspects caused by hydrodynamic flow, important modifications also arise from the fact that in biological systems active processes happen which transform chemical energy into mechanical motion. The physics of such active membranes is reviewed in [73].
These few and necessarily pretty scattered remarks on some of the perspectives of the work described in this lecture demonstrate that vesicle research is not only a vivid topic in statistical physics. In retrospect in a couple of decades, this research may well be seen as having contributed first steps towards a quantitative understanding of cell biology. Within such a vision, the canon of physics, shaped in the twentieth century by fields like elementary particle physics, nuclear physics, atomic physics and solid state physics, will include not only soft matter physics, which is presently being established, but may even comprise a field ultimately called "cell physics".

Appendices

A Differential geometry for vesicles

The membrane configuration $\mathbf{R}(s_1, s_2)$, parametrized by internal coordinates (s_1, s_2), is embedded in the three dimensional space. This space will be parametrized by Cartesian (x, y, z) or spherical (r, θ, ϕ) coordinates as $\mathbf{r} = x_\alpha \mathbf{e}_\alpha = r\mathbf{e}_r$ where $\alpha = x, y, z$. Summation over double indices is implied throughout the article. There are two tangential vectors

$$\mathbf{R}_i \equiv \partial_i \mathbf{R}(s_1, s_2) \quad \text{for} \quad i = s_1, s_2, \tag{47}$$

from which one obtains the metric tensor

$$g_{ij} \equiv \mathbf{R}_i \cdot \mathbf{R}_j. \tag{48}$$

Its determinant, $g \equiv \det(g_{ij})$, yields the area element

$$dA = \sqrt{g}\, ds_1 ds_2. \tag{49}$$

The normal vector, $\mathbf{n}(s_1, s_2)$, is given by

$$\mathbf{n} = \frac{\mathbf{R}_1 \times \mathbf{R}_2}{|\mathbf{R}_1 \times \mathbf{R}_2|}. \tag{50}$$

Finally, the mean and the Gaussian curvature follow from the curvature tensor

$$h_{ij} \equiv (\partial_i \partial_j \mathbf{R}) \cdot \mathbf{n} \tag{51}$$

as

$$H \equiv \frac{1}{2} h^i{}_i \tag{52}$$

and

$$K \equiv det(h^i{}_j), \tag{53}$$

where $h^i_j \equiv g^{ik} h_{kj}$ and g^{ij} are the matrix elements of the matrix inverse of (g_{ij}). Following the convention used in differential geometry, a sphere with the usual spherical coordinates ($s_1 = \theta, s_2 = \phi$) has $H < 0$.

B Steric of fluctuation potential

In this appendix, the derivation of the steric of fluctuation potential $V_\mathrm{f}(l)$ for a membrane under tension is sketched according to [30]. For a tensionless membrane, $V_\mathrm{f}(l)$ has been derived much earlier [35].

We consider a membrane pushed by a linear potential $+pl$ towards a rigid wall at $l = 0$. Its free energy functional becomes

$$F = \int dA \left\{ \frac{\kappa}{2} (\nabla^2 l)^2 + \frac{\Sigma}{2} (\nabla l)^2 + pl \right\} \simeq \int dA \left\{ \frac{\kappa}{2} (\nabla^2 l)^2 + \frac{\Sigma}{2} (\nabla l)^2 + pl + V_\mathrm{f}(l) \right\} \tag{54}$$

with configurations restricted to $l(\mathbf{r}) > 0$. For the second step, we have assumed that this geometrical constraint can be replaced by a smooth unknown repulsive potential $V_\mathrm{f}(l)$. Expanding around the minimum l_0 with $p = -V_\mathrm{f}'(l_0)$ and $h(\mathbf{r}) \equiv l(\mathbf{r}) - l_0$, we get after a Fourier transformation

$$F = \frac{1}{2} \int \frac{d^2 q}{(2\pi)^2} \left(\kappa q^4 + \Sigma q^2 + V_\mathrm{f}''(l) \right) |h_q|^2 \tag{55}$$

where we dropped the index 0 at l_0 again. On small length scales, $q > 1/\xi$, the harmonic energy (55) is dominated by the elastic part $\kappa q^4 + \Sigma q^2$; on large scales $q < 1/\xi$ the fluctuations feel the steric potential. The correlation length ξ can be defined precisely by

$$\kappa/\xi^4 + \Sigma/\xi^2 = V_\mathrm{f}''(l), \tag{56}$$

where ξ depends implicitly on l (the mean distance from the substrate), or l on ξ.

The confinement leads to a free energy density $f = b(kT)/\xi^2$. This form basically follows from a dimensional analysis, assuming that ξ is the only relevant length scale. The numerical prefactor b is of order 1. In the present context, f corresponds to $V_\mathrm{f}(l)$, which implies with (56) the non-linear differential equation

$$\kappa \left(\frac{V_\mathrm{f}(l)}{bkT} \right)^2 + \Sigma \left(\frac{V_\mathrm{f}(l)}{bkT} \right) = V_\mathrm{f}''(l) \tag{57}$$

for the fluctuation potential. The unique solution of this differential equation which diverges for

$l \to 0$ and vanishes for $l \to \infty$, is

$$V_f(\Sigma; l) = \left(\frac{6b^2(k_BT)^2}{\kappa l^2}\right)\left(\frac{y^2}{\sinh^2 y}\right) \quad \text{with} \quad y \equiv (\Sigma/bT)^{1/2} l/2, \tag{58}$$

given in eq. (19).
For tensionless membranes, where $\Sigma = 0$, the second bracket becomes unity. This form corresponds to the well-known Helfrich potential [35]. The numerical prefactor $6b^2 \simeq 0.1$ can be estimated by Monte-Carlo simulations [34]. In the presence of a lateral tension Σ, which pulls at the membrane, the power law decay of the first bracket is cut-off by the second factor at a length-scale $l_\Sigma \equiv (4bT/\Sigma)^{1/2}$ [34]. The best choice for b in this case is $b = 1/2\pi$.

C The Oseen-tensor

In this appendix, we sketch the derivation of the Oseen-tensor according to [56]. The Stokes equations (31) read after Fourier transformation for all quantities with $e^{i\mathbf{kr}}$

$$\begin{aligned} i\mathbf{k}p + \eta k^2 \mathbf{v} &= \mathbf{f} \\ \mathbf{kv} &= 0 \end{aligned} \tag{59}$$

Multiplying the first one with $-i\mathbf{k}$ and using the second leads to the pressure

$$p = -\frac{i\mathbf{k} : \mathbf{f}}{k^2} \tag{60}$$

being uniquely determined by the force. Here, : signifies the dyadic product $[(\mathbf{k} : \mathbf{f})_{\alpha\beta} = k_\alpha f_\beta]$. The first equation of (59) is now solved by

$$\mathbf{v} = \frac{1}{\eta k^2}\left(1 - \frac{\mathbf{k} : \mathbf{k}}{k^2}\right)\mathbf{f} \tag{61}$$

In real space, this relation becomes

$$\mathbf{v}(\mathbf{r}) = \int \mathcal{O}(\mathbf{r}, \mathbf{r}') f(\mathbf{r}') d^3 r' \tag{62}$$

with the Oseen-tensor

$$\mathcal{O}(\mathbf{r}, \mathbf{r}') \equiv \mathcal{O}(\rho) \equiv \frac{1}{\eta} \int \frac{d^3k}{(2\pi)^3} \frac{1}{k^2}\left(1 - \frac{\mathbf{k} : \mathbf{k}}{k^2}\right) e^{-i\mathbf{k}\rho} \tag{63}$$

Since the rhs is a tensor which depends only on $\rho \equiv |\boldsymbol{\rho}|$ with $\boldsymbol{\rho} \equiv \mathbf{r} - \mathbf{r}'$, its Cartesian matrix elements can be written as

$$\mathcal{O}_{\alpha\beta}(\rho) = A\delta_{\alpha\beta} + B\rho_\alpha \rho_\beta / \rho^2 \tag{64}$$

The scalars A and B are determined from the two equations

$$\mathcal{O}_{\alpha\alpha} = 3A + B \quad (\text{ summation over } \alpha \text{ understood }), \tag{65}$$

$$\mathcal{O}_{\alpha\beta}\rho_\alpha\rho_\beta/\rho^2 = A + B, \tag{66}$$

i. e.

$$3A + B = \frac{1}{(2\pi)^3} \int d^3k \frac{2}{\eta k^2} e^{-i\mathbf{k}\boldsymbol{\rho}} = \frac{1}{2\pi\eta\rho} \tag{67}$$

and

$$A + B = \frac{1}{(2\pi)^3} \int d^3k \frac{(1 - (\mathbf{k}\boldsymbol{\rho})^2/k^2\rho^2) e^{-i\mathbf{k}\boldsymbol{\rho}}}{\eta k^2} = \frac{1}{4\pi\eta\rho} \tag{68}$$

This leads to

$$A = B = \frac{1}{8\pi\eta\rho} \tag{69}$$

and hence the form (33) given in the main text.

References

[1] U. Seifert, Adv. Phys. **46**, 13 (1997).

[2] K. Berndl *et al.*, Europhys. Lett. **13**, 659 (1990).

[3] J. Käs and E. Sackmann, Biophys. J. **60**, 825 (1991).

[4] W. Wintz, H.-G. Döbereiner, and U. Seifert, Europhys. Lett. **33**, 403 (1996).

[5] X. Michalet and D. Bensimon, Science **269**, 666 (1995).

[6] X. Michalet and D. Bensimon, J. Phys. II France **5**, 263 (1995).

[7] P. B. Canham, J. Theoret. Biol. **26**, 61 (1970).

[8] W. Helfrich, Z. Naturforsch. **28c**, 693 (1973).

[9] U. Seifert, K. Berndl, and R. Lipowsky, Phys. Rev. A **44**, 1182 (1991).

[10] E. Evans, Biophys. J. **14**, 923 (1974).

[11] S. Svetina and B. Zeks, Eur. Biophys. J. **17**, 101 (1989).

[12] L. Miao, U. Seifert, M. Wortis, and H.-G. Döbereiner, Phys. Rev. E **49**, 5389 (1994).

[13] M. Jaric, U. Seifert, W. Wintz, and M. Wortis, Phys. Rev. E **52**, 6623 (1995).

[14] H. J. Deuling and W. Helfrich, J. Physique **37**, 1335 (1976).

[15] U. Seifert, Z. Phys. B **97**, 299 (1995).

[16] F. Jülicher, U. Seifert, and R. Lipowsky, J. Phys. II France **3**, 1681 (1993).

[17] F. Jülicher, U. Seifert, and R. Lipowsky, Phys. Rev. Lett. **71**, 452 (1993).

[18] H. Engelhardt, H. P. Duwe, and E. Sackmann, J. Physique Lett. **46**, L 395 (1985).

[19] F. Schwabl, *Statistische Mechanik* (Springer, Berlin Heidelberg, 2000).

[20] M. A. Peterson, H. Strey, and E. Sackmann, J. Phys. II France **2**, 1273 (1992).

[21] M. Wortis, M. Jaric, and U. Seifert, J. Mol. Liq. **71**, 195 (1997).

[22] V. Heinrich, F. Sevsek, S. Svetina, and B. Zeks, Phys. Rev. E **55**, 1809 (1997).

[23] W. Haeckl, U. Seifert, and E. Sackmann, J. Phys. II France **7**, 1141 (1997).

[24] P. Meleard, C. Gerbeaud, T. Pott, and L. Fernandez-Puente, Biophys. J. **72**, 2616 (1997).

[25] H.-G. Döbereiner *et al.*, Phys. Rev. Lett. **91**, 048301 (2003).

[26] A. Zilker, H. Engelhardt, and E. Sackmann, J. Physique **48**, 2139 (1987).

[27] J. O. Rädler, T. J. Feder, H. H. Strey, and E. Sackmann, Phys. Rev. E **51**, 4526 (1995).

[28] U. Seifert and R. Lipowsky, Phys. Rev. A **42**, 4768 (1990).

[29] R. Lipowsky and U. Seifert, Mol. Cryst. Liq. Cryst. **202**, 17 (1991).

[30] U. Seifert, Phys. Rev. Lett. **74**, 5060 (1995).

[31] R. M. Servuss and W. Helfrich, J. Phys. France **50**, 809 (1989).

[32] E. Evans and W. Rawicz, Phys. Rev. Lett. **64**, 2094 (1990).

[33] S. M. Bailey, S. Chiruvolu, J. N. Israelachvili, and J. A. N. Zasadzinski, Langmuir **6**, 1326 (1990).

[34] R. Lipowsky, in *Structure and dynamics of membranes - from cells to vesicles*, Vol. 1B of *Handbook of biological physics*, edited by R. Lipowsky and E. Sackmann (Elsevier, Amsterdam, 1995), pp. 521–602.

[35] W. Helfrich, Z. Naturforsch. **33a**, 305 (1978).

[36] U. Seifert, in *Giant vesicles*, edited by P. Luisi and P. Walde (J. Wiley and Sons, Europe, 1999), pp. 71–91.

[37] R. Lipowsky and S. Leibler, Phys. Rev. Lett. **56**, 2541 (1986).

[38] M. Mutz and W. Helfrich, Phys. Rev. Lett. **62**, 2881 (1989).

[39] W. Helfrich, in *Structure and dynamics of membranes - from cells to vesicles*, Vol. 1B of *Handbook of biological physics*, edited by R. Lipowsky and E. Sackmann (Elsevier, Amsterdam, 1995), pp. 691–722.

[40] M. Kraus, U. Seifert, and R. Lipowsky, Europhys. Lett. **32**, 431 (1995).

[41] H. G. Döbereiner *et al.*, Phys. Rev. E **55**, 4458 (1997).

[42] R. Bar-Ziv and E. Moses, Phys. Rev. Lett. **73**, 1392 (1994).

[43] J. Happel and H. Brenner, *Low Reynolds number hydrodynamics* (Noordhoff International Publishing, Leiden, 1973).

[44] F. Brochard and J. F. Lennon, J. Physique **36**, 1035 (1975).

[45] E. Evans, A. Yeung, R. Waugh, and J. Song, in *The Structure and Conformation of Amphiphilic Membranes*, Vol. 66 of *Springer Proceedings in Physics*, edited by R. Lipowsky, D. Richter, and K. Kremer (Springer, Berlin, 1992), pp. 148–153.

[46] U. Seifert and S. A. Langer, Europhys. Lett. **23**, 71 (1993).

[47] J. Prost, J.-B. Manneville, and R. Bruinsma, Eur. Phys. J. B **1**, 465 (1998).

[48] U. Seifert, Eur. Phys. J B. **8**, 405 (1999).

[49] Z. C. Ou–Yang and W. Helfrich, Phys. Rev. A **39**, 5280 (1989).

[50] L. Miao, M. A. Lomholt, and J. Kleis, Eur. Phys. J. **9**, 143 (2002).

[51] T. Pott and P. Meleard, Europhys. Lett. **59**, 87 (2002).

[52] H.-G. Döbereiner, E. Evans, U. Seifert, and M. Wortis, Phys. Rev. Lett. **75**, 3360 (1995).

[53] H.-G. Döbereiner and U. Seifert, Europhys. Lett. **36**, 325 (1996).

[54] M. Kraus, W. Wintz, U. Seifert, and R. Lipowsky, Phys. Rev. Lett. **77**, 3685 (1996).

[55] K. H. de Haas *et al.*, Phys. Rev. E **56**, 7132 (1997).

[56] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, Oxford, 1986).

[57] B. Lorz, R. Simson, J. Nardi, and E. Sackmann, Europhys. Lett. **51**, 468 (2000).

[58] M. Abkarian, C. Lartigue, and A. Viallat, Phys. Rev. Lett. **88**, 068103 (2002).

[59] S. Sukumaran and U. Seifert, Phys. Rev. E **64**, 011916/1 (2001).

[60] I. Cantat and C. Misbah, Phys. Rev. Lett. **83**, 880 (1999).

[61] P. Olla, J Phys. II France **7**, 1533 (1997).

[62] U. Seifert, Phys. Rev. Lett. **83**, 876 (1999).

[63] J. R. Blake, Proc. Camb. Phil. Soc. **70**, 303 (1971).

[64] T. Baumgart, S. Hess, and W. Webb, Nature **425**, 821 (2003).

[65] H.-G. Döbereiner, Current Opinion in colloid & interface science **5**, 256 (2000).

[66] G. Lei and R. MacDonald, Biophys. J. **85**, 1585 (2003).

[67] G. Zeck and P. Fromherz, Langmuir **19**, 1580 (2003).

[68] A.-S. Smith, E. Sackmann, and U. Seifert, Europhys. Lett. **64**, 281 (2003).

[69] T. Biben and C. Misbah, Phys. Rev. E **67**, 031908 (2003).

[70] H. Lim, M. Wortis, and R. Mukhopadhyay, Proc. Nat. Acad. Sci. **99**, 16766 (2003).

[71] N. Gov, A. Zilman, and S. Safran, Phys. Rev. Lett. **90**, 228101 (2003).

[72] S. K. Bathia, M. R. King, and D. A. Hammer, Biophys. J. **84**, 2671 (2003).

[73] S. Ramaswamy and M. Rao, C. R. Acad. Sci. **2**, 817 (2001).

D 4 Physics of Multicomponent Membranes

R. Zorn
Institut für Festkörperforschung
Forschungszentrum Jülich GmbH

Contents

1 Preface 2

2 Introduction 2

3 Lipid Inhomogeneity 4

4 Physical Description 6
4.1 Strong Segregation Limit 7
4.2 Weak Segregation Limit 13

5 Summary and Outlook 18

1 Preface

The physics of multicomponent membranes is a research field which is still under development. Current textbooks of membrane physics [1, 2, 3] usually only contain a couple of pages about this subject. Therefore, this lecture is mostly based on recent journal articles. Because in the literature a vast amount of theoretical models of multicomponent membranes is devised it is only possible to rephrase those which for the author seem to reflect the most important tendencies at the moment (2004). Therefore, if confronted with a specific problem the reader is encouraged to research the literature for the most appropriate model and not only to rely on those presented here.

2 Introduction

The membrane of a single species of cell usually contains in the order of one hundred different lipids. The lipids differ in the type of hydrophilic headgroup as well as the hydrocarbon chains. One of the central questions of cell biology is what the reason is for this diversity.
Of course, there are ways how the chemical properties of the lipids may come into play, e.g. [1]

- Some lipids are important as regulatory agents, e.g. the derivatives of phosphatidylinositol in the plasma membranes of eukaryotic cells [4].
- Some lipids may participate in biosynthetic pathways, e.g. in *E. coli* phosphatidylglycerol provides the glycerol phosphate moiety in the biosynthesis of periplasmic oligosaccharides [5].
- Specific lipids may be required for optimal enzyme activity of particular enzymes [6].

This lecture will *not* deal with these mechanisms but rather discuss the *physical* properties which originate mainly from the influence of the shape of the lipid molecules on the shape of the membrane.
The physical properties of lipids important in this context are shown in Figure 1:

1. Length and area determine the thickness of the membrane and the density of lipids in the membrane.
2. Depending on the ratio between headgroup area and cross-section of the tails lipids may be more or less conical. Especially, lipids with one or three hydrocarbon chains tend to exhibit conicity. This shape clearly influences the preferred curvature of the membrane.
3. There may be an interaction between the head groups, in most cases due to charges. This interaction may give rise to an additional curvature.

There are several rôles lipids with special physical properties may play in a biological membrane, e.g. (Figure 2)

- The size of a protein molecules may not match the average thickness of a membrane in which it is supposed to be embedded. In this case shorter (or longer) lipids may concentrate close to the protein in order to adapt the membrane thickness to the protein size.

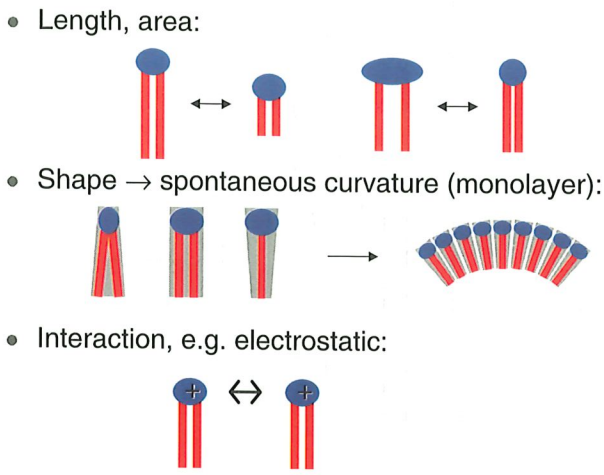


Fig. 1: *Physical properties of lipids which are relevant for the shape of the membranes formed.*

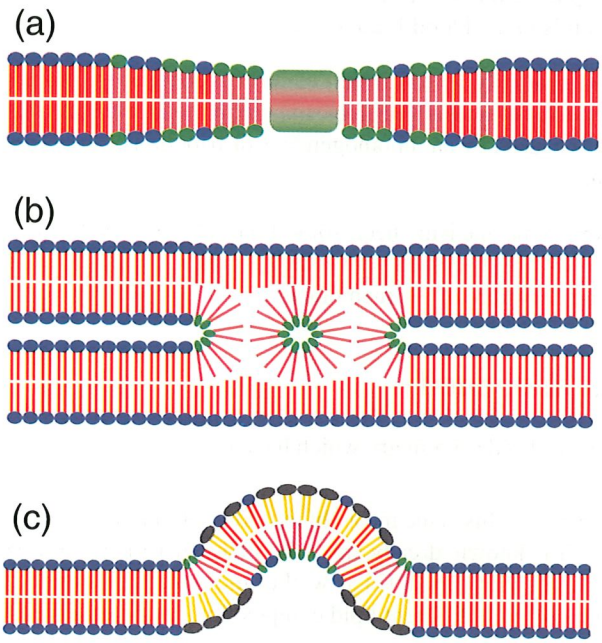


Fig. 2: *Physical lipid functions. (a) Adaptation of membrane thickness to a protein, (b) Local non-lamellar phase in tight junction, (c) Assistance of curvature in endo-/exocytosis.*

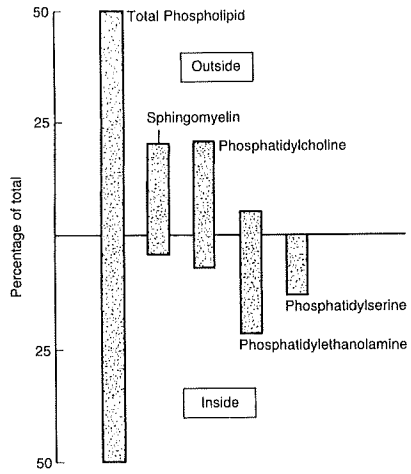


Fig. 3: *Representation of the relative distribution of the major lipids in the human red erythrocyte membrane (from [1]).*

- Lipids with a high conicity often form non-lamellar phases, e.g. hexagonal tubular phases. These phases may make some sense for biological systems locally. The most interesting possible example are the 'tight junctions' which membranes form in order to prevent the passage of liquids in the blood-brain barrier.
- Although it is generally assumed that clathrin is the main actor for creating buds in a membrane (endo-/exocytosis) mechanisms without clathrin can be conceived for which the curvature is caused by an inhomogeneous distribution of lipids with different conicity or interaction.

It is clear that all those mechanisms need some lipid inhomogeneity in order to work. Here, mainly the last topic will be discussed with some models which desccribe the coupling of lipid inhomogeneity to the curvature of the membrane.

3 Lipid Inhomogeneity

There are three kinds of lipid asymmetry which have to be considered:

Lateral Inhomogeneity This is the most obvious kind of inhomogeneity which occurs within a single monolayer. It is known already since the first experiments on mixed monolayers in a Langmuir trough. There, with optical [7] as well as more recently atomic force microscopy [8] μm-sized phase domains with different lipid composition can be observed.

Transmembrane Asymmetry Because the biological membrane is a *bi*layer there may also be a difference in the lipid composition between the inside and the outside. Experiments on natural cell membranes show that this is rather the rule than the exception [9]. The best and longest studied example is the red blood cell (Figure 3) [10]. But even artificial model vesicles

show such an asymmetry [11] which means that there is no active mechanism necessary to support it.
Experimental methods to detect this asymmetry have to be inside/outside selective, e.g.

- Phospholipases are able to sever the headgroup of a lipid from its hydrocarbon tail. If such an enzyme is added to a vesicle solution it attacks only the outside. From the lysis products the composition of the outer membrane can be determined. The risk of this method is that the membrane may be ruptured.
- Low-molecular agents as trinitrobenzenesulfonic acid bind selectively to certain lipid species, in this case phosphatidylethanolamines. From the concentration reduction of the agent by binding to the outside of the vesicles the concentration of the corresponding lipids can be determined. One has to make sure that this binding process does not distort the distribution.
- Phospholipid exchange proteins (PLEPs) are able to transfer lipids from the outside of a cell membrane to that of a probe vesicle. The transfer can be followed by radioactive marking. This method is rather slow and also the composition may be changed.
- In nuclear magnetic resonance (^{31}P or ^{13}C) the addition of paramagnetic ions as Mn^{2+} quenches the signal of lipids on the side where the ions are added. Even this method is not completely safe from artifacts. In some cases the ions may bind to the lipid headgroups thus changing the properties of the membrane.

Coupling of Lateral Inhomogeneity between Monolayers In the case of lateral inhomogeneity for bilayers the question arises whether the domains are at the same place on both sides, i.e. whether there is a transmembrane coupling. To answer this question is not possible with the methods explained before because they are either position sensitive or inside/outside sensitive but not both.
Only recently it has become possible to do such an experiment by selective staining of microphases in giant unilamellar vesicles [12, 13]. Figure 4 shows a confocal microscopy picture of a vesicle of dilauroylphosphatidylcholine/dipalmitoylphosphatidylcholine. The lipids separate into two phases: an ordered DPPC-rich and a fluid DLPC-rich one. Now it is possible to use a 585 nm-fluorescent dye (DiI-C_{20}, represented as red in the figure) preferring the ordered phase and a 522 nm dye (Bodipy-PC, represented as green) for the fluid phase. From the picture it can be seen that there are green and red domains but only thin yellow lines between them. Because the microscope is not able to resolve the thickness of the membrane (about 10 nm) the mixture of colours (yellow) should be seen where different phase are opposed in the membrane. The absence of larger yellow regions demonstrates that the microdomains on both sides of the bilayer are strongly correlated.

Finally, a word has to be said about the mobility of lipids which determines the speed with which inhomogeneities can build up or decay. For the lateral motion the diffusion is quite fast in the fluid phases (e.g. L_α): $D \sim 10^{-8}\,\mathrm{cm}^2/\mathrm{s}$. This means that within about one second the concentrations will be equilibrated in a vesicle of $1\,\mu$m size. For the gel phases (e.g. L_β) the diffusion is about two orders of magnitude slower. Much slower is the transmembrane motion connected to the concentration asymmetry. The reason is mainly that on its way to the other side the lipid headgroup has to cross the hydrophobic membrane core. The time scale in pure

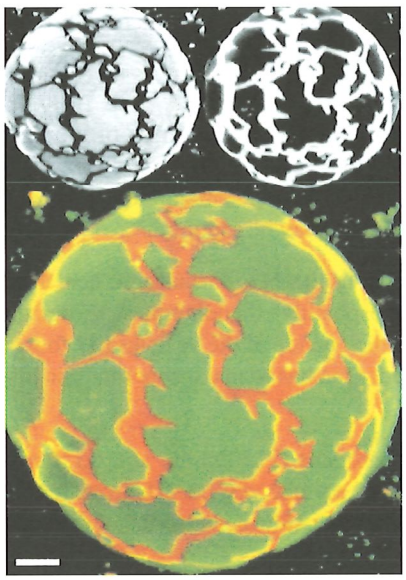


Fig. 4: *Confocal fluorescence microscopy images of a giant unilamellar vesicle. Lipid composition:* DPPC/DLPC $= 1/1$. *The upper right image is the fluorescence emission through a 585 nm filter from DiI-C_{20}. The upper left image is the emission through a 522 nm filter from Bodipy-PC. The bottom image is colour-merged, representing the former emission in red the latter in green (from [13]).*

model bilayers will amount to days. In natural membranes the process may be assisted by proteins (flippases) which provide a hydrophilic side to facilitate the transition of the headgroup to the other side. From these, hypothetical translocases have to be distinguished which would perform an active (ATP fuelled) transport from one monolayer to the other. It is also known that uncharged lipids flip faster for obvious reasons.

4 Physical Description

There are two fundamentally different approaches for physical models of the lipid inhomogeneity and its interaction with membrane curvature. (1) In the *strong segregation limit* it is assumed that the lipids segregate into several microphases which form domains of nanometer through micrometer size. In this case the free energy of the membrane consists of the curvature energy of the domains and the line energy of the one-dimensional interface. The shape of the membrane as a whole is determined by minimising this total free energy. (2) On contrary the *weak segregation limit* assumes that there is only a fluctuation of the lipid concentrations around its average value expected from the total lipid composition of the membrane. Here the curvature energy (in turn depending on the local lipid composition) is also one part of the free energy. But it is compensated by the mixing entropy which avoids a complete separation of lipids.

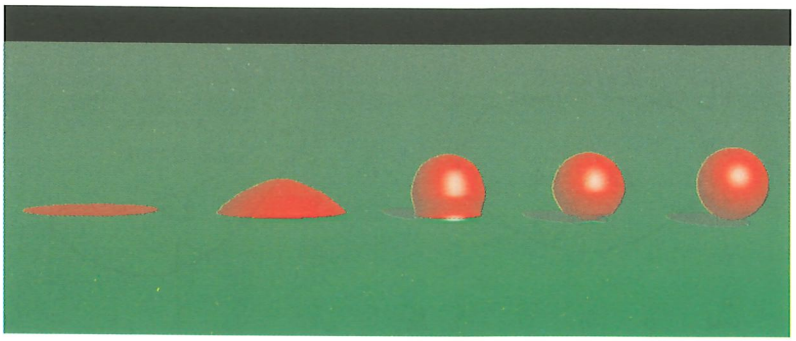


Fig. 5: *Schematics of bud formation in the strong segregation limit. The colours red and green indicate the different phases. The curvature of the "red" domain increases from left to right and so does the curvature energy. At the same time the interface gets shorter and its line energy is reduced.*

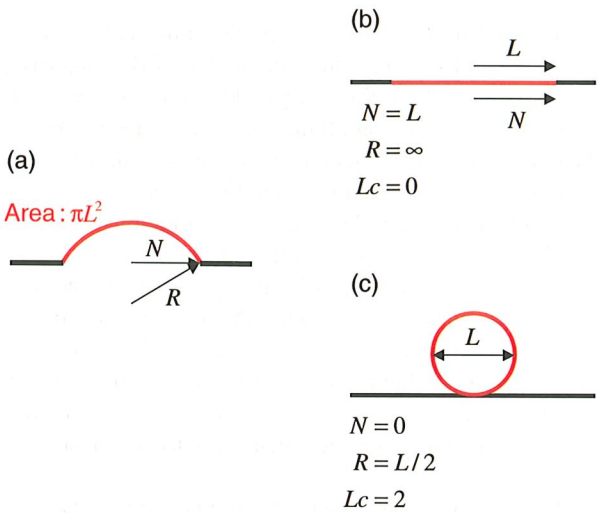


Fig. 6: *Simple model for the strong segregation limit. (a) Definition of parameters, (b) Limiting case of flat b-domain, (c) Limiting case of detached b-vesicle.*

4.1 Strong Segregation Limit

The basic principle of all theoretical models in the strong segregation limit is the balance of line energy of the microphase boundary and curvature energy of the microphase domains. Figure 5 is a qualitative sketch of a membrane with several segregated domains (red) showing increasing curvature. It is immediately clear that with increasing curvature the perimeter of the "neck" where the two microphases touch shortens. This means that the line energy can be reduced by "budding" until it vanishes for a completely separated vesicle. On the other hand the curvature energy increases. Depending of the curvature elasticity and the line tension there will be an optimal situation where the total energy is minimised.

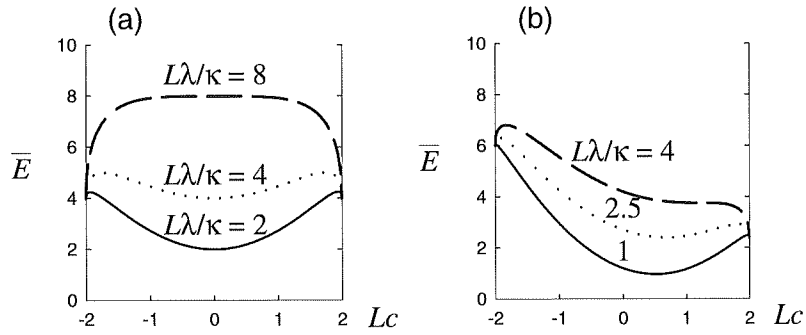


Fig. 7: *Dimensionless energy of the bud in the simple model of Figure 6. (a) no spontaneous curvature, (b) spontaneous curvature* $c_{\rm sp} = 0.45/L$. *Continuous line: stable incomplete bud, dotted line: metastable incomplete bud with same energy as complete bud, dashed line: instable incomplete bud.*

The probably simplest model in the strong segregation limit is that of a single microdomain of one phase (called 'b' in the following) in a membrane of the majority phase 'a' [14]. It is immediately clear that reduction of the line energy of the interface leads to a circular perimeter of the b-domain. If there is a curvature at all the curvature of the domain will be spherical due to optimisation of the curvature energy. Of course there may or may not be a curvature of phase a but if the area is assumed to be much larger than that of phase b its energy change due to a shape change of the b-domain is negligible. Therefore, without loss of generality it is sufficient to consider a spherical b-cap on a flat a-plane (Figure 6).
The shape of the cap can be fully described by two parameters: the area and the (constant) radius of curvature R. Assuming incompressibility the former is constant and can be expressed by the radius L of the b-domain in the limit of flatness: πL^2 (Figure 6(b)). The curvature is $c = 1/R$ and it is convenient to define a dimensionless co-ordinate of "budding", $Lc = L/R$. The range of values of this parameter is $-2 \ldots +2$ where $Lc = 0$ corresponds to a completely flat domain while $Lc = \pm 2$ denotes a b-vesicle detaching to the outside (Figure 6(c)) or inside depending on the sign.
With the bending rigidity κ of the membrane the bending energy of the bud can be expressed as

$$E_{\rm bend} = \pi L^2 \frac{\kappa}{2} (2c - 2c_{\rm sp})^2 = 2\pi\kappa \, (Lc - Lc_{\rm sp})^2 \, . \tag{1}$$

Here, $c_{\rm sp}$ is the spontaneous curvature of the membrane which it assumes in the absence of forces, e.g. by an asymmetric distribution of conical lipids. Obviously, reducing the energy favours a low curvature (or one close to the spontaneous).
The edge energy can be calculated with the line tension λ of the a-b interface:

$$E_{\rm edge} = \lambda \, 2\pi N = 2\pi\lambda L \sqrt{1 - (Lc/2)^2} \, . \tag{2}$$

Here, the energy can be reduced by shortening the interface line which can be achieved by a more curved bud.
Introducing the dimensionless energy $\bar{E} = (E_{\rm bend} + E_{\rm edge})/\, 2\pi\kappa$ the universal form for the energy of the bud is

$$\bar{E} = (Lc - Lc_{\rm sp})^2 + (L\lambda/\kappa) \sqrt{1 - (Lc/2)^2} \, . \tag{3}$$

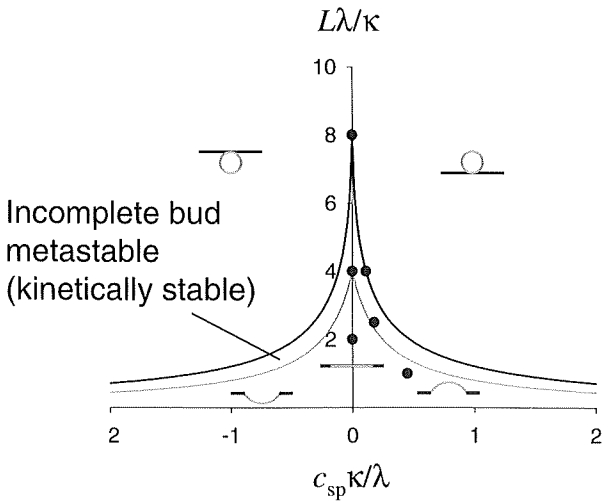


Fig. 8: *Phase diagram of the simple model of the strong segregation limit of Figure 6. The ordinate is a dimensionless parameter expressing domain size, membrane flexibility, and/or interfacial tension. The abscissa denotes the dimensionless spontaneous curvature. The black line indicates where the incomplete buds become unstable. The grey metastability line results from an approximation. The dots correspond to the situations of Figure 7.*

This equation expresses the counteraction of the curvature energy trying to flatten the bud and the line tension acting towards a reduction of the "neck" radius. Figure 7 shows this expression for different values of the dimensionless parameter $L\lambda/\kappa$ which comprises domain size, stiffness, and line tension. The two cases of absence of spontaneous curvature and $c_{\mathrm{sp}} = 0.45/L$ are plotted.

It can be seen that for different values of $L\lambda/\kappa$ qualitatively different situations arise. If $L\lambda/\kappa$ is small (small domains, rigid membrane, and/or low line tension) the absolute minimum of the energy is attained for the flat membrane in the absence of spontaneous curvature. In the presence of a spontaneous curvature it is close to Lc_{sp}. So there may be some curvature but not sufficient to detach the bud (incomplete budding). On the other side for large $L\lambda/\kappa$ the energy minima are located at $Lc = \pm 2$. In this case it is optimal to avoid the line energy completely by detaching a highly curved b-vesicle from the membrane (complete budding). In between there is a region of $L\lambda/\kappa$ where three relative minima are present. This would correspond to a coexistence situation where the non-absolute minima are metastable situations. But if one calculates the energy barrier with realistic values of κ being an order of magnitude larger than $k_{\mathrm{B}}T$ it turns out to be insurmountable. Therefore, the incomplete bud will be stable until coming very close to the vanishing of the central energy minimum.

The scenario in the case of non-vanishing spontaneous curvature only differs in one respect from that of intrinsically flat membranes: If $L\lambda/\kappa$ is gradually increased the latter show a discontinuous change from $Lc = 0$ to $Lc = \pm 2$ where the direction of budding is undetermined. For existing spontaneous curvature there is a continuous increase from $Lc \approx Lc_{\mathrm{sp}}$ to $Lc = \pm 2$ where now the sign is determined by that of c_{sp}.

Figure 8 shows the diagram of stability of incomplete and complete buds, the "phase diagram",

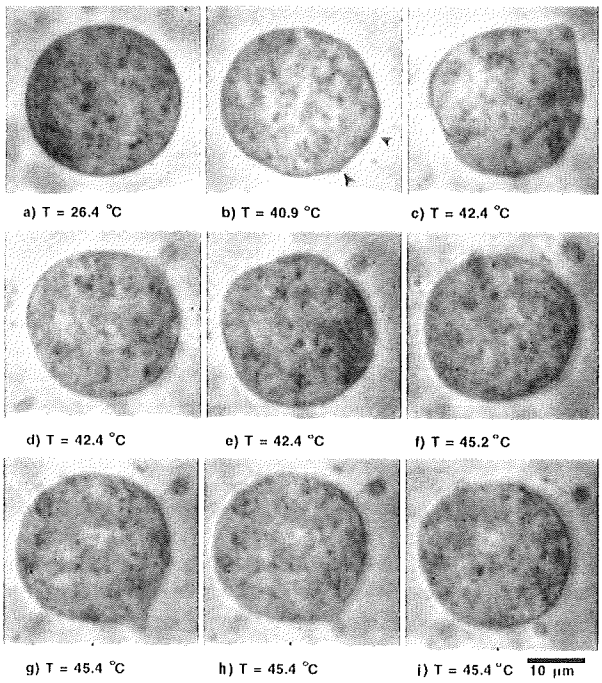


Fig. 9: *Experimental observation of spontaneous bud formation in a multicomponent lipid vesicle (from [15]).*

with respect to the parameters $L\lambda/\kappa$ and $c_{\mathrm{sp}}\kappa/\lambda$.

From this model a qualitative prediction of the kinetics of budding in a multicomponent membrane is possible. The primarily formed domains may be too small to form complete buds. Therefore, in a first step their size increases until $L\lambda/\kappa$ becomes sufficiently large to enter the complete budding regime in Figure 8. The critical size can be estimated from typical bending rigidities and line tensions to 80 nm. Possible mechanisms for the domain growth are diffusion, coalescence ($L \propto t^{1/2}$), or evaporation-condensation ($L \propto t^{1/3}$). The absolute time scale for this step can be estimated from the lateral diffusion constant of lipids in the membrane to 10^{-2} s. As soon as the domains are sufficiently big, the transition into the energetically favoured state of a detached vesicle will occur very fast. The time scale is determined by the stretching of the a-phase due to the extraction of the b-domain; typical values for lateral compressibility and mass density of the membrane yield 10^{-8} s.

Figure 9 shows an experiment which qualitatively confirms this model [15]. A large vesicle of natural sphingomyelin (bovine brain) is observed under a microscope. At the same temperature where calorimetry shows the onset of a phase transitions—which in a multicomponent mixture probably implies phase coexistencies—spontaneous formation of buds starts. Despite the similarity of the behaviour it has to be mentioned that the experimentally observed buds are by one magnitude larger than expected from the model.

From this experiment using standard optical microscopy it can only be indirectly concluded that the bud is really a domain of a different phase. The direct observation requires the identification of phases by phase selective dyes (see section 3). The pictures of Figure 10 were obtained

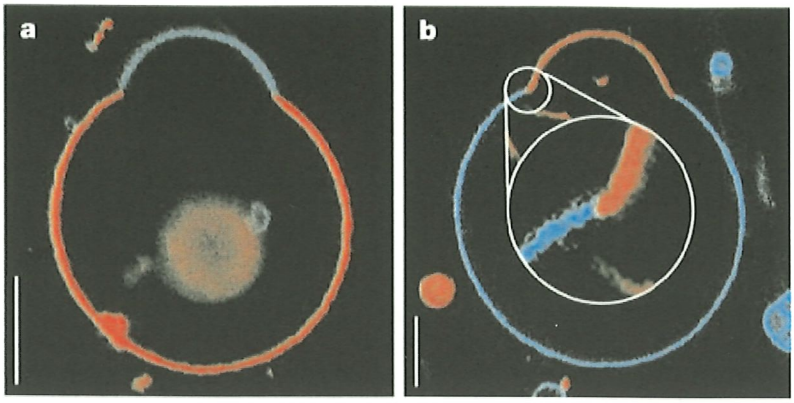


Fig. 10: *Two-photon microscopy images showing equatorial sections of Sphingomyelin/DOPC/Cholesterol vesicles with a coexistence of lamellar phases. One phase is labelled with perylene (blue) the other with rho-DPPE (red). Scale bar: 5 μm (from [16]).*

by observing a selectively stained Sphingomyelin/DOPC/Cholesterol vesicle with two-photon microscopy [16]. They show equatorial sections of the vesicle. From the colouration it is clear that the (incomplete) bud is made of another phase than the mother vesicle. The resolution even allows to see the structure of the interface line which is a narrow region of negative curvature.
The assumption of a single domain of microphase b in an infinite a-membrane is an unrealistic idealisation if the two microphases have similar proportions. A more realistic model has to take into account that domains arrange in the 2D plane of the membranes leading to *modulated phases* [17]. In order to model this situation the membrane curvature has to be treated as a function of the 2D co-ordinate τ (e.g. $\tau = [x, y]$ for a flat membrane, $\tau = [\phi, \theta]$ for a near-spherical vesicle).
The bending energy has now to be calculated by integration over the membrane surface $\mathcal{T}$:

$$E_{\text{bend}} = \frac{\kappa}{2} \int_{\mathcal{T}} \mathrm{d}^2\tau \sqrt{G(\tau)} \left(c^{(1)}(\tau) + c^{(2)}(\tau) - c_{\text{sp}}(\tau) \right)^2 . \tag{4}$$

$c^{(1)}(\tau)$ and $c^{(2)}(\tau)$ denote the principal curvatures at point τ, $c_{\text{sp}}(\tau)$ the spontaneous curvature which may depend on the phase identity at τ. $G(\tau)$ is the determinant of the metric tensor (e.g. $R^4 \sin^2\theta$ for a near-spherical vesicle).
The edge energy will be obtained by a line integral over the perimeter of the b-domains $\partial\mathcal{T}^{\text{b}}$:

$$E_{\text{edge}} = \lambda \oint_{\partial\mathcal{T}^{\text{b}}} \mathrm{d}l \, . \tag{5}$$

Finally, in this model also the total change of the area of the membrane has to be taken into account with the surface tension σ:

$$E_{\text{surf}} = \sigma \int_{\mathcal{T}} \mathrm{d}^2\tau \sqrt{G(\tau)} \, . \tag{6}$$

The integrals can be evaluated for specific situations, e.g. for near-flat membranes by using the

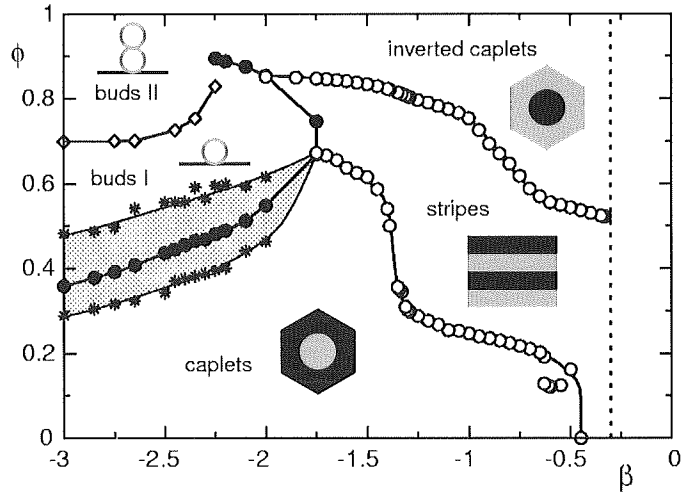


Fig. 11: *Phase diagram of the 2D modulated phase model described in the text.* $\lambda = 0.5\kappa|c_{\rm sp}^{\rm a}|$, $\sigma = 0.4\kappa(c_{\rm sp}^{\rm a})^2$. *Open and closed circles show the points of equal energy for the coexisting phases. The asterisks indicate the boundaries of the coexistence region (grey) of caplets and one-bead buds. Other two-phase regions are narrow and not shown. For* $\beta > -0.3$ *macroscopic phase separation occurs. (from [17]).*

Monge representation[1] the bending energy can be written as

$$E_{\rm bend} = \frac{\kappa}{2}\int_{\mathcal{T}} {\rm d}^2\tau \left(\left(\nabla^2 z(\tau) - c_{\rm sp}(\tau)\right)^2 + \frac{1}{2} c_{\rm sp}(\tau)^2 \left(\nabla z(\tau)\right)^2 \right) . \tag{7}$$

The total energy $E_{\rm bend} + E_{\rm edge} + E_{\rm surf}$ will now be minimised using a Fourier expansion ansatz for $z(\tau)$. This will be

$$z(x,y) = \sum_{l=1}^{N} a_l \cos\left(\frac{2\pi l}{\Lambda} x\right) \tag{8}$$

for striped phases and

$$z(r,\phi) = \sum_{l=1}^{N} a_l \cos\left(\frac{2\pi l}{\rm P} r\right) \tag{9}$$

for hexagonal phases. Note that the latter expression is approximative because it replaces the hexagonal symmetry by circular.

Figure 11 shows the phase diagram of this model (actually only the cut for $\lambda = 0.5\kappa|c_{\rm sp}^{\rm a}|$, $\sigma = 0.4\kappa(c_{\rm sp}^{\rm a})^2$) with respect to the parameters $\beta = c_{\rm sp}^{\rm b}/c_{\rm sp}^{\rm a}$ and the area ratio $\phi = A^{\rm b}/(A^{\rm a}+A^{\rm b})$. For positive β, i.e. identically directed spontaneous curvatures, the gain in curvature energy is too small to overcome the energy cost due to line and surface tension and the system phase-separates macroscopically. For intermediate negative values $\beta = -0.3 \ldots -1.8$ the striped phase is preferred for balanced area ratios while the hexagonal caplet phases are favourable if one phase is present in larger excess. For even lower values $\beta < -1.8$ spontaneous budding occurs or even multiple budding into pearl-necklace-like structures. Figure 12 shows that the

[1] In the Monge representation the membrane shape is described by the height $z(\tau)$ above a certain nominal surface (plane or sphere).

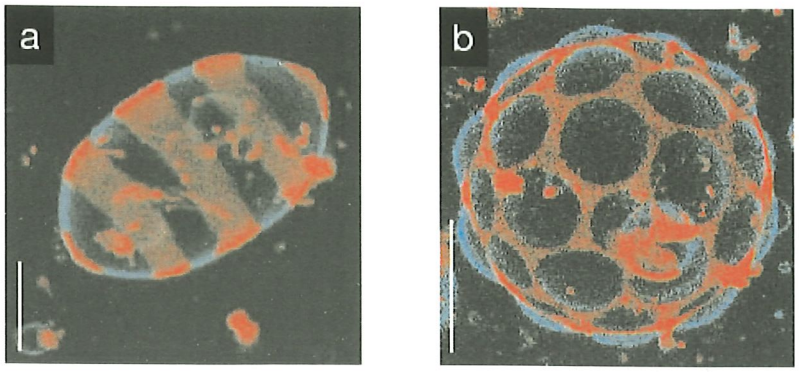


Fig. 12: *Two-photon microscopy images (hemispherical projections) of modulated phases. (a) striped phase, (b) hexagonal phase. Scale bar:* 5 μm *(from [16]).*

striped phase as well as the hexagonal array of buds can be observed experimentally [16] by two-photon microscopy of selectively stained membranes.

4.2 Weak Segregation Limit

In the weak segregation limit it is assumed that no microphase separation takes place but a coupling between curvature and lipid composition may facilitate bending the membrane. In response to a curvature the lipids may redistribute between the sides of the bilayer changing the spontaneous curvatures in a way that they are closer to the actual membrane curvature.
Again we start with a simplified model where the curvature is treated as constant over the surface [18]. Taking into account that the inner and outer monolayer may have different lipid compositions the curvature energy per unit area is

$$f_{\text{bend}} = \frac{\kappa}{2}\left((c + c_{\text{o}})^2 + (c - c_{\text{i}})^2\right) . \tag{10}$$

Here c is the actual curvature of the membrane and c_{o} and c_{i} are the spontaneous curvatures of the outer and inner monolayers respectively. Because the lipids in the two monolayers are oriented oppositely, $c = c_{\text{i}}$ is the preferred curvature while it is $c = -c_{\text{o}}$ for the outer. If the composition on both sides is the same $c_{\text{i}} = c_{\text{o}}$ holds and the energy is minimised by $c = 0$. This is the obvious fact that for symmetric bilayers the uncurved lamellar state is optimal.
In order to derive the consequences of coupling between lipid composition and curvature energy one has to assume a relation between lipid composition and spontaneous curvature. In the simplest interaction-free case this is just a linear superposition, for two lipid species:

$$c_{\text{i}} = (1 - \psi_{\text{i}})c_1 + \psi_{\text{i}} c_2 \tag{11}$$
$$c_{\text{o}} = (1 - \psi_{\text{o}})c_1 + \psi_{\text{o}} c_2 \tag{12}$$

Here ψ_{i} denotes the concentration of lipid 1 in the inner monolayer and ψ_{o} that in the outer. c_1 and c_2 are the spontaneous curvatures of monolayers made up of pure lipid 1 or 2 respectively.
It will later turn out that it is crucial to assume an interaction term which is proportional to the

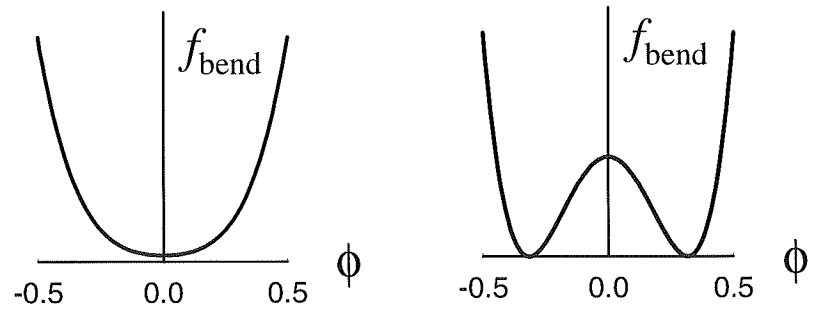


Fig. 13: *(Free) energy of bending, simple model of the weak segregation limit. Left, $\beta\bar{c} < 0$: one minimum for vanishing composition asymmetry and curvature. Right, $\beta\bar{c} > 0$: two minima for finite composition asymmetry ϕ^*.*

probability $\psi(1-\psi)$ that different lipid molecules are neighboured:

$$c_{\text{i}} = (1-\psi_{\text{i}})c_1 + \psi_{\text{i}}c_2 + \beta\psi_{\text{i}}(1-\psi_{\text{i}}) \quad (13)$$

$$c_{\text{o}} = (1-\psi_{\text{o}})c_1 + \psi_{\text{o}}c_2 + \beta\psi_{\text{o}}(1-\psi_{\text{o}}) \quad (14)$$

The influence of the interaction is cast into β, expressing the difference between the interactions of same lipid molecules and different lipid molecules. β is positive if due to an attractive interaction between the headgroups of different lipids the spontaneous curvature is increased in a mixed monolayer compared to a pure monolayer.
We now introduce two new composition parameters: $\psi = \frac{1}{2}(\psi_{\text{o}}+\psi_{\text{i}})$ be the total composition of the membrane which is constant. $\phi = \frac{1}{2}(\psi_{\text{o}} - \psi_{\text{i}})$ be the concentration difference (asymmetry) of the membrane which may change in response to curvature. With these variables equations (13) and (14) can be re-expressed as

$$c_{\text{i}} = \bar{c} + \alpha\phi - \beta\phi^2 \quad (15)$$

$$c_{\text{o}} = \bar{c} - \alpha\phi - \beta\phi^2 \quad (16)$$

where

$$\bar{c} = (1-\psi)c_1 + \psi c_2 + \beta\psi(1-\psi) \quad (17)$$

is the average spontaneous curvature of a monolayer assumed in the absence of asymmetry ($\phi = 0$) and

$$\alpha = c_1 - c_2 - (1-2\psi)\beta \quad (18)$$

is the first order coefficient of the influence of the asymmetry onto the spontaneous curvature. Substituting ψ and ϕ into equation (10) yields the simple result

$$f_{\text{bend}} = 4\kappa\left(\left(\bar{c} - \beta\phi^2\right)^2 + \left(\alpha\phi - c\right)^2\right). \quad (19)$$

In order to minimise this expression both square terms have to be minimised separately: The second term can be reduced to zero by choosing $c = c^*$ with

$$c^* = \alpha\phi\,. \quad (20)$$

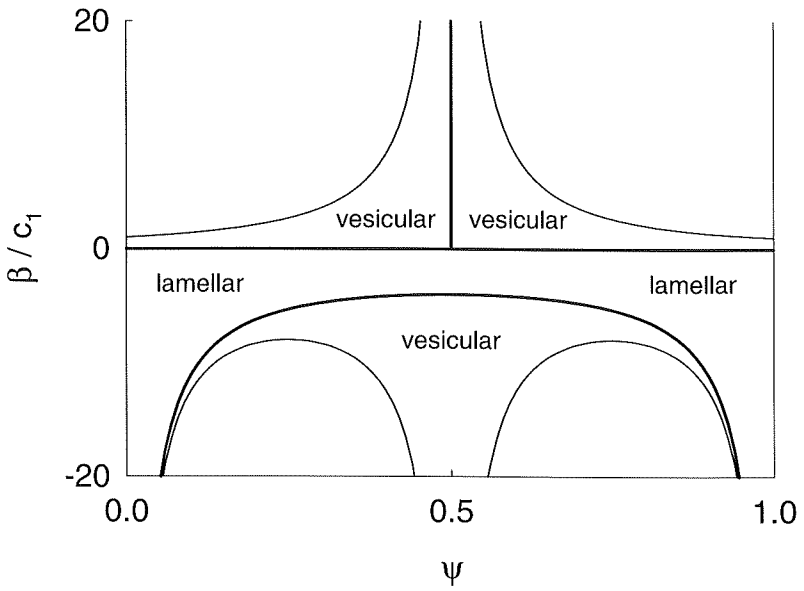


Fig. 14: *Phase diagram of the simple model in the weak segregation limit ($c_1 = c_2 > 0$). The thick lines denote the boundaries between lamellar and vesicular phase. The thin lines demarcate the regions where $|\phi^*| > \min(|\psi|, |1-\psi|)$, i.e. one lipid species is totally depleted from one monolayer.*

This is obviously the relation between curvature and energy which gives the curvature resulting from a certain lipid composition asymmetry or vice-versa the composition asymmetry resulting from an imposed curvature. It can be seen that for $\alpha = 0$ (identical individual spontaneous curvatures *and* no interaction) the flat membrane is optimal.
For minimising the first term one has to distinguish two cases (Figure 13): (1) $\beta\bar{c} < 0$: In this case $(\bar{c} - \beta\phi^2)^2$ has *one* minimum at $\phi^* = 0$ with the value $f_{\text{bend}} = 4\kappa\bar{c}^2$. From (20) follows that $c^* = 0$. So the optimal situation is that of a flat bilayer with symmetric lipid composition. The residual energy results from the frustration of outer and inner monolayer having the same spontaneous curvature but in opposite directions. (2) $\beta\bar{c} > 0$: Here, the first term can be reduced to zero by setting $\phi = \phi^*$ with

$$\phi^* = \pm\sqrt{\bar{c}/\beta}\,. \tag{21}$$

This means that an appropriate value of the asymmetry leads to a spontaneous curvature of the inner and outer monolayer which exactly match the actual curvature of the membrane $c^* = \pm\alpha\sqrt{\bar{c}/\beta}$. With that rearrangement of lipids between the two sides of the membrane the frustration vanishes completely. The ambiguity of the sign in (21) results from the symmetry that inner and outer monolayer may be exchanged if the curvature changes sign at the same time.
Figure 14 shows the phase diagram resulting from this simple model for positive $c_1 = c_2$. In that case it is clear that also $\bar{c} > 0$ and $\beta\bar{c} > 0$ is fulfilled for positive β. Nevertheless, there is also a region with negative β where $\beta\bar{c} > 0$ because $\bar{c}$ decreases with negative β via (17). This means that the repulsive interaction between different lipids may compensate or even reverse the sign of the spontaneous curvature. In all cases of vesicular (curved membrane) phases one has to

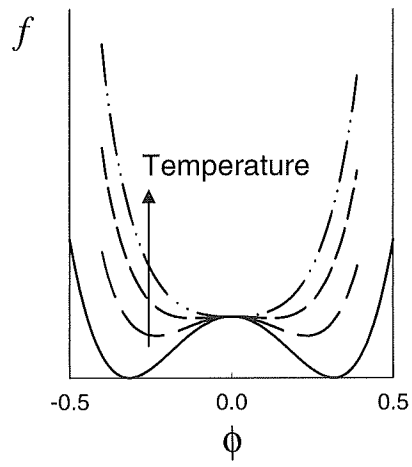


Fig. 15: *Influence of the temperature on the free energy considering the mixing entropy.* $\phi^* = 1/\sqrt{10}$. *Black: limit* $T = 0$. *Temperatures:* $T = 0$ *(continuous line),* $k_BT/A\kappa\bar{c}^2 = 8$ *(long dashed),* 16 *(short dashed), and* 24 *(dash-dotted).*

observe that $|\phi| \leq \min(|\psi|, |1-\psi|)$ because the most asymmetric situation is that where one lipid species depletes completely from one monolayer. This means that high absolute values of ϕ^* cannot be attained but only the limit of total depletion $\pm\min(|\psi|, |1-\psi|)$. In these regions the curvature does not increase with $|\beta|$ any more but remains constant at $\pm\alpha|\psi|$ or $\pm\alpha|1-\psi|$. It is interesting that this model does not require a difference between c_1 and c_2 what would be naïvely expected (e.g. Figure 2c) from lipids of different conicity redistributing in order to achieve a certain overall curvature. On the other hand the interaction is essential. It is easy to see that minimising the energy (10) leaves ϕ undetermined if $\beta = 0$.
In the above treatment the entropy of mixing was neglected. Therefore, the phase diagram in Figure 14 corresponds to the situation in the limit $T \to 0$. The free energy of mixing (per unit area) which has to be added to the bending energy is for ideal mixing:

$$f_{\text{mix}} = \frac{k_BT}{A}\left(\psi_i \ln\psi_i + (1-\psi_i)\ln(1-\psi_i) + \psi_o \ln\psi_o + (1-\psi_o)\ln(1-\psi_o)\right) . \qquad (22)$$

Here, A is the surface area occupied by a single lipid molecule. Using the sum and difference variables ψ and ϕ and the expression (10) the total free energy is

$$\begin{aligned} f \;=\; & 4\kappa\left(\left(\bar{c}-\beta\phi^2\right)^2 + (\alpha\phi - c)^2\right) + \\ & \frac{k_BT}{A}((\psi-\phi)\ln(\psi-\phi) + (\psi+\phi)\ln(\psi+\phi) + (1-\psi-\phi)\ln(1-\psi-\phi) \\ & \quad +(1-\psi+\phi)\ln(1-\psi+\phi)) . \end{aligned} \qquad (23)$$

The influence of the mixing entropy term is shown qualitatively in Figure 15. As expected it favours symmetrical composition bilayers ($\phi = 0$). For high temperatures the double-minimum property of f may get lost. In consequence for higher temperatures the lamellar region is more extended in the phase diagram.
Evidence of formation of highly curved vesicles by (electrostatic) interaction between different lipids can be found in mixtures of anionic and cationic lipids. While cetyl trimethylammonium

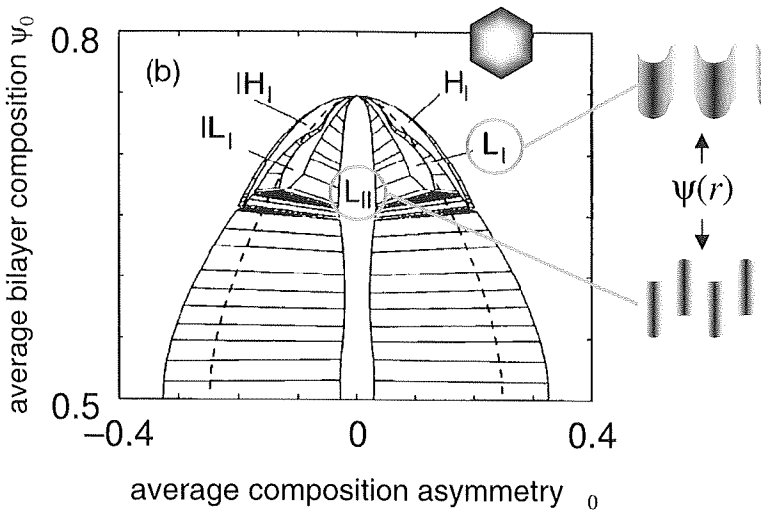


Fig. 16: *Phase diagram for the modulated phase model of the weak segregation limit. Parameters:* $a' = -2.7$, $b' = 0$, $(C^2 - b_1'\sigma)^2/8\kappa C^2 = 0.24$, $k_BT = 1.2$. H_I *denotes a hexagonally modulated phase,* L_I *and* L_{II} *are lamellar phases described in the text. The prefix* I *means "inverse", i.e. reversal of lipid species. The dashed line is the spinodal of the metastable* L_{II} *phase. (from [20])*

tosylate (CTAB, cationic) and sodium dodecylbenzene sulfonate (SDBD, anionic) alone do not form vesicles, mixtures of both spontaneously form vesicles with radii of 30–80 nm [19].
Also for the weak segregation limit models of modulated phases have been devised, e.g. [20]. This specific model is characterised by (1) opposite spontaneous curvature of the component lipids, (2) inclusion of mixing entropy, (3) in-layer interaction, (3) transbilayer interaction, (4) chemical potentials controlling the concentrations of lipids. From these assumptions a free energy functional of the concentration variables $\phi(\tau)$ and $\psi(\tau)$ (defined as before but as functions of the 2D co-ordinate τ) and the Monge height $z(\tau)$ can be constructed:

$$
\begin{aligned}
F[\phi,\psi,z] = & \\
\int \mathrm{d}^2\tau \; \Big(& \frac{\kappa}{2}\left(\nabla^2 z\right)^2 + \frac{\sigma}{2}(\nabla z)^2 + C\phi\nabla^2 z && (24)\\
& + \frac{b_1'}{2}(\nabla\phi)^2 + \frac{b_2'}{2}(\nabla\psi)^2 + \frac{a'}{2}\phi^2 + \frac{b'}{2}\psi^2 && (25)\\
& + \frac{k_BT}{2}\big(\, (1+\psi+\phi)\ln(1+\psi+\phi) + && (26)\\
& \qquad (1-\psi-\phi)\ln(1-\psi-\phi) + \\
& \qquad (1+\psi-\phi)\ln(1+\psi-\phi) + \\
& \qquad (1-\psi+\phi)\ln(1-\psi+\phi) \, \big) \\
& - \eta\phi - \Delta\psi \; \Big). && (27)
\end{aligned}
$$

The term (24) is the curvature energy where C is the coupling coefficient between concentration asymmetry and spontaneous curvature. Term (25) represents the interaction between lipids

where b'_1, b'_2, a', and b' are linear combinations of the basic interaction energies of the model J_1 (in-plane), J_2 (interplane), and J_3 (diagonal interplane). The mixing entropy contribution is expressed by (26). Finally the control of the concentrations by chemical potentials is contained in (27) where η and Δ are the difference and sum of chemical potentials of the two lipid species. The free energy is—like in the strong segregation limit—solved by a Fourier ansatz. The resulting phase diagram for one certain combination of parameters is shown in Figure 16. A large number of different phases have to be taken into account because there are various possibilities how the curvature correlates with the composition fluctuation. E.g. Figure 16 shows two striped phases L_I and L_{II} which have the same curvature modulation. Nevertheless the concentration modulation has the double spatial frequency for the L_{II} case. There, one lipid species concentrates at the positions of highest curvature while for the L_I phase it concentrates at the height minima.

5 Summary and Outlook

We have seen that there are two fundamentally different approaches to the description of the interaction of lipid distribution and curvature of the membrane. In the strong segregation limit the equilibrium state results from a balance of curvature energy and line energy of the microphase interface. In the weak segregation limit a dependence of the spontaneous curvature on lipid composition is considered.
Several experimental findings can be explained by models based on one of the two principles: The spontaneous creation of vesicles may either be due to microphase separation or interaction-induced changes of the spontaneous curvature. The fact that also artificial vesicles show an inside-outside asymmetry can be explained by a coupling of curvature to lipid composition. Etc.
There is currently a large number of theoretical models and no single 'standard model' of the multicomponent membrane. This is due to the fact that there are many parameters which may or may not be included in a certain model: spontaneous curvature, lipid interaction, entropy of mixing, transmembrane coupling, polydispersity of vesicles ... All these models explain certain experimental results but sometimes in strongly different ways.
Also on the experimental side the situation is still in flux. While earlier methods only could observe qualitatively phenomena predicted by the models (e.g. the occurrence of spontaneous budding) new methods give access to microscopic quantities (e.g. the fact that the buds are indeed separate microphases). It can be expected that a further development of experimental techniques will lead to more stringent test of the models. In the future it should be possible to find out the properties of the membrane lipids which are essential to their physical function.

References

[1] R. B. Gennis, *Biomembranes* (Springer, New York 1989)

[2] D. Hoekstra ed., *Cell Lipids* (Academic Press, San Diego 1994)

[3] P. L. Yeagle, *The Membranes of Cells* (Academic Press, San Diego 1993)

[4] P. W. Majerus, D. B. Wilson, T. M. Connolly, T. E. Bross, E. J. Neufeld, Science **234**, 1519 (1986)

[5] E. A. Dawidowicz, Ann. Rev. Biochem. **56**, 43 (1987)

[6] Chapter 6 in ref. 1

[7] P. Rice, H. M. McConnell, Proc. Natl. Acad. Sci. USA **86**, 6445 (1989)

[8] C. Yu, J. Furlong, P. Burgos, L. J. Johnston, Biophys. J. *82*, 2526 (2002)

[9] J. A. F. Op den Kamp, Ann. Rev. Biochem. **48**, 47 (1979)

[10] A. J. Verkleij, R. F. A. Zwaal, B. Roelofsen, P. Comfurius, D. Kastelijn, L. L. M. Van Deenen, Biochim. Biophys. Acta **323**, 178 (1973)

[11] A. Kumar, C. M. Gupta, Biochemistry **24**, 5157 (1985)

[12] J. Korlach, P. Schwille, W. W. Webb, G. W. Feigenson, Proc. Natl. Acad. Sci. USA **96**, 8461 (1999)

[13] G. W. Feigenson, J. T. Buboltz, Biophys. J. **80**, 2775 (2001)

[14] R. Lipowsky, J. Phys. II France **2**, 1825 (1992)

[15] H.-G. Döbereiner, J. Käs, D. Noppl, I. Sprenger, E. Sackmann, Biophys. J. **65**, 1396 (1993)

[16] T. Baumgart, S. T. Hess, W. W. Webb, Nature **425**, 821 (2003)

[17] W. T. Góźdź, G. Gompper, Europhys. Lett. **55**, 587 (2001)

[18] S. A. Safran, P. Pincus, D. Andelman, Science **248**, 354 (1990)

[19] E. W. Kaler, A. K. Murthy, B. E. Rodriguez, J. A. N. Zasadzinski, Science **245**, 1371 (1989)

[20] P. B. Sunil Kumar, G. Gompper, R. Lipowsky, Phys. Rev. E **60**, 4610 (1999)

D 5 Lipid rafts in multicomponent membranes

K. Simons
Max-Plack-Institute of Molecular Cell
Biology and Genetics
Dresden

Contents

1 Introduction 2

2 What do we learn from model systems? 3

3 How does cholesterol interact with neighbouring phospholipids ? 4

4 How do the model membrane studies relate to cell membranes? 10

5 Protein interactions with lipid rafts 14

6 What does detergent resistance tell us about lipid domains ? 16

7 Perspectives 18

8 Postscript concerning Terminology 19

A Appendix: Abbreviations 21

1 Introduction[1]

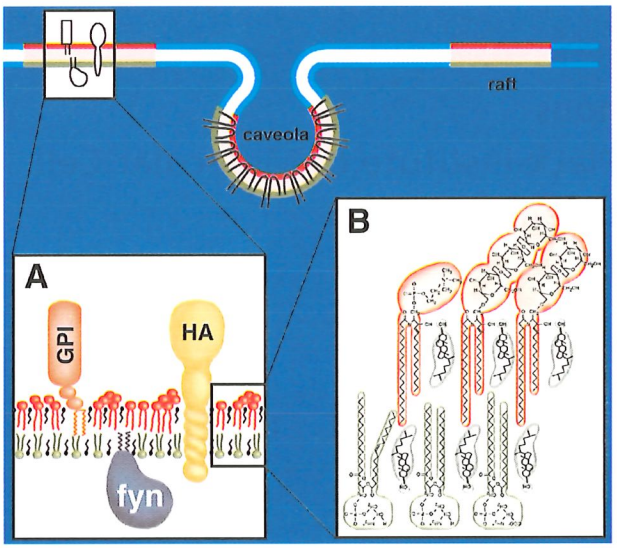


Fig. 1: *Lipid raft*

The lipid bilayer that forms cell membranes is a two dimensional liquid the organization of which has been the object of intensive investigations for decades by biochemists and biophysicists. Although the bulk of the bilayer has been considered to be a homogeneous fluid, there have been repeated attempts to introduce lateral heterogeneities, lipid microdomains, into our model for the structure and dynamics of the bilayer liquid (22, 37, 38, 89). The identification of boundary lipids around proteins created excitement in the 1970s but disappeared into the ESR realm of research (48) when it was shown that on the time-scale of NMR experiments the boundary lipids could not be observed.

The realization that epithelial cells polarize their cell surfaces into apical and basolateral domains with different protein and LIPID compositions in each of these domains, initiated a new development that led to the "lipid raft" concept which has stirred up commotion and a lively controversy in the field (80, 78). The concept of assemblies of sphingolipids and cholesterol functioning as platforms for membrane proteins was promoted by the observation that these assemblies seemed to survive detergent extraction (8). This operational breakthrough caused a flood of papers where raft-association was equated with resistance to Triton-X100 extraction at 4°C. The addition of another criterion, depletion of cholesterol using methyl-β-cyclodextrin (33, 72), leading to loss of detergent resistance, added even more papers to the field. But at the same time these new developments added further controversies. Do lipid rafts exist in membranes or are they just artifacts of detergent extraction? Despite, or

[1] This manuscript appears in similar form in: K. Simons, W.L.C. Vaz, Annu.Rev.Biophys.Biomol.Struct. **33**, 2004

maybe because of all this controversy, the field has matured and many new methods have added substance to the definition of lipid domains in cell membranes and have significantly improved our understanding of heterogeneity in membrane systems. There is now increasing support for a role of lipid assemblies in regulating numerous cellular processes including cell polarity, protein trafficking and signal transduction. As Edidin (16) aptly put it: "Despite great reservations about the interpretation of classical operational definition of lipid raft components and function, we are left with . . . . the stubborn insistence by cells that raft lipids can be organized and segregated into membrane domains". In this review we will try to integrate what we have learned about phase immiscibility in model systems with what we know of lipid domain organization in cell membranes today.

2 What do we learn from model systems?

Cell membranes are two-dimensional liquids. Thus, lateral heterogeneity implies liquid-liquid immiscibility in the membrane plane. It has been well known that hydrated lipid bilayers undergo phase transitions as a function of temperature. These transitions, which occur at defined temperatures for each lipid species, always involve some change in the order of the system. The most important of these transitions is the so-called "main" or "chain-melting" transition in which the bilayer is transformed from a highly ordered quasi-two dimensional crystalline solid to a quasi-two dimensional liquid. It involves a drastic change in the order of the systems, in particular of the translational (positional) order in the bilayer plane and of the conformational order of the lipid chains in a direction perpendicular to this plane. Translational order is related to the lateral diffusion coefficient in the plane of the membrane and conformational order is related to the trans/gauche ratio in the acyl chains. The main transition has been described as an ordered-to-disordered phase transition, so that the two phases may be labeled as solid-ordered (so) below the transition temperature and liquid-disordered (ld) above that temperature. An important advance was the realization that cholesterol and phopholipids could form a liquid-ordered (lo)) phase that could coexist with a cholesterol-poor liquid-disordered (ld) phase thereby permitting phase coexistence in wholly liquid phase membranes (34,35). Sterols do so as a result of their flat and rigid molecular structure, which is able to impose a conformational ordering upon a neighboring aliphatic chain (68), when the sterol is the nearest neighbor of the chain, without imposing a corresponding drastic reduction of the translational mobility of the lipid (56). Due to the fact that the sterol does not fit exactly in the crystalline lattice of an so (gel) lipid bilayer phase it will, if it dissolves within this phase, disrupt the crystalline translational order without significantly perturbing the conformational order. Thus, cholesterol at adequate molar fractions can convert ld or so lipid bilayer phases to liquid-ordered (lo) phases. This conceptual and experimental progress in model membrane research set the stage for detailed analysis of the molecular basis for liquid-liquid immiscibility.

The degree of translational freedom (lateral mobility) of the lipid molecules in an lo phase is similar to that in an ld phase, the lateral diffusion coefficient being only reduced by a factor of about 2–3 in the former compared to the latter (2). The conformational order of the lipid hydrocarbon chains in a lo phase is, however, more similar to that of the so phase (21). Exactly by how much the conformational order of an lo phase differs from that of an so or an ld phase has yet to be defined for most lipid bilayers that can serve as models for biological membranes. This is particularly true for the interaction of cholesterol with lipids containing unsaturated acyl chains where it is not clear whether cholesterol configurationally orders the

cis-unsaturated chain. POPC, for example, forms an lo phase with cholesterol only at a molar fraction of ≥40 mol % cholesterol (49) and the translational diffusion coefficient in this phase is about twofold lower than in the ld phase formed from pure POPC (95). The same effect is achieved in bilayers formed from DMPC, DPPC or SpM (above the transition temperature) at ~25–30 mol % cholesterol (1, 67, 91). It also remains to be defined in exactly what respects lo phases formed from saturated and unsaturated lipids are different from or similar to each other. Inspite of these uncertainties, there is broad agreement that liquid-ordered and liquid-disordered domains can coexist in hydrated bilayers containing cholesterol and saturated phospholipids and that cholesterol plays a key role in lo phase formation (1, 67, 68, 69, 91).

There has been considerable discussion in the literature as to how large these domains can be. Feigenson and Buboltz (18) have shown that the domains can be large enough to be visible under the light microscope (i.e., larger than ≥500 nm across) or have dimensions that are below the limit of resolution of the optical microscope. There is no fundamental requirement that any phase in a heterogeneous system not be divided into several parts or domains, although their exact thermodynamic description becomes unreliable when the domains become too small. When phases coexist, there is a mismatch of properties at the interface where two phases meet. This mismatch results in a tension (a line tension in two-dimensional systems like lipid bilayers, and a surface tension in three-dimensions) that works to reduce the interface of mismatch and, as a result, causes a macroscopic phase separation. In lipid bilayers, the dipole moments of the lipid molecules in each monolayer are, due to geometric constraints of the system, oriented in the same direction so that a repulsive interaction between the lipid molecules results. This repulsion works against the line tension at the phase boundary so that the domain size and shape is a balance between these two contradictory forces (84). Additionally, effects like "surfactancy" (the ability of surface-active agents to reduce the interfacial surface tension at the interface between two phases as is commonly seen in emulsions) can reduce the interfacial tension to the point that it no longer drives the system to a macroscopic phase separation. The extreme complexity of the chemical composition of biological membranes makes "surfactancy" at the domain boundaries a very likely possibility. Many of the protein and lipid components of biological membranes might actually preferentially partition into domain boundaries.

3 How does cholesterol interact with neighbouring phospholipids ?

McConnell and coworkers, in a series of papers published over the last four years [reviewed in (51, 52)] have suggested that cholesterol can form reversible oligomeric chemical complexes with phospholipids with a fundamental stoichiometry of between 3 phospholipids per 2 cholesterols and 2 phospholipids per cholesterol in a cooperative manner with a cooperativity factor of between 4 and 12. Thus a single complex, depending upon the exact fundamental stoichiometry and the cooperativity factor, could have between a total of 60 molecules of which 24 would be cholesterol and a minimum of 12 molecules of which 4 would be cholesterol. The formation of these complexes has been proposed to lead to the "condensing" effect of cholesterol, by which the surface area occupied by the phospholipid is decreased by interaction with cholesterol. The condensed complex is proposed to be stable at temperatures below the phase transition temperature of the phospholipids and to dissociate endothermically above this temperature resulting in the broad endothermic process seen in calorimetric studies and generally attributed to a transformation of an lo phase to an ld phase.

Effectively, in this view, the micelle-like cholesterol-lipid aggregates would exist in an lo phase below the phase transition temperature of the lipid. These conclusions are derived from studies on mixed monolayers, which show phase diagrams with two upper miscibility critical points, and a proportional increase in the chemical potential of cholesterol above the complex stoichiometry. It is assumed that the behavior of monolayers can be extrapolated to bilayer membranes.

Huang and Feigenson (32) have presented another model of cholesterol-phospholipid mixing, based on multibody interactions instead of pair-wise additivity of nearest neighbor interactions. They propose a model where nonpolar cholesterol relies on shielding by the lipid polar head groups to avoid the energetically unfavorable contact with water. The phospholipid head groups effectively act as "umbrellas" for the cholesterol molecules below them. As the concentration of cholesterol increases, acyl chains and cholesterol become more tightly packed because they share the limited space under the phospholipid head groups. When the head groups cannot cover additional cholesterol molecules, the solubility limit is reached and cholesterol precipitates to form a separate cholesterol monohydrate crystalline phase (31). The interactions of glycerophospholipids with cholesterol decreases in the following order: phosphatidylcholine > phosphatidylserine > phosphatidylethanolamine. It is also important to note that cholesterol has a preference for interaction with lipids that have fully saturated aliphatic chains when compared with lipids that have one or more unsaturated chains [reviewed in (75)]. The former can be more conformationally ordered by the molecular flatness of the rigid sterol ring structure than the latter. Also, it is to be expected that the position of a double bond in an aliphatic chain and its configuration will be of significance with regard to the interaction of this chain with cholesterol. Unsaturation beyond the fourteenth to fifteenth carbon atom in the chain is likely to have very little if any effect upon interaction with cholesterol since the chain conformation beyond this position is not likely to enter in conflict with the rigid and flat ring structure of the sterol.

Multiple unsaturation of the fatty acyl chain significantly decreases the interaction with cholesterol in both bilayers and monolayers. This can be observed by comparing the areas of phospholipids molecules at the air-water interface in the absence and presence (at a molar ratio of about 1:1) of cholesterol in monolayers at pressures equivalent to those in cell membranes. POPC condenses from an area of 0.61 nm2 to an area of 0.42 nm2; saturated DPPC condenses from an area of 0.46 nm2 to an area of 0.39 nm2; 18:1-SpM condenses from an area of 0.57 nm2 to an area of 0.40 nm2; and egg-SpM condenses from an area of 0.49 nm2 to an area of 0.40 nm2 (81).

A considerable amount of data demonstrates that cholesterol interacts more favorably with SpM than with other phospholipids such as PC in both bilayers and monolayers (9, 57, 63). The rate of cholesterol desorption from SpM-containing bilayers is slower than desorption from membranes with saturated and acyl chain-matched PCs. Cholesterol preferentially abolishes the phase transition of SpM in binary mixtures with other phospholipids, indicating a preferential interaction with the SpM molecules. The water permeability is lower in SpM/cholesterol bilayers than in PC/cholesterol membranes, indicative of denser lateral packing in the former system. In monolayers, the oxidation of cholesterol by cholesterol-oxidase was reduced in SpM-containing monolayers as compared to PC-containing monolayers. Another sensitive indicator of SpM-cholesterol interactions is the measurement of the interfacial elasticity in monolayers (9). Recent studies (82) demonstrate that the measured elasticity-reducing effect of cholesterol on SpM is significantly stronger than on PCs with fully saturated and matched acyl chains.

The reason for the preference of cholesterol for SpM as compared to PC lies in the structural differences of these lipid molecules. SpM is a derivative of sphingosine (d-erythro-2-amino-trans-4-octadecene-1,3-diol). A fatty acid is attached to the 2-amino group of sphingosine by formation of an amide. In mammalian sphingolipids this fatty acid is variable but is generally a 16–24 carbon saturated chain with a small fraction of 24:1Δ15cis chains depending upon the source tissue. Thus, in terms of the apolar part of the molecules, these lipids present rather long fully saturated aliphatic chains (with a trans unsaturation in the sphingosine base) and when a cis-unsaturation appears it is located deep down toward the bilayer midplane. Monounsaturated PCs, the major species in mammalian cell membranes, usually have a cis double bond in the Δ9 position of one of the fatty acids and may have more double bonds lower down the acyl chain. Therefore, in terms of what was discussed earlier with regard to cholesterol interactions with lipids, sphingolipids present ideal partners for interaction with cholesterol. The polar portions of the molecules are also of interest from this perspective. At the membrane-water interface the sphingosine-based lipids have a hydroxyl group and an amido nitrogen, both of which can act as hydrogen bond donors as well as acceptors. These groups can, together with the fatty acid carbonyl functions, hydrogen bond to water and to the other lipids. In comparison the PCs, having only hydrogen bond acceptors in the form of the ester carbonyl functions and oxygen atoms, have a lower versatility in hydrogen bonding with the water molecules in the aqueous interface and are not able to directly hydrogen bond with their lipid neighbors in the bilayer. If hydrogen bonding is assumed to be a structure-stabilizing interaction, it must be concluded that sphingolipids have a greater propensity for ordered structures in the membrane. The most abundant sphingolipid in membranes is SpM, which has a phosphorylcholine head group attached to the 1-hydroxyl group of the acylamidosphingosine (ceramide). This phosphorylcholine head group, analogous to the phosphorylcholine head group in PCs, can eventually serve as a very effective "umbrella" in terms of favoring the solubilization of cholesterol in a SpM bilayer.

Systematic alteration of the functional groups in the SpM molecule have identified the amide linkage to be important for the interaction with cholesterol, possibly due to hydrogen bonding to the cholesterol 3-OH group. Recent infrared spectroscopic studies on egg SpM showed that the amide-I band was shifted to lower wave numbers in the presence of cholesterol. This was interpreted to be indicative of a change in the hydrogen-bonding pattern of the SpM amide group, possibly connecting to the hydroxyl group of cholesterol (90). However, results from NMR (24) and fluorescence (30) spectroscopies indicate that there may be no chemical complex formed between SpM and cholesterol.

The differential interaction of cholesterol with SpM and unsaturated PC is convincingly reflected in the recent ternary phase diagram of palmitoyl-SpM, POPC, and cholesterol (11). This is the first phase diagram of lipid bilayers formed from the ternary mixture of the three major lipid components of the exoplasmic leaflet of mammalian plasma membranes and, although rather simple in compositional terms compared to a biological membrane it is a good starting point for our understanding of the physical complexity of the exoplasmic leaflet of the plasma membrane. The phase diagram (**Fig.1**) shows that only ld and lo phases coexist at concentrations mimicking the composition of the outer leaflet of cell plasma membranes. Both coexisting phases contain all three chemical constituents of the system, the lo phase being richer in SpM and the ld phase richer in POPC. Other studies, notably those of Silvius (76) and Feigenson (18) have also shown liquid-liquid (lo/ld) immiscibilities in ternary mixtures containing cholesterol and two PCs, one of which had long saturated acyl chains. In the latter study, the liquid-liquid coexistence was shown to form domains that were visible in

the light microscope in certain parts of the phase diagram but domains of dimensions below the limit of resolution of optical microscopy in other regions of the diagram. Using fluorescent probes that partition preferentially into one of the coexisting liquid phases and optical microscopy, lo/ld phase coexistence has been visualized in monolayers at the air-water interface, supported monolayers and bilayers, planar lipid bilayers, and giant unilamellar vesicles containing SpM, unsaturated PCs and cholesterol (13, 14, 66). Even lipid extracts from epithelial brush border membranes form monolayers or bilayers, which display liquid-liquid immiscibilities (13). Liquid-liquid immiscibility and the formation of domains are clearly dependent upon cholesterol concentration. Treatment of such membrane systems with methyl-β-cyclodextrin abolishes the domains. Dietrich et al. (14) demonstrated that two raft components, the ganglioside GM1 and the GPI-anchored protein Thy-1 significantly partitioned into the raft-like liquid-ordered domains in supported monolayers. In similar monolayers the partitioning of monomers and antibody-cross-linked dimers of the fluorescein-labeled lipids, F-DPPE (saturated acyl chains) and F-DOPE (unsaturated acyl chains), between domains of an lo and an ld phase were compared. Both probes partitioned preferentially into the ld phase. The estimates for the equilibrium partition coefficients between the lo and ld phases were 0.07 for F-DOPE and 0.14 for F-DPPE. A report from the same laboratory (13) indicated that F-DPPE partitioned about equally between lo and ld domains in supported bilayers. Dimerizing the probes with antifluorescein antibody resulted in a preferential partitioning of the F-DPPE into the lo domains whereas the F-DOPE remained in ld domains (14). This is an interesting finding because it could be interpreted to mean that cross-linking (dimerization or multimerization) of some membrane constituents can drive them into rafts. Silvius and coworkers (92) used a fluorescence-quenching assay to study the partitioning of lipidated peptide species into liquid-ordered domains in lipid bilayers composed of DPPC, a spin-labeled unsaturated PC, and cholesterol. They demonstrated that peptides incorporating isoprenyl groups, or multiple unsaturated acyl chains showed a low affinity for lo domains. However, peptides containing multiple S- and/or N-acyl chains or a cholesterol residue plus an N-terminal palmitoyl chain displayed significant partitioning into lo domains under the same conditions.

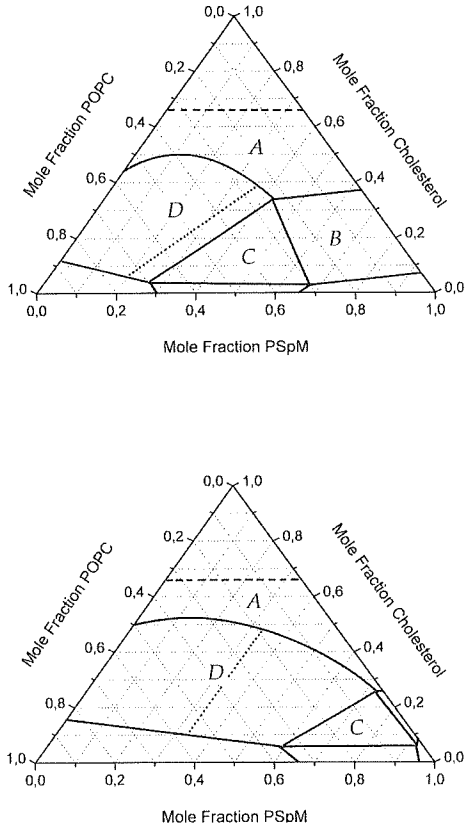


Fig. 2: *Phase diagrams for fully hydrated lipid bilayer membranes prepared from ternary mixtures of palmitoyl-SpM (PSpM), POPC, and cholesterol (adapted, with permission, from reference (11). An experimentally obtained phase diagram is shown for 23°C (upper panel) and a hypothetical phase diagram is shown for 37°C (lower panel). In each panel the dashed line at 66 mol % cholesterol is the hypothetical limit of cholesterol solubility in the bilayer phase. Above 66 mol %, excess cholesterol separates out as a cholesterol monohydrate crystalline phase (32). To avoid cluttering, the major regions in the phase diagram have been labeled A through D. Starting from the region A, which corresponds to a single l_o phase, and proceeding clockwise, the regions along the axes of the diagram correspond to a two-phase region with s_o + l_o phase coexistence (marked B), a single s_o phase region (unmarked), a two-phase region with s_o + l_d phase coexistence (unmarked), a single l_d phase region (unmarked), and a two-phase region with l_d + l_o phase coexistence (marked D). The central tie-triangle (marked C) is a three-phase region with l_d + l_o + s_o phase coexistence. In the region marked D a possible tie-line passing through the 1:1:1 composition is indicated by a dotted line in both panels. For further details the reader is referred to the original work (11).*

Another interesting tool for studying lipid domains is atomic force microscopy since it displays the surface landscape of a supported monolayer or bilayer with and without proteins. Using this technique it becomes possible to visualize l_o/l_d phase coexistence as well as to study the partitioning of gangliosides and GPI-anchored proteins in bilayers that show this phase coexistence. An l_o domain in a lipid bilayer composed of $C_{18:0}$-SpM and cholesterol should have a thickness of 4.6 nm while a bilayer containing di-$C_{18:1}$-PC is 3.5 nm thick. Inclusion of cholesterol increases the thickness of the di-$C_{18:1}$-PC bilayer to 4.0 nm. In contrast, cholesterol does not affect the thickness of a SpM bilayer. Studies using AFM on supported bilayers have been able to visualize domain formation by the height difference of l_d and l_o domains. Rinia et al. (64) using 1:1 mixtures of DOPC and SpM saw domains increasing in size and area coverage with increasing concentration of cholesterol. The height difference between the ordered domains and the surrounding fluid bilayers decreased from 1 nm in the absence of cholesterol (s_O phase SpM domains) to 0.8 nm at 50 mol % cholesterol (l_O phase SpM-colesterol domains). Milhiet et al. (54) used POPC and SpM (1:3) and observed microscopic separation up to 25 mol % cholesterol with height differences of the domains of 0.3–0.4 nm. Both these studies were performed on compositionally symmetric bilayers. Yuan et al. (96) used bilayers of SpM/DOPC/cholesterol 1:1:1 to which 1% of the ganglioside GM1 was added. In this case, a successive deposition of two monolayers onto the support, formed an asymmetric supported bilayer in which the first (lower) monolayer was of DPPE, and the upper monolayer was prepared from the SpM/DOPC/cholesterol mixture. They saw little height difference between the ordered SpM/cholesterol-rich and the disordered DOPC-rich phases in the bilayer. The GM1 molecules could be visualized as small islands of 200–300 nm in diameter and about 2nm above the bilayer matrix. Thus, ganglioside formed small aggregates presumably within the ordered phase not detected by AFM. Saslowsky et al. (70) studied the incorporation of GPI-anchored placental alkaline phosphatase (PLAP) into supported bilayers with AFM. As in the previous studies they could see the formation of domains in equimolar mixtures of SpM, DOPC, and cholesterol that had a height difference of 0.7 nm. When PLAP was reconstituted into liposomes made from this mixture and these liposomes were transferred to a mica surface to form a supported bilayer, lipid domain formation was seen with AFM and PLAP was seen protruding out of the elevated liquid-ordered domains. PLAP mostly occupied a molecular volume corresponding to a dimer. However, also larger protruding particles were seen that could correspond to PLAP oligomers. In all the AFM studies domain formation was also seen with the SpM/PC mixtures without cholesterol. These domains probably correspond to s_o/l_d phase coexistence in which the s_o phase was rich in SpM and the l_d phase was rich in DOPC.

AFM studies and optical microscopy on model bilayers have also demonstrated that the outer and the inner leaftlets in the liquid-ordered domains are coupled in bilayers prepared from mixtures containing SpM, unsaturated PCs, and cholesterol. However, in cell membranes the inner leaflet is not composed of the same lipids as the outer leaflet. The inner leaflet contains phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine and phosphoinositides. Wang and Silvius (93) studied mixtures of inner leaflet lipids together with cholesterol and could not detect formation of segregated liquid-ordered domains employing a fluorescence-quenching assay. Keller et al. (40), on the other hand, could see phase segregation when small amounts of SpM were present. This is interesting because although SpM is asymmetrically distributed in cell membranes, a fraction is known to be in the inner leaflet.

4 How do the model membrane studies relate to cell membranes?

The studies with model monolayers and bilayers have demonstrated that mixtures of lipids mimicking the composition of the outer leaflet of plasma membranes exhibit liquid-liquid immiscibility and segregate into l_o and l_d domains. SpM, which carries mostly saturated hydrocarbon chains, preferably partitions with cholesterol into l_o phase domains, segregating from unsaturated PCs, which are the major constituents of domains of the l_d phase. The size of l_o domains seems to vary greatly depending on temperature, pressure and composition. The domains can be large enough to be visible in the optical microscope or smaller than the limit of resolution of light microscopy (18). Cell membranes on the other hand have an extremely complex lipid composition and the fact that lipids are continuously being added to or removed from the membrane does not simplify the situation. A typical mammalian cell plasma membrane is constituted from about 8 major classes of lipids including cholesterol (94). In each of these classes, with the exception of cholesterol, there is a great variation in the acyl chain composition so that the total number of chemical species that compose the lipid bilayer of these membranes is extremely large. Recent mass spectrometric analysis demonstrated that only for PC there were close to hundred molecular species present in MDCK cells.(16b) The binary and ternary lipid systems studied so far can only provide the boundary conditions within the framework of which we may attempt to extrapolate towards understanding the complexity of cellular membranes.

Rafts have been proposed as lipid platforms of a special chemical composition (rich in sphingomyelin and cholesterol in the outer leaflet of the cell membrane) that function to segregate membrane components within the cell membrane. As discussed later in this review, they are understood to be small (estimates of size vary considerably) but they can be coalesced under certain conditions. An estimate of the number of "rafts" in a cell plasma membrane (based upon lipid composition, total membrane size and size of the raft domains) would be on the order of 10^5 to 10^6 (77). Their specificity with regard to lipid composition is reminiscent of phase separation behavior in heterogeneous model membrane systems. In fact, many of their properties with regard to chemical composition and detergent solubility are similar to what is observed in model systems composed of ternary mixtures of an unsaturated phosphatidylcholine, sphingomyelin (or a long-chain saturated phosphatidylcholine), and cholesterol (11). Rafts could be considered domains of a l_o phase in a heterogeneous l phase lipid bilayer composing the plasma membrane. What the other coexisting phase (or phases) is (or are) is not clear at this time. There is consensus that the biological membrane is a liquid, so s_o phase coexistence may be ignored for most cases. Whether the other phase (phases) is (are) l_d or l_o phases will depend upon the chemical identity of the phospholipids that constitute this phase (these phases) and the molar fraction of cholesterol in them. At this point it would appear most correct ***to equate rafts with a liquid-ordered phase*** and refer to the rest of the membrane as the nonraft liquid phase.

Although there is increasing consensus that lipid rafts do exist in cell membranes (7, 16, 46, 75, 79, 45) this field of research is still in its infancy. There is an ongoing debate on the size and life time of rafts (20, 54 b). This is not understood even in model systems but there is evidence that domain sizes of l_o phases may be dependent upon their composition. When proteins are incorporated into these domains, still another level of complexity is introduced. In fact, there are contradictory views of how rafts organize themselves. One view maintains that raft proteins act as nucleation sites for raft domains. Anderson and Jacobson (4)

suggested that raft proteins organize "shells" composed of maximally 80 molecules of cholesterol and sphingolipid around them. Assuming a GPI anchor for the raft protein, the shell would have about 4–5 layers of lipid around the anchor. The shells are postulated to be what facilitates the proteins to eventually integrate into large rafts. This hypothesis raises certain fundamental questions: First, the association of the first layer of lipids with the protein membrane anchor would require a rather strong interaction of the shell lipids (which presumably are the same as those typically encountered in rafts) with the anchor. No such specificity has yet been demonstrated and all that is known about "boundary" lipids hardly indicates that such a strong interaction is probable. Boundary lipid-protein interaction, limited on the average to less than a millionth of a second, has not been shown, with the exception of a very few cases, to show any very significant specificity (48). Second, formation of the succeeding layers in the shell would require that the lipids of the shell have a rather strong tendency to associate. If this were the case, one wonders why the raft protein would be needed at all. An alternative is to view the proposed "shell" as a solvation shell around the raft protein in which case its composition should reflect the chemical composition of the entire lipid bilayer in which the protein is inserted. It is not clear in this case why this shell would have any particular affinity for larger rafts.

Another view is based on studies by Kusumi and co-workers who have developed high-speed video microscopy to study lipid and protein movement in the plasma membrane of living cells. The diffusive walk of 40nm beads coupled by antibodies to lipids and to GPI-anchord lipids was monitored at the time resolution of 25 μs. Kusumi et al. propose that raft proteins such as GPI-anchored proteins form a small "raft" containing only a few raft lipids with a life time of less than 1 ms. These dynamic entities can coalesce and cluster to stabilized or to signalling rafts (86).

A third view of raft organization maintains that lipid rafts are manifestations of the thermodynamic properties of the lipid bilayer of the membrane, namely, that this lipid bilayer is a heterogeneous system (78). The phase coexistence is manifested as the existence of domains of the coexisting phases. A subset of these domains, corresponding to one particular phase with defined physical and chemical properties are the rafts. These domains are proposed to be not just condensed complexes of cholesterol and sphingolipids around proteins but are larger. How large? The use of different techniques has given variable size estimates (<700nm) for lipid rafts in membranes (41, 61, 62, 74, 15). However, if one analyses the different size determinations, it becomes obvious that they reflect different states of rafts. The large 700 nm size was obtained from measurements using single molecule tracking on the plasma membrane of human aortic smooth muscle cells (74). Rafts of this size could not be seen in other cell types and might have been due to raft clustering into caveolar or other large raft domains (G. Schütz, personal communication). An estimate of 200 nm raft size was obtained by single particle tracking. This work, however, used beads to which multiple antibodies were bound and multiple contacts with GPI-anchored proteins on the cell surface could have caused raft aggregation (15). So what could the intrinsic raft size in the resting state (non-clustered) be? Using another method, photonic force microscopy (29b) a larger raft size was obtained (61). This method employs a laser trap to confine the motion of a bead bound by antibodies to raft and nonraft proteins into a small area (< 100 nm) and used high-resolution single particle tracking to estimate the local viscous drag upon the particle. Multiple attachment of the bead to the plasma membrane was carefully avoided by titration with soluble antigen so that one single antibody was assumed to bind to the surface antigen. The viscous drag measured for the three different raft-associated proteins, two with a GPI-

anchor and one transmembrane protein was independent of the type of membrane anchor, and was significantly larger than the local viscous drag of the nonraft proteins. After cholesterol depletion by methyl-β-cyclodextrin the viscous drag upon the raft-associated proteins decreased to the level of the nonraft proteins. In contrast, these latter proteins did not change their behavior after cholesterol depletion. The mean diameter of the raft protein assembles obtained from these measurements was 52 ± 26 nm, the platform diffusing as a single entity for minutes. Recently, Prior et al. (62), using a completely different method based on expression of a lipid raft marker partitioning into the inner leaflet of plasma membrane rafts and on immunoelectron microscopy, coupled with spatial point pattern analysis. Their results indicated a diameter of 44nm for the rafts, occupying 35% of the plasma membrane area. A platform of this size would contain around 3000 lipid molecules and probably not more than 10–20 protein molecules. Kusumi's measurements referred to above give an even smaller size. A very small raft size that results in a platform which accomodates less than two copies of the same protein species might explain the lack of FRET between GPI-anchored proteins in the experiments reported by Kenworthy et al (41). Mayor et al. (87,74b) have, on the other hand, observed FRET between GPI-anchored folate receptors using a more sensitive detection method and this may be due to a minority population, containing clusters of receptors. Also Saslowsky et al. (70) saw some PLAP oligomers in their AFM experiments with reconstituted proteoliposomes.

Whatever the intrinsic size of rafts in cell membranes may be, the important question is what prevents the raft from coalescing to a microscopically visible phase separated by the smallest possible boundary from the other coexisting phase or phases. A part of the answer may lie in the dimensionality of the system. The line tension to mass ratio in a two-dimensional domain may not be quite as high as the surface tension to mass ratio in a three-dimensional liquid. It is the line tension in two-dimensions and the surface tension in three-dimensions that promote spontaneous domain coalescence. As discussed earlier, another part of the answer could lie in "surface-active" effects at the interface between coexisting domains boundaries in lipid membranes. These effects, in which the role of proteins and minority lipids that may accumulate at the domain boundaries are quite important, would effectively reduce domain size somewhat analogously to what happens in a three-dimensional micro-emulsion. The raft phase composition may also constitute another reason. For example, unsaturated phospholipids are components of the raft phase. These lipids prefer the nonraft phase and this may limit the domain size when the probability of encountering them in the raft domain perimeter exceeds a certain value.

Although there is disagreement on the size and the lifetimes of the lipid assemblies in which raft proteins are embedded while floating in the nonraft liquid bilayer phase, there is consensus that lipid rafts can cluster into larger domains that are visible in the light microscope after oligomerization or cross-linking. There are numerous examples of such a raft clustering process, one of which is the formation of caveolae (3, 43). Caveolae are flask-shaped surface invaginations of 50–100 nm diameter, containing raft lipids and the protein caveolin. The caveolins form oligomers and formation of caveolae is probably driven by protein polymerization in the membrane phase.

Another process involving raft clustering is influenza virus envelope formation. The membrane of this virus is acquired by budding through the host cell plasma membrane. Two glycoproteins, hemagglutinin and neuraminidase, are tightly packed in the viral membrane. Both are raft-associated, hemagglutinin was shown to be in 50 nm rafts by photonic force microscopy (61, 71). The viral membrane is enriched in cholesterol and SpM compared to the

host cell membrane from which it was formed. The budding process is thought to be driven by association of the cytoplasmic tails of the viral membrane glycoproteins with the influenza virus matrix M1 protein that drives the budding process (23, 97). The working model is that the binding of the M1 protein to the cytoplasmic tails protruding from the inner leaflet of the plasma membrane facilitates docking of the M1 protein into an assembly complex that then interacts with other M1 protein subunits, associated with the ribonuclein particles to form a complete virus particle. Clustering of rafts containing the viral glycoproteins is assumed to be accomplished by M1 protein polymerizing with itself. Other rafts and raft proteins are excluded from the bud because they do not interact with the M1 protein lattice and are sterically hindered from entering the growing viral envelope. An alternative mechanism could involve association of the M1 protein with the domain on the inner leaflet of the raft domain in much the same way in which spectrin associates with phosphatidylserine on the inner surface of the erythrocyte membrane. Since this protein has a propensity to polymerize with itself, this polymerization leads to an aggregation of the inner leaflet domain of rafts and a consequent aggregation of rafts in the outer leaflet. The viral glycoproteins, hemaglutinin and neuraminidase, partition favorably into rafts and, therefore, get concentrated in this aggregated domain. A favorable interaction between the cytoplasmic tails of these proteins and the aggregated M1 matrix attached to the cytoplasmic leaflet further stabilizes the entire structure. Association of the viral ribonuclein particle with the M1 polymer matrix will induce a curvature of the membrane which, together with the mismatch at the boundary of the patched raft phase with the other coexisting phase or phases, will favor the budding process (44).

The specificity of raft clustering is nicely illustrated by experiments where raft clustering is induced by cross-linking with specific antibodies on the surface of living cells. If two raft proteins are cross-linked by antibodies, they will form overlapping patches in the plasma membrane. However, simultaneous patching of a raft protein and a nonraft protein leads to the formation of segregated patches. Copatching of the two raft components is dependent on the simultaneous addition of antibodies to both raft proteins to the cells (28). If antibodies are added sequentially, segregated patches predominate. Notably, the copatching behavior is dependent on cholesterol. Before cross-linking both the raft and the nonraft components were completely dispersed over the cell surface. These examples illustrate an important property of raft domains in cell membranes. Whatever the size of the resting state, for rafts to function in cellular processes, they are usually specifically clustered together by different means such as oligomerization or scaffolding by other proteins. For two different raft proteins to meet in a raft they have to be brought together. Clustering of rafts by antibodies to a GPI-anchored or transmembrane cell surface antigens also brings protein attached to the inner leaflet of lipid rafts, such as doubly acylated src-kinases into the raft cluster (28). The "passive" dragging of raft residents during raft coalescence will depend on concentration of the resident in the membrane and in rafts.This is not easy to explain by the "lipid shell model" of raft protein behavior. However, it is easily understood by clustering small pre-existing rafts and protein partition behavior (including that of proteins associated with the inner leaflet). This assumes coupling of outer and inner leaflets of individual small rafts, presumably stabilized by transmembrane raft proteins, or perhaps by the interdigitating long-chains of the sphingolipids between the two leaflets of the membrane. The asymmetric distribution of SpM would leave a fraction on the cytoplasmic side (~15%), which could preferentially be raft-associated. It should be remembered, however, that very little is known about the properties of the inner leaflet surface interactions with the cytoplasmic milieu. Interactions with cytoplasmic proteins

and electrolytes may promote heterogeneity in the inner monolayer of the cell membrane that might not be observable in model membranes of the same lipid composition.

How rafts cluster to perform functions in membrane trafficking, signal transduction and cell polarization is still poorly understood. But there are now so many instances of such processes in cellular life, both in health and disease, so that there is bound to be progress just by the sheer volume of research that is being carried out.

5 Protein interactions with lipid rafts

One key issue is how proteins associate with rafts. Lipid rafts are known to contain specific sets of proteins (7, 79). These include GPI-anchored proteins, doubly acylated proteins such as tyrosine kinases of the src family, G_α subunits of heteromeric G-proteins and endothelial nitric oxide synthase, the cholesterol- and palmitate-linked hedgehog protein and other palmitate-linked proteins, and transmembrane proteins. It is fairly straightforward to understand that proteins with attached saturated acyl chains and cholesterol can be associated with liquid-ordered raft domains. Studies with model membranes have confirmed that peptides containing such lipid modifications associate with liquid-ordered domains (92). It should be noted that the GPI anchors differ in their fatty acid composition. Some GPI anchors are known to contain unsaturated acyl chains and how these interact with lipid rafts remains to be studied. Transmembrane proteins, however, do pose a problem. It could be imagined that since they cross the bilayer they should disrupt the packing of the liquid-ordered domain but it must be recalled that the l_o phase is a liquid phase and, therefore, does not have long-range order in the membrane plane. Association of proteins with lipid rafts can be viewed as a simple solubility problem described by an equilibrium partition coefficient for partitioning of the protein between two coexisting phases, or can be understood to require some chemical affinity for raft lipids. Several proteins are known to interact with cholesterol. Caveolin is the prime example (55). There are also examples of receptor proteins interacting with glycosphingolipids including gangliosides (26). A structural protein motif has been identified for binding to sphingolipids (47). Recent results also demonstrate that proteins can exist in different states depending on the membrane environment. Glutamate receptors, which are G-protein-coupled heptahelical transmembrane proteins, are in a low affinity state when reconstituted into membranes lacking cholesterol. The receptor changes its conformation in liquid-ordered cholesterol-containing membranes and now binds its ligand with high affinity (17). The EGF receptor is activated by interaction with the ganglioside GM3 and inactivated by cholesterol depletion (54a). The receptor seems to depend on the lipid environment for high affinity binding capability.

One way to view this differential behavior (88) would be to consider the protein as a solute in the bilayer solvent of the membrane. If the lipid bilayer has two phases, each phase is a different solvent. The protein has a conformation that is dependent upon its environment and therefore dependent upon the bilayer solvent phase in which it is dissolved. So one can expect that in a nonraft domain, it will have one conformation and in the raft domain it will have another. The receptor activation would be dependent upon the partition coefficient between the different lipid domains in the bilayers and upon phase coexistence. Another issue is the length of the transmembrane domains the protein, since a liquid-ordered bilayer is thicker than a liquid-disordered one. These parameters have been shown to play a role in protein sorting to the cell surface (5). But how precisely the trans-membrane domains should be matched with the thickness of the bilayer, is an open issue. So far, no detailed analysis has

been carried out of how different transmembrane proteins having different trans-membrane domain lengths partition into liquid-ordered and liquid-disordered domains. It is generally known that the transmembrane domains of single-span transmembrane proteins is usually longer than the transmembrane domains of proteins that reside in the Golgi complex or in the endoplasmic reticulum (5). It is also worth considering how oligomerization of proteins affects their affinities for a raft domain. It can be shown that if the oligomerization does not affect the partition coefficient of each monomer in the oligomer, the partition coefficient of the oligomer is a product of the partition coefficients for the monomers that constitute it. A partition coefficient that is weakly in favor of a raft phase (say $K_P \approx 2$) will, in the case of homo-oligomerization, lead to a substantially increased affinity of the oligomer for this phase ($K_P \approx 4$ for a dimer, 8 for a trimer, 16 for a tetramer, etc.). This effect is shown in **Figure 2 (see color insert)**. The increased affinity of the raft-partitioning proteins after oligomerization can also be a driving force for aggregation of raft domains and their stabilization. Only careful reconstruction studies will be able to throw light on conformation and/or solubility in lipid phases and what structural characteristics will define the principles for raft partitioning.

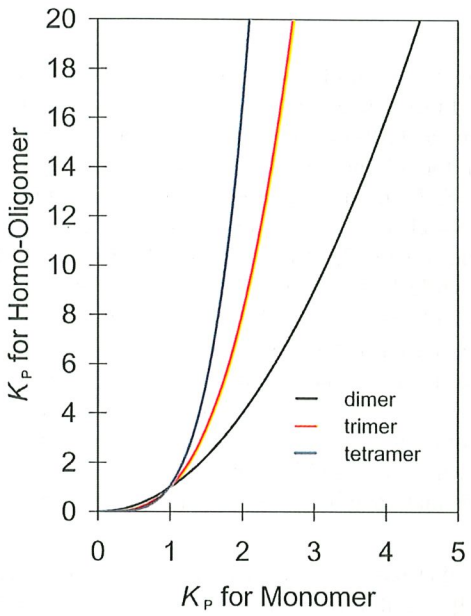


Fig. 3: *The effect of cross-linking upon the partitioning of membrane components between domains of coexisting phases. Homo-oligomerization is assumed and it is also assumed that the cross-linking does not alter the individual partition coefficents of each monomer in the oligomer.*

Nevertheless, the problem remains of how transmembrane proteins will disturb the packing in liquid-ordered domains. Will the protein be located at the raft boundary where they could work as surfactants and decrease the line tension and consequently raft size? Would raft size in cell membranes depend on their raft protein constituents? As discussed earlier, there can be

a substantial interfacial energy associated with the interface between liquid phases that results in a line tension at the domain interfaces. An interface with a large line tension would be a trap for impurities; thus, McConnell and Vrljic (52) postulate that if there were liquid-liquid interfaces in cell membranes, these would be decorated by specific proteins and/or lipids.

One interesting class of proteins is the group of GAP43-like proteins, which include GAP43, LAP23, MARCKS and MacMarcks (10, 53). These are highly hydrophilic and contain a basic domain that binds acidic phospholipids including PI (4,5)P2. When binding to the inner leaflet at sites containing PI(4,5)P2, these proteins seem to organize lipid raft-like domains, visible in the light microscope. These sites recruit WASP and ERM proteins to promote actin polymerization and filament assembly. The GAP43-like proteins are myristoylated or palmitoylated and it has been postulated that the association of multiple GAP43-like proteins through their binding to PI(4,5)P2 and the intercalation of multiple saturated fatty acyl chains leads to nucleation of a raft domain.

6 What does detergent resistance tell us about lipid domains ?

One of the stickiest issues in the lipid domain field is the issue of how detergent resistance can be used to define lipid raft composition. There is no doubt that the finding of Brown and Rose demonstrating that a GPI-anchored protein became resistant for Triton X-100 solubilization at 4°C in the Golgi complex during transport to the cell surface was a breakthrough in the field (8, 7, 46). The detergent-resistant membrane fraction (DRM) floated to low density and was enriched in sphingolipids and cholesterol. The subsequent work of Brown and London [reviewed in (7, 46)] attributed the property of detergent-resistance to liquid-ordered domains, while liquid-disordered bilayers were solubilized by Triton X-100. This set in motion a wave of research in which DRMs were used to define raft association. This was coupled with cholesterol depletion using methyl-β-cyclodextrin; solubilization of raft proteins after this treatment became a standard tool for specifying whether a protein was involved in raft processes or not. The usefulness of this methodology is attested to by the fact that most proteins that have a substantial fraction in DRMs have indeed been found to associate with rafts also when analyzed by other methods. However, it should be noted that the number of raft proteins that have been carefully investigated by additional methods, is still rather small.

There are several caveats that have to be kept in mind when using DRM methods to analyze raft association. First of all, not all proteins that are in rafts are in DRMs. One such example is the vesicular stomatitis virus G protein. It is Triton X-100 soluble but copatches with raft markers using antibody cross-linking (28). The G protein is also involved in phenotypic mixing, a process based on mingling of raft-associated proteins in virus envelope formation during budding from the host cell plasma membrane (6, 58). Furthermore, recent studies have demonstrated that Triton X-100 may promote liquid-ordered domain formation in model membranes (29). The addition of a detergent to a membrane results in its initial insertion into that membrane. When sufficient detergent has been inserted, the detergent becomes an additional chemical constituent of the system thereby altering its phase behavior. It is, in fact, surprising that the detergent-resistance criterion has turned out to be as useful as it has. This may be related to the kinetics and/or thermodynamics of mixing of detergent with l_o phase bilayer and speaks in favor of a pre-existence of domains of this phase in the membrane.

Hence, DRMs should not be expected to extract lipid rafts from cell membranes precisely reflecting their detailed composition. Triton X-100 addition may not only enhance liquid-ordered domain formation, the detergent also leads to fusion of existing rafts to large confluent membrane aggregates. The kinetics and thermodynamics of detergent interaction with membranes that contain an l_o phase, analogous to the raft phase in biomembranes, will have to be studied in detail in order to answer some of these questions.

A number of other mild detergents have been introduced to extract lipid domains from cell membranes. Several of these detergents were compared in a recent study (73) which demonstrated that detergents of the Lubrol and the Brij series extracted DRMs with different lipid and protein composition when compared to that of TritonX-100-DRMs. CHAPS behaved more like Triton X-100. This study emphasizes that the use of different detergents to extract DRMs is only a first step in defining possible membrane domains and should be followed up by other methods to define domain existence and composition. Song et al. (85) introduced a detergent-free method based on sonication, pH 11 treatment, and density gradient centrifugation to isolate what they called the caveolar membrane fraction. This fraction is highly enriched in caveolin protein but it also contains membranes from other cellular compartments. A recent proteomics study concluded that 75% of the proteins detected in this membrane fraction did not correspond to raft components as defined by inclusion in Triton X-100 DRMs and sensitivity to cyclodextrin treatment (19). 703 proteins were identified in DRMs and 585 in the pH 11 resistant fraction. Out of the 703 proteins 392 were quantifiable and revealed that 241 were sensitive to cholesterol depletion. Analyzing the proteins found, both the DRMs and the pH 11 fractions contain proteins that should not be there such as ribosomal proteins, only demonstrating that care has to be taken when using these criteria for defining raft association. Smart and Anderson (83) introduced another method in which a crude plasma membrane fraction was first isolated. This fraction was sonicated to release lipid rafts (and caveolae) which were isolated by flotation in a continuous Optiprep gradient. The detailed protein composition of this fraction has not yet been analyzed. This fraction has been analyzed for lipid content by mass spectrometry (59). In contrast to Triton X-100 DRMs, these membranes are enriched in arachidonic acid-containing ethanolamine plasmalogens. These lipids with polyunsaturated fatty acids would not be expected to partition into liquid-ordered domains. On the other hand, myelin, which is a specialized raft-like phase contains high amounts of ethanolamine plasmalogens, mostly composed of two 18:1 fatty acyl chains (65). The plasmalogens have not yet been carefully studied in model systems and their propensity for forming liquid-ordered domains should be explored.

The use of cyclodextrins to deplete cholesterol is also not without pitfalls. Treatment of living cells with methyl-β-cyclodextrin not only leads to dissociation of raft components, but has side effects as well. Cholesterol depletion affects the function of the plasma membrane by decreasing its permeability and changing its behavior in unpredictable ways (25, 60). Thus, effects of cholesterol depletion alone cannot be used to define raft function. There are also membranes such as the apical membrane of epithelial cells that remain resistant to cyclodextrin extraction in living cells (73). However, when a membrane fraction, derived from epithelial cells, was treated with methyl-β-cyclodextrin, release of raft-associated proteins from the apical membrane could be demonstrated after Triton X-100 extraction. The apical membrane is especially resistant to intervention probably because of its high raft lipid concentration, the membrane being practically covered by rafts. An interesting new approach for isolating clustered rafts in cell membranes is the immuno-isolation of activated T-cell

receptors and associated signaling molecules in plasma membrane subdomains (27). Receptor-activating antibodies attached to magnetic beads were used to cluster rafts on cell membranes which were then disrupted by nitrogen cavitation. The clustered raft fraction was isolated using a magnet.

7 Perspectives

The lesson that we have learnt from studies of model membranes prepared from binary and ternary lipid mixtures is that there can be liquid-liquid immiscibilities in lipid monolayers and bilayers. This observation conforms to the general idea that more ordered liquid domains rich in sphingolipids and cholesterol segregate from less ordered liquid domains composed of mainly unsaturated phospholipids. The size and behavior of these domains are sensitive to the chemical composition of the membranes as well as to pressure and temperature. Cell membranes on the other hand, although they contain hundreds of lipid species, also exhibit the formation of ordered liquid domains – rafts – which are rich in sphingolipids and cholesterol, and segregate from less ordered liquid domains composed of unsaturated phospholipids (and cholesterol). It is quite amazing how well results obtained from the model systems studied correlate with postulated and experimentally verified properties of sphingolipid-cholesterol rafts. The study of model systems prepared from lipid mixtures relevant to cell membranes are providing increasing support for the existence of liquid-liquid immiscibility in cell membranes which manifests itself as a segregation of liquid-ordered raft domains in a fluid bilayer. However, the detailed lipid composition of lipid raft domains is not known. Neither is the coupling between outer and the inner leaflets understood. The cell membranes are nonequilibrium systems, in which lipids and proteins are constantly being removed and added by intracellular transport. The plasma membrane is also constantly releasing cholesterol to lipoproteins in the extracellular medium. One possibility is that the liquid-liquid immiscibilities are undergoing constant fluctuations due to these perturbations. But, how could such metastable behavior be regulated? Is it compatible with the functions that cell membranes have?

We favor the view that the complex lipid composition of cell membranes has evolved to be buffered against such fluctuations. The little evidence that exists (12, 39, 50) indicates that multiphase bilayers respond extremely slowly to perturbations of phase equilibria. Perhaps the cell membrane needs such a large number of lipids precisely to avoid rapid responses to perturbations of its state. The complexity serves the need to pack lipids such that stable but fluid lipid rafts of defined size can coexist in a fluid matrix. Also, the liquid disordered matrix in cell membranes might be less permeable and more tightly packed than a simple 2-component unsaturated phospholipid-cholesterol model membrane. Israelachvili postulated early on (36) that lipid complexity is required to fill the holes the integral proteins create in membrane bilayers. Perhaps lipid complexity also helps to construct liquid membranes in which the bilayer space is tightly fitted to make it impermeable. However, at the same time the lipid composition is such that liquid-ordered-like domains can be formed. These move about as individual rafts of discrete size but they can be clustered by protein-protein cross-linking to form raft clusters. As a consequence, proteins must also have evolved to either function in the fluid matrix or within the rafts. Movement of proteins in and out of rafts demands structures capable of raft inclusion through simple partitioning, conformational changes leading to preferential partitioning or vice-versa, or by scaffolding interactions with resident raft proteins. There are significant advantages in terms of enhanced/reduced

bimolecular interactions to be had from dynamic compartmentalization of the lipid bilayer (42, 88).

How proteins interact with lipid rafts can only be decided by careful reconstitution studies in model systems that are designed to mimic cell membranes in an increasingly authentic fashion. One outstanding parameter is lipid asymmetry. If model systems are to give more detailed insights into how biological membranes work, it becomes urgent to develop methods that permit bilaterally asymmetric model membranes to be prepared and studied. The rapid progress in identifying and characterizing lipid flippases may provide new tools to construct asymmetric membranes. The characterization of such reconstituted membranes by new sophisticated tools such as solid state NMR will provide answers as to how proteins interact with lipids in membranes. Despite the perplexing complexity facing us, there is no escape from accepting that membrane biophysics should take its inspiration from cellular systems. This means that our community has to support attempts to overcome the technical difficulties involved. We are optimistic. The influx of new methodologies and young blood both from the physics and from the biology community is a sign of hybrid vigor. The broadening horizon of membrane research is opening up exciting vistas for the future.

8 Postscript concerning Terminology

A considerable amount of confusion in the field is created by the somewhat arbitrary use of terminology. Much of the terminology used in describing physical phenomena that occur in biological membranes is borrowed from the physical/chemical description of model membrane systems. Many of these terms have a precisely defined meaning in the context of the physics/chemistry of model membranes and may not be directly applicable to biomembranes. Other terms do not have precise physical definitions and care must be taken in their utilization. A lipid bilayer "phase" is a physical state of a lipid bilayer characterized by structural and dynamic properties. The structure of an isolated and fully hydrated single lipid bilayer can be characterized by the positional and orientational order of the lipid molecules in the plane of the layer (defined in a laboratory coordinate system as the xy plane) and by the thickness of the layer in a direction perpendicular to this plane (defined as parallel to the z axis of the same coordinate system). The dynamics of the bilayer have to do with the temporal behavior of the position of the lipid molecules (translational order), their orientation (rotational order) and the configuration of the chemical bonds in each lipid molecule (configurational or conformational order). The translational order is measured in terms of the hopping frequency (the number of times per second that a lipid molecule exchanges places with its neighbours) and is related to the translational diffusion coefficient. Rotational order is measured in terms of an angular velocity, usually for gyration around the long axis of the molecules which is perpendicular to the bilayer plane, and is related to the rotational diffusion coefficient. "Flip-flop" is a special case of rotational dynamics around the x and y axes of the coordinate system with a limited translational step in the z direction. The configurational order results from *trans-gauche* isomerism and is measured in terms of an order parameter. In structural terms, the lipids in biomembranes are in a lamellar lipid bilayer phase. In dynamic terms the bilayer membrane can exist as an ordered solid, s_o, or liquid, l, phase. An s_o phase is characterized by high degrees of translational (translational diffusion coefficient on the order of 10^{-11} cm^2s^{-1} or lower) and configurational (high *trans/gauche* ratio in the acyl chains) order whereas an l phase is characterized by lower translational (translational diffusion coefficient on the order of about 10^{-8} cm^2s^{-1}) and configurational (low *trans/gauche* ratio in the acyl chains) order. An l phase with a high molar fraction of cholesterol (varies from about

0.25 to about 0.5 depending upon the chemical identity of the other lipids in the bilayer) has been classified as being in a "liquid-ordered" phase, l_o. This l_o phase has been shown to have a slightly increased translational order compared to a cholesterol-free l phase (translational diffusion coefficient about 2 times lower) and a configurational order that is comparable to that of an s_o phase. To distinguish the l_o phase from the l phase bilayer without cholesterol, the latter has been termed a "liquid-disordered" phase, l_d. Using the criteria of translational and configurational order, l phase bilayers with a low molar fraction of cholesterol also qualify as l_d phases. There are, however, no exact definitions in terms of the hopping frequencies or order parameters that rigidly distinguish between l_d and l_o phase bilayers. Due to the increased *trans/gauche* ratio in a l_o phase, this phase (and its domains) are thicker than a l_d phase.

Within the framework of thermodynamics, a phase is always a macroscopic system, consisting of large number of molecules. However, in lipid bilayers the phases often tend to be fragmented into small domains (often only a few thousand molecules) each of which, per se, may not have a sufficient number of molecules to strictly satisfy the thermodynamic definition of a phase. In the absence of a better description for this sort of mesoscopic states, and assuming that there are a large number of domains in a given system, the domains may be treated as if they were a part of a macroscopic phase so that the same properties are attributed to the domains that would describe the phase. This definition is probably adequate as long as the domains do not get too small. The liquid-ordered raft phase thus comprises all the domains (small or clustered) of the raft phase in the membranes. The rest of the membrane surrounding the rafts, the liquid phase, may be a homogeneous percolating liquid phase or may be further subdivided into liquid domains not yet characterized.

Acknowledgments

We would like to thank G. Schutz for allowing us to cite his unpublished work. We also thank M. Prieto for supplying us with a preprint of his work reporting the ternary phase diagram for the SpM/POPC/cholesterol system, and P. Kinnunen for preprint of his paper on SpM-cholesterol interactions. D. Brown, E. Melo, T. Kurzchalia, G. Schutz, and the membranes of the Simons laboratory are gratefully acknowledged for their useful comments on the manuscript. This work was supported in part by grants (to WLCV) from the Fundacao para a Ciencia e a Tecnologia (FCT) of the Portuguese Ministry for Higher Education and Scientific Research through the POCTI program.

A Appendix: Abbreviations

AFM,	atomic force microscopy
DOPC	1,2-dioleoylphosphatidylcholine
DPPC	1,2-dipalmitoylphosphatidylcholine
DPPE	1,2-dipalmitoyl phosphatidylethanolamine
ESR	electron spin resonance
F-DOPE	N-fluoresceincarboxamido-1,2-dioleoylphosphatidylcholine
F-DPPE	N-fluoresceincarboxamido-1,2-dipalmitoylphosphatidylcholine
FRET	fluorescence resonance energy transfer
NMR	nuclear magnetic resonance
PC	phosphatidylcholine
SpM	sphingomyelin
POPC	1-palmitoyl-2-oleoylphosphatidyl choline

References

[1] Almeida PF, Vaz WL, Thompson TE. 1992. Lateral diffusion in the liquid phases of dimyristoylphosphatidylcholine/cholesterol lipid bilayers: a free volume analysis. *Biochemistry* 31:6739

[2] Almeida PF, Vaz WL, Thompson TE. 1993. Percolation and diffusion in three-component lipid bilayers: effect of cholesterol on an equimolar mixture of two phosphatidylcholines. *Biophys. J.* 64:399

[3] Anderson RG. 1998. The caveolae membrane system. *Annu. Rev. Biochem.* 67:199)

[4] Anderson RG, Jacobson K. 2002. A role for lipid shells in targeting proteins to caveolae, rafts, and other lipid domains. *Science* 296:1821

[5] Bretscher MS, Munro S. 1993. Cholesterol and the Golgi apparatus. *Science* 261:1280

[6] JA, Wilk T, Fuller SD. 2003. Do lipid rafts mediate virus assembly and pseudotyping? *J. Gen. Virol.* 84:757

[7] Brown DA, London E. 2000. Structure and function of sphingolipid- and cholesterol-rich membrane rafts. *J. Biol. Chem.* 275:17221

[8] Brown DA, Rose JK. 1992. Sorting of GPI-anchored proteins to glycolipid-enriched membrane subdomains during transport to the apical cell surface. *Cell* 68:533

[9] Brown RE. 1998. Sphingolipid organization in biomembranes: what physical studies of model membranes reveal. *J. Cell Sci.* 111 (Pt 1):1

[10] Caroni P. 2001. New EMBO members' review: actin cytoskeleton regulation through modulation of PI(4,5)P(2) rafts. *EMBO J.* 20:4332

[11] de Almeida RF, Fedorov A, Prieto M. 2003. Sphingomyelin/ Phosphatidylcholine/ Cholesterol phase diagram: boundaries and composition of lipid rafts. *Biophys. J.* in press

[12] de Almeida RF, Loura LM, Fedorov A, Prieto M. 2002. Nonequilibrium phenomena in the phase separation of a two-component lipid bilayer. *Biophys. J.* 82:823

[13] Dietrich C, Bagatolli LA, Volovyk ZN, Thompson NL, Levi M, et al. 2001. Lipid rafts reconstituted in model membranes. *Biophys. J.* 80:1417

[14] Dietrich C, Volovyk ZN, Levi M, Thompson NL, Jacobson K. 2001. Partitioning of Thy-1, GM1, and cross-linked phospholipid analogs into lipid rafts reconstituted in supported model membrane monolayers. *Proc. Natl. Acad. Sci. USA* 98:10642

[15] Dietrich C, Yang B, Fujiwara T, Kusumi A, Jacobson K. 2002. Relationship of lipid rafts to transient confinement zones detected by single particle tracking. *Biophys. J.* 82:274

[16] .b Ekroos K, Ejsing Ch S, Bahr U, Karas M, Simons K, Shevchenko A. in press. Charting Molecular Composition of Phosphatidylcholines by Fatty Acid Scanning and Ion Trap MS^3 fragmentation. *Journal of lipid Research*

[17] b Ekroos K, Ejsing Ch S, Bahr U, Karas M, Simons K, Shevchenko A. in press. Charting Molecular Composition of Phosphatidylcholines by Fatty Acid Scanning and Ion Trap MS^3 fragmentation. *Journal of lipid Research*

[18] Eroglu C, Brugger B, Wieland F, Sinning I. 2003. Glutamate-binding affinity of *Drosophila* metabotropic glutamate receptor is modulated by association with lipid rafts. *Proc. Natl. Acad. Sci. USA* in Press

[19] Feigenson GW, Buboltz JT. 2001. Ternary phase diagram of dipalmitoyl-PC/dilauroyl-PC/cholesterol: nanoscopic domain formation driven by cholesterol. *Biophys. J.* 80:2775

[20] Foster LJ, De Hoog CL, Mann M. 2003. Unbiased quantitative proteomics of lipid rafts

reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. USA* 100:5813
[21] Fujiwara T, Ritchie K, Murakoshi H, Jacobson K, Kusumi A. 2002. Phospholipids undergo hop diffusion in compartmentalized cell membrane. *J. Cell Biol.* 157:1071
[22] Gally HU, Seelig A, Seelig J. 1976. Cholesterol-induced rod-like motion of fatty acyl chains in lipid bilayers a deuterium magnetic resonance study. *Hoppe Seylers Z Physiol Chem.* 357: 1447
[23] Glaser M. 1993. Lipid domains in biological membranes. *Curr. Opin. Struct. Biol.* 3:475
[24] Gomez-Puertas P, Albo C, Perez-Pastrana E, Vivo A, Portela A. 2000. Influenza virus matrix protein is the major driving force in virus budding. *J. Virol.* 74:11538
[25] Guo W, Kurze V, Huber T, Afdhal NH, Beyer K, Hamilton JA. 2002. A solid-state NMR study of phospholipid-cholesterol interactions: sphingomyelin-cholesterol binary systems. *Biophys. J.* 83:1465
[26] Haines TH. 2001. Do sterols reduce proton and sodium leaks through lipid bilayers? *Prog. Lipid Res.* 40:299
[27] Hakomori SI. 2002. Inaugural Article: The glycosynapse. *Proc. Natl. Acad. Sci. USA* 99:225
[28] Harder T, Kuhn M. 2000. Selective accumulation of raft-associated membrane protein LAT in T cell receptor signaling assemblies. *J. Cell Biol.* 151:199
[29] Harder T, Scheiffele P, Verkade P, Simons K. 1998. Lipid domain structure of the plasma membrane revealed by patching of membrane components. *J. Cell Biol.* 141:929
[30] Heerklotz H. 2002. Triton promotes domain formation in lipid raft mixtures. *Biophys. J.* 83:2693
[31] Hörber JKH, Miles J.2003 Scanning Probe Evolution in Biology. *Science.* 302:1002
[32] Holopainen J, Metso AJ, Mattila JP, Jutila A, Kinnunen PKJ. 2003. Evidence for the lack of a specific interaction between cholesterol and sphingomyelin. *Biophys. J.* in Press
[33] Huang J, Buboltz JT, Feigenson GW. 1999. Maximum solubility of cholesterol in phosphatidylcholine and phosphatidylethanolamine bilayers. *Biochim. Biophys. Acta* 1417:89
[34] Huang J, Feigenson GW. 1999. A microscopic interaction model of maximum solubility of cholesterol in lipid bilayers. *Biophys. J.* 76:2142
[35] Ilangumaran S, Hoessli DC. 1998. Effects of cholesterol depletion by cyclodextrin on the sphingolipid microdomains of the plasma membrane. *Biochem. J.* 335 (Pt 2):433
[36] Ipsen JH, Karlstrom G, Mouritsen OG, Wennerstrom H, Zuckermann MJ. 1987. Phase equilibria in the phosphatidylcholine-cholesterol system. *Biochim. Biophys. Acta* 905:162
[37] Ipsen JH, Mouritsen OG, Zuckermann MJ. 1989. Theory of thermal anomalies in the specific heat of lipid bilayers containing cholesterol. *Biophys. J.* 56:661
[38] Israelachvili JN. 1977. Refinement of the fluid-mosaic model of membrane structure. *Biochim. Biophys. Acta* 469:221
[39] Jacobson K, Vaz WL. 1992. Special issue dedicated to "Domains in Biological Membranes". *Comments Mol. Cell. Biophys.* 8
[40] Jain MK, White HB, 3rd. 1977. Long-range order in biomembranes. *Adv. Lipid Res.* 15:1
[41] Jorgensen K, Klinger A, Biltonen RL. 2000. Non-equilibrium lipid domain growth in the gel-fluid two phase region of a DC16PC-DC22PC lipid mixture investigated by Monte Carlo computer simulation, FTIR, and fluorescence spectroscopy. *J. Phys. Chem.* 104:11763

[42] Keller S, Pitcher WH, Huestis WH, McConnell HM. 1998. Red blood cells form immiscible liquids. *Phys. Rev. Lett.* 81:5019
[43] Kenworthy AK, Petranova N, Edidin M. 2000. High-resolution FRET microscopy of cholera tocin B-subunit and GPI-anchored proteins in cell plasma membranes. *Mol. Biol. Cell.* 11 (5):1645
[44] Kholodenko BN, Hoek JB, Westerhoff HV. 2000. Why cytoplasmic signalling proteins should be recruited to cell membranes. *Trends Cell Biol.* 10:173
[45] Kurzchalia TV, Parton RG. 1999. Membrane microdomains and caveolae. *Curr. Opin. Cell Biol.* 11:424
[46] Lipowsky R. 2002. Domains and rafts in membranes - hidden dimensions of self-organization. *J. Biol. Phys.* 28:195
[47] London E. 2002. Insights into lipid raft structure and formation from experiments in model membranes. *Curr. Opin. Struct. Biol.* 12:480
[48] London E, Brown DA. 2000. Insolubility of lipids in triton X-100: physical origin and relationship to sphingolipid/cholesterol membrane domains (rafts). *Biochim. Biophys. Acta* 1508:182
[49] Mahfoud R, Garmy N, Maresca M, Yahi N, Puigserver A, Fantini J. 2002. Identification of a common sphingolipid-binding domain in Alzheimer, prion, and HIV-1 proteins. *J. Biol. Chem.* 277:11292
[50] Marsh D, Horvath LI. 1998. Structure, dynamics and composition of the lipid-protein interface. Perspectives from spin-labelling. *Biochim. Biophys. Acta* 1376:267
[51] Mateo R, Acuna A, Brochon JC. 1995. Liquid-crystalline phases of cholesterol/lipid bilayers as revealed by the fluorescence of trans-parinaric acid. *Biophys. J.* 68:978
[52] McConnell HM. 1996. Equilibration rates in lipid monolayers. *Proc. Natl. Acad. Sci. USA* 93:15001
[53] McConnell HM, Radhakrishnan A. 2003. Condensed complexes of cholesterol and phospholipids. *Biochim. Biophys. Acta* 1610:159
[54] McConnell HM, Vrljic M. 2003. Liquid-Liquid Immiscibility in Membranes. *Annu. Rev. Biophys. Biomol. Struct.*:
[55] McLaughlin S, Wang J, Gambhir A, Murray D. 2002. PIP(2) and proteins: interactions, organization, and information flow. *Annu. Rev. Biophys. Biomol. Struct.* 31:151
[56] Milhiet P, Vie V, Giocondi M, Le Grimellec C. 2001. AFM characterization of model rafts in supported bilayers. *Single Mol.* 2:119
[57] Miljan EA, Bremer EG. 2002. Regulation of growth factor receptors by gangliosides. *STKE* 160:15
[58] Munro S, 2003. Lipid Rafts: Elusive or Illusive? *Cell* 115:377
[59] Murata M, Peranen J, Schreiner R, Wieland F, Kurzchalia TV, Simons K. 1995. VIP21/caveolin is a cholesterol-binding protein. *Proc. Natl. Acad. Sci. USA* 92:10339
[60] Nielsen M, Miao L, Ipsen JH, Zuckermann MJ, Mouritsen OG. 1999. Off-lattice model for the phase behavior of lipid-cholesterol bilayers. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 59:5790
[61] Ohvo-Rekila H, Ramstedt B, Leppimaki P, Slotte JP. 2002. Cholesterol interactions with phospholipids in membranes. *Prog. Lipid Res.* 41:66
[62] Pickl WF, Pimentel-Muinos FX, Seed B. 2001. Lipid rafts and pseudotyping. *J. Virol.* 75:7175
[63] Pike LJ, Han X, Chung KN, Gross RW. 2002. Lipid rafts are enriched in arachidonic acid and plasmenylethanolamine and their composition is independent of caveolin-1 expression: a quantitative electrospray ionization/mass spectrometric analysis.

Biochemistry 41:2075
[64] Pizzo P, Giurisato E, Tassi M, Benedetti A, Pozzan T, Viola A. 2002. Lipid rafts and T cell receptor signaling: a critical re-evaluation. *Eur. J. Immunol.* 32:3082
[65] Pralle A, Keller P, Florin EL, Simons K, Horber JK. 2000. Sphingolipid-cholesterol rafts diffuse as small entities in the plasma membrane of mammalian cells. *J. Cell Biol.* 148:997
[66] Prior IA, Muncke C, Parton RG, Hancock JF. 2003. Direct visualization of Ras proteins in spatially distinct cell surface microdomains. *J. Cell Biol.* 160:165
[67] Ramstedt B, Slotte JP. 2002. Membrane properties of sphingomyelins. *FEBS Lett.* 531:33
[68] Rinia HA, Snel MM, van der Eerden JP, de Kruijff B. 2001. Visualizing detergent resistant domains in model membranes with atomic force microscopy. *FEBS Lett.* 501:92
[69] Rodemer C, Thai TP, Brugger B, Kaercher T, Werner H, et al. 2003. Inactivation of ether lipid biosynthesis causes male infertility, defects in eye development and optic nerve hypoplasia in mice. *Hum. Mol. Genet* 12:1881
[70] Samsonov AV, Mihalyov I, Cohen FS. 2001. Characterization of cholesterol-sphingomyelin domains and their dynamics in bilayer membranes. *Biophys. J.* 81:1486
[71] Sankaram MB, Thompson TE. 1990. Interaction of cholesterol with various glycerophospholipids and sphingomyelin. *Biochemistry* 29:10670
[72] Sankaram MB, Thompson TE. 1990. Modulation of phospholipid acyl chain order by cholesterol. A solid-state 2H nuclear magnetic resonance study. *Biochemistry* 29:10676
[73] Sankaram MB, Thompson TE. 1991. Cholesterol-induced fluid-phase immiscibility in membranes. *Proc. Natl. Acad. Sci. USA* 88:8686
[74] Saslowsky DE, Lawrence J, Ren X, Brown DA, Henderson RM, Edwardson JM. 2002. Placental alkaline phosphatase is efficiently targeted to rafts in supported lipid bilayers. *J. Biol. Chem.* 277:26966
[75] Scheiffele P, Rietveld A, Wilk T, Simons K. 1999. Influenza viruses select ordered lipid domains during budding from the plasma membrane. *J. Biol. Chem.* 274:2038
[76] Scheiffele P, Roth MG, Simons K. 1997. Interaction of influenza virus haemagglutinin with sphingolipid- cholesterol membrane domains via its transmembrane domain. *EMBO J.* 16:5501
[77] Schuck S, Honsho M, Ekroos K, Shevchenko A, Simons K. 2003. Resistance of cell membranes to different detergents. *Proc. Natl. Acad. Sci. USA* 100:5795
[78] Schutz GJ, Kada G, Pastushenko VP, Schindler H. 2000. Properties of lipid microdomains in a muscle cell membrane visualized by single molecule microscopy. *EMBO J.* 19:892
[79] Sharma P, Sabharanjak S, Satyajit M, 2002. Endocytosis of lipid rafts: an identity crisis. *Sem. in Cell and Dev. Biology* 13:205
[80] Silvius JR. 2003. Role of cholesterol in lipid raft formation: lessons from lipid model systems. *Biochim. Biophys. Acta* 1610:174
[81] Silvius JR, del Giudice D, Lafleur M. 1996. Cholesterol at different bilayer concentrations can promote or antagonize lateral segregation of phospholipids of differing acyl chain length. *Biochemistry* 35:15198
[82] Simons K, Ehehalt R. 2002. Cholesterol, lipid rafts, and disease. *J. Clin. Invest.* 110:597
[83] Simons K, Ikonen E. 1997. Functional rafts in cell membranes. *Nature* 387:569
[84] Simons K, Toomre D. 2000. Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell Biol.* 1:31

[85] Simons K, van Meer G. 1988. Lipid sorting in epithelial cells. *Biochemistry* 27:6197
[86] Smaby JM, Brockman HL, Brown RE. 1994. Cholesterol's interfacial interactions with sphingomyelins and phosphatidylcholines: hydrocarbon chain structure determines the magnitude of condensation. *Biochemistry* 33:9135
[87] Smaby JM, Kulkarni VS, Momsen M, Brown RE. 1996. The interfacial elastic packing interactions of galactosylceramides, sphingomyelins, and phosphatidylcholines. *Biophys. J.* 70:868
[88] Smart EJ, Ying YS, Mineo C, Anderson RG. 1995. A detergent-free method for purifying caveolae membrane from tissue culture cells. *Proc. Natl. Acad. Sci. USA* 92:10104
[89] Smorodin V, Melo E. 2001. Shape and dimensions of gel domains in phospholipid bilayers: a theoretical study. *J. Phys. Chem.* 105:6010
[90] Song KS, Li S, Okamoto T, Quilliam LA, Sargiacomo M, Lisanti MP. 1996. Co-purification and direct interaction of Ras with caveolin, an integral membrane protein of caveolae microdomains. Detergent-free purification of caveolae microdomains. *J. Biol. Chem.* 271:9690
[91] Subczynski WK, Kusumi A. 2003. Dynamics of raft molecules in the cell and artificial membranes: approaches by pulse EPR spin labeling and single molecule optical microscopy. *Biochim. Biophys. Acta* 1610:231
[92] Varma R, Mayor S. 1998. GPI-anchored proteins are organized in submicron domains at the cell surface. *Nature* 394: 798
[93] Vaz WL. 1996. Consequences of phase separations in membranes. In *Handbook of Non-Medical Applications of Liposomes*, ed. Y Barenholz, D Lasic, pp. 51. Boca Raton, FL: CRC Press
[94] Vaz WL, Almeida P. 1993. Phase topology and percolation in multi-phase lipid bilayers: is the biological membrane a domain mosaic? *Curr. Opin. Struct. Biol.* 3:482
[95] Veiga MP, Arrondo JL, Goni FM, Alonso A, Marsh D. 2001. Interaction of cholesterol with sphingomyelin in mixed membranes containing phosphatidylcholine, studied by spin-label ESR and IR spectroscopies. A possible stabilization of gel-phase sphingolipid domains by cholesterol. *Biochemistry* 40:2614
[96] Vist MR, Davis JH. 1990. Phase equilibria of cholesterol/dipalmitoylphosphatidylcholine mixtures: 2H nuclear magnetic resonance and differential scanning calorimetry. *Biochemistry* 29:451
[97] Wang TY, Leventis R, Silvius JR. 2000. Fluorescence-based evaluation of the partitioning of lipids and lipidated peptides into liquid-ordered lipid microdomains: a model for molecular partitioning into "lipid rafts". *Biophys. J.* 79:919
[98] Wang TY, Silvius JR. 2001. Cholesterol does not induce segregation of liquid-ordered domains in bilayers modeling the inner leaflet of the plasma membrane. *Biophys. J.* 81:2762
[99] White DA. 1973. The phospholipid Composition of Mammalian Tissues. In *Form and Function of Phospholipids*, ed. G Ansell, J Hawthrown, R Dawson. New York: Elsevier
[100] Wu ES, Jacobson K, Papahadjopoulos D. 1977. Lateral diffusion in phospholipid multibilayers measured by fluorescence recovery after photobleaching. *Biochemistry* 16:3836
[101] Yuan C, Furlong J, Burgos P, Johnston LJ. 2002. The size of lipid rafts: an atomic force microscopy study of ganglioside GM1 domains in sphingomyelin/DOPC/cholesterol membranes. *Biophys. J.* 82:2526
[102] Zhang J, Pekosz A, Lamb RA. 2000. Influenza virus assembly and lipid raft

microdomains: a role for the cytoplasmic tails of the spike glycoproteins. *J. Virol.* 74:4634

D 6 The Mechanical Strength of Bioadhesion

R. Merkel
Institut für Schichten und Grenzflächen
Forschungszentrum Jülich GmbH

Contents

1 Foreword 2

2 Biological Aspects 2
2.1 Specificity of Bioadhesion 2
2.2 The Scene of Action 4
2.3 The Major Families of Adhesion Molecules 6

3 A Biophysical Approach to Specific Adhesion 9
3.1 Model Membrane Studies of Adhesion 9
3.2 Mechanical Strength of Single Bonds 17

4 Conclusion and Outlook 25

1 Foreword

Adhesion is "the molecular attraction exerted between the surfaces of bodies in contact" [1]. Albeit this definition originates from a non-scientific source, it describes the essential features of adhesion quite clearly. **Bio**adhesion in turn is adhesion within a biological context. Obviously there is an extremely large number of "bodies" in contact within a living organism. Examples range all the way from macromolecules, cell organelles, cells, parts of tissues to biominerals. In some instances adhesion is of vital importance, e.g. consider the connection between tendons and bone or between cells and the connective tissue. In others it would have fatal consequences. For example, imagine your red blood cells would adhere to each other.
These simple considerations show that bioadhesion must in general be very selective and also under tight biological control. Accordingly, bioadhesion is a topic with very many aspects. It is entirely impossible to present even the very basic aspects of bioadhesion within two hours of lectures. Therefore this contribution will focus on the mechanical strength of bioadhesion. However, you should be aware that bioadhesion is a very active field of research. It ranges from basic technology (e.g. controlling the development of biofilms in the water supply tubes of your house), all the way to biophysics, biochemistry, and cell biology. Moreover, there are major implications of bioadhesion in medical issues like the development of tumors, the entry of pathogenic cells into the body, and the attack of parasites on our skin. The best I can hope for is that this short lecture will make you curious to learn more about this fascinating subject.

2 Biological Aspects

In this section, I will briefly discuss some aspects of adhesion in a biological context. Readers who want to go beyond the very limited scope of this presentation should consult one of the excellent textbooks on molecular cell biology, e.g. [2, 3].

2.1 Specificity of Bioadhesion

As mentioned above adhesion is the cumulative effect of short range interactions between surfaces of bodies in contact. We encounter adhesion in our daily live by using articles like adhesive tape, band aid, glue, and sticky dough. These articles will stick to almost every surface irrespective of its molecular composition. This type of adhesion is called *nonspecific adhesion*. Nonspecific adhesion is caused by surface interactions of physical nature. Examples are van der Waals forces, electrostatic attraction or polymer-induced attraction forces. These interactions can be described by potentials acting between the surfaces. Most often these potentials are additive enabling a very elegant treatment of complicated situations where several interactions act simultaneously. A famous example for the power of this approach is the DLVO-theory for the stability of colloidal suspensions, see for example [4, 5, 6]. In a biological context nonspecific adhesion occurs mostly in the formation of biofilms, i.e. relatively thick layers of bacteria which adhere to many surfaces. A polymeric matrix of polysaccharides secreted by the bacteria endows these layers with adhesion to surfaces and with cohesion in itself [7]. An example for such a biofilm is presented in Figure 1.
Surprisingly, nonspecific adhesion seems to play no or a very minor role in higher cells. In general, eucaryotic cells are very specialized. Many of them are part of multicellular organisms. Let us begin with a look at one specific event in embryonic development, cf. Figure 2. This

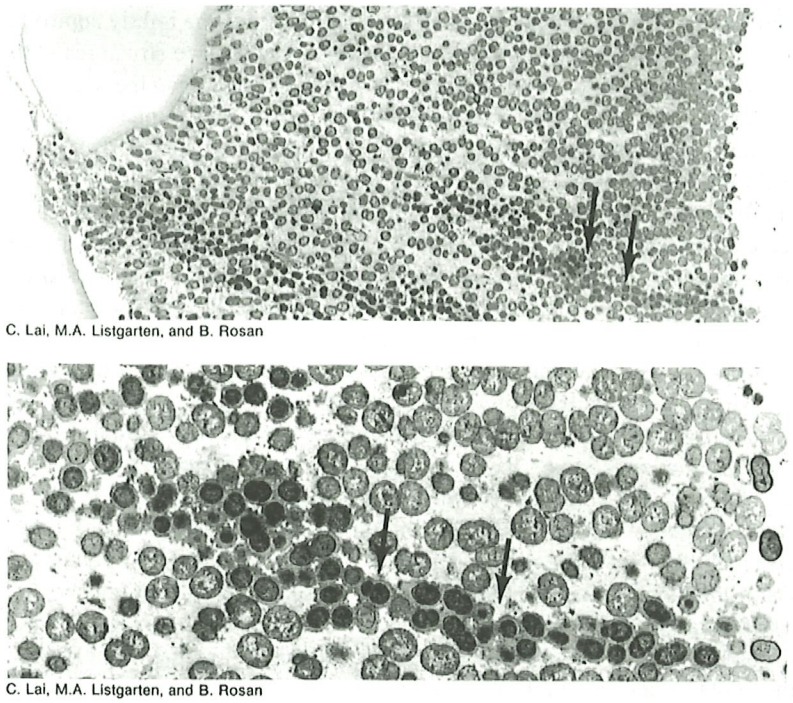


Fig. 1: *Dental plaque as an example for nonspecific adhesion in biology. The plaque layer exhibits an overall thickness of 50 μm. It is formed by bacteria (the dark spots) embedded in a polysaccharide matrix. In these electron micrographs (thin sections of real plaque layers) bacteria of the species* Streptococcus sobrinus *are labelled and therefore appear darker. This species converts sugar into lactic acid with deleterious and painful consequences for the habitat of the bacteria, i.e. the owner of the tooth. Top: Low magnification, the tooth surface is at the left side of the image. Bottom: Higher magnification. Pictures taken from [7].*

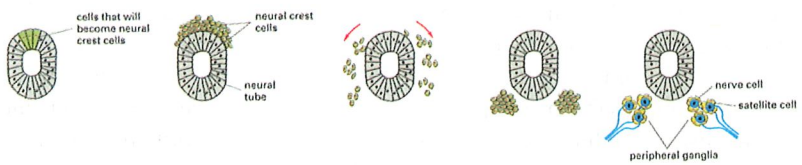


Fig. 2: *A sequence of events occurring during tissue formation in embryogenesis. Time proceeds from left to right. The early embryo displays a neural tube originating from epithelium (first stage). Some cells within this tube reduce their adhesiveness to neighboring cells and disengage from the neural tube. Subsequently these cells form a "crest" on top of the neural tube (2nd stage). Later on the same cells start to move actively through the embryonic tissue. This motion is directed, i.e. the cells must possess the ability to sense and follow certain molecular surface structures through the tissue (3rd stage). Upon arrival at the correct location the cells alter their adhesive behavior and aggregate again (4th stage). Finally, the cells differentiate to form specialized nervous structures within the body. Drawings taken from [2].*

example shows in a very conclusive way that cell adhesion must be tightly controlled. In other words, adhesiveness must occur exclusively to very specific surface structures. Moreover, the adhesion strength and the type of target structure must be varied during the development. These rather strict demands cannot be met by a nonspecific mechanism as the one presented for biofilm formation. Nature's design solution for this problem is the so-called *specific bond* which will be described in the following.

Specific bonds occur in biomacromolecules only. I will describe the case of proteins, but the same concepts hold also for DNA and RNA. On a molecular scale proteins are large objects. They exhibit molecular weights ranging from 10^4 to 10^6 g/mol and thus consist of thousands of atoms which adopt a well defined three-dimensional structure, see for example [8, 9, 10, 11]. Protein folding is an exciting scientific question worthwhile some careful consideration. However, for our purposes it is sufficient to accept that proteins exhibit a well defined surface topography. Moreover, the surface of a protein carries a pattern of chemical groups which are able to take part in numerous weak physicochemical interactions. Most prominent are hydrogen bonding, electrostatic interaction, van der Waals forces, and hydrophobic interactions. All these interactions exhibit a very short range of typically 1 Å. Please note that even the electrostatic interaction is screened in physiological buffer solution and has a range of at most 1 nm. Most interactions exhibit not only a very short range of action but are also dependent on the relative orientations of the target groups. A well known example for an angular dependent interaction is hydrogen bonding [9]. Moreover, all these interactions contribute binding enthalpies on the order of the thermal energy (0.2 to 3 k_BT).[1] Thus *no single* of these interactions is able to stabilize the contacts between molecules that always occur during random encounters. Taken together these considerations imply that the chance of bonding between arbitrarily selected proteins is vanishingly small because in order to adhere to each other proteins must exhibit extended surface areas that fit in shape and interaction to each other. The overall bond exhibits a rather low binding enthalpy (several k_BT) in comparison to covalent bonds ($\sim$100 k_BT). As protein conformations exhibit a certain flexibility, binding sites may alter their conformation upon ligand binding. However, the binding induced displacements rarely exceed 1 Å. Specific bonding occurs between proteins that were shaped to this end by evolution. By this astonishing strategy (multiple weak and short range interactions superimpose to form the overall bond) Nature is able to achieve tremendous specificity. For example, the immune system is able to sense antigens in the blood plasm at concentrations as low as 10^{-12} mol/liter. As blood plasm exhibits a protein concentration in the mM range this implies a specificity of at least $1 : 10^9$! The message to remember from this paragraph is: *Specificity is the ability to distinguish different molecules.* Nature achieves this by a superposition of several weak and short range interactions. It comes at the cost of binding enthalpy. Thus *specificity and affinity are two very different features of bonds* that should be not confused. The concept of specific bonds or molecular recognition is central in biology and by no means restricted to bioadhesion. More information can be found in textbooks of molecular cell biology [2, 3], for more detailed monographes consult [12, 13].

2.2 The Scene of Action

Adhesion occurs between surfaces. Therefore specific bioadhesion is mediated by molecular recognition (i.e. specific binding) between complementary cell adhesion molecules on the sur-

[1] Please note that in contrast to the weakness of the attractive interactions, strong repulsions exist for the case of steric mismatch.

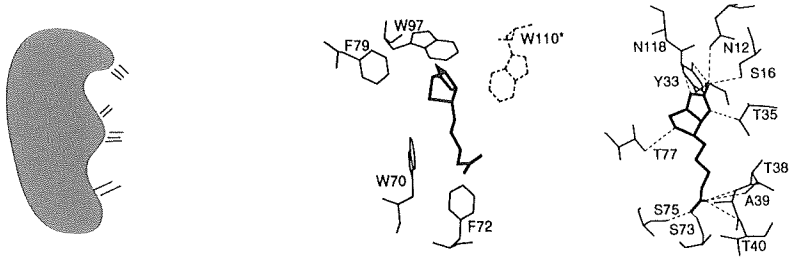


Fig. 3: *Specific bonds. Left: Principle of bonding. Both proteins exhibit a nontrivial surface topography as well as a unique interaction pattern on this surface. The individual interactions (indicated by black lines) are short range and often strongly dependent on direction. Moreover, their strength of bonding is on the order of the thermal energy. Therefore a stable complex is formed exclusively if several of these microscopic interactions contribute. This never happens by chance, i.e. the bond is specific. The other figures show the bond between avidin and biotin. This is a very well known specific bond because of its remarkable affinity on the order of* $10^{15}\ Mol^{-1}$. *In these figures biotin is drawn in bold lines. The central picture displays hydrophobic amino acids in close contact to biotin (abbreviations: F phenylalanine, W tryptophane). The right hand picture shows hydrogen bonds between biotin and the protein (abbreviations: A alanine, N asparagine, S serine, T threonine, Y tyrosine). Please note the substantial number of microscopic interactions that take part in the overall specific bond. The structures of the avidin biotin complex were taken from [14].*

face of the adherent cells. In general animal cells adhere exclusively at very specific target positions, for an example see Figure 2. In order to achieve the necessary level of distinction between the many possible surfaces within a body Nature "uses" very many different types of cell adhesion molecules. Here we will only sketch some very basic facts on these receptors, good starting points for further reading would be [2, 3, 15].

Contacts between cells are often organized in supramolecular assemblies at the cell surface. Here many different cell adhesion molecules (CAM) aggregate and cooperatively connect the adherent cells. These cell junctions are most prominent in permanent tissues like the epithelium sheets lining the gut. Historically cell junctions were first described by electron microscopy as parts of the ultrastructure of cells. At present research focuses on characterizing the molecules that take part in cell junctions and on unravelling their specific roles in these assemblies. Here we will not go into further detail but ask the reader to remember for the rest of this lecture that cell adhesion molecules in general perform their task in a collective way and rarely ever act as isolated molecules.

Cell-cell adhesion as well as cell-matrix adhesion is mediated by CAMs. Within the cell CAMs are linked to the cytoskeleton. In most cases many accessory proteins form a complex connecting the cytoplasmatic domain of CAMs to the cytoskeleton, cf. Fig. 4. The latter structure is formed by filaments of proteins within the cell. The cytoskeleton is of paramount importance for cell mechanics. It controls passive mechanic properties (elasticity, plasticity, and viscosity) and enables mechanical cell activities like locomotion or contraction. It is only the linkage of CAMs to the cytoskeleton that enables adherent cells to withstand significant mechanical forces. Unfortunately a plethora of molecules take part in these linkages which often occur as part of cell junctions. At the present moment the issue of CAM-cytoskeleton linkage is one of

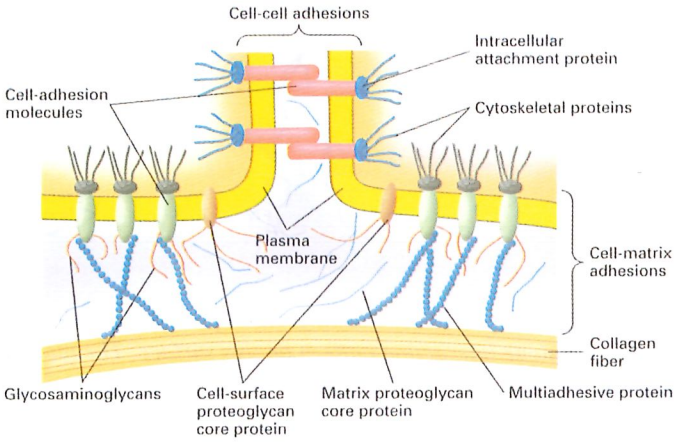


Fig. 4: *Schematics of bioadhesion of animal cells. Denoted are cell adhesion molecules (CAMs), which are in general integral membrane proteins. Please note the connections between CAMs and the cytoskeleton within the cell. Bioadhesion occurs between cells and between cells and substrates. For animal cells the adhesion substrate is the extracellular matrix, here denoted by collagen fibers, soluble multiadhesive proteins, and proteoglycans. For details see text. The artwork was taken from [3].*

unsolved questions of cell biology and biomechanics.
Bioadhesion occurs as cell cell adhesion and as cell matrix adhesion, cf. Figure 4. In the latter case "matrix" denotes the extracellular network of polymeric filaments which is the major architectural element of the connective tissue. This network can organize into a three dimensional gel as in the connective tissue of cartilage. In this case cells are embedded into the network. Alternatively, this network can assume the shape of basal membranes, i.e. two dimensional sheets of crosslinked polymers which often form a stable base for cell layers. An example is the basal membrane underneath the monolayer of endothelial cells which seals venules, cf. Figure 13. In the extracellular matrix these crosslinked polymeric networks are able to withstand tensile forces. However, organisms cannot evade compressive forces (i.e. pressure). To withstand compression the extracellular network contains proteoglycans, that is proteins decorated with large polysaccharides. The latter are called glycosaminoglycans and contain many electric charges that are screened by counterions. It is the osmotic pressure of those counterions that enables the extracellular matrix to withstand compression. The adhesion between cells and the extracellular matrix is facilitated by soluble multiadhesive proteins that link structural elements of the extracellular matrix to cell adhesion molecules. Please note, that the same architectural principles are also found in man-made composite materials. For example, in glass fiber reinforced plastics the fibers support tensile stress, the polymer matrix compression. Adhesion between fibers and plastic matrix has to be enhanced by surface coatings of the fibers.

2.3 The Major Families of Adhesion Molecules

Several hundred types of cell adhesion molecules (CAMs) are known. According to similarity in function and molecular sequence they are grouped in several families. Here we will briefly

discuss the major groups.

Members of the **Cadherin** family are among the most important CAMs in cell-cell adhesion. At the present moment more than 40 members of this family are known. They all mediate predominantly homophilic adhesion. "Homophilic" in this respect describes the fact that adhesion molecules bind to CAMs of the same molecular species but located at the surface of the opposing cell. Cadherin molecules are transmembrane proteins composed of about 750 amino acids. Their cytoplasmatic domain (i.e. the part of the protein inside the cell) is known to bind to the cytoskeleton. The extracellular domain has several binding sites for Ca^{2+} ions. Isolated cadherin molecules do not act as adhesion molecules, they must dimerize to become active. As the dimerization is dependent on Ca^{2+} ions, cadherin mediated cell adhesion can be abolished by calcium depletion of the medium. Interestingly, the distributions of the diverse cadherin molecules are very tissue specific. For example, E-cadherin is characteristic for the pre-implantation embryo and for non-neural epithelial tissue, whereas N-cadherin is mostly present on nerve cells. Cell types expressing different cadherins can be isolated, mixed, and subsequently cultured. As time passes, the initially randomly distributed cells sort themselves and form cell patches consisting of only one type of cells. This sorting process can be shown to depend on cadherins. In accord with this observation cadherins play an important role in tissue differentiation and cell homing.

Nerve Cell Adhesion Molecules, N-CAMs. Like cadherins N-CAMs mediate primarily homophilic cell adhesion. However, the adhesion potential of N-CAMs does *not* depend on Ca^{2+} ions. Moreover, all N-CAMs derive from one single gene. N-CAM diversity originates from different gene splicing, i.e. removal of sequence sections that occurs on the level of RNA. The splicing variants N-CAM 180, N-CAM 140, and N-CAM 120 vary in their cytoplasmic domain which is large in N-CAM 180, medium in N-CAM 140 and replaced by a lipid anchor in N-CAM 120. All these different N-CAMs exhibit an attachment site for polysialic acid. The latter substance is a highly charged polysaccharide. Different lengths of polysialic acid are found. Varying the length of this polyelectrolyte seems to be a mechanism for cellular control of N-CAM adhesiveness. For example, 25% of the total molecular weight of N-CAMs from embryonic brain consists of polysialic acid, whereas this fraction is only 8% in the adult brain. This is in accord with the plasticity of the embryonic brain which requires a rapid kinetics of cell contacts, i.e. weak adhesion.[2]

Selectins are selective lectins[3], that is molecules that bind certain oligosaccharides with high selectivity. Selectins are mostly known for their role in inflammation response. In the course of this process leukocytes first adhere to the endothelial cells lining the inside of blood vessels, whereupon they transform from relatively passive to highly active cells and finally pass through the endothelium to enter the extracellular matrix. There they attack pathogenic microorganisms. The initial adhesion between white blood cells and the endothelium is mediated by the specific bond between P-selectin on the surface of endothelial cells and the oligosaccharide moiety "sialyl Lewis-x" which is abundant in leukocyte glycoproteins and glycolipids. Later on, other selectins come into play. During the initial phase of adhesion leukocytes undergo a rolling motion which is caused by a rapid sequence of bond formation and tensile failure processes between selectins and glycoproteins. This process can be studied on the level of single adhesion events which is at present an active field of research, cf. [16]. As in the case of cadherins,

[2]Nota bene: Despite being derived from one single gene there is a wide variety of N-CAMs. This puts the human genom project in the proper perspective. There is much more to understanding organisms than just compiling the gene sequences.

[3]Per definition lectins are proteins that bind oligosaccharides.

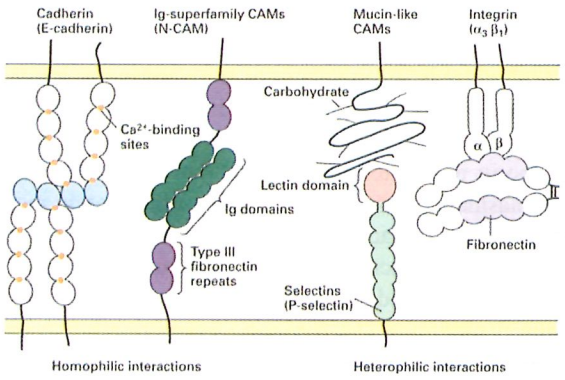


Fig. 5: *This schematics summarizes the salient features of the cell adhesion molecule (CAM) families.* ***Cadherins*** *form Ca^{2+}-linked dimers that mediate homophilic cell adhesion.* ***N-CAMs*** *(alternatively termed Ig-superfamily CAMs) are also homophilic cell adhesion molecules, yet they are independent of Ca^{2+}.* ***Selectins*** *bind carbohydrates with their lectin domains. Carbohydrates taking part in cell adhesion are called "mucin like CAMs". In contrast to all other molecules in this scheme* ***integrins*** *mediate cell matrix adhesion. They are heterodimers and bind to molecules of the extracellular matrix (here fibronectin is shown as example). The artwork was taken from [3].*

selectin mediated adhesion is Ca^{2+} dependent.

Integrins mediate cell matrix adhesion. The integrins are a very versatile family of CAMs. All members of this family are heterodimers, composed of an α and a β protein. At present 17 α units and 8 β units are known. Many of the 17×8 heterodimers exist. The binding site of integrins is composed from *both* subunits. Therefore integrins differing in only one subunit might exhibit selectivities for very different molecules. The best known integrin is most likely $\alpha_{IIb}\beta_3$. This integrin is present at the surface of blood platelets where it is responsible for binding to various molecular species during blood clotting. Integrins also play a major role for cell locomotion which usually occurs on structures of the extracellular matrix. Integrins exhibit affinities in the range of 10^{-6} to 10^{-8} mol/liter which is about 3 orders of magnitude lower than the affinities of typical receptors (10^{-9} to 10^{-11} mol/liter). This low affinity is compensated by the fact that integrins usually function in clusters and rarely ever as isolated molecules. Integrin research still poses many puzzles. Besides the large number of integrins with very pronounced tissue specificity an enigmatic process of integrin activation was described. It appears to be clear that many integrins increase their affinity upon target molecule binding. Moreover, the linkage of these molecules to the cytoskeleton is enhanced upon bond formation. Activation might be a process of the integrin molecule itself. Alternatively, it might depend on processes in the supramolecular adhesion complexes. At present this is one of the puzzles in integrin research.

3 A Biophysical Approach to Specific Adhesion

In the previous part of this lecture, I attempted to give a very short but unbiased overview of the basic aspects of bioadhesion. In contrast, the remainder of this Lecture is extremely biased as I have selected two very specialized experimental approaches to bioadhesion. In a textbook on bioadhesion this section would be just one chapter among a full dozen or so. However, it would be impossible to cover such a wealth of topics in just one lecture. Thus I had to settle for few and closely related aspects. The Reader might have guessed already that this selection was dictated by my personal research interests.

3.1 Model Membrane Studies of Adhesion

Even in my short and very simplified description of bioadhesion in living cells it became clear that many different molecules, structures, and processes contribute to "real" bioadhesion. Given this complexity it seems hopeless to strive for a rigorous quantitative description of the process. One way to overcome this problem is using *biomimetic* systems that on one hand display key processes of bioadhesion but are on the other hand reduced to the bare essentials. By this strategy one can obtain insight into basic physical phenomena in cells. However, a careful design of the biomimetic system is of utmost importance.

In the previous chapters of this book (D1 through D5) biomembranes and model membranes were already discussed. In the laboratory it is possible to prepare closed shell lipid membranes that are many micrometers in diameter but just one lipid bilayer in thickness. These objects are called "giant unilamellar vesicles" or in short GUVs. They can be observed *in situ* and real time by light microscopy. Employing video cameras and digital image processing [17] one can quantify the conformation of the membrane in space and time with high resolution. For this reason GUVs are uniquely suited for studies of biomembrane dynamics and interactions (cf. Chapters D2, D3 and D4 in this book).

In the first part of this section, I will briefly describe a model membrane study of bioadhesion [18]. Here a cell adhesion protein, termed "contact site A" or in short "csA", from the amoeba *Dictyostelium discoideum* was studied. This molecule mediates homophilic cell adhesion and is independent of Ca^{2+}. It is heavily glycosylated, has a molecular weight of 80 kDa, and exhibits a lipid anchor for membrane attachment, see Fig. 6. After purification from amoebae [19] csA forms small micelles, i.e. aggregates of some 10 csA molecules where the lipid anchors form a hydrophobic core surrounded by the hydrophilic protein moieties. Upon incubation with GUVs some of these csA micelles fuse with the membrane. Thus csA is relatively easily incorporated into GUVs. This is rather unusual because most CAMs are transmembrane proteins with a cytoplasmic and an extracellular domain. Such molecules barely incorporate into lipid membranes because of the high energy barrier for the passage of a hydrophilic protein domain through the hydrophobic core of the membrane.

CsA containing membranes are adhesion competent, i.e. they adhere to each other. One of the key features of csA is its high mobility within the membrane. Therefore csA molecules will migrate into the contact zone. CAM aggregation in contact zones was also reported from living cells [20]. In order to study this process in model membranes we used the micropipette aspiration technique [21]. In this technique a cylindrical glass tube of micron-sized diameter (the micropipette) is hydraulically connected to a water reservoir[4] and mounted in a micromanipulator. Manipulation is done in the field of view of a light microscope. Using this technique

[4]The height of this reservoir relative to the measurement chamber defines the aspiration pressure of the pipette.

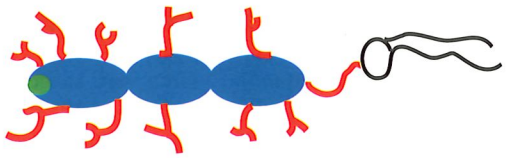


Fig. 6: *Scheme of the cell adhesion molecule* ***csA*** *from the amoeba* Dictyostelium discoideum. *This molecule mediates homophilic cell adhesion and is independent of* Ca^{2+}. *Black: lipid anchor, red: oligosaccharides, green: binding site for opposing csA molecules, blue: domains of the tertiary protein structure.*

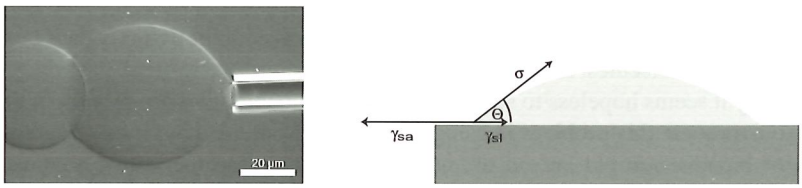


Fig. 7: ***Left:*** *Two adhering giant unilamellar vesicles held in opposing micropipettes. Only the right one is visible. The left vesicle is held under very high suction pressure, resulting in a spherical shape. The right one is much less aspirated and thus assumes a shape which is dictated by the minimization of free energy.* ***Right:*** *Illustration of the Young-Dupre equation of wetting, Eq. 1. See text for details.*

one can aspirate a GUV, control its membrane tension via the suction pressure, and position it wherever necessary. To perform vesicle adhesion experiments one obviously needs two micropipettes in opposition, cf. Figure 7. This type of experiment was introduced by Evans [22]. In order to simplify the geometry, one vesicle is held under very high suction pressure. Thus it assumes a spherical shape. The other vesicle is much more flaccid. The shape of this "test" vesicle is not spherical and dictated by the minimization of the free energy.

Now we are faced with the problem to extract the free energy of adhesion per unit area, γ, from the data. Let us first consider the well known case of a liquid droplet on a solid surface (cf. Figure 7). In this simple example we have to consider three different energy contributions: (i) the free energy density of the solid surface in contact with air, γ_{sa}, (ii) the free energy density of the solid surface in contact with liquid, γ_{sl}, and finally (iii) the free energy density of the liquid surface in contact with air, σ. This latter quantity is just the familiar surface tension of the liquid. Considering the balance of forces at the rim of the droplet one easily derives the familiar equation of Young-Dupre.

$$\gamma = \gamma_{sl} - \gamma_{sa} = \sigma(1 - \cos\theta) \tag{1}$$

Here we introduced the contact angle, θ, of the droplet, cf. Figure 7. Please note that in this situation the "adhesion" energy density, γ, is given by the difference in free energy between dry and wetted solid surface. Thus for a liquid droplet measuring γ is simple: just determine θ and look up the surface tension of the liquid in a table.

However, lipid membranes contain a *fixed* number of lipid molecules. Therefore the surface

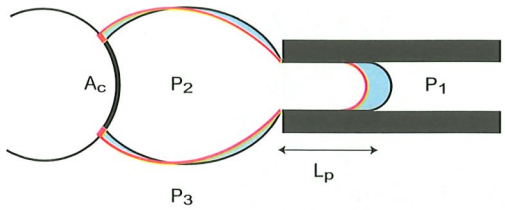


Fig. 8: *Energy change during a virtual displacement of the adherent vesicle. Indicated are the hydrostatic pressures within the pipette, P_1, within the vesicle, P_2, and in the measurement chamber, P_3. The virtual displacement results in increased aspiration length of the vesicle in the pipette, L_p, as well as altered adhesion area, A_c. Volume changes are marked blue.*

area of a lipid vesicle is restricted to a fixed value[5] in contrast to the surface of a liquid droplet. This leads to a variable surface tension. The underlying reason for this is that surface area and surface tension are thermodynamically conjugated quantities. In order to fix the area to a specific value the tension has to change. It is very difficult to calculate the surface tensions for two vesicles held as in Figure 7. The reason for this unfortunate fact is the presence of axial forces that are transmitted via the vesicles from one pipette to the other. Moreover, extraction of the contact angle of both membranes from data like Figure 7 is very uncertain. Thus one cannot use the Young-Dupre equation in a naive way.
Yet there is a way out of this predicament. The vesicles are in mechanical and thermal equilibrium. Therefore the free energy of the configuration must be an extremum, that is the energy change due to virtual displacements must vanish. Consider the energy changes during a virtual displacement, cf. Figure 8. The total energy change, δW, is given by

$$\delta W = \gamma \cdot \delta A_c - (P_3 - P_2) \cdot \delta V_o - (P_1 - P_2) \cdot \delta V_p \tag{2}$$

Here V_p stands for the change of volume of the vesicle part within the pipette and V_o for the one outside the pipette. The first term describes energy changes due to adhesion, while the other two describe hydraulic work. The total volume enclosed within the vesicle is conserved, therefore

$$\delta V_p + \delta V_o = 0$$

Thus we can simplify Eq. 2 to

$$\begin{aligned} \delta W &= \gamma \cdot \delta A_c + (P_3 - P_1) \cdot \delta V_p \\ &= \gamma \cdot \delta A_c + \Delta P \cdot \pi R_p^2 \cdot \delta L_p \end{aligned} \tag{3}$$

where R_p denotes the inner diameter of the pipette, and ΔP the pressure difference between the inside of the pipette and the chamber. This pressure difference is easily determined in the experiment. Mechanical equilibrium implies

$$\delta W = 0$$

[5]To be exact, the surface area of a vesicle can be changed by elastic expansion of the membrane. However, upon stretching by as little as 2% the membrane fails.

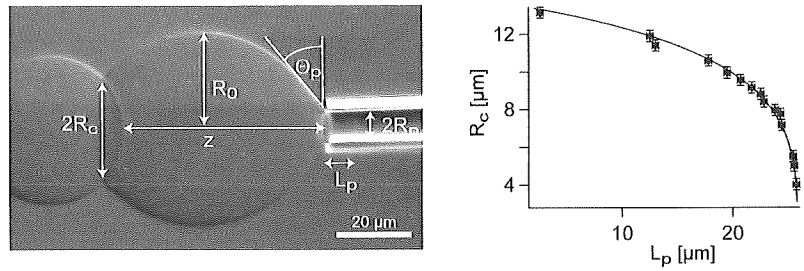


Fig. 9: *Measuring the adhesion energy density using a micropipette.* ***Left:*** *Geometry of an adherent vesicle. R_c denotes the radius of the rim of the adhesion area, R_o the maximum radius of the vesicle, L_p the length of the aspirated nipple, θ_p the angle between membrane tangent and symmetry axis at the mouth of the pipette, R_p the radius of the pipette, and z is the distance between the pipette and the apex of the second vesicle. Please note that z and R_p are fixed quantities, whereas R_c, R_o, and θ_p vary in the course of an experiment.* ***Right:*** *A plot of the calculated relation between the radius of the adhesion zone, R_c, and the projection length, L_p. Superimposed are measured data points. The system is driven along the $R_c(L_p)$ relation by variation of the aspiration pressure.*

This yields the final equation [22]:

$$\gamma = -\Delta P \pi R_p^2 \frac{dL_p}{dA_c} \tag{4}$$

This equation relates the free energy density for adhesion, γ, with the equilibrium shape of the adherent vesicle via the derivative of the projection length with respect to the adhesion area. The next task is to determine this equilibrium shape. The accuracy of light microscopy is not sufficient to use measured values of L_p and A_c. However, the problem can be solved by calculation. The geometry of the adherent vesicle is in fact quite simple. Within the pipette it consists of a hemispherical end cap and a cylindrical part. At the outside there is a cut through a sphere where the vesicle is in contact to the opposing one. The only problematic part is the free segment of the membrane. This part exhibits cylindrical symmetry. Its surface is a surface of constant mean curvature because (i) isotropic tension is the only deformation mode,[6] (ii) due to the absence of external shear forces isotropic tension is constant within the membrane, and (iii) the hydrostatic pressure difference across the membrane is constant because of the absence of fluid flow. In such a situation the Laplace equation is valid. It connects the mean curvature of a surface, $\bar{c}$, and the isotropic membrane tension with the pressure difference across it. It is often encountered in the textbook problems of the internal pressure of soap bubbles or the vapor pressure of little droplets.

$$\begin{aligned} P_3 - P_2 &= \bar{c} \cdot \sigma \\ \bar{c} &= 1/R_1 + 1/R_2 \end{aligned} \tag{5}$$

[6]There are only three deformation modes for membranes: isotropic stretch, in plane shear, and out of plane bending. Shear plays no role because the lipid is in its fluid phase. Per definition fluid materials are unable to support static shear forces. Membrane are so soft for out of plane bending that we can neglect bending stress in comparison to isotropic stretch for the purpose of this experiment.

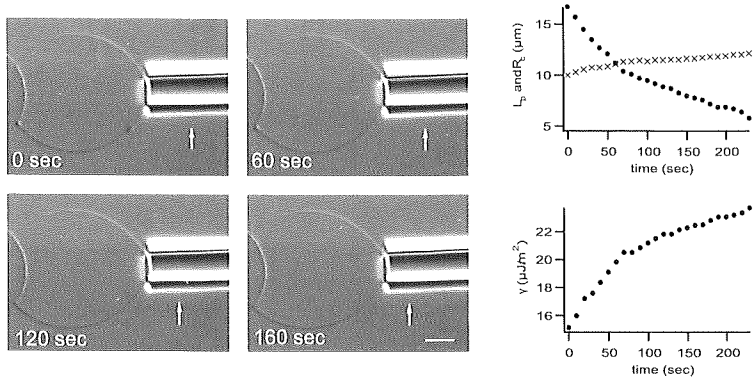


Fig. 10: ***Left:*** *Evolution of the adhesion zone. Aspiration pressure was constant at 5 Pa during the whole experiment.* ***Right Top:*** *Evolution of the geometric parameters, crosses R_c, filled circles L_p.* ***Right Bottom:*** *Adhesion energy density, γ, as derived from the above data.*

Here R_1 and R_2 are the two principal radii of curvature of the surface. It follows from Equation 5 together with considerations (i) through (iii) that $\bar{c}$ must be constant all over the free part of the membrane. Calculating an axially symmetric surface of constant mean curvature is a tedious exercise in differential geometry that neither poses fundamental problems nor gives further insights. The main point in this calculation is that such a surface that connects the mouth of the pipette with the opposing vesicle and exhibits a maximum of the radius in between can be parameterized by just four parameters: R_p, θ_p, R_o, and R_c, cf. Figure 9. From these parameters R_p is a fixed quantity, the others are variable. However, the length of the aspirated nipple within the pipette, L_p, is an additional variable. In this problem there are three conserved quantities: membrane area, enclosed volume, and the distance, z, between the pipette and the opposing vesicle. In other words, we have four unknown parameters (θ_p, R_o, R_c, and L_p) in three equations. Thus it is possible to arbitrarily select one parameter as control parameter and to calculate the other three ones as function of the latter. However, solution is only possible by numerical methods and will not be described here. Readers who need to go into further detail should consult [22, 23, 24]. For the purpose of this Lecture it is sufficient to state that the radius of the adhesion zone, R_c, can be calculated as function of the projection length, L_p, as long as the vesicle is not leaky. In the actual experiments the adherent vesicle is driven through the different states by changing the suction pressure. At each pressure R_c and L_p are determined. After the experiment the function $R_c(L_p)$ is numerically generated and compared with the data, see Fig. 9. By this method leakiness of the vesicle is easily detected. For each data point the adhesion energy density is calculated by Equation 4 employing the numerically determined derivative of L_p with respect to the adhesion area A_c. Using this technique membrane adhesion due to depletion forces or van der Waals attraction were determined [22].

We used this technique on the containing membranes and found an intriguing result. Upon establishment of membrane contact, the adhesion zone spread over minutes, see Figure 10. This is in striking contrast to the aforementioned studies where adhesion was due to colloidal interactions and equilibrated instantaneously. Please note that this observation implies that the adhesion energy density *increases* with time. This result is very surprising at a first glance.

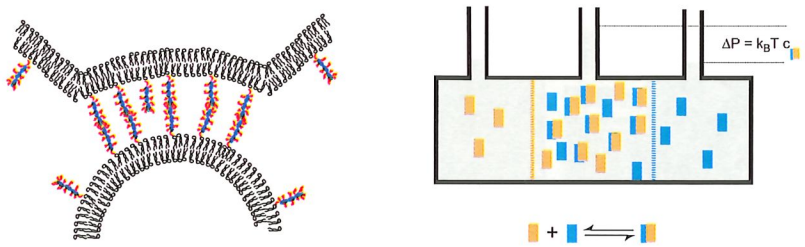


Fig. 11: ***Left:*** *Scheme of the final equilibrium state of csA mediated membrane adhesion. The vast majority of csA molecules forms bound dimers which are in equilibrium with few unbound molecules. Bound dimers are confined to the contact zone.* ***Right:*** *A thermodynamic analogue of this process. Macromolecules (blue and brown rectangles) are separated by semipermeable membranes. Within the commonly accessible volume bonding between both species may occur. Due to this process most molecules migrate into the commonly accessible volume where they increase the osmotic pressure.*

However, it can be understood by considering the molecular processes during formation of the contact zone [25]. Within the contact zone of both membranes csA molecules can form complexes whereas they cannot do this outside the contact zone. Thus the bound complexes are confined to the contact zone. For this reason csA molecules that diffuse into the contact zone will become trapped. Within the contact zone a "gas" of bound receptor pairs forms whose density increases with time. We expect a typical time scale, t_o, of this diffusion limited process of $t_o \sim R_o^2/D$ where R_o denotes the radius of the test vesicle and D the diffusion constant of the molecules. Inserting typical values (R_o 10 μm and D 1 $\mu m^2/sec$) we estimate a typical time scale of this process of about 1 minute. This agrees well with the observation, cf. Figure 10. Please note, that in the final equilibrium the contact zone will contain predominantly bonded receptor pairs plus very few free csA molecules. The concentrations of both species are controlled by the law of mass action.

$$\frac{(f_M x_M)\cdot(f_M x_M)}{f_D x_D} = \exp\left(-\frac{\Delta G_o}{k_B T}\right) \tag{6}$$

Here x_M denotes the molar fraction of monomeric csA molecules and x_D the molar fraction of bonded csA molecules within the membrane, f_M and f_D the respective activity coefficients and ΔG_o the standard free enthalpy for the chemical reaction. Please note that ΔG_o is defined with respect to the chemical reaction of membrane bound receptors. At present there are many techniques to measure the reaction enthalpies of *soluble* molecules. Yet there is no feasible approach to determine the respective value for membrane bound csA molecules. Because the entropic contributions to ΔG_o depend on the membrane anchoring Equation 6 cannot be used to calculate the concentrations.[7]

The contact zone spreads due to the osmotic pressure of the bound receptor pairs, cf. the thermodynamic analogue system in Figure 11 and references [25]. Thus we identify the adhesion energy density, γ, with the osmotic pressure of the bound receptor pairs within the contact zone.

[7] As soon as measurements of the respective concentrations in the membrane become feasible Eq. 6 can serve as a starting point for the measurement of ΔG_o. However, care has to be taken because Eq. 6 contains the implicit assumption of infinite reservoirs, i.e. fixed concentrations, which is not correct here.

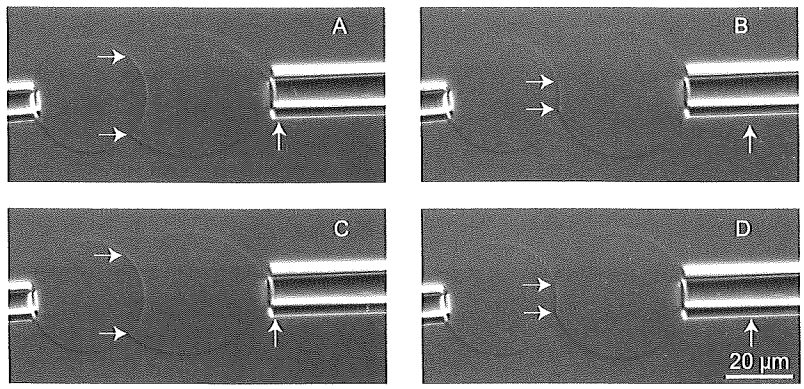


Fig. 12: *A sequence of pressure jumps.* ***A:*** *Final equilibrium at an aspiration pressure of 4 Pa. Radius of the adhesion zone 16.5 µm, $\gamma \cdot A_c$=7 fJ.* ***B:*** *Aspiration pressure 15 Pa. Radius of the adhesion zone 4 µm, $\gamma \cdot A_c$=0.7 fJ.* ***C:*** *Aspiration pressure 5 Pa. Radius of the adhesion zone 16.5 µm, $\gamma \cdot A_c$=9 fJ.* ***D:*** *Aspiration pressure 15 Pa. Radius of the adhesion zone 4.1 µm, $\gamma \cdot A_c$=0.8 fJ.*

In most three dimensional solutions the osmotic pressure of a system follows the van t'Hoff law

$$\gamma = k_B T \cdot c \tag{7}$$

where c denotes the volume density of the molecules. However, it is not at all clear if the same would hold in two dimensions. In order to investigate this question we performed experiments where we brought csA containing vesicles to adhesion, waited for the adhesion zone to equilibrate and finally varied the aspiration pressure in a jump like fashion. The results of such an experiment are shown in Figure 12. From Equation 7 and the finite number of receptors which are basically confined to the adhesion zone, we would expect that the product of adhesion area and adhesion energy density were constant. In the above experiment the adhesion area was reduced to approximately 1/8. At the same time $\gamma \cdot A_c$ is reduced to 10 %. This is by no means constant. Thus we must conclude that *the bound receptors are no ideal gas but interact with each other.* This interaction is attractive. In another study csA molecules without membrane anchors were prepared. These molecules form dimers that do not aggregate any further, i.e. there is no detectible attraction between the csA dimers. Therefore we conclude that *the attractive interaction between the bound receptor pairs is mediated by the membrane.* There are alternative theoretical explanations of this finding [26]. Unfortunately we were unable to measure the concentrations of csA molecules within the membranes directly. Without this information it is impossible to discern between the different theoretical approaches.

In the second and last part of this section, I will discuss another study of membrane adhesion exploiting the possibilities of model membranes [27]. Here we mimicked the adhesion of endothelial cells to the basal membrane of venules. Mechanically a venule has a very simple layout. It contains a pressurized medium and is hydraulically sealed by a monolayer of endothelium cells. The internal pressure in the cylindrical blood vessel has to be balanced by a circumferential force which is supported by a lining of polymers of the extracellular matrix. Thus the layout resembles a bicycle tire where the inner tube contains the pressurized air and the outer case carries the tensile stress necessary to balance this pressure. Obviously, adhesion

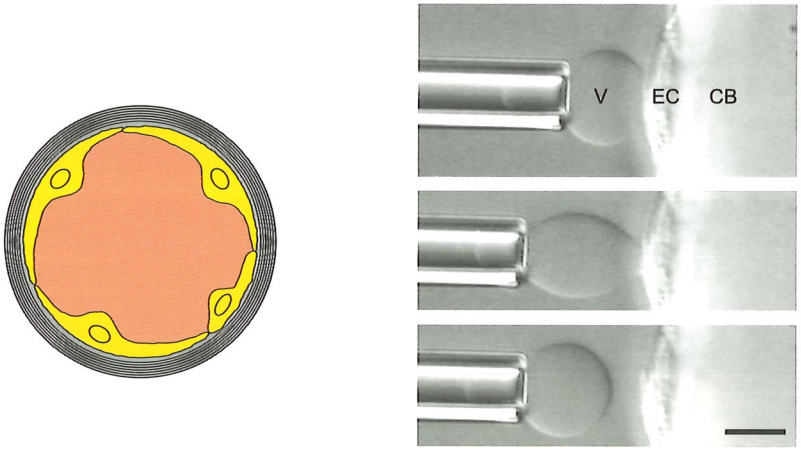


Fig. 13: ***Left:*** *Schematic structure of a venule. Red: blood, yellow: endothelial cells, grey: extracellular polymers.* ***Right:*** *A typical experiment. A functionalized vesicle (V) in contact with an endothelium cell (EC) grown on a collagen bead (CB). The vesicle was aspirated with a pressure of 200 Pa. Top: unloaded state. Middle: The pipette was retracted with a speed of 0.5 μm/sec resulting in vesicle deformation. A force of 4 $\pm$ 1 nN is acting. Bottom: Immediately after failure. The rate of force application was 600 $\pm$ 150 pN/sec. Scale bar, 10 μm.*

between extracellular matrix and endothelium cells is very important. In this case adhesion is predominantly caused by specific binding between the cellular integrin $\alpha_V\beta_3$ and vitronectin, a major constituent of the extracellular matrix.

In order to simplify the problem we replaced the relatively stiff extracellular matrix by a giant unilamellar vesicle (GUV). This vesicle was biofunctionalized by incorporation of a synthetic lipid whose polar group was a RGD peptide. It is well known that integrins recognize certain motifs on the surface of their target molecules that contain the sequence RGD[8]. Therefore integrins bind as well to soluble peptides containing the RGD motif in their sequence [28]. While linear RGD-peptides are recognized rather unspecifically by all integrins, certain circular peptides exhibit a pronounced selectivity for specific integrins. We used the cyclic pentapeptide cyclo[Arg-Gly-Asp-D-Phe-Lys] which is highly specific for $\alpha_V\beta_3$ [29]. This peptide was linked to a normal lipid by a flexible molecular moiety which made it accessible as adhesion motif on the surface of the vesicles.

Endothelial cells from human umbilical cord (HUVEC) were cultivated on the surface of collagen beads. These cells present the integrin $\alpha_V\beta_3$ in a density of approximately 100 molecules per μm^2. In our experiments we brought functionalized vesicles in contact with these HUVECs, see Fig. 13. Indeed we observed RGD-dependent adhesion between cells and vesicles. However, the adhesion zone did not spread at constant aspiration pressure. Moreover, the shape of the vesicles did not change. Adhesion became only apparent upon the attempt to separate vesicle and cell. The absence of spreading indicates that the bound integrins do not constitute an osmotic pressure. The underlying reason for this is the linkage of the integrin molecules to the cytoskeleton of the cell. This linkage immobilizes the bound molecules and thus abol-

[8]R arginine, G glycine, D aspartic acid.



Fig. 14: ***Left:*** *Schematic geometry of the adherent vesicle. To the left intracellular structures are sketched. In contrast to Figure 9 the adherent membrane part is essentially flat, otherwise the analogy is perfect.* ***Middle:*** *Definition of the contact angle θ at the rim of the contact zone.* ***Right:*** *Schematic force distribution across the adhesion zone. For details see text.*

ishes osmotic pressure. Thus we are not able to measure an adhesion energy density of any kind with the usual experiment. Instead we retracted the pipettes with controlled velocities (0.1 μm/sec, 0.5 μm/sec, 1.25 μm/sec, and 2.5 μm/sec). Membrane adhesion became apparent by pronounced vesicle deformation, cf. Figure 13 right, middle. Interestingly, adhesive failure occurred instantaneously, i.e. instead of slow peeling we observed a sudden breakage of adhesion. The main quantity to determine in these experiments was the yield force.
This force could be determined by an extension of the techniques presented above. Again, the free part of the vesicle membrane assumes a shape of constant mean curvature. In essence, the calculation is analogous as before. The only differences are that z is still known but increases with time, and that R_c is fixed in addition to the other conserved quantities (area and volume). For the definition of the variables see Figure 9. Thus we have now three equations (distance between pipette and apex of the membrane, conserved volume, conserved area) and three unknown parameters (R_o, θ_p, and L_p). This system of equations can be uniquely solved. From this solution we determine the membrane tension and the contact angle at the border of the adhesion zone. This yields the tensile force at the rim of the adhesion zone via

$$F_t = 2\pi R_c \cdot \sigma \cdot \sin\theta \tag{8}$$

where F_t denotes the tensile force. The definition of θ can be found in Figure 14. Please note that the pressure within the vesicle exceeds both the pressure within the pipette and the pressure within the chamber. Therefore the membrane is gently pressed unto the cell within the interior of the contact zone. However, at the rim of the contact zone a tensile force, Eq. 8, is acting. This rim is not a mathematical line but a smoothly curved contour. The width of this curved zone is approximately $\sqrt{\kappa/\sigma}$, where κ denotes the bending rigidity of the membrane. From typical values of these parameters ($\kappa \sim 50\ k_BT$ and $\sigma \sim 0.1$ mN/m) we estimate an extension of this bending zone on the order of 0.2 μm. From this width of this bending zone and the density of receptors we estimate that in our experiments between 40 and 400 bonds shared the mechanical load. Repeating these experiments many times we clearly find a marked dependence of the yield forces on the rate of force application, see Figure 15. In this case a rigorous theoretical analysis is not yet possible, however there are first very promising results which connect these results with the behavior of single bonds under load [30], see the next section.

3.2 Mechanical Strength of Single Bonds

In the previous section we saw that the yield force of adherent cells depends on the rate of force application. Due to the size of the acting forces the experiments of the preceding section were

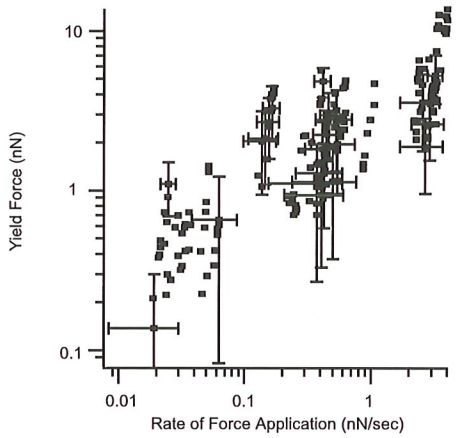


Fig. 15: *Results of many experiments as shown in Figure 13. Error bars are presented for 10% of all data points which were randomly chosen.*

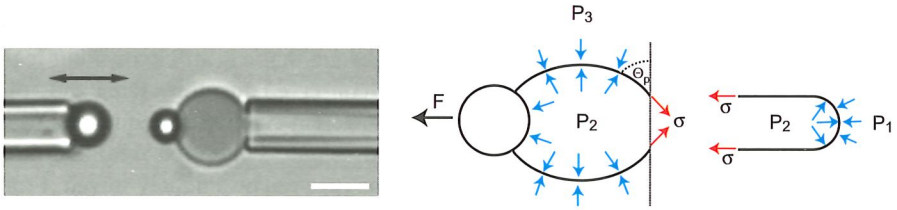


Fig. 16: ***Left:*** *The micropipette based force transducer. A red blood cell is aspirated by the right micropipette. A microbead is glued onto its apex. This microbead carries the receptor molecules and a biochemical "glue" to red blood cell membranes, whereas the ligand molecule is attached to the opposing microbead. Scale bar, 5 μm.* ***Right:*** *Free body diagram of the force transducer. Blue: hydrostatic pressures, red: membrane tension, black: external force.*

relatively straightforward. However, theoretical analysis is very complicated because many bonds share the mechanical load. It is indeed a formidable task to model the distribution of loads between the bonds, to estimate the possible interactions between different bonds, and to extract the chemical kinetics of the individual bonds. In this respects the technically very demanding experiments on single bonds under load proved to be much simpler to interpret.

Several techniques were used to probe the mechanical strength of single specific bonds. Among these are hydrodynamic shear on adherent microparticles, surface force microscopy (AFM), optical tweezers, and a micropipette based technique. Here I will only describe the micropipette based technique. For a review of more techniques and results see [31]. In the micropipette based technique a red blood cell is aspirated by a micropipette and a microbead is glued to its apex, cf. Figure 16. The microbead carries two different types of proteins covalently attached. One type serves to glue the bead to the red blood cell, the other one is the receptor under study. A second microbead carrying the ligand molecule is held directly in a micropipette. Both beads are gently brought into contact and separated again by a piezoelectric transducer. In cases where bonds

were formed during these encounters a finite force is necessary to separate the beads. This force has to be transmitted via the red blood cell. The latter serves as tuneable, ultrasoft spring which will be described in the following paragraph.
Here it is important to note that red blood cells are mechanically extremely simple cells. In essence they consist of a biomembrane enclosing a dense solution of proteins. Mechanically the red blood cell membrane is closely related to the model membranes as described before. It is extremely soft with respect to out of plane bending and barely extensible by isotropic tension. However, it possesses a small but finite elasticity with respect to in plane shear. Yet for large enough tensions (exceeding 25 μN/m) in plane shear plays no role. Thus for the purpose of this experiment we can analyze the red blood cell membrane like any other vesicle. The force transducer exhibits a rather simple geometry, comparable to the membrane adhesion experiments described before, compare Figures 16, 14, and 9. Therefore we are able to analyze the acting forces by a free body diagram [32]. The laws of mechanical equilibrium imply that the sum of all forces that are transmitted across any plane through a body vanishes. Here we consider a plane perpendicular to the symmetry axis located directly at the pipette entrance, cf. Figure 16, and consider the axial forces. The right hand side of this diagram yields:

$$\pi R_p^2(P_2 - P_1) = 2\pi R_p \sigma \tag{9}$$

From the left hand side follows:

$$F + \pi R_p^2(P_2 - P_3) = 2\pi R_p \sigma \sin\theta_p \tag{10}$$

Taken together these two equations yield

$$\begin{aligned} \sigma &= \frac{\Delta P R_p}{2(1-\sin\theta_p)}(1-\hat{f}) \\ \hat{f} &= F/(\pi R_p^2 \Delta P) \end{aligned} \tag{11}$$

This equation relates the membrane tension with the axial force transmitted by it. For the further analysis we need the Laplace equation 5

$$P_2 - P_3 = \bar{c}\sigma$$

Inserting this equation into Eq. 10 we arrive at

$$F + \pi R_p^2 \sigma \bar{c} = 2\pi R_p \sigma \sin\theta_p \tag{12}$$

Finally we insert Eq. 11 into the previous formula to arrive at

$$\hat{f} = \frac{\bar{c} - 2\sin\theta_p/R_p}{\bar{c} - 2/R_p} \tag{13}$$

This equation deserves a closer look. First of all it connects the force acting on the transducer with its shape, i.e. its mean curvature $\bar{c}$. Please note, that force occurs exclusively normalized by the aspiration pressure, ΔP. Therefore the force extension relation of the force transducer will scale directly with the aspiration pressure, i.e. this is indeed a tuneable force transducer. Now consider the case of a spherical red blood cell. In this case $\bar{c} = 2/R_o$ and $\sin\theta_p = R_p/R_o$. Therefore $\bar{c} - 2\sin\theta_p/R_p = 0$, i.e. the force vanishes for a spherical force transducer. In

other words, this force transducer is based on the deviation of a tensed membrane capsule from spherical shape.

For a rigorous calculation of the force elongation relation we have to come back to our discussion of the vesicle adhesion experiment in the previous section. As detailed there the free part of the membrane is a shape of constant mean curvature. Such a shape has only four parameters R_p, θ_p, R_o, and R_c, cf. Figure 9 and the discussion there. In addition, the length of the aspirated part, L_p, can vary. Of these free parameters two are fixed, namely R_p and R_c. Again we have two conserved quantities, volume and surface area. Thus we can again select one parameter as control parameter and calculate the other ones. Here the most practical way to approach the problem is to select a mean curvature $\bar{c}$.[9] From this selection and the equations of volume and area conservation we can calculate the extension of the force transducer as well as R_o, θ_p, and L_p. Moreover, the acting force results directly from the mean curvature via Eq. 13. The response of this force transducer is non-linear and can be only generated numerically. However, for usual parameters (R_{oo}=3 μm and R_p=1 μm) and extensions below 300 nm we can give an approximate equation for the transducer stiffness.

$$k = \frac{R_p \Delta P}{2(1 - R_p/R_{oo})} \frac{2\pi}{\ln\left(\frac{4R_{oo}^2}{R_c R_p}\right) - (1 - R_p/(4R_{oo}) - 3/8 \cdot (R_p/R_{oo})^2 + (R_c/R_{oo})^2)} \tag{14}$$

Thus we find that for small extensions the force probes exhibits a stiffness which is given by membrane tension of the unloaded transducer (the first fraction) multiplied by a geometric factor. For typical values of the geometrical parameters this factor is about 2.4.

This calculation is accurate for a fluid membrane. However, the red blood cell membrane exhibits shear stiffness. Therefore test experiments were performed to explore the validity of this approach [24]. We found that the theory was correct for membrane tensions as low as 20 μN/m. Thus we can achieve transducer stiffnesses as low as 50 μN/m. This is two orders of magnitude lower as the softest AFM cantilevers that are commercially available at present. Moreover, we found that the stiffness of the force transducer can be calculated to an accuracy of 14%. Please note that we are aiming at forces in the low piconewton range where this level of accuracy is indeed remarkable.

In order to measure forces we need not only a spring, we must also measure its extension. For this purpose we use digital image processing to track the microbead at the red blood cell. As can be seen in Figure 16 the light microscopic image of the microbead exhibits a prominent maximum of the intensity within the microbead. This maximum is *not* influenced by the proximity of a second bead. Thus it can serve as marker for the bead position. The light microscopic image is observed with a CCD camera. The present set-up uses a standard TV norm camera. This implies 25 Hz image repetition rate and 50 Hz repetition rate for half-images. Half-images are digitized and transferred to the RAM of a PC. The maximum of the intensity distribution is determined by fitting a Gaussian shape to the distribution of grey scale values. By this method we are able to extract the bead position at a rate of 50 Hz with a standard deviation of 3 nm even so the pixel size of the camera corresponds to 100 nm.

This standard deviation of position corresponds to a noise in the force levels of 0.15 pN. However this is below the thermal noise level of the force transducer. This is given by the equipartition theorem of statistical physics which states that every harmonic degree of freedom is ther-

[9] The mean curvature increases during loading of the force transducer from $2/R_{oo}$ to $1/R_p$ where R_{oo} stands for the radius of the unloaded membrane capsule. Typical values are R_{oo}=3 μm and R_p=1 μm.

mally with half the thermal energy [33]:

$$k/2 < x^2 >= k_B T/2 \tag{15}$$

This corresponds to a fluctuation of the measured forces of

$$\sqrt{< F^2 >} = \sqrt{k k_B T} \tag{16}$$

If we insert the lowest possible stiffness of the force transducer, 50 μN/m, and the thermal energy at room temperature, 4 10^{-21} J, we arrive at a root mean square fluctuation of the force of 0.5 nN. Please note that the bead position exhibits thermal drift. The time scales of drift and Brownian fluctuations are in general way to close to employ temporal averaging. Thus we cannot easily overcome the limits given by Eq. 15. In the praxis we arrive at a level of force detection of 3 pN which is already fairly close to the thermal fluctuations.
Up to now, I described the problems of measuring forces. These are in essence questions of mechanics and optics. However, for these experiments binding partners must be covalently attached to the surfaces. This requires chemical techniques. As this School is focused on students of physics I will not go into detail here. However, the reader should bear in mind that the results of the experiments depend in a very sensitive way on the surface chemistry of the beads. This has several reasons. First of all, proteins tend to change their conformation in close proximity to solid surfaces. This is called denaturation and severely impairs protein function. Moreover, we attempt to measure the interactions due to one single specific bond. However, the contact area of both beads is about 50 000 nm^2.[10] Thus the colloidal interactions between the beads sampled over an area 1000 times larger than one single molecule must be negligible. These are very stringent requirements indeed. In our experiments this problem was solved by separating the proteins from the bead surface by a swollen polymer matrix [34]. All studies on single specific bonds that were published up to now employed some sort of flexible spacer to separate the molecules under investigation from the surface of the solid.
In our study we investigated the bond between protein A and immunoglobulin G (IgG) [34, 35, 36]. Experimental results are presented in Figure 17. Up to now, I described how we are able to measure Piconewton forces and how to immobilize proteins. However, we still have to know how many bonds are probed in one event as shown in Figure 17. Our method rests on the intuitive assumption that several bonds loaded in parallel should resist higher forces than a single isolated bond. The number of available binding sites was varied in the experiment by adding a small amount of one binding partner to the solution, in our case we chose protein A. These dissolved protein A molecules bind to the IgG molecules on the surface of the microbead. The fraction, u, of still available IgG molecules is given by the Langmuir isotherm

$$u = \frac{1}{1 + c_p/k_a} \tag{17}$$

where c_p stands for the concentration of protein A within the solution and k_a is the affinity constant of the interaction between protein A and IgG. In Figure 17 the cumulative distributions of yield forces are shown for different concentrations of protein A. We find the following medians of yield forces 35.5 pN, 30.0 pN, 20.5 pN, 19.5 pN, and 13 pN at protein A concentrations of 0 nM, 0.75 nM, 1.5 nM, 3, nM, and 1 μM. The corresponding probabilities of bond formation

[10]This value follows from the surface roughness ($\sim$ 5 nm) and the radius (1.5 μm) of the beads used here. Similar roughnesses were reported for all types of available microbeads.

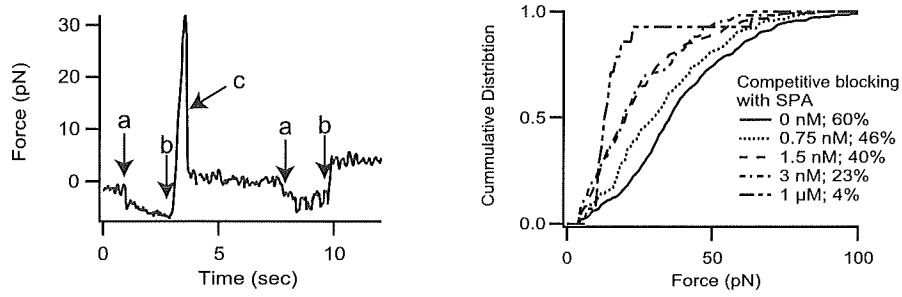


Fig. 17: ***Left:*** *An example of bond formation, here the bond between protein A and immunoglobulin G (IgG) was tested. At* ***a*** *both beads touch and at* ***b*** *piezo retraction starts. In the first encounter there is an adhesive interaction indicated by the retraction of the bead and sudden failure* ***c****.* ***Right:*** *Cumulative distributions of yield forces. To control the density of available binding sites soluble protein A was added to the buffer. This blocks a certain fraction of the binding sites (IgG). Indicated are the concentrations of protein A and the probabilities of bond formation during bead encounters.*

during a bead encounter were 60%, 46%, 40%, 23%, and 4%. Obviously, the yield forces of the adhesive contacts are reduced in going from 0 nM to 0.75 nM and 1.5 nM. However, going from 1.5 nM to 3 nM the yield forces do not change significantly. It is only the frequency of bond formation that drops. Thus we conclude that at concentrations of protein A exceeding 1.5 nM we detect predominantly single molecular interactions. At a concentration of 1 μM basically all binding sites are occupied. Thus at this concentration we detect the background of nonspecific interaction which exhibits a low frequency (4%) and a low median (13 pN).

Finally, we are at a point where we are able to load isolated specific bonds with mechanical forces and to measure their yield forces. The results of these experiments are summarized in Figure 18. Here we find several interesting facts. First of all, the yield forces exhibit a broad distribution. The width of this distribution substantially exceeds the experimental uncertainty. Thus we are lead to the conclusion that *bond failure is a stochastic process*. Moreover, we find that the distributions change markedly upon altered rate of force application, $\dot{f}$.

In order to explain these remarkable facts we have to consider some basic facts about the kinetics of chemical reactions [37].[11] First of all, any chemical reaction can be rationalized as a motion of the thermodynamic system along a one dimensional path in phase space. This path is called the "reaction path". It connects the bound state with the dissociated state of the system. Both states are local minima of the free enthalpy, G. The reaction path is uniquely defined as the path requiring the lowest possible excitation energy. The reaction coordinate is simply the arc length along this path. The microscopic kinetics of complex dissociation can be treated by two alternative approaches, the theory of the activated complex and the theory of Kramers. Both theories agree that the dissociation rate, d_{diss}, is dominated by the height of the activation energy, i.e. the highest value G has to assume along the reaction path, cf. Figure 19.

$$k_{diss} = \alpha \cdot \exp\left(-\frac{\Delta G}{k_B T}\right) \tag{18}$$

[11] The following theory was developed by Evans [38] on the basis of an intuitive approach of Bell [39].

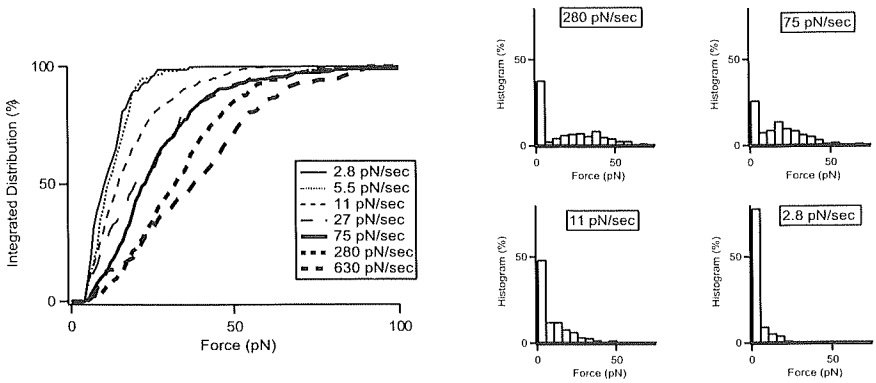


Fig. 18: *Distributions of yield forces of isolated bonds between protein A and IgG determined at different rates of force application, $\dot{f}$.* ***Left:*** *Cumulative distributions.* ***Right:*** *Histograms.*

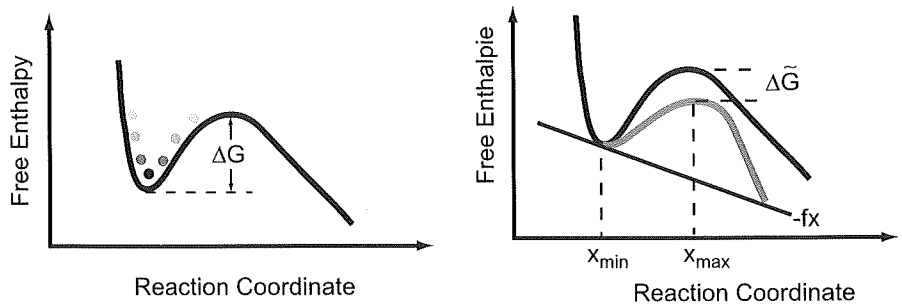


Fig. 19: ***Left:*** *Scheme of the energetics of a dissociation reaction.* ***Right:*** *A mechanical force f tilts the energy landscape in phase space. This reduces the activation energy ΔG by the amount $\Delta\tilde{G}$.*

Here only α is model dependent. This factor is only weakly dependent on the details of the free enthalpy. Thus it may be assumed constant during the rest of this Lecture. A mechanical force, f, acting on the bond is assumed to simply tilt the energy landscape in phase space. Please note, that this is a far reaching assumption which cannot be generally true. However, this assumption will yield results in agreement with observation. In this model the force reduces the activation energy barrier. Thus we get

$$\begin{aligned} k_{diss}(f) &= \alpha \cdot \exp\left(-\frac{\Delta G(f)}{k_B T}\right) \\ &= \alpha \cdot \exp\left(-\frac{\Delta G(f=0)}{k_B T}\right) \cdot \exp\left(+\frac{\Delta\tilde{G}}{k_B T}\right) \\ &= k_o \cdot \exp\left(\frac{f(x_{max} - x_{min})}{k_B T}\right) \end{aligned} \tag{19}$$

Here we introduced the abbreviation k_o for the dissociation rate in the absence of force[12]. The other definitions can be found in Figure 19.
Equation 19 predicts the dissociation rate of a bond held at *constant* force. However, during bond loading we increase the force. This problem can be easily solved. We have to consider the chemical kinetics of the bound complexes. Consider the probability, $p(t)$, that one bond existing at time zero survives up to time t. This probability follows a simple chemical kinetics of first order.

$$\begin{aligned} \frac{dp}{dt} &= -k_{diss}(t) \cdot p(t) + (1 - p(t)) \cdot k_{on} \\ p(t=0) &= 1 \\ k_{diss}(t) &= k_{diss}(f = \dot{f}t) \end{aligned} \tag{20}$$

Here we introduced the rate of bond formation, k_{on}, and the rate of force application, $\dot{f}$, which is constant. The rate of bond formation vanishes as soon as a finite force is applied, i.e. a broken bond cannot be formed again. Please note that this is not true for the case of several bonds acting in parallel. In this case the beads stay in contact even though one bond might break. For this reason the chemical kinetics of force-induced single bond dissociation is very simple whereas the case of multiple bonds cannot be solved analytically.
Using the assumption $k_{on} = 0$ and transforming the variable from time t to force f we arrive at

$$\begin{aligned} \frac{dp}{df} &= -k_{diss}(f)/\dot{f} \\ p(f=0) &= 1 \end{aligned} \tag{21}$$

which has the solution

$$p(f) = \exp\left[-1/\dot{f} \int_0^f k_{diss}(\phi) d\phi\right] \tag{22}$$

Please note, that this equation contains the loading rate in an explicit way. This shows that the distributions of yield forces depend on loading rate, in accord with our observation, cf. Figure 18. Finally we insert Eq. 19 into Eq. 22 and derive the value of the median of yield forces, $F_{1/2}$

$$F_{1/2} = f_o \ln\left(1 + \frac{\dot{f}}{k_o f_o} + \ln 2\right) \tag{23}$$

Here we have introduced the abbreviation $f_o = k_B T/(x_{max} - x_{min})$. This equation predicts that the median of yield forces depends *logarithmically* on the loading rate. This prediction can be tested from our data, cf. Figure 20. Obviously the agreement between theory and experiment is very good.
Despite the surprising success of this simple model there are still many open questions in the field of force-induced dissociation of isolated bonds. For example, the measured distributions consistently and significantly exceed the predictions of Eq. 22. Moreover, the limits of the simple model as described in Figure 19 remain to be explored. Up to now it was remarkable successful in describing experimental results.

[12]Please note that k_o still depends on the presence of the force transducer because during a bond formation the Brownian motion of the force probe are frozen in. This process costs entropy which by itself changes ΔG [36].

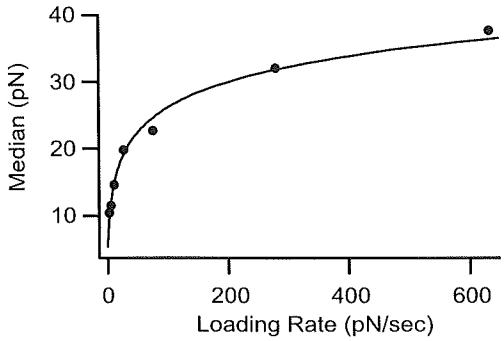


Fig. 20: *Fitting of Eq. 23 to the measured medians of yield forces, cf. Fig. 18. Fit results are* $k_o = 0.12\ sec^{-1}$ *and* $f_o = 5.6\ pN$.

4 Conclusion and Outlook

Here I have presented an overview of bioadhesion from membrane adhesion to force-induced dissociation of single specific bonds. I am well aware of one discrepancy in this article: On one hand I stressed the fact that in living cells cell adhesion molecules almost always act as parts of supramolecular complexes. On the other hand, in the experimental section I went from colloidal interactions between membranes to isolated molecular bonds. You will wonder about the further course of bioadhesion research. Will we proceed to ever more detailed investigations of molecular particulars? I believe there is still a lot to be learned along those lines. However, in my opinion the most pressing task on the way to a physical understanding of bioadhesion is studying the interplay of cell adhesion molecules[13] within the supramolecular cell adhesion complexes of living cells. Unravelling the complex interplay of interactions, molecular species, and mechanical effects at the interface between living cells and their environment is a challenging task for the next decades. I would be happy if some of you would join this exciting field of biophysics.

Acknowledgement

In this Lecture I have described several of my research projects. I am grateful to my students Doris Simson, Markus Strigl, Melanie Nguyen-Duong, Katrin Prechtel, and Christian Meier for their dedicated and gifted work and the pleasure I took from working with them. These projects were funded by the Deutsche Forschungsgemeinschaft and the Alexander von Humboldt Foundation. My research is deeply influenced by my scientific teachers Erich Sackmann and Evan Evans.

[13] Obviously, including their linkages to the cytoskeleton and the extracellular matrix.

References

[1] F. C. Mish (editor) *Merriam Webster's Collegiate Dictionary* (Merriam-Webster, Springfield, 1994)

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Garland, New York, 2002)

[3] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology* (Freeman, New York, 1999)

[4] J. N. Israelachvili *Intermolecular and Surface Forces* (Academic Press, London, 1991)

[5] D. F. Evans and H. Wennerström, *The Colloidal Domain* (Wiley-VCH Verlag, Weinheim, 1999)

[6] D. Leckband and J. Israelachvili, Quarterly Reviews of Biophysics **34**, 105 (2001)

[7] T. D. Brock and M. T. Madigan, *Biology of Microorganisms* (Prentice-Hall, Englewood Cliffs NJ, 1991)

[8] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry, Part I* (Freeman, New York, 1980)

[9] T. E. Creighton, *Proteins: Structures and Molecular Properties* (Freeman, New York, 1993)

[10] C. Chothia, Annual Reviews in Biochemistry **55**, 537 (1984)

[11] C. I. Bränden, and J. Tooze, *An Introduction to Protein Structure* (Garland, New York, 1991)

[12] N. Voiculetz, I. Motoc, and Z. Simon, (editors) *Specific Interactions and Biological Recognition Processes* (CRC Press, Boca Raton, 1993)

[13] M. Delaage (editor) *Molecular Recognition Interactions* (VCH Verlag, Weinheim, 199 1)

[14] D. Livnah, E. A. Bayer, M. Wilchek, and J. L. Sussman, Proc. Natl. Acad. Sci. USA **90**, 5076 (1993)

[15] A. G. Lee (editor) *Biomembranes, Volume 3, Receptors of Cell Adhesion and Cellular Recognition* (Jai Press, Greenwich, Conn, 1996)

[16] R. Alon, S. Chen, K. D. Puri, E. B. Finger, and T. A. Springer, Journal of Cell Biology **138**, 1169 (1997); K. D. Puri, S. Chen, and T. A. Springer, Nature **392**, 930 (1998)

[17] D. B. Murphy *Fundamentals of light microscopy and electronic imaging* (Wiley-Liss, New York, 2001)

[18] C. W. Maier, A. Behrisch, A. Kloboucek, D. A. Simson, and R. Merkel, European Physical Journal E **6**, 273 (2001)

[19] A. Kloboucek, A. Behrisch, J. Faix, and E. Sackmann, Biophysical Journal **77**, 2311 (1999)

[20] M. McCloskey and M.-M. Poo, Journal of Cell Biology **102**, 2185 (1986); M. L. Dustin, Journal of Biological Chemistry **272**, 15782 (1997)

[21] J. M. Mitchison and M. M. Swann, Journal of Experimental Biology **31**, 443 (1954)

[22] E. A. Evans, Biophysical Journal **31**, 425 (1980); E. A. Evans and M. Metcalfe, Biophysical Journal **46**, 423 (1984)

[23] E. A. Evans and R. Skalak *Mechanics and Thermodynamics of Biomembranes* (CRC Press, Boca Raton, FL, 1980)

[24] D. A. Simson, F. Ziemann, M. Strigl, and R. Merkel, Biophysical Journal **74**, 2080 (1998)

[25] G. Bell, M. Dembo, and P. Bongrand, Biophysical Journal **45**, 1051 (1984); E. Evans, Biophysical Journal **48**, 175 (1985); A. Tozeren, Biophysical Journal **58**, 641 (1990)

[26] R. Bruinsma, M. Goulian, and P. Pincus, Biophysical Journal **67**, 746 (1994); D. Zuckerman and R. Bruinsma, Phys. Rev. E **57**, 964 (1998); T. Weikl, R. Netz, and R. Lipowsky, Phys. Rev. E **62**, 45 (2000); S. Komura and D. Andelman Europ. Phys. J. E **3**, 259 (2000); R. Bruinsma, A. Behrisch, and E. Sackmann Phys. Rev. E **61**, 4253 (2000)

[27] K. Prechtel, A. R. Bausch, V. Marchi-Artzner, M. Kantlehner, H. Kessler, and R. Merkel, Phys. Rev. Lett. **89**, 028101 (2002)

[28] M. D. Pierschbacher and E. Ruoslahti, Nature **309**, 30 (1984); J. W. Tamkun, D. W. DeSimone, D. Fonda, R.S. Patel, C. Buck, A. F. Horwitz, and R. O. Hynes, Cell **46**, 271 (1986)

[29] R. Haubner, R. Gratias, B. Diefenbach, S. L. Goodman, A. Jonczyk, and H. Kessler, J. Am. Chem. Soc. **118**, 7461 (1996); M. Pfaff, K. Tangemann, B. Müller, M. Gurrath, G. Müller, H. Kessler, R. Timpl, and J. Engel, J. Biol. Chem. **269**, 20233 (1994)

[30] U. Seifert, Europhys. Lett. **58**, 792 (2002)

[31] R. Merkel, Physics Reports **346**, 343 (2001)

[32] A. K. C. Yeung *Mechanics of Inter-Monolayer Coupling in Fluid Surfactant Bilayers* (PhD thesis, The University of British Columbia, Vancouver, 1994)

[33] R. Becker *Theorie der Wärme* (Springer Verlag, Berlin, 1985)

[34] M. Strigl, D. Simson, C. Kacher, and R. Merkel, Langmuir **15**, 7316 (1999)

[35] D. Simson, M. Strigl, M. Hohenadl, and R. Merkel, Phys. Rev. Lett **83**, 652 (1999)

[36] M. Nguyen-Duong, K.-W. Koch, and R. Merkel, Europhysics Letters **61**, 845 (2003)

[37] P. W. Atkins, *Physical Chemistry* Oxford University Press, Oxford, 1998

[38] E. Evans and K. Ritchie, Biophysical Journal **72**, 1541 (1997)

[39] G. Bell, Science **200**, 618 (1978)

D7 Ion Channels and Pumps

R. Seifert
Institut für biologische Informationsverarbeitung
Forschungszentrum Jülich GmbH

Contents

1 Introduction 2

2 Biological Membranes 2
2.1 Ion Permeability of the Lipid Bilayer 2

3 ATP-driven Transport 5
3.1 Na^+-K^+-ATPase 5
3.2 Ca^{2+}-ATPase 6

4 Ion Channels 9
4.1 Membrane Potential 11
4.2 What Determines the Resting Membrane Potential? 11
4.3 The Action Potential 14
4.4 Propagation of the Action Potential 15
4.5 Neurotransmitter Release 17

5 Molecular Structure of Ion Channels 19
5.1 Ion Channel Families 19
5.2 The Acetylcholine Receptor (AChR) 19
5.3 K^+ Channels 23

6 Conclusion 30

1 Introduction

Electrical pulses generated by neurons, skeletal muscle, and heart muscle cells play an essential role for the coordination of most physiological functions and for information processing, learning and memory in the central nervous system. In this tutorial, the molecules underlying electrical excitability will be introduced.

The text and Fig.s of this review are, in part, based on the text book "Principles of Neural Sciences" by Kandel et al. [1] (Chapters 2, 6-9) and the text books "Molecular Neurobiology" by Hall [2] and "Neuroscience" by Purves et al. [3]. The interested reader is also referred to other textbooks and articles cited at the end of the text [4-8]. Moreover, structures of the Ca^{2+}-ATPase, the nicotinic acetylcholine receptor and various K^+ channels are from references [9-11].

2 Biological Membranes

Cells and subcellular organelles are separated from their environment by membranes that separates its internal contents from the surrounding. The major constituent of the cellular membranes are phospholipids. In a phospholipid molecule, a fatty acid is attached to each of two hydroxyl groups of a glycerol, and a phosphoric acid residue is attached to the third hydroxyl group, instead of a third fatty acid, as in neutral fat (see Fig. 1). Thus all phospholipids thus have a hydrophobic "tail" consisting of two fatty acid chains and a hydrophilic "head" made up of a negatively charged phosphoric acid residue to which a charged or hydrophilic group is attached.

Therefore, phospholipids are amphipathic molecules; they encompass both hydrophobic and hydrophilic groups. Because of their amphipathic nature, phospholipids can interact with both water and non-aqueous phases. In aqueous solutions, phospholipids assemble into *lipid bilayers* (see Fig.1), wherein the fatty acid chains are in contact with each other and the hydrophilic head is in contact with the water phase. This molecular arrangement eliminates unfavourable contact between water and the hydrophobic chains, yet permits the solvation of the polar head groups. Lipid bilayers are roughly 6 nm thick, as measured by electron microscopy and X-ray diffraction techniques. The lipid bilayer presents a formidable energetic barrier to the passage of polar and charged molecules. The ability of lipid bilayers to separate two aqueous compartments is of fundamental importance for all biological membranes. These membranes organize biological processes by compartmentalization.

The low permeability of lipid bilayers to polar and charged molecules allows the solute concentration on each side of the membrane to differ dramatically. Thus, a bilayer-bounded compartment can maintain the complement of large and small molecules required to carry out metabolic processes. However, many biological processes demand that certain molecules and ions pass through the lipid bilayer. Specific transport proteins are incorporated into the membrane that are either able to actively transport molecules across membranes or, alternatively, open water-filled pores that allow the passive flux of ions. In this review some important transport proteins, namely "ion channels" and "pumps" will be described.

2.1 Ion Permeability of the Lipid Bilayer

The water molecule represents a dipole with a partial negative charge on the oxygen and a partial positive charge on the two hydrogens. Therefore, water molecules are strongly attracted by ions in aqueous solutions. In fact, water molecules build a hydration shell around ions (see Fig.1). How strong are the interactions between ions and water molecules? We can estimate the interaction energy from the "heat of hydration" of an ion. It is defined as the increase in enthalpy as one mole of free ion in a vacuum is dissolved in a large volume of water.

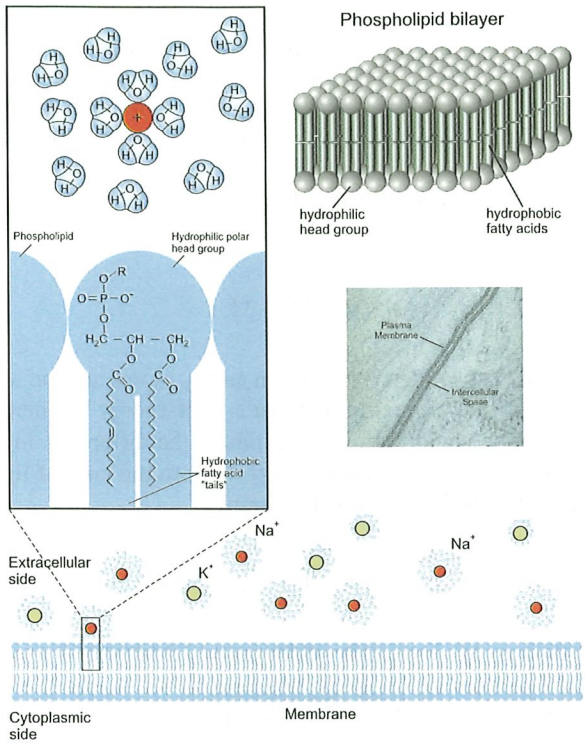


Fig. 1: *(Left) Phospholipids form bilayers that are the basis for all cellular membranes. Phospholipids have a hydrophilic head and a hydrophobic tail. The hydrophobic tails join to exclude water and ions, while the polar hydrophilic heads face the aqueous environment of the extracellular fluid and cytoplasm. Ions in solution are surrounded by a cloud of water molecules. In the illustration, a positively charged ion attracts the electronegative oxygen atoms of the surrounding water molecules. This cloud is carried along by the ion as it diffuses through solution. It is energetically unfavourable, and therefore improbable, for the ion to leave this polar environment to enter the nonpolar environment of the lipid bilayer. The inset also shows the structure of a phospholipid. It is composed of a backbone of glycerol in which two of its -OH groups are linked by ester bonds to fatty acid molecules. The third -OH group of glycerol is linked to phosphoric acid. The phosphate group is further linked to one of a variety of small, polar head groups (R). (Top right) Assembly of phospholipids to bilayer structures. (Bottom right) Electron microscopic picture of a bilayer membrane (Fig. from http://www.willamette.edu/~stas/physiology/labs/lab1).*

The interaction energy can be calculated, for the components of a salt, as the sum of the enthalpy of assembling the salt crystal from the gaseous ions plus the heat of dissolving the crystal in water:

$$\Delta H_{\text{hydration}} = \Delta H_{\text{gaseous ions} \rightarrow \text{solution}} = \Delta H_{\text{gaseous ions} \rightarrow \text{salt}} + \Delta H_{\text{salt} \rightarrow \text{solution}} \qquad (1)$$

Because the "enthalpy of solution" of salts is relatively small, i.e. the temperature of a solution changes only little when a salt is dissolved, the ion-water interaction energies are as large

as the energies that hold a crystal together. For example, the enthalpy of assembling the NaCl crystal in vacuum is –787.3 kJ/mol, the enthalpy of solution of the salt is only 3.8 kJ/mol, and the hydration energy for the pair $Na^+ + Cl^-$ is, therefore, -783.5 kJ/mol.
The classical calculation of hydration energy is based on continuum theory. Born treated water as a homogeneous dielectric, polarized by a charged sphere, the ion. The hydration energy represents the energy required to place the charge into the dielectric. Born calculated from electrostatics that the free energy of transfer of an ion from the vacuum to a medium with dielectric constant ε is:

$$E_b = z^2 q_e^2 / (8\pi\varepsilon\varepsilon_0 r) \quad (2)$$

where r is the ionic radius, z the valence of the ion and q_e represents the unit charge. This difference in free energy often is referred to as Born's "self-energy", sometimes also called dielectric "self-energy". This equation allows to estimate the difference in free energy ΔG for the transfer of an ion from an aqueous medium ($\varepsilon_1 = 80$) to the core of the lipid bilayer ($\varepsilon_2 = 4$), simply by calculating the difference in Born energy:

$$\Delta G = z^2 q_e^2 / (8\pi\varepsilon_0 r) * (1/\varepsilon_1 - 1/\varepsilon_2)) \quad (3)$$

For a K^+ ion (r = 0.138 nm), ΔG is roughly 48 kT, or 117.2 kJ/mol. The probability P that the K^+ ion is located in the phase with dielectric constant ε_2, i.e. within the lipid bilayer, can be calculated from the Boltzmann probability distribution:

$$P_{(\text{lipid bilayer})} = \exp(-E_{b2}/kT) / (\exp(-E_{b2}/kT) + \exp(-E_{b1}/kT)) \quad (4)$$

The probability that the K^+ ion enters the lipid phase is almost zero ($P = 6.5 * 10^{-22}$). Therefore, by all practical means, it is almost impossible for an ion to traverse the lipid bilayer. Table 1 lists the experimentally determined permeability of pure phosphatidyl serine membranes for different ions.

Ion	Permeability (cm/s)
sodium	$<1.6 \times 10^{-13}$
potassium	$<9 \times 10^{-13}$
chloride	$<1.5 \times 10^{-11}$

Table 1: *Ion permeability of pure phosphatidyl serine vesicles.*

However, the permeabilities of biological membranes are several orders of magnitude (10^3-10^6 times) larger (Table 2).

Ion	Permeability (cm/s)
potassium	5×10^{-7}
sodium	5×10^{-9}
chloride	1×10^{-8}

Table 2: *Ion permeability of a mammalian cell membrane.*

How can we account for the significantly larger permeabilities of ions in mammalian cell membranes compared to pure lipid vesicles? The permeabilities are due to active and passive transport processes that are carried out by integral membrane proteins.

3 ATP-driven Transport

Active transport is an energy consuming process that, in most cases, is coupled to the hydroysis of ATP:

$$ATP \rightarrow ADP + P_i \; (\Delta G = -31 kJ/mol) \tag{5}$$

Enzyms that hydrolyze ATP are often called ATPases. The elucidation of the mechanism by which the chemical energy stored in form of ATP is used to drive a mechanical process such as transport has been a challenging problem. In this section we will review the action of some membrane-bound ATPases that translocate cations.

3.1 Na^+-K^+-ATPase

One of the best studied active transport systems is the Na^+-K^+-ATPase in the plasma membrane of higher eukaryotes. This membrane protein consists of two type of subunits: a 110 kDa non-glycosylated α subunit that contains the enzyme's catalytic activity and ion binding sites, and a 55 kDa glycoprotein β subunit of unknown function. Sequence analysis suggests that the α subunit has eight transmembrane segments and two large cytoplasmic domains. The β subunit has a single transmembrane helix and a large extracellular domain. The protein may function as an $(\alpha\beta)_2$ tetramer *in vivo*.

The Na^+-K^+-ATPase is often called Na^+-K^+-pump because it pumps Na^+ out of and K^+ into the cell; the transport is coupled to the hydrolysis of intracellular ATP. The overall stoichiometry of the reaction is:

$$3Na^+(in) + 2K^+(out) + ATP + H_2O \leftrightarrow 3Na^+(out) + 2K^+(in) + ADP + P_i \tag{6}$$

Thus the Na^+-K^+-ATPase separates charges across the membrane, because three positive charges exit the cell for two charges that enter the cell. The resulting concentration gradient of ions is responsible for the electrical excitability of nerve and muscle cells. In fact, all cells spend a large fraction of their ATP (30-70 %) to maintain their cytosolic Na^+ and K^+ concentrations.

Key to the understanding of the Na^+-K^+-ATPase activity is the phosphorylation of a specific aspartate residue. ATP phosphorylates this residue only in the presence of Na^+, whereas the aspartyl phosphate residue is subject to hydrolysis only in the presence of K^+. This suggests that the Na^+-K^+-ATPase has two distinct conformational states (called E1 and E2) with different structures, different catalytic activities, and different ion-binding specificities. The protein is thought to operate in the following manner:

1. The transporter in the E1 state binds three Na^+ ions inside the cell and then binds ATP to yield an E1·ATP·3 Na^+ complex.
2. ATP hydrolysis produces ADP and a "high energy" aspartyl phosphate intermediate E1 ~ P · 3 Na^+ ("~" indicates a "high energy" bond).
3. This "high energy" intermediate relaxes to its " low energy" conformation, E2 — P · 3 Na^+, and releases bound Na^+ to the external side of the cell.

4. E2 — P binds two K^+ ions from the external side of the cell to form the E2 – P · 2 K^+ complex.
5. The phosphate ester bond is then hydrolyzed, yielding E2 · 2 K^+.
6. E2 · 2 K^+ changes conformation to E1 · 2K^+ , releases two K^+ ions to the cytosolic side of the cell, and takes up three Na^+ ions, thereby completing the transport cycle.

Fig.2 shows a sketch of the operation of the Na^+-K^+-ATPase. Although each of the reaction steps are reversible, under physiological conditions the reaction cycle proceeds only in the clockwise direction. This is because ATP hydrolysis and ion transport are coupled, vectorial (unidirectional) processes. The vectorial nature of the reaction cycle results from alternating steps of the exergonic ATP hydrolysis reaction (Steps 1,2 and Step 5) with some of the steps of the endergonic ion transport process (Steps 3,4 and Step 6). Thus, neither process can be completed unless the other one is also completed.

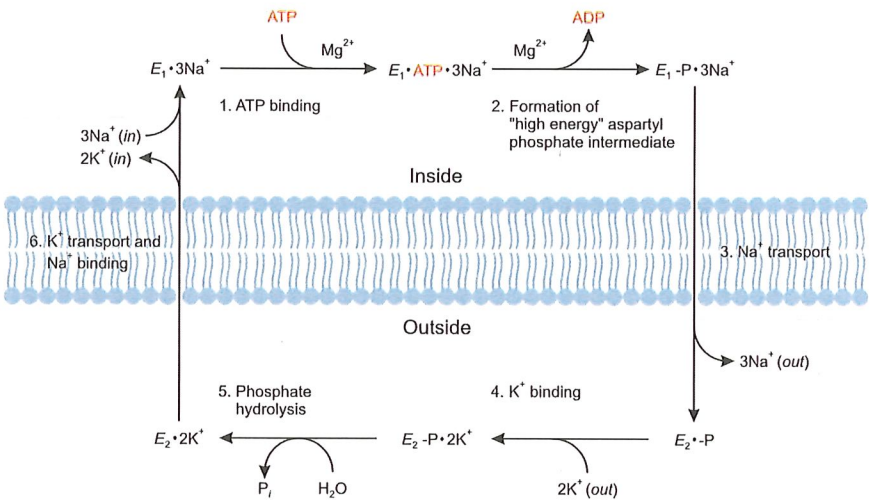


Fig. 2: *A kinetic scheme for the active transport of Na^+ and K^+ by the Na^+-K^+-ATPase. (Adapted from [27]).*

3.2 Ca^{2+}-ATPase

A transient increase in the cytosolic Ca^{2+} concentration triggers numerous cellular responses including muscle contraction and neurotransmitter release. The Ca^{2+} concentration in the cytosol is exceptionally low (~100 nM), more than four orders of magnitude lower than in the extracellular space (1.5 mM). This large concentration gradient is maintained by the active transport of Ca^{2+} across membranes by a Ca^{2+}-ATPase, which actively pumps Ca^{2+} out of the cytosol at the expense of ATP hydrolysis. Fig.3 shows schematically the principle of the operation of the Ca^{2+}-ATPase. Active Ca^{2+} transport has been postulated to be achieved by changing the enzyme structure from state E1 to state E2 (see Fig. 3). In E1, cytoplasmic Ca^{2+} binds to high-affinity sites of the protein; in E2, the binding sites have low affinity and are facing the extracellular side. Transfer of bound Ca^{2+} from the inside to the outside is thought to take place between E1P and E2P.

Recently the crystal structure of the Ca^{2+}-ATPase was solved by X-ray crystallography [24,25]. It provides a framework for the interpretation of functional data and provides, for the

first time, a detailled view of an ion-pumping molecule (Fig. 4). The crystal structure of the Ca^{2+}-ATPase reveals ten transmembrane α–helices, three globular domains on the cytoplasmic side of the membrane and small loops on the extracellular side. The cytoplasmic domains are linked to the transmembrane domain by an α-helical stalk region. The stalk region consists of extensions of transmembrane α–helices M2–M5 beyond the expected membrane surface. The stalk region of the α–helices differs from the transmembrane region in being much less hydrophobic and containing a high proportion of charged residues, particularly glutamic acid and aspartic acid. The centre of the three cytoplasmic domains is the phosphorylation (P) domain. As in the Na^+-K^+-ATPase, a negatively charged aspartate (D351) embodies the phosphorylation site. The P domain is made up of two parts that are distantly located in the linear amino-acid sequence: an N-terminal region comprising the amino acids from about N330 to N359 and a larger C-terminal region from K605 to D737. The nucleotide- (ATP-) binding domain (N) is the largest of the three cytoplasmic domains, extending roughly from Q360 to R604. The N domain is flanked by the two regions making up the P domain. The N domain is linked to the P domain by two loops.

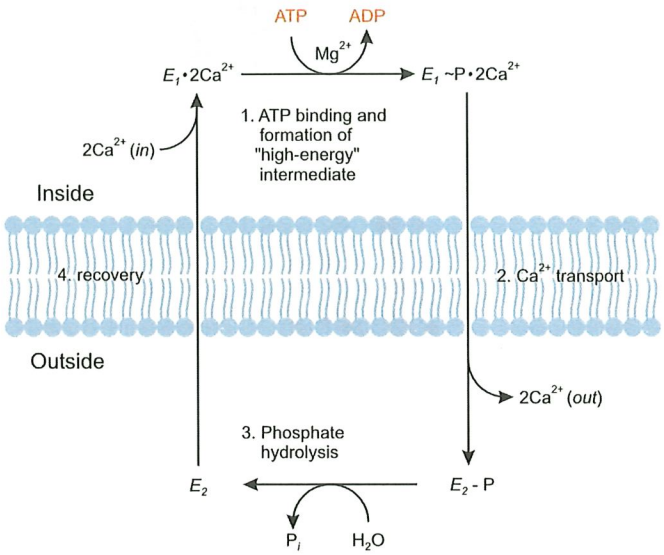


Fig. 3: *The kinetic mechanism of the Ca^{2+}-ATPase. Here (in) refers to the cytosol and (out) refers to the outside of the cell. (Adapted from [27]).*

This should give the N domain enhanced flexibility with respect to the rest of the protein structure. The residues between stalks S2 and S3, together with the N-terminal 50 residues, form what has been called the A domain. This domain is almost isolated from the rest of the structure, being connected to the transmembrane region by three long loops. The function of this domain is, at present, unclear. The Ca^{2+}-bound ATPase structure shows two Ca^{2+} ions in the transmembrane region of the Ca^{2+}-ATPase (Fig. 4); this is in remarkably good agreement with predictions based on site-directed mutagenesis. The two Ca^{2+} ions are bound between transmembrane α–helices M4, M5, M6 and M8. Binding of Ca^{2+} to the ATPase is cooperative and binding of the second Ca^{2+} ion to the ATPase blocks the release of the first Ca^{2+} ion. Studies of the effects of mutation of residues involved in Ca^{2+} binding suggested that one of the two Ca^{2+} ions (the first to bind) binds to site I in the crystal structure, and the other Ca^{2+}

ion (the second to bind) binds to site II. An intriguing feature of the structure is that no obvious aqueous pathway exists from the cytoplasmic surface to the Ca^{2+}-binding sites. It has been proposed that Ca^{2+} ions enter from the cytoplasmic surface between helices M4 and M6. However alternative pathways are also possible. This issue needs further experimental testing. The ATP-binding site of the protein was identified using a chemical analog of ATP (TNP-AMP). TNP-AMP binds to the surface of the N-domain, more than 25 Å away from D351, the residue that is phosphorylated by ATP. Therefore, large changes in the conformation of the cytoplasmic domains must occur to permit phosphorylation of the ATPase. How is phosphorylation of D351 linked to the transport step? Two small helices in the stalk region, P1 and P2, make contacts to the loop between M6 and M7 (see Fig. 5). P1 is close to D351, the phosphorylated residue. In the unphosphorylated state, D351 forms a salt bridge with K684. Phosphorylation of D351 disrupts this salt bridge. Mutation of K684 has significant effects on the rate of phosphorylation of the ATPase, emphasizing its functional importance.

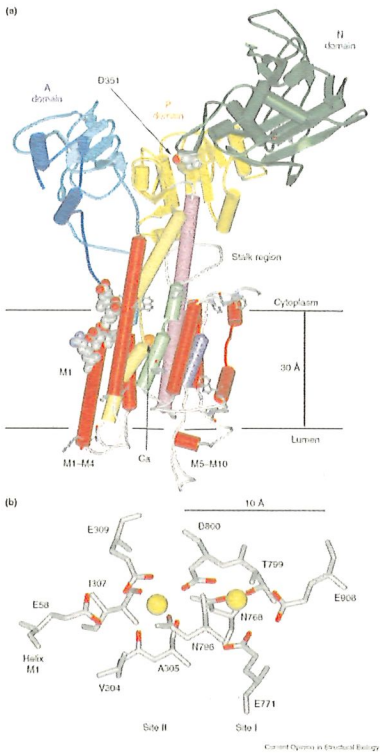


Fig. 4: *The structure of the Ca2+-ATPase. (a) The three cytoplasmic domains A, P and N are coloured blue, yellow and green, respectively. The N-terminal region is coloured dark blue. The four transmembrane α-helices that contribute residues to the high-affinity pair of binding sites are coloured as follows: M4, yellow; M5, lilac; M6, green; M8, blue. Only one of the two bound Ca2+ ions (orange) is visible in this view. Charged residues in M1 are shown in space-fill representation. (b) The residues contributing to the two high-affinity Ca2+ ion binding sites are shown. Also shown is E58 in transmembrane α-helix M1 (PDB code 1EUL, adapted from [9]).*

Movement of D351, upon phosphorylation, could move P1 relative to the M6–M7 loop, and thereby also move the M6–M7 loop itself. This could result in a change in the packing of M6 and M7 and possibly other transmembrane segments. This could convert the cytoplasmic binding sites from a state of high Ca^{2+} affinity to a state of low affinity that is exposed to the extracellular side. As fascinating as this structure is, many questions remain unresolved. However, this structural template should stimulate future experiments to advance our mechanistic understanding of these important transport molecules.
In conclusion, ion-pumping ATPases generate an ion gradient across cellular membranes. Table 3 lists the concentrations of the most important ions in the cytoplasmic and the extracellular medium. The chemical gradients of Na^+, K^+, Cl^- and Ca^{2+} ions are the basis for electrical excitability.

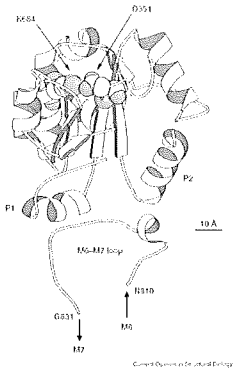


Fig. 5: *The link between the P domain and the transmembrane domain. The Fig. shows how helices P1 (L336 to C344) and P2 (E606 to A617) in the stalk region of the ATPase contact the loop between transmembrane α-helices M6 and M7. Also shown is the interaction between D351, the residue in the ATPase that is phosphorylated by ATP, and K684 (adapted from [9]).*

The advent of the first high-resolution structure of the Ca^{2+}-ATPase provided a wealth of novel information that will greatly enhance our understanding of this class of proteins.

	intracellular (mM)	extracellular (mM)
potassium	140	5
sodium	5-15	145
chloride	4	110
calcium	0.0001	1.5
organic anions	138	9

Table 3: *Concentration of important ions in the intracellular and extracellular medium.*

4 Ion Channels

Ion channels control the electrical activity of literally every cell. They are particularly important for the generation of action potentials and their propagation along the axons of neurons, as well as transmission of action potentials across synapses. The advent of molecular biology,

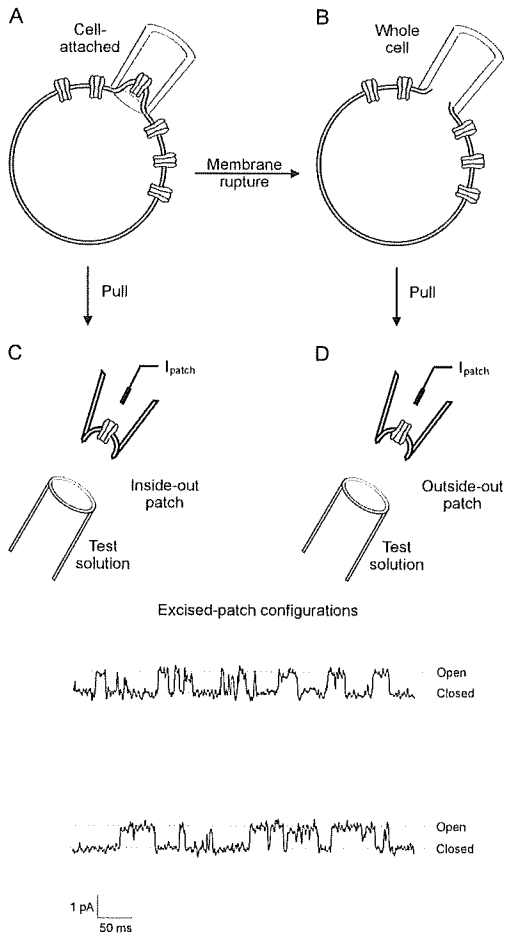


Fig. 6: *(Top) Principle of the patch-clamp technique. A clean glass micropipette can form a high resistance (GΩ) contact when attached to a cellular membrane. Different recording configurations are possible: (A) Cell-attached, (B) Whole-cell, (C) Inside-out and (D) Outside-out. (Bottom) Patch-clamp record of the current flowing through a single cyclic nucleotide-gated ion channel [12].*

in conjunction with advances in electrophysiological recording techniques (patch clamp) that permit registration of single-channel events in microscopically small patches of the cell membrane (see Fig.6), have uncovered a bewildering diversity of structures and mechanisms of activation and modulation of ion channels. Broadly speaking, ion channels fall into two groups: channels that are activated by a change in membrane voltage (*voltage-gated channels*) and channels that become activated by binding of a small ligand to a receptor site on the channel polypeptide (*ligand-gated channels*). Ligand-gated channels fall into two subgroups: channels that bind ligands on the extracellular side (mostly receptors for neurotransmitters) and channels that are activated by *intracellular messengers*, such as Ca^{2+} ions, adenosine

3',5'-cyclic monophosphate (cAMP), guanosine 3',5'-cyclic monophosphate (cGMP), and inositol-1,4,5-trisphosphate (IP_3).

Channels are set apart from each other by their ion selectivity. Most voltage-gated channels are ion selective in that they sharply discriminate between cations and anions and also between different cations. However, selectivity among alkali ions or for monovalent versus divalent cations is not perfect. Relative ion permeabilities P_x/P_y (for example: $x = Na^+$, $y = K^+$) are a common measure of selectivity and range roughly between 1 : 10 and 1 : 1000. In addition, some ion channels even carry mixed currents, i.e. Na^+/Ca^{2+}- or Na^+/K^+ currents. We refer to this class of channels as *non-selective (cation) channels.*
In the following chapter we will introduce the membrane potential and disuss how the activity of ion channels may lead to various forms of electrical activity in neurons.

4.1 Membrane Potential

The activity of the electrogenic Na^+-K^+-ATPase leads to ion gradients across the cellular membrane that are maintained due to the physico-chemical properties of the phospholipid membrane. At rest, a neuron has an excess of positive charges on the outside of the membrane and an excess of negative charges on the inside. The charge separation gives rise to a difference in electrical potential, or voltage, across the cell membrane, called the *membrane potential* (V_m). It is defined as:

$$V_m = V_{in} - V_{out} \tag{7}$$

where V_{in} is the potential on the inside of the cell and V_{out} the potential on the outside. The membrane potential of a cell at rest is called the *resting membrane potential* (V_r). Since, by convention, the potential outside the cell is defined as zero, the resting potential V_r is equal to V_{in}. In neurons, V_r typically ranges from -60 mV to -70 mV. All electrical signalling in a neuron involves brief changes from the resting membrane potential. This is achieved by the transient opening and closing of ion channels leading to the flow of electrical current across the cell membrane. The electric current that flows into and out of the cell is carried by positively charged (*cations*) and negatively charged (*anions*) ions. The direction of current flow is conventionally defined as the direction of *net* movement of *positive charge*. Thus, in an ionic solution cations move in the direction of the electric current, anions in the opposite direction. Whenever there is a net flow of cations or anions into or out of the cell, the charge separation across the membrane is disturbed, altering the polarization of the membrane. A reduction of charge separation, leading to a less negative membrane potential, is called *depolarization*. An increase in charge separation, leading to a more negative membrane potential, is called *hyperpolarization*. Changes in membrane potential that do not lead to the opening of ion channels, are called *electrotonic potentials* and are said to be passive responses of the membrane. Hyperpolarizing responses are almost always passive, as are small depolarizations. However, when the depolarization of the neuron reaches a critical level, called the *threshold*, the cell responds actively with the opening of voltage-gated ion channels and an *action potential* is produced.

4.2 What Determines the Resting Membrane Potential?

The unequal distribution of ions raises an important question. How do ionic gradients contribute to the resting membrane potential? We shall answer this question by considering two examples of membrane permeability: a cell, which is permeable to only one species of ions, and a cell, which is permeable to at least three species of ions.

The overall selectivity of a membrane for different ion species is determined by the fraction of the various types of ion channels that are open. Consider the simple case of a cell that has a membrane that is exclusively permeable to K^+ ions. Because K^+ ions are present at much higher concentration inside the cell compared to the outside, K^+ ions diffuse down their chemical concentration gradient from the inside to the outside. As a result, the outside of the membrane becomes more positive due to the slight excess of K^+, whereas the inside becomes more negative, because of the decrease of the number of K^+ ions and the resulting slight excess of organic anions. The diffusion of K^+ out of the cell is self-limiting. The separation of charge resulting from the diffusion of K^+ gives rise to an electrical potential difference: positive outside, negative inside of the cell. The more K^+ ions leave the cell, the more charges will be separated and the larger will be the potential difference. This potential difference opposes the further efflux of K^+. Thus, ions are subject to different forces driving them across the membrane: (1) a *chemical driving force* that depends on the concentration gradient of the ion across the membrane and (2) an *electrical driving force* that depends on the electrical potential difference across the membrane. Once K^+ diffusion has proceeded to a certain point, a membrane potential is reached at which the electrical force driving K^+ into the cell exactly balances the chemical force driving K^+ ions out of the cell. That is, the outward movement of K^+ (driven by its concentration gradient) is equal to the inward movement of K^+ (driven by the electrical potential difference across the membrane). This potential is called the potassium *equilibrium potential*, E_K. In a cell that is permeable to K^+ ions only, E_K determines the resting membrane potential.

The equilibrium potential for any ion X can be calculated from an equation derived in 1888 from basic thermodynamic principles by the German physical chemist Walter Nernst:

$$E_X = \frac{RT}{zF} \ln \frac{[X]_o}{[X]_i} \qquad \text{Nernst equation} \tag{8}$$

where R is the gas constant, T the temperature (in degrees Kelvin), z the valence of the ion, F the Faraday constant, and $[X]_o$ and $[X]_i$ are the concentrations of the ion outside and inside of the cell, respectively. (To be precise, chemical activities should be used rather than concentrations.)

Since RT/F is 25 mV at 25 °C, and the factor for converting from natural- to base 10 logarithms is 2.3, the Nernst equation can also be written as:

$$E_X = \frac{58 \text{ mV}}{z} \log \frac{[X]_o}{[X]_i} \tag{9}$$

Using the K^+ concentrations given in Table 3:

$$E_k = -84\text{mV} \tag{10}$$

The Nernst equation can be used to calculate the equilibrium potential of any ion that is present on both sides of the cell and can permeate the cell membrane (see Table 4).

In contrast to most other cells of our body, neurons at rest are not only permeable to K^+ ions but to Na^+ and Cl^- ions as well. How do the three ionic gradients interact to determine the cell's resting membrane potential?

To answer this question, it will be easiest to examine first the diffusion of K^+ and Na^+. Let us return to the simple example of a cell having only K^+ channels and concentration gradients for

	intracellular (mM)	extracellular (mM)	E (mV)
potassium	140	5	-84
sodium	from 5 to 15	145	from +85 to +57
chloride	from 4 to 15	110	from -83 to -50
calcium	0.0001	1.5	120
organic anions	138	9	impermeable

Table 4: *Concentration of important ions and their equilibrium potentials (E).*

K^+, Na^+, Cl^-, and A^- as shown in Table 4. Under these conditions the resting membrane potential, V_r, is determined solely by the K^+ concentration gradient and will be equal to E_K (- 84 mV).
Now consider what happens if a few Na^+ channels are added to the membrane, making it slightly permeable to Na^+. Two forces drive Na^+ into the cell. First, the Na^+ concentration is higher outside the cell than inside and, therefore, Na^+ will flow into the cell down its chemical concentration gradient. Second, Na^+ is also driven into the cell by the negative membrane potential. The influx of positively charged Na^+ depolarizes the cell from the K^+ equilibrium potential (-84 mV). However, the new membrane potential does not come even close to the Na^+ equilibrium potential of +57 mV, because the number of open K^+ channels exceeds the number of open Na^+ channels in the membrane.
As soon as the membrane potential begins to depolarize from the value of E_K (-84 mV), K^+ flux is no longer in equilibrium across the membrane. The reduction in the electrical force driving K^+ into the cell will result in a net efflux of K^+ from the cell, tending to counteract the Na^+ influx. The more the membrane potential is depolarized and moves away from the K^+ equilibrium potential, the larger is the electrochemical force driving K^+ out of the cell and consequently the larger is the K^+ efflux. Eventually, the membrane potential reaches a new resting potential at which the efflux of K^+ balances the influx of Na^+. This balance point (usually -60 mV) is far from the Na^+ equilibrium potential (+57 mV) and is only slightly more positive than the equilibrium potential for K^+ (-84 mV).
To understand how the membrane potential arises, bear in mind that the magnitude of the flux of an ion across a cell membrane is the product of its *electro-chemical driving force* (the sum of the electrical and the chemical driving forces) and the conductance of the membrane to the ion:

ion flux = (electrical driving force + chemical driving force) x membrane conductance.

At rest, a cell has relatively few Na^+ channels open, so the conductance to Na^+ is low. Therefore, despite the large chemical and electrical forces driving Na^+ into the cell, the influx of Na^+ is small. Since there are many more K^+ channels open at rest, the membrane conductance of K^+ is relatively large. As a result, the small outward force acting on K^+ at the resting membrane potential is enough to produce a K^+ efflux equal to the Na^+ influx.
Because both Na^+ and K^+ fluxes determine the value of the resting potential, V_r is neither equal to E_K nor to E_{Na}. As a general rule, when V_m is determined by two or more species of ions, the influence of each species is determined not only by the concentration gradients but also by the ease with which the ions cross the membrane. In terms of electrical current flow, the membrane's conductance provides a convenient measure of how readily an ion crosses the membrane. Another convenient measure is the permeability (P) of the membrane to an ion in units of velocity (cm/s, see tables 1 and 2). This measure is similar to that of a diffusion con-

stant, which describes movement in solution. The dependence of membrane potential on ion permeability and concentration is described by the Goldman-Hodgkin-Katz (GHK) equation:

$$V_m = \frac{RT}{F} \ln \frac{P_K\left[K^+\right]_o + P_{Na}\left[Na^+\right]_o + P_{Cl}\left[Cl^-\right]_i}{P_K\left[K^+\right]_i + P_{Na}\left[Na^+\right]_i + P_{Cl}\left[Cl^-\right]_o} \quad \text{GHK Equation} \tag{11}$$

This equation applies only when V_m is constant. It states that the larger the concentration of a particular ion species and the larger its membrane permeability, the larger its role in determining the membrane potential. When the permeability to an ion species is exceptionally high, the GHK equation reduces to the Nernst equation for that ion. So, if $P_K \gg P_{Cl}$ and P_{Na}, the equation becomes:

$$V_m \cong \frac{RT}{F} \ln \frac{\left[K^+\right]_o}{\left[K^+\right]_i} \tag{12}$$

For further reading, we recommend Benedek and Villars [4] and Hille [5].

4.3 The Action Potential

The action potential represents a brief reversal of electric polarization of the membrane of a neuron or a muscle cell (see Fig.7). In neurons, the action potential constitutes the nerve pulse, in muscle cells it produces contraction. Sometimes the action potential is called a propagated potential because a wave of excitation is actively transmitted along the nerve or muscle fibre. The action potential is conducted at speeds that range from 1 to 150 m/s, depending on the properties of the fibre and its environment. How is the action potential generated?

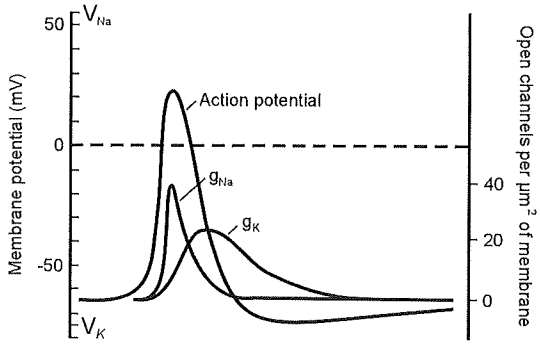


Fig.7: *Shape of an action potential. The sequential opening of* Na^+ *and* K^+ *channels and the resulting conductance changes* (g_{Na}, g_K) *determine the shape of the action potential (adapted from [1]). The change in conductance is expressed as number of open* Na^+ *and* K^+ *channels per* μm^2 *of membrane on the ordinate at the right hand of the panel.*

Usually when a cell is stimulated by a neurotransmitter or by a sensory input, non-selective ion channels open and sodium diffuses into the cell leading to a small depolarization of that part of the membrane (local potential). If the depolarization reaches a critical value (called the threshold potential), then voltage-gated sodium channels open. Due to the large electro-chemical driving force, an inward Na^+ current develops. This Na^+ current causes further depolarization of the cell. More Na^+ channels will open (i.e. g_{Na} becomes larger) and the Na^+ inward current increases further. This positive feedback cycle eventually drives the membrane potential to the peak of the action potential. If the membrane conductance at the peak of the action potential would be solely determined by Na^+ ions ($g_{Na} >> g_K$), then V_m would be equal to E_{Na} (+ 57 mV, see Table 4).
The depolarization not only leads to the opening of Na^+ channels, but also causes them to close again, a process called *inactivation*. The inactivated state of the Na^+ channel is different from its closed state – although both states are non-conductive. The Na^+ channel must proceed from the inactivated to the closed state in order to be able to open again. When Na^+ channels become inactivated, no more Na^+ current can enter the cell. At about the time when Na^+ channels inactivate, the delayed opening of voltage gated K^+ channels occurs. K^+ ions will leave the cell according to their electro-chemical driving force. Because positively charged K^+ ions leave the cell and Na^+ ions cease to enter the cell, the membrane potential returns to its resting value (*repolarization*). Often, the action potential is followed by a transient hyperpolarizing afterpotential. This is caused by K^+ channels that stay open during the late phase of the action potential. The K^+ efflux from the cell is larger than during the resting state and V_m becomes slightly more negative than V_r. Fig.7 shows an action potential and the timing of the changes in the Na^+ and K^+ conductances, g_{Na} and g_K. The first phase of the action potential is dominated by the increase in g_{Na}, while the second phase is almost completely determined by g_K.
When a neuron has just generated an action potential, the cell is unable to generate another one within a certain time span. This *refractory phase* lasts a few milliseconds and is caused by the time it takes for the Na^+ channels to recover from inactivation.

4.4 Propagation of the Action Potential

A fascinating phenomenon is the propagation of an action potential along the nerve or muscle fibre without decrement. This form of signal propagation can be explained by two distinct forms of current flow, first the Na^+ inward current through voltage-gated Na^+ channels and second the passive flow of current along the axon. A depolarizing stimulus – either a postsynaptic potential or a receptor potential – locally depolarizes the fibre, thus opening Na^+ channels in that region. The resulting Na^+ inward current depolarizes the cell even further and opens still more Na^+ channels, generating an action potential at that site. Because the excited segment of the fibre adopts for a brief period of time a potential that is different from that of the rest of the cell, some of the local current produced by the action potential will flow passively down the axon and depolarize the membrane potential in the adjacent regions thereby opening new Na^+ channels (see Fig.8). The continuing cycle spreads until the action potential reaches the end of the axon.
The velocity of action potential propagation depends on both kinds of current flow. For example, the conduction velocity increases with the square root of the diameter of the fibre, because the internal resistance to passive current flow decreases with increasing fibre diameter. This presumably explains why many invertebrates such as squids have evolved giant axons with diameters of up to 1 mm. Another means to enhance the passive flow of electric current in neurons is to insulate the axonal membrane, reducing the ability of current to leak out of the axon. Insulation therefore increases the distance along the axon that a given local current can flow passively. This strategy has resulted in the evolution of *myelination*, a process by

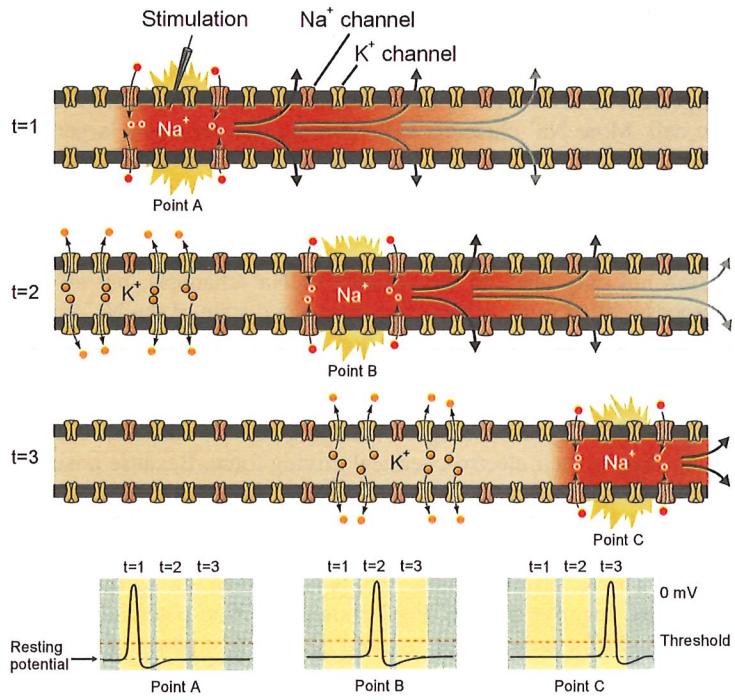


Fig. 8: *Mechanism of propagation of the action potential along the axon. Conduction requires both active and passive current flow. Active current flow occurs through ion channels across the membrane; passive or electrotonic flow of current occurs inside the axon. Depolarization at one point along an axon opens Na^+ channels locally and produces an action potential in this region (point A) of the axon (time point (t)=1). The resulting inward current flows passively along the inside of the axon, depolarizing the adjacent region (point B) of the axon. At a later time (t=2), the depolarization of that adjacent region has opened Na^+ channels and produced an action potential at this site. More inward current is produced that again spreads passively to the adjacent region (point C) further along the axon. At a still later time (t=3), the action potential has propagated even further. This cycle of events continues along the full length of the axon until the action potential reaches the synaptic terminal. As the action potential propagates, the membrane potential repolarizes due to the delayed opening of K^+ channels and the inactivation of Na^+ channels, leaving a „wake" of refractoriness behind the action potential that prevents its backward propagation. (Adapted from [3]).*

which specific glia cells wrap the axon in myelin, an insulating sheet consisting of multiple layers of flattened glia membrane (see Fig.9). By acting as an electrical insulator, myelin greatly speeds up the conduction of the action potential.
Conduction velocities of "naked" unmyelinated axons range from about 0.1 to 10m/s, whereas myelinated axons can conduct at velocities up to 150 m/s. However, if the entire surface of an axon were insulated, there would be no place for current to flow in or out of the axon and action potentials could not be generated. The axon, therefore, is not completely insulated; the myelin sheet is repetitively interupted at certain points along the axon, called *nodes of Ranvier.* At these sites, the density of Na^+ and K^+ channels is particularly high, whereas no or

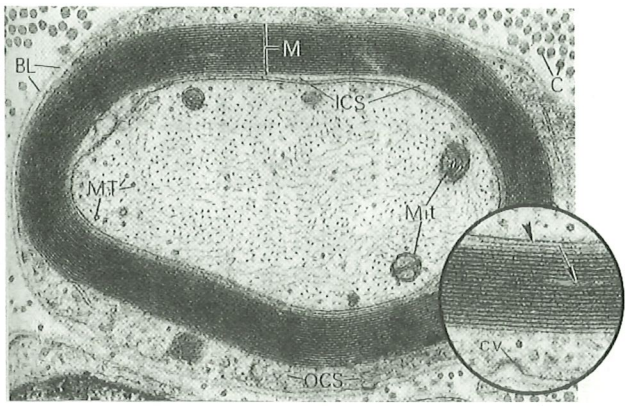


Fig. 9: *Electron microscopic image of the cross-section of a myelinated nerve axon. The myelin membrane (M) wraps around the axon many times and, therefore, creates an insulating shield. This is the basis of saltatory action potential propagation (Fig. from http://www.willamette.edu/~stas/physiology/labs/lab1/).*

only few channels are found underneath the insulating myelin. The current generated by an action potential at a *node of Ranvier* flows passively within the myelinated segment until the next node is reached. Because the action potential is generated only at the nodes, this type of signal propagation is often called "saltatory", meaning that the action potential jumps from node to node. The major reason for the larger conduction velocity is that the time-consuming process of action potential generation is retricted to small segments along the axon.

4.5 Neurotransmitter Release

Different types of ion channels serve distinct functions in a neuron. For example, the action potential is generated by the activity of Na^+ and K^+ channels. An important question is: how is the electrical signal transmitted from a neuron to a neighboring cell? Neurons have developed a system of chemical signal transmission (see Fig.10).

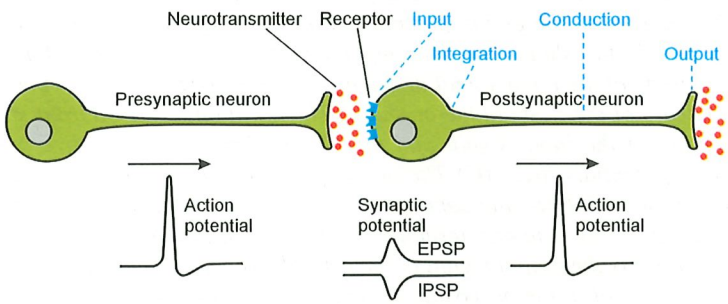


Fig. 10: *Communication between neurons. Neurons convey information by a combination of chemical and electrical signals. At chemical synapses, an action potential causes release of neurotransmitters onto the postsynaptic cell. The neurotransmitters bind to specific receptors and evoke electrical signals (synaptic potentials) in the postsynaptic cell. These synaptic potentials eventually trigger an action potential which travels down the axon to the nerve terminal and causes neurotransmitter release onto the next cell (adapted from [2]).*

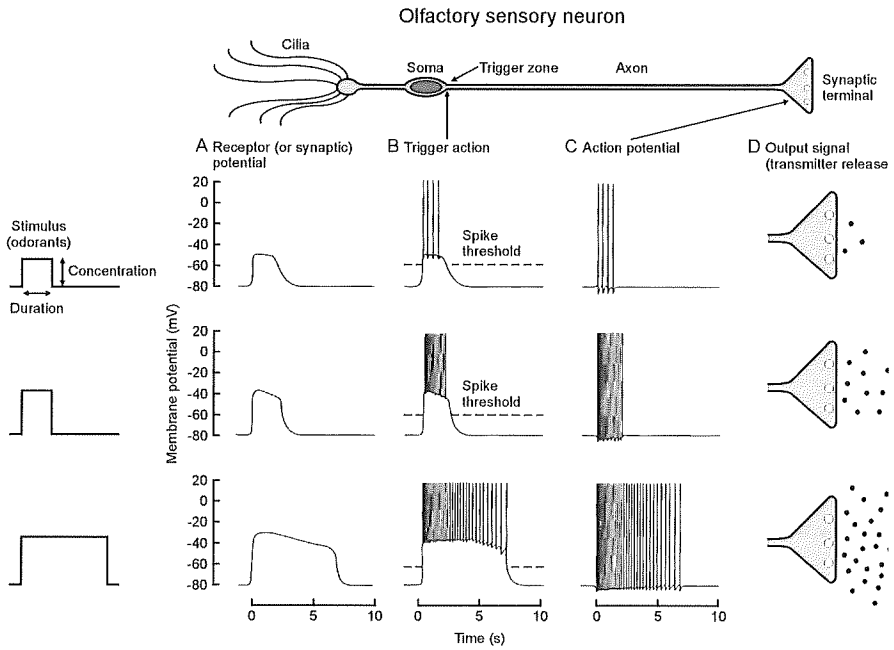


Fig. 11: *An olfactory neuron transforms a chemical stimulus into electrical activity in the cell. Sensory neurons in the olfactory epithelium which lines the nasal cavity consist of a cell body, a small dendrite, and a long axon that projects to the olfactory bulb in the brain. Approximately 10-15 thin cylindrical cilia protrude from the swollen end of the dendrite. The cilia harbour the molecular machinery to register odorants and transduce the stimulus into an electrical signal.* ***(A)*** *The input signal is a change in the resting potential – called the receptor potential – elicited by stimulation of the cell with an odorant. The receptor potential is graded in amplitude and duration. The relation between stimulus strength (odorant concentration) and receptor potential amplitude can be very steep due to several stages of biochemical amplification inside the cilia.* ***(B)*** *The trigger zone integrates the input signal and produces an action potential. The action potential is only generated if the receptor potential is greater than a certain threshold. Once the input signal surpasses this threshold, any further increase in amplitude of the input signal augments the frequency with which the action potentials are generated, not their amplitude or shape. The duration of the odorant stimulus determines the duration of the input signal and thereby the length of the active period, i.e. the number of action potentials fired.* ***(C)*** *The action potentials are invariant in size and shape. Thus every action potential has the same amplitude, duration and waveform.* ***(D)*** *When the action potential enters the synaptic terminal, a chemical messenger or neurotransmitter is released that represents the output signal of the cell. The released neurotransmitter molecules elicit a synaptic potential in the postsynaptic cell. The synaptic potential represents the input signal to that cell.*

The action potential stimulates the release of small organic molecules (*neurotransmitters*; acetylcholine, Ach; γ-amino butyric acid, GABA; glutamate; glycine; 5-hydroxytryptamin, serotonin etc.) from the synaptic terminal into the synaptic cleft. Neurotransmitter release is a complex process that involves several consecutive, highly regulated biochemical reactions.

Neurotransmitters are stored in small organelles called *synaptic vesicles*. When the action potential enters the synaptic terminal, the vesicles fuse with the plasma membrane. The fusion process requires Ca^{2+} to flow into the synaptic terminal. The action potential causes opening of voltage-gated Ca^{2+} channels which mediate Ca^{2+} influx into the cell. These channels are located in close proximity to the fusion and release sites of the synaptic vesicles (*active zone*). Only a few vesicles will fuse and release their content per action potential. Within a couple of milliseconds this process, also called *exocytosis*, has been completed, and the vesicles are retrieved by *endocytosis* from the plasma membrane and refilled with neurotransmitter. What happens to the released neurotransmitters? The neurotransmitters freely traverse the synaptic cleft and bind to receptors on the surface membrane of the postsynaptic cell. Activation of these receptors (ligand-gated ion channels) leads to either depolarizing (ACh-, glutamate receptor) or hyperpolarizing (Glycin-, GABA receptor) currents mediated by Na^+/Ca^{2+} or Cl^- ions, respectively. The target cell will respond with a small depolarization (*epsp*; excitatory postsynaptic potential) or hyperpolarization (*ipsp*; inhibitory postsynaptic potential). The electrical properties of the activated receptors thus determine the electrical response of the postsynaptic cell.

The sequence of events from a receptor potential to synaptic signaling is illustrated in Fig.11.

5 Molecular Structure of Ion Channels

5.1 Ion Channel Families

Fig.12 gives an overview of different ion channel families. They differ significantly in their membrane topology and subunit stoichiometry. While many neurotransmitter-gated channels assemble as heteropentamers, the voltage-gated channels are usually tetrameric heteromultimers like the members of the superfamily of K^+ channels or "pseudotetrameric" multimers like voltage-gated Ca^{2+} and Na^+ channels. Many voltage-gated channels have additional nonporeforming subunits that are either important for targeting of the channel to the correct subcellular structure in a cell, or that are involved in the modulation of the gating properties of a channel. In the next section a few well studied examples of ion channels will be discussed in more detail.

A third family of K^+ channels has a characteristic subunit structure corresponding to two repeats of the inward-rectifier K^+ channel architecture, with two P-regions in tandem. The subunit composition of these channels is not known.

5.2 The Acetylcholine Receptor (AChR)

The nicotinic acetylcholine (ACh) receptor at the nerve–muscle synapse is the best understood neurotransmitter-gated ion channel. These channels are located in the postsynaptic membrane of the muscle cell. They open transiently on binding of ACh released from the nerve terminal into the synaptic cleft, allowing sodium and potassium ions to flow. The resulting depolarization of the postsynaptic membrane leads to the activation of voltage-gated channels and ultimately to muscle contraction. Many basic biochemical and physical properties of the ACh receptor have become firmly established as a result of intense research over the past 25 years. In bare details, the AChR is a 290-kDa glycoprotein composed of two ACh - binding subunits(α) and three other subunits (β, γ, and δ) which have significant sequence similarity to the α subunits. These five subunits form a ring spanning the membrane and encircling a gated pathway for the ions. Recently, enormous amount of information about the structure of the AchR has been obtained by electron microscopy of two-dimensional pro-

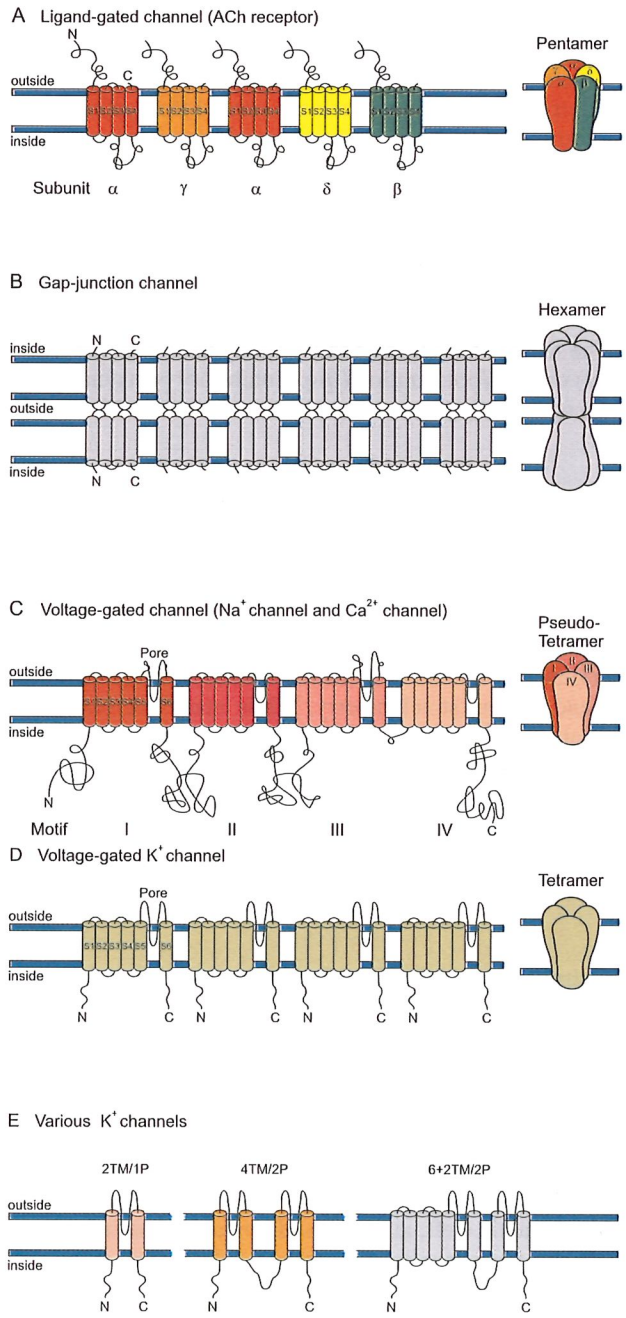


Fig. 12: *Families of ion channels:* ***(A)*** *Several ligand-gated channels, including the nicotinic acetylcholine (Ach) receptor-channel, have five subunits, and each subunit consists of four transmembrane segments (S1-S4). Each cylinder represents a single transmembrane α-helix. A three-dimensional model of the channel is shown on the right.* ***(B)*** *The gap-junction channel, found at electrical synapses, is formed from a pair of hemichannels in the pre- and post-*

synaptic membranes that join in the space between two cells. Each hemichannel is made of six subunits, each with four transmembrane segments. A three-dimensional model of the two hemichannels is illustrated on the right. ***(C)*** *The voltage-gated Na^+ channel is formed from a single (α) polypeptide chain that contains four homologous domains or repeats (motifs I-IV), each with six α-helical transmembrane segments (S1 to S6) and one P-region thought to line the pore. The Fig. on the right shows the hypothetical model of the channel.* ***(D)*** *Voltage-gated K^+ channels are composed of four polypeptide chains. Each subunit corresponds to one repeated domain of a voltage-gated Na^+ or Ca^{2+} channel, with six transmembrane segments and a P-region.* ***(E)*** *Inward-rectifier K^+ channels are composed of four polypeptide subunits. Each subunit has only two transmembrane segments, connected by a P-region loop.*

tein crystals [23] and by X-ray crystallography of a bacterial ACh-binding protein [26] that bears high homology to the binding fold of the eukaryotic receptor.

Each α subunit has a bipartite structure with an N-terminal ligand-binding domain and a C-terminal membrane-spanning domain. The ligand-binding domains are organised around two sets of β–sheets packed into a curled β–sandwich and joined through a disulfide bridge. The encircling chains create a long central vestibule that is ~20 Å diameter (see Fig.13). The membrane-spanning part of the subunits is made from four α-helical segments (M1 to M4), and short connecting loops. The segments M1 to M4 are arranged symmetrically, forming an inner ring of helices (M2), which constitutes a narrow pore, and an outer shell of helices (M1, M3, and M4), which coil around each other and shield the inner ring from the lipids. As shown in Fig.13, the polypeptide chains of the ligand-binding domain are superimposed on the outer helices, leaving only the inner helices exposed directly to the ions.

The inner pore-lining helices (M2) are 40 Å long (including portions outside the membrane), and are tilted toward the central axis; in the middle of the membrane, the M2 helices are slightly kinked. At this level, in the closed pore, the set of M2 helices come together separately from the set of outer helices, creating intervening water-filled spaces (Fig.13). These intervening spaces play an important role in the gating mechanism (see below), by providing room for the M2 helices to move relative to the outer lipid-facing protein wall.

The closed pore has two important structural features in the middle of the membrane. First, it is maximally constricted in this region due to the small separation of the M2 helices and to the presence of bulky hydrophobic side chains. Second, it is essentially symmetrical in this region, due to equal side-to-side hydrophobic interactions between equivalent surfaces of homologous residues. The contacts are at two levels: one involving leucine (at α L251) with the neighbouring alanine (or serine) side chains (at α S252), and the other phenylalanine (at α F256) with the neighbouring valine (or isoleucine) side chains (at α V255).

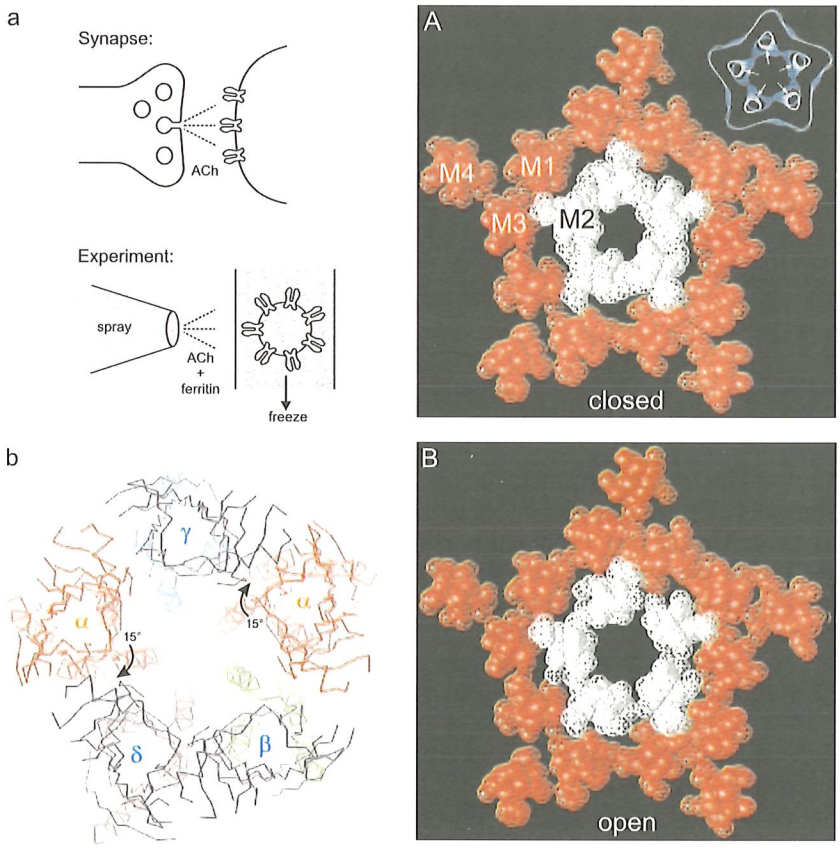


Fig. 13: *(Left) Electron microscopy of the ACh receptor.* ***(A)****: Channel-opening event at the synapse (above) and the experiment to capture this event in vitro (below). In the experiment, the spraying of ACh-containing droplets onto an aqueous film containing isolated postsynaptic membranes mimics the release of neurotransmitter onto the target cell, and rapid freezing traps the channels in the open state. Ferritin marker particles, included in the spray solution, reveal the regions of the film where the droplets have impinged and spread, and so identify the membranes containing open channels.* ***(B)****: The polypeptide chains (α-carbon traces) fitted to the electron image density maps, as viewed from the synaptic cleft. The ligand-binding domain (uppermost) is composed mainly of β-sheet, while the domain forming the membrane-spanning pore (underneath) has an α-helical fold. Binding of ACh to the ligand-binding domain initiates rotational movements in the α subunits (arrows) which are communicated to the inner helices shaping the pore.*
(Right) Cross-sections at the gate in middle of the membrane, showing van der Waal's surfaces of the atoms encircling the closed (A) and the open (B) pore. The open pore is modelled by applying 15° clockwise rotations to each of the inner helices (coloured white). The change in structure, involving a widening of the pore by ~3 Å, is consistent with the changes observed experimentally at 9 Å resolution; these changes are shown in the inset (blue: closed channel; white and brown: open channel). The view is from the synaptic cleft; individual helices, M1–M4, are identified on one of the subunits (adapted from [10]).

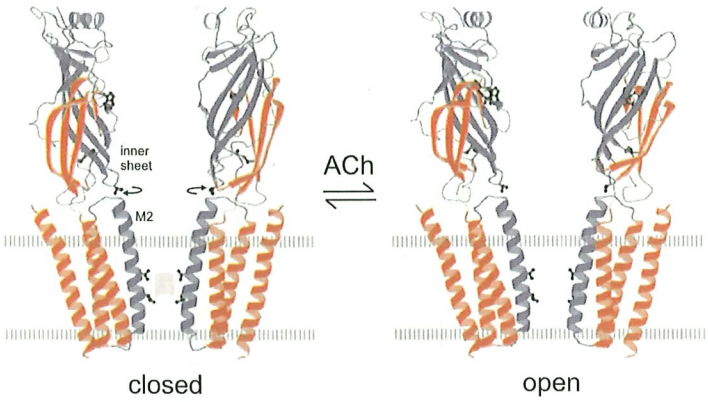


Fig. 14: *Conformations of the closed (left) and open (right) channels. Binding of ACh opens the channel by initiating rotational movements (arrows) of the inner β–sheets of the α subunits in the ligand-binding domain. These movements are communicated to the inner (M2) helices lining the pore and break apart the gate – a hydrophobic girdle in the middle of the membrane – so that ions can flow. A tryptophan side chain in the ligand-binding domain identifies the ACh-binding region; a valine side chain links the inner sheet to the inner helix; leucine and valine side chains on the inner helices make the gate (pink patch); the locations of the membrane surfaces are indicated by broken lines; the relevant moving parts are in blue. Most of the loop regions in the ligand-binding domain are modelled on the structure of AChBP (adapted from [10]).*

These symmetrical side-to-side interactions bring together the side chains on the neighbouring helices to form a tight hydrophobic girdle around the pore (see Fig.14). This girdle creates an energetic barrier to ion permeation across the lipid bilayer because the ~6 Å diameter pore is too narrow for a Na^+ or K^+ ion to pass through while retaining its first hydration shell, and the ion cannot readily lose part of this shell in the absence of polar surfaces that would substitute for water. Since there is no other barrier or protein occlusion that would block the flow of ions, the hydrophobic girdle most probably corresponds to the gate of the pore.
The binding of ACh initiates a rotational movement of the inner β–sheets of both α subunits (see Fig.14). These rotations are communicated to the inner, pore-lining helices and cause them to switch cooperatively to another conformation. The change in conformation destabilises the gate and opens the pore.

5.3 K^+ Channels

Selectivity

All known K^+ channels are members of a large protein family. They are found in bacterial, archeal, and eukaryotic cells – both plants and animals – and their amino-acid sequences are easy to recognize because K^+ channels contain a highly conserved segment called the K^+ channel signature sequence (Fig.15). This sequence forms a structural element known as the selectivity filter, which prevents the passage of Na^+ ions but allows K^+ ions to pass across the membrane at rates approaching the diffusion limit.

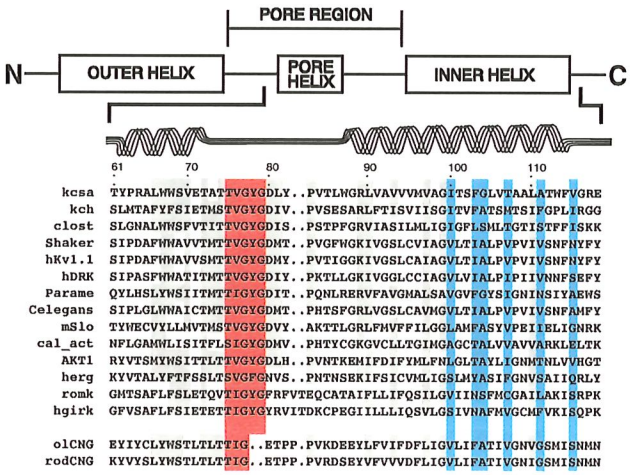


Fig. 15: *Sequence alignment of selected K^+ channels and cyclic nucleotide-gated channels. The numbering and secondary structural elements for the Streptomyces lividans K^+ channel (KcsA) is given above the sequences. Selectivity filter, red; lining of the cavity and inner pore, blue; residues in which the nature of the side chain is preserved (>50% similarity), grey. The last two sequences, separate from the rest, are from cyclic nucleotide-gated channels, which are not K^+ selective (adapted from [14]).*

This is the hallmark of K^+ channels: nearly perfect selectivity for K^+ ions over Na^+ ions and very high K^+ conduction rates. Fig.15 shows the sequence alignment of the pore region of various K^+ channels. A short stretch of amino acids is invariant in all known K^+ channels, T(V/I)G(Y/F)G (one letter code). From mutational studies it is known that these amino acids can not be exchanged for others without losing the exquisite K^+ selectivity. Indeed, two members of the cyclic nucleotide-gated channel family listed at the bottom of the table (olCNG, rodCNG) deviate exactly in this important structural domain; these channels are not K^+ selective; rather they belong to the group of non-selective cation channels.

How is K^+ selectivity achieved? The atomic radius of K^+ is 1.33 Å and that of Na^+ is 0.95 Å. Despite this small difference in ionic radius, K^+ channels manage to select for K^+ over Na^+ by a factor of >1000. Even more significant, this strong selectivity for K^+ is achieved without compromising on the conduction rates that approach the diffusion limit. Atomic structures provided a comprehensive and detailled understanding of K^+ selectivity and the ion conduction process [14, 15, 16, 18].

The K^+ channel pore is comprised of four identical subunits that form with four-fold symmetry a central ion conduction pathway. Two of the four subunits of a bacterial K^+ channel (KcsA) are shown in Fig.16. Each subunit contains two membrane-spanning α–helices termed inner (nearest the ion pathway) and outer (nearest the membrane) helices, and a shorter tilted pore helix (red, Fig.16) that runs half way through the membrane. Its C-terminal negatively charged end points toward the ion pathway. Near the centre of the membrane, the ion pathway is almost 10 Å in diameter, forming a central water-filled cavity. In the crystal-structure a hydrated K^+ ion is located in the center of this cavity. The presence of a water-filled cavity and oriented α–helices helps to explain an important property of the K^+ channel: in order to catalyze high conduction rates, the channel must overcome the electrostatic repulsion that a K^+ ion would normally experience when moving from bulk water into the low di-

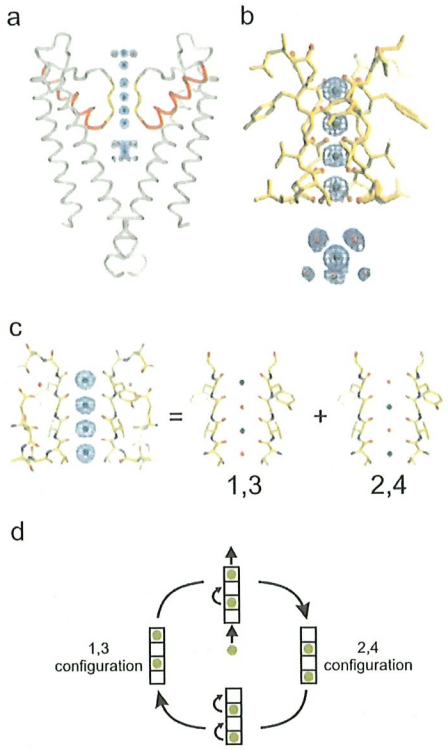


Fig. 16: *The ion conduction pore of K^+ channels. a: Two of the four subunits from the KcsA pore are shown with the extracellular side on top. Each subunit contains an outer helix close to the membrane, an inner helix close to the pore, a pore helix (red) and a selectivity filter (gold). Blue mesh shows electron density for K^+ ions and water along the pore. b: Close-up view of the selectivity filter with dehydrated K^+ ions at positions 1 through 4 (external to internal) inside the filter and a hydrated K^+ ion in the central cavity below the filter. c: Electron density in the filter corresponds to two configurations of K^+ ions (green) alternating with water molecules (red): 1,3 contains K^+ at positions 1 and 3, 2,4 contains K^+ at positions 2 and 4. d: Through-put cycle for K^+ conduction invoking 1,3 and 2,4 configurations (adapted from [11]).*

electric environment of the membrane. By allowing the K^+ ion to stay hydrated at the membrane centre, and by orienting the C-terminal end of the pore helices toward the ion pathway, a K^+ ion is stabilized at the centre of the membrane.

Potassium selectivity occurs in the selectivity filter, located in the extracellular third of the ion pathway, between the central cavity and the extracellular solution (Fig.16, gold). The signature sequence of K^+ channels, conserved throughout the tree of life, forms the selectivity filter. Here, conducting K^+ ions interact with four evenly spaced layers of carbonyl oxygen atoms and a single layer of threonine hydroxyl oxygen atoms. These layers create four K^+-binding sites, numbered 1 to 4 from the extracellular to the intracellular side (Fig.16). At these sites, K^+ ions bind in a dehydrated state, surrounded by eight oxygen atoms from the protein, four `above' and four `below' each ion in Fig.16. The arrangement of oxygen atoms that form binding sites in the selectivity filter is very similar to the arrangement of water molecules

around the hydrated K^+ ion in the central cavity. Thus, the crystal structure provides an explicit and beautiful demonstration of the selectivity filter's role to create a queue of K^+-binding sites that mimic the water molecules in the hydration shell of a K^+ ion. It is for these reasons that K^+ ions are able to diffuse from water into the selectivity filter where the energetic cost of dehydration is compensated. Sodium ions, on the other hand, do not seem to enter the selectivity filter in crystal structures even when Na^+ is present in vast excess. In fact, reduction of K^+ to 3 mM in the presence of 150 mM Na^+ causes the selectivity filter to collapse.

High selectivity and high conduction rates present an apparent paradox: why does the precise ion coordination required for high selectivity not involve tight binding of the ions and thereby prevent their rapid diffusion through the pore? Two mechanisms may reconcile high conduction rates with high selectivity. First, the selectivity filter hosts more than one ion at a time; the repulsion between closely spaced ions helps to overcome the intrinsic affinity that each ion has for its binding site. On average, two K^+ ions occupy two of the four positions of the selectivity filter. The two preferred configurations are shown in Fig.16. In each configuration, 1-3 and 2-4, two K^+ ions are separated by one water molecule; the crystal structure represents an average of these configurations.

The 1-3 and 2-4 configurations form the final stage of a simple conduction cycle (Fig.16). The cycle corresponds to an ion crossing the filter: the ion enters the queue from one side of the filter, while a different ion exits from the opposite side. The direction of flow is determined by the electrochemical gradient of K^+. This picture, based on crystallographic studies, is in good agreement with the experiments of Hodgkin and Keynes, which led them to propose in 1955 that the K^+ conduction mechanism involves the single-file movement of two or three ions across the membrane.
The second reason for high conduction rates in the face of high selectivity is that the selectivity filter's structure depends on the presence of K^+ ions. Specifically, when the K^+ concentration is lowered well below normal intracellular levels, the filter's occupancy in the crystal drops to one K^+ ion and a specific conformational change occurs. The conducting conformation of the filter shown in Fig.16 requires the presence of two K^+ ions, and it is the entry of the second K^+ ion that induces the conformational change [18]. This observation has a simple thermodynamic consequence: some fraction of the ion-binding energy is used to change the filter's structure, and as a result the ions bind less tightly than if a conformational change did not occur. Weak binding is a prerequisite for high conduction rates. In summary, K^+ channels conduct K^+ ions specifically because the selectivity filter contains multiple binding sites that mimic the hydration shell of a K^+ ion. Potassium channels achieve high conduction rates by exploiting electrostatic repulsion between closely spaced ions and by coupling the conformation of the selectivity filter to ion binding within the filter.

Gating
What are the conformational changes that underlie pore opening in K^+ channels? Structural comparison of KcsA with another K^+ channel, MthK, begins to answer this question. Whereas KcsA was crystallized under conditions that favour the closed conformation, MthK, a Ca^{2+}-gated K^+ channel, was crystallized in the presence of Ca^{2+} which opens the pore in membranes [22]. The pores of KcsA and MthK are shown in Fig.17. Aside from the obvious difference in the length of the extracellular `turret' loops connecting the outer and pore helices, the major structural difference between these two channels is the position of their inner helices on the intracellular side of the selectivity filter.

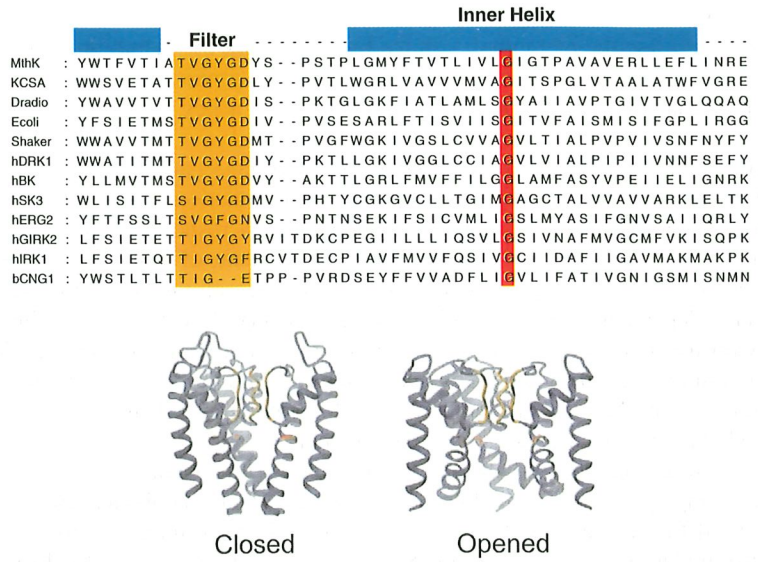


Fig. 17: *Closed and opened conformations of the pore. Sequences of ligand- and voltage-gated K^+ channels are from [Y. Jiang et al.. Nature **417** (2002), pp. 523–526]. The selectivity filter sequence is highlighted orange and the glycine gating hinge red. Three subunits of the closed pore conformation of KcsA (left) and the opened pore conformation of MthK (right) are shown with the selectivity filter and gating hinge colored to match the sequence highlights (adapted from [11]).*

In KcsA, the inner helices are straight and form a helix bundle near the intracellular membrane surface. At the bundle crossing in KcsA, the pore constricts to about 3.5 Å in diameter and is lined with hydrophobic amino acids, creating a barrier for the flow of K^+ ions. In contrast, the MthK inner helices are bent in a hinge region and splayed open so that the central cavity becomes connected with the cytoplasm, allowing free ion flow between the cytoplasm and the selectivity filter. From measurements of gating activity in membranes, the conditions under which the KcsA and MthK channels were crystallized are expected to yield closed and opened K^+ channels, respectively. The crystal structures are in good agreement with these expectations.

The KcsA and MthK structures are probably representative of the closed and open pore structures of many different K^+ channels, irrespective of the mechanism by which the pore opens. This conclusion is suggested by the conservation of a glycine residue at the inner helix hinge region of most K^+ channels (Fig.17, red). A glycine gating hinge would allow the inner helices to switch between a closed KcsA-like conformation and an open MthK-like conformation. The conclusion is also supported by pharmacological and mutational studies of voltage-gated and G-protein-gated K^+ channels, which are consistent with the closed and opened pore conformations shown in Fig.17. In summary, large conformational changes within the membrane underlie pore opening in K^+ channels. Inner helices obstruct the pore in the closed state and expand its intracellular diameter in the open state. Different gating mechanisms – ligand binding and voltage sensing – appear to bring about a similar conformational change in the pore of K^+ channels.

Voltage Sensing

Ion channels determine the membrane voltage by controlling the membrane's permeability to ions. For voltage-gated channels the membrane voltage determines the open probability, and therefore these channels provide a means for feedback – a key requirement for generating electrical pulses.

The conformational rearrangements that occur during voltage gating must involve the redistribution of charged residues. It was shown by numerous mutational studies that a highly conserved domain in a putative transmembrane segment was responsible for voltage sensing. This domain is located in the fourth transmembrane segment of voltage-gated channel subunits, also called the S4 segment. Fig.18 shows a sequence alignment of the S4 segment of various voltage-dependent channels. Quite strikingly, every third position in this segment is occupied by positively charged amino acid residues (R or K). Between these residues, mostly hydrophobic amino acids are found. This structure was thought to constitute an amphipathic helix that moves from the inside to the outside of the membrane when the membrane voltage is depolarized, thereby opening the pore of the channel.

Thus, these "gating charges" couple electrical work to the opening process. The "gating charge" of a voltage-gated K^+ channel has been measured and is equivalent to almost 14 electrons moving all the way across the transmembrane voltage difference. Most of this gating charge is attributable to four of the arginine residues per subunit, i.e. a total of 16 positive charges in a tetrameric channel. Each arginine accounts for approximately one electron charge. The results of many mutational studies in combination with modern spectroscopic techniques led to the view that the S4 segment is located in a "gating canal" with most of the gating charges accessible from the intracellular side, and that the S4 helix rotates around its length axis and, thereby, transfers all the charges to the extracellular domain.

The crystal structure of the voltage-gated K^+ channel KvAP has changed our view of the voltage sensor [19, 20]. The structure contains the typical pore of K^+ channels and α-helical voltage sensors (Fig.19). In the crystal, the voltage sensors are fixed by monoclonal Fab fragments, which in the absence of a membrane environment probably forces the sensors into a non-native conformation (Fig.19). Nevertheless, the structure offers a simple possibility for how the voltage sensors might transfer gating charges through the membrane electric field: the important arginine residues are located in hydrophobic helix-turn-helix structures – or voltage sensor paddles – that appear as if they could move within the membrane at the protein–lipid interface when the channel opens.



Fig. 18: *Alignment of the S4 segment of various voltage-gated channels. Positively charged residues (R and K, blue background) are regularly spaced at every third position interspersed by largely hydrophobic (yellow background) amino acids.*

The movement of voltage-sensor paddles was tested in lipid membranes by tethering biotin to various positions and assessing whether avidin can access the biotin from either side of the membrane. Biotin, tethered to certain positions on one side of the paddle is actually dragged all the way across the membrane from the intracellular side in the closed conformation to the extracellular side in the open conformation (Fig.19). That is, biotin is accessible to avidin

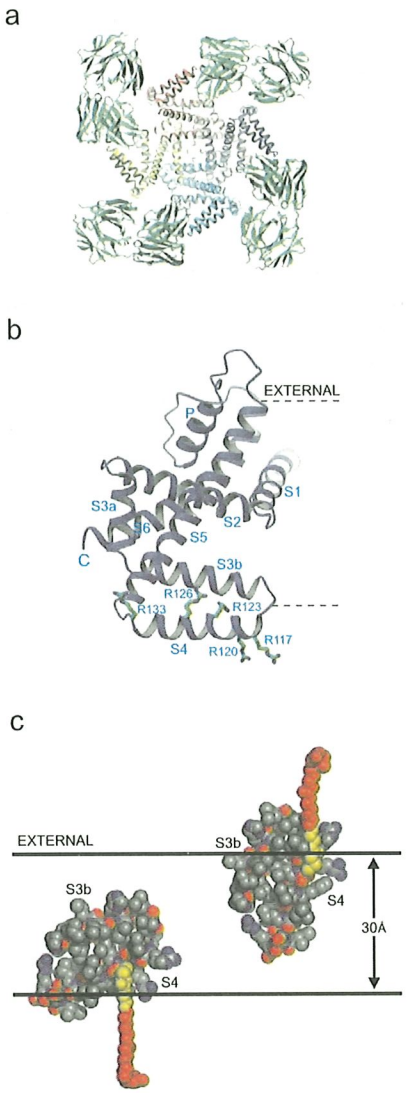


Fig. 19: *The voltage-gated K^+ channel KvAP. a: A tetramer from the crystal viewed down the four-fold axis from the cytoplasm. The channel is α–helical with four subunits shown in different colors. Monoclonal Fab fragments (green) used to crystallize KvAP are attached to the voltage sensors. b: A single subunit of KvAP is shown with the extracellular side on top. Helical elements S3b and S4 form a hydrophobic `voltage sensor paddle' with gating charge arginine residues. In the crystal, the Fab is attached to the voltage sensor paddle and has extended it toward the cytoplasm. c: CPK model of the voltage sensor paddle shown in approximate closed (left) and opened (right) depths relative to the 30 Å hydrophobic core. Tethered biotin (red and yellow) attached to position 121 is accessible to internal avidin when the channel is closed and to external avidin when the channel is opened. Avidin engulfs the red segment when it binds to biotin (adapted from [11]).*

from the intracellular side when the channel is closed and from the extracellular side when it is opened. Given the length of the tether to which biotin is attached, and constraints for avidin binding, this finding requires that the paddles move a large distance. Biotin, attached to other positions, support the same conclusion: motions large enough to transfer the four arginine gating charges most of the way across the membrane.

6 Conclusion

Electrical activity of cells is a fascinating phenomenon that has been studied by physiologists for more than 50 years. The cloning of the genes that code for transport proteins opened the opportunity to study them in isolation in heterologous systems. The first high-resolution structures of channels and pumps have recently appeared. These snapshots facilitate both the design and the interpretation of studies on protein function. The next challenge will be to illuminate the dynamic processes that underlie the inner workings of these fascinating miniature machines. The power of the multi-disciplinary approach, combining experimental manipulations, high-resolution structures, and computational studies, will provide answers to these exciting questions.

References

[1]*Principles of Neural Sciences.* Eds: E.R. Kandel, J.H. Schwartz, T.M. Jessel, 4th Edition, McGraw-Hill Companies 2000.

[2]*Molecular Neurobiology.* Z.W. Hall, Sinauer Associates Inc. 1992.

[3]*Neuroscience.* Eds.: D. Purves, G.J. Augustine, D. Fitzpatrick, L.C. Katz, A.-S. LaMantia, J.O. McNamara, Sinauer Associates Inc. 1997.

[4]*Physics.* G.B. Benedek, F.M.H. Villars, Vol. 3: Electricity and Magnetism. Addison-Wesley Publishing Company 1979.

[5]*Ionic Channels of Excitable Membranes.* B. Hille, 3rd Edition, Sinauer Associates Inc. 2001.

[6]*Biophysics of Computation.* C. Koch, Oxford University Press, Oxford, 1999.

[7]*Ion Channels and Disease.* F.M. Ashcroft, Academic Press, San Diego, 2000.

[8]U.B. Kaupp & A. Baumann. *The Molecular Basis of Electrical Excitability of Nerve Cells.* Nanoelectronics and Information technology, Wiley-VCH, 2003.

[9]A.G. Lee. *A calcium pump made visible.* Curr Opin Struct Biol. 2002 Aug;12(4):547-54.

[10]N. Unwin. *Structure and action of the nicotinic acetylcholine receptor explored by electron microscopy.* FEBS Letters 555 (2003)91 -95

[11]R. MacKinnon. *Potassium Channels.* FEBS Letters 555 (2003)62 -65

[12]U.B. Kaupp, R. Seifert, *Cyclic nucleotide-gated ion channels*, Physiol. Reviews - (2002) in press.

[13]R.S. Molday, U.B. Kaupp, *Ion channels of vertebrate photoreceptors.* Handbook of Biological Physics, Vol. 3: Vision (Part I) eds: D.G. Stavenga, W.J. de Grip, E.N. Pugh, Jr., Elsevier North-Holland, 2000.

[14]D.A. Doyle, J.M. Cabral, R.A. Pfuetzner, A. Kuo, J.M. Gulbis, S.L. Cohen, B.T. Chait, R. MacKinnon, *The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity.* Science, 280, 69-77 (1998).

[15]J.H. Morais-Cabral, Y. Zhou,R. MacKinnon, *Energetic optimization of ion conduction rate by the K^+ selectivity filter.* Nature, 414, 37-42 (2001).

[16]Y. Zhou, J.H. Morais-Cabral, A. Kaufman, R. MacKinnon, *Chemistry of ion coordination and hydration revealed by a K^+ channel-Fab complex at 2.0 Å resolution.* Nature, 414, 43-48 (2001).

[17]U.B. Kaupp, R. Seifert, *Molecular diversity of pacemaker ion channels.* Annu. Rev. Physiol. 63, 235-237, 2001.

[18]Y. Zhou, R. MacKinnon. *The occupancy of ions in the K+ selectivity filter: charge balance and coupling of ion binding to a protein conformational change underlie high conduction rates.* J Mol Biol. 2003 Nov 7;333(5):965-75.

[19]Jiang Y, Ruta V, Chen J, Lee A, MacKinnon R. *The principle of gating charge movement in a voltage-dependent K+ channel.* Nature. 2003 May 1;423(6935):42-8.

[20]Jiang Y, Lee A, Chen J, Ruta V, Cadene M, Chait BT, MacKinnon R. *X-ray structure of a voltage-dependent K+ channel.* Nature. 2003 May 1;423(6935):33-41.

[21]Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R. *The open pore conformation of potassium channels.* Nature. 2002 May 30;417(6888):523-6.

[22]Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R. *Crystal structure and mechanism of a calcium-gated potassium channel.* Nature. 2002 May 30;417(6888):515-22.

[23]Miyazawa A, Fujiyoshi Y, Unwin N. *Structure and gating mechanism of the acetylcholine receptor pore.* Nature. 2003 Jun 26;424(6943):949-55.

[24]Toyoshima C, Nakasako M, Nomura H, Ogawa H. *Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution.* Nature. 2000 Jun 8;405(6787):647-55.

[25]Toyoshima C, Nomura H, Sugita Y. *Crystal structures of Ca2+-ATPase in various physiological states*. Ann N Y Acad Sci. 2003 Apr;986:1-8.
[26]Brejc K, van Dijk WJ, Klaassen RV, Schuurmans M, van Der Oost J, Smit AB, Sixma TK. *Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors*. Nature. 2001 May 17;411(6835):269-76.
[27]Voet & Voet. Biochemistry. John Wiley & Sons, Inc., 1995.

D 8 Membrane receptors

Martin Heck[§] and Klaus Peter Hofmann[§$]
§ Institut für Medizinische Physik und Biophysik
Universitätsklinikum Charité Campus Mitte
$ Zentrum für Biophysik, Humboldt Universität, Berlin

Contents

1 Introduction 2

2 G-protein coupled signal transduction 3

3 Rhodopsin: visual pigment and G-protein coupled receptor 4

4 The dark ground state 7

5 Activation by light 8

5.1 Retinal isomerization and early intermediates 8
5.2 Rhodopsin can be activated without retinal isomerization 10
5.3 The meta states 10
5.4 Microstructural changes during rhodopsin activation 11
5.5 Probing chromophore-protein interaction in the active state 12

6 Catalytic interaction between rhodopsin and transducin 14

6.1 Loci of interaction 14
6.2 Time ordered sequence of interactions 14
6.3 Role of lipid 15

7 General considerations 16

1 Introduction

To perform their functions in organisms or to adapt to their environment, living cells react to external stimuli. These stimuli are realized in physical or chemical signals, which are in most cases detected by specialized membrane receptor proteins. Membrane receptors are membrane spanning proteins, which bind or absorb the signal and make the free energy of ligand binding available for the cell. Because membrane receptors connect to both sides of the plasma membrane or specialized membranes of an organellum, signal induced alterations of the receptor´s structure can carry the signal to the cell interior, thus generating intracellular responses. Intracellular proteins, so-called effectors (including channels or enzymes) mobilize then second messengers. These are intracellular substances that are regulated in their concentration, e.g. cyclic guanosine monophosphate (cGMP) or Ca^{2+} ions. Receptor and effector can form a functional unit, as for example in certain ion channels which are regulated by ligands. In this case, the extracellular signal acts directly on the activity of the effector. The second possibility is that the receptor, in a first step, activates intracellular transmitter proteins, which in turn activate an effector protein. Membrane-bound protein-protein interactions are key processes in this mode of signal transduction. The biomembrane reduces the dimensionality of the interactions and provides a barrier for the soluble components involved. Transduction of the signal through several steps opens possibilities of regulation and amplification which are not possible in a single step mechanism.

Guanine nucleotide binding regulatory proteins (G-proteins) are an important and ubiquitary class of transmitter proteins. The large family of the respective G-protein-coupled receptors (GPCRs) comprises ca. 1000 members or 3% of the human genome. GPCRs transmit their signal by coupling to heterotrimeric so-called large G-proteins, which consist of α, β and γ subunits. In the activated form, the cytoplasmic surface of a GPCR becomes competent for binding a G-protein, leading to subsequent catalytic nucleotide exchange on the α-subunit and G-protein activation. G-protein coupled systems are involved in cellular signal transduction processes from light, taste, and odor perceptions to the vast variety of processes of intraorganismal regulation via hormones and neurotransmitters. They are of high significance for basic research as well as for medical application, since more than 50% of the targets for pharmacological intervention are currently based on GPCRs. GPCRs share the common structural motif, predicted by their amino acid sequences, of seven transmembrane α-helical segments, and thus belong to an even larger family of proteins that serve a broad variety of functional properties, from ion translocation (e.g. bacteriorhodopsin, see the contribution by Georg Bueldt, in this issue) to signal transduction. In spite of the structural similarities between the GPCR receptor apoproteins, the ligands can be exceedingly diverse, including biogenic amines, amino acids, lipids, peptides and polypeptides, nucleosides and nucleotides. The effector proteins, to which G-proteins bind to transmit the signal, include enzymes (such as adenylate cyclases or phospholipases) and ion channels. The majority of ligands reaches GPCRs by diffusion and bind to a site near its extracellular surface. In photoreceptors, the physical light signal is received by a fixed chromophoric ligand.

An archetype of a photoreceptor and GPCR is rhodopsin, the visual pigment of the retinal rod cell. Biophysical methods have much contributed to the elucidation of rhodopsin's structure and function, and of the related signal cascade. With the vertebrate rod outer segment, nature has provided us with a model system for the study of ligand-protein and protein-protein interactions in signal transduction. It contains the relevant proteins in abundance, and is accessible to biophysical as well as biochemical techniques. The analysis is simpler than in most related systems, since the rod photoreceptive apparatus is basically specialized to just one purpose, the monochromatic detection of low light levels. This article will focus on

rhodopsin and discuss some of the more recent developments. For more details, the reader is referred to review articles from this [1],[2],[3] and other [4],[5] laboratories.

2 G-protein coupled signal transduction

Fig.1 shows the basic scheme of G-protein coupled signal transduction, for two examples (Fig.1A,B). In the case of photoreception, the receptor contains a chromophore as a fixed prostethic group, which becomes an activating agonist by absorption of a photon and subsequent photochemical reactions (Fig.1A shows the *cis/trans* isomerisation of retinal, see below).

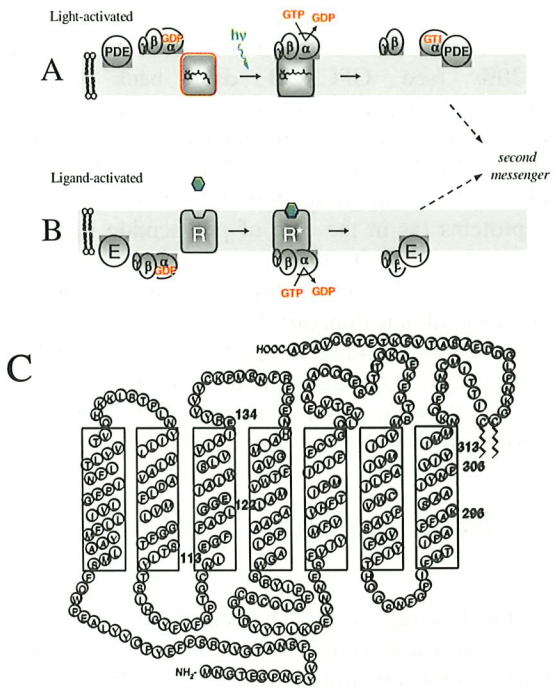


Fig. 1: *G-protein coupled signal transduction.*
(A) *Activation by light. The receptor contains a covalently bound chromophore as prostethic group, which becomes an agonist by absorption of a light quantum (green arrow) and subsequent photochemical reactions. Shown is the example of retinal cis/trans-isomerisation. The photochemically activated state is analogous to R* (cytoplasmic side upwards, following the convention for retinal proteins).* ***(B)*** *Activation by a diffusible ligand, e.g. a hormone. By binding an activating ligand (agonist, green, the receptor R is converted into an active state R*. In both cases, the active receptor binds the heterotrimeric G-Protein ($G\alpha\beta\gamma$) and catalyses the exchange of GTP for bound GDP. After dissociation from R*, the activated G protein subunits (α and the $\beta\gamma$ complex) can bind an effector (e.g. E_1), which mobilizes second messengers in the cell cytoplasm. In the case of the visual cascade, E is a phosphodiesterase (PDE) und the second messenger is cGMP.* ***(C)*** *Seven helix structure of rhodopsin, as first derived from amino acid analysis [33],[34].*

The case of an activating diffusible ligand (an agonist, e.g. certain hormones) is shown in Fig.1B (agonist is shown in *green*). The interaction with an agonist leads to changes in the tertiary structure of the receptor, which are recognized by the G-protein and also by other regulatory proteins, which include GPCR receptor kinases and arrestins and serve to deactivate the receptor. Common to both signaling pathways is the receptor catalyzed nucleotide exchange: bound GDP is replaced with GTP, which activates the G-protein. In many cases, activation of the G-protein leads to dissociation of the GTP-bound α-subunit from the $\beta\gamma$-complex which both can than bind to different effectors, thus mobilizing second messengers (e.g. cGMP and Ca^{2+}). In the case of vertebrate photoreception, there is only one effector, a cGMP-specific phosphodiesterase (PDE), which is disinhibited by the GTP-bound α-subunit.

The large family of GPCRs comprises five classes of proteins with a sequence homology in each class of ca. 20% (see GPCRDB data bank for more information, http://www.cmbi.kun.nl/7tm/). Rhodopsin is the archetype of the class I comprising 90% of the GPCRs. The optimization of the receptors for their respective ligand was reached in an evolution that started some 750 million years ago. There was a development of the proteins "around" existing ligands (such as amino acids or the retinal) or a co-evolution between ligands and the receptor proteins (as in the case of polypeptide ligands). Although GPCRs bind so diverse ligands, their global architecture appears to be quite uniform. Based on amino acid analysis, one assumes that almost all GPCRs span their membrane seven times in form of predominantly α-helical transmembrane domains (Fig.1C).
Many drugs, including the "classical" opium, heroin, LSD, hashish, and a large percentage of the therapeutic drugs bind to GPCRs, either activating or inhibiting. Thus understanding structure and function of GPCRs is not only a scientifically challenging task, it may also provide valuable insights for drug design and/or the search for further therapeutic targets.

3 Rhodopsin: visual pigment and G-protein coupled receptor

Rhodopsin is the visual pigment of the rod cell. This retinal cell shares with all sensory cells a bipartite outer/inner segment structure (Fig.2A) with a thin connecting cilium. The inner segment bears the nucleus, mitochondria, and the endoplasmic reticulum, while all components needed for transduction of the light signal are contained in the outer segment (ROS). Rhodopsin is contained as an integral membrane protein in the flat disc vesicles, which fill the ROS like a stack of coins. The cellular response, a hyperpolarization of the plasma membrane, is caused by the closure of cyclic GMP dependent ion channels. The second messenger function of cyclic GMP arises from its hydrolysis through the cyclic GMP phosphodiesterase (PDE) effector, which is anchored to the outer face of the disc membrane (Fig.2B) and activated by coupling to the G-protein transducin (G_t). The G-protein is catalytically converted into the active, GTP-bound form by interaction with light-activated rhodopsin (R).
At this point, we only mention that all the activation brought about by the light stimulus must be reset on a time scale of ca. 200ms, to allow repeated activation of the visual system. After a short photoexcitation of the dark-adapted rod, the membrane potential returns to the dark level within 1-2 seconds. The necessary restoration of the cyclic GMP dark level is satisfyingly explained by a feedback loop that activates a guanylate cyclase via calcium-binding protein(s). A complete reset of the transduction system also implies the deactivation of the cascade on all its stages, namely, receptor, G-protein and effector. G and PDE activities

are thought to be deactivated simultaneously by the intrinsic GTPase activity of the G-protein. Shut off of rhodopsin's activity toward the G-protein involves phosphorylation by rhodopsin kinase and subsequent binding of arrestin. Slow regeneration of rhodopsin with fresh, metabolically supplied 11-*cis*-retinal closes the rhodopsin cycle (see Fig.4 and Refs. [2],[6]).

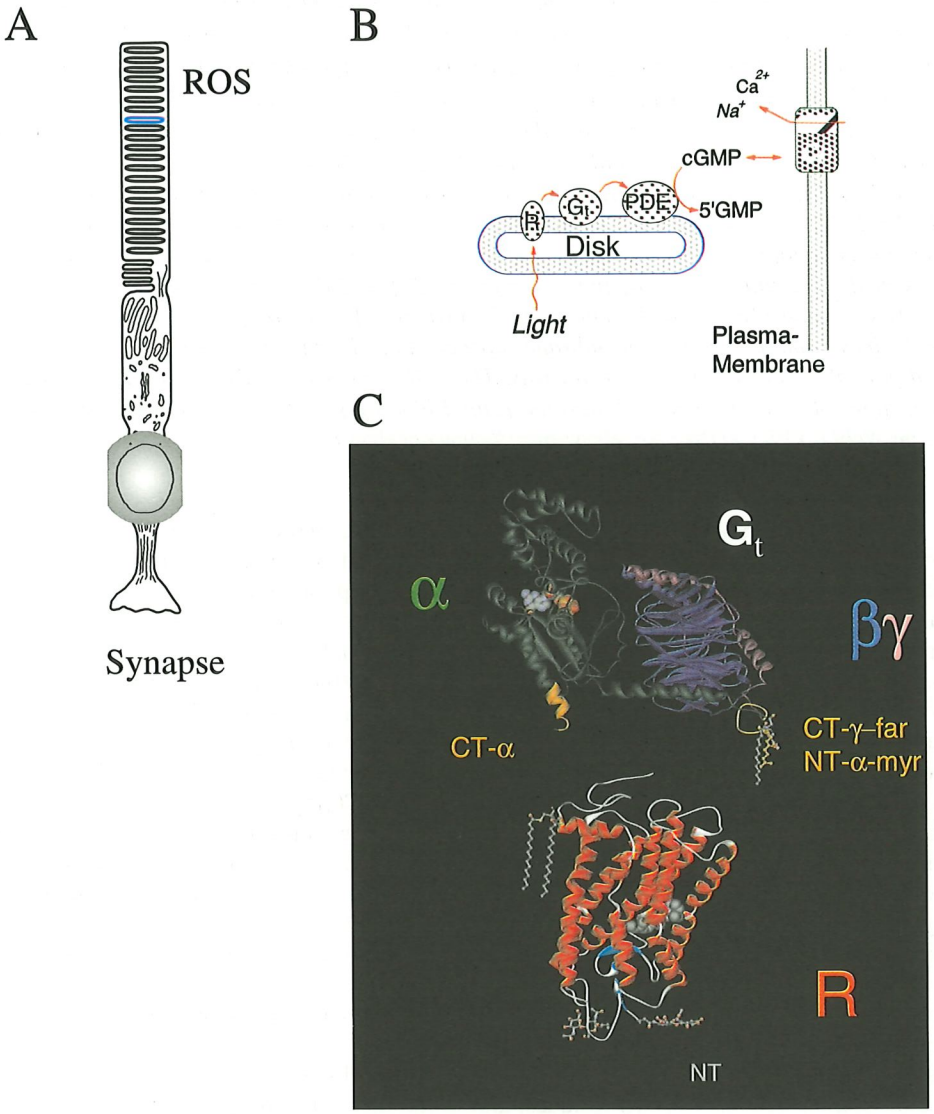


Fig. 2: *Rhodopsin and the visual cascade.*
*(**A**) Schematic representation of a rod cell. The rod consists (from below) of the synaptic body, the nucleus, the inner segment (with the metabolic machinery) and, connected by a cilium, the outer segment (ROS), which contains a stack of discs. (**B**) The key components of the visual cascade rhodopsin (R), transducin (G_t) and cGMP phosphodiesterase (PDE) are located on the disc membrane. Active PDE hydrolyzes cGMP to GMP which in turn leads to the closure of cGMP dependent ion channels in the plasma membrane of the rod cell. (**C**) Tertiary structures of transducin (G_t) and rhodopsin (R). Top: G_t - heterotrimer, $G_t\alpha\beta\gamma$(PDB-ID: 1GOT). The G_t subunits are shown in green (α), blue (β) and magenta (γ). The regions not resolved in the X-ray structure, namely the C-termini (orange) of $G_t\alpha$ (CT-α, amino acids 340-350) and $G_t\gamma$ (CT-γ, 50-71) and the membrane anchors CT-γ-far and α-myr were modeled. CT-α is shown in the putative receptor bound conformation (PDB-ID: 1AQG). GDP in the nucleotide binding site of $G_t\alpha$ is shown as a space filling model. Bottom: Rhodopsin (R) (PDB-ID: 1F88) with bound 11-cis-retinal (grey). See also Fig.3. The interface to G_t, with four cytoplasmic loops, points towards G_t. Unresolved regions were modeled, including some regions in the cytoplasmic loops, the C-terminus (CT) and the palmitoyl ester cystein side chains (Cys^{322} and Cys^{323}) at the end of helix VIII, which runs perpendicular to the other seven helices (and parallel to membrane surface, not shown). The extracellular densely packed part of rhodopsin points downwards. The E-II loop (see Fig.3) forms an antiparallel β-sheet (light blue) and folds back into the retinal binding site. Drawing and modeling with program WebLab ViewerPro; Accelrys; http://www.accelrys.com.*

As a visual pigment, rhodopsin consists of the apoprotein (called opsin) and a light sensitive chromophore as the prostethic group. Retinal is the aldehyde of vitamin A and is present in the dark ground state in its 11-*cis*-conformation. After absorption of a photon, it isomerizes to the all-*trans*-conformation, thus starting the activation process (Fig.2C and below). Cryo electron microscopy on rhodopsin was for a long time the only source of information on the tertiary structure of a GPCR, yielding correct information on the arrangement of the helices in the membrane [7]. Since 2000, several X-ray structures have been published [1],[8]. Fig.2C illustrates how the retinal (grey) fits in the 348 amino acids long opsin apoprotein. The extracellular (intradiscal) and intracellular regions of rhodopsin (Figs.2C and Fig.3) each consist of three interhelical loops (E(xtracellular) and C(ytoplasmic), respectively), and terminal (COOH- or NH_2-) tail regions. A fourth cytoplasmic loop is formed by anchoring the C-terminal tail to the membrane via two Cys residues (Fig.2C), which carry palmitates in the native structure. It contains the short helix H-VIII, adjacent to H-VII, which runs parallel to the cytoplasmic surface. Although the mass distribution to the extracellular and cytoplasmic regions is comparable, they are very different in their packing: the four extracellular loops associate tightly with each other, while only a few interactions are observed among the cytoplasmic domains. The E-II loop is connected to H-III via a disulfide bridge and fits into a limited space inside the bundle of helices (Fig.3).
The transmembrane helices of rhodopsin vary in length, and there are also differences in the degree of bending around Gly/Pro residues, and in the tilt angles to the expected membrane surface. In places, there are also strong distortions of the α-helices, for example around Pro^{267} in H-VI. H-VII shows a considerable elongation in the region around the retinal attachment site Lys^{296}. It contains Pro^{291} and Pro^{303}, the latter is part of the highly conserved Asn/Pro/X/X/Tyr (NPxxY) motif.

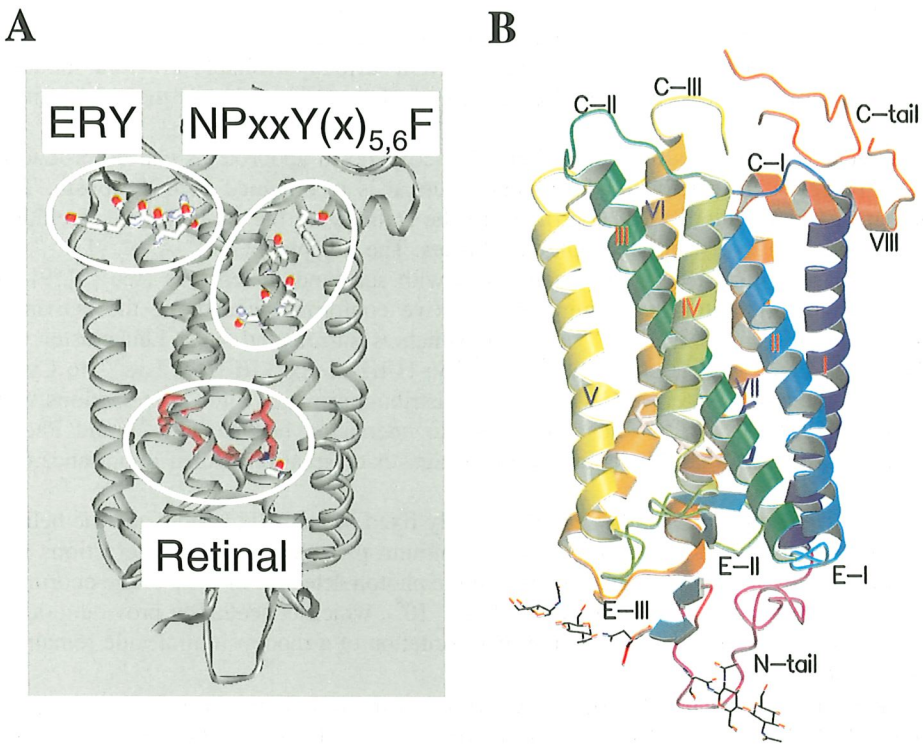


Fig. 3: *Functionally important modules and topology of rhodopsin.*
(A) *The retinal binding site and the two cytoplasmic domains ERY and NPxxY(x)$_{5,6}$F region are functionally linked in the activation process of rodopsin. See text for details.* ***(B)*** *Topography of rhodopsin; the molecule is shown in a similar view as in Figs.2C and 3A.*

4 The dark ground state

Retinal has a conjugated π-electron system forming an efficient photon absorber. By interaction with the apoprotein, its absorption in the near UV range is shifted into the green range of the UV/Vis spectrum. The covalent retinylidene linkage between chromophore and apoprotein is provided by a Schiff base, which arises from a condensation reaction between aldehyde of the retinals and the ε-amino group of Lys^{296}. In the dark ground state, the nitrogen of the Schiff base on 11-*cis*-retinal is protonated, leading to an enhanced π-electron density and a bathochromic shift of the absorption maximum from 380nm to 440nm. An additional so-called opsin shift results from interactions in the chromophore binding pocket. It generates the 500nm absorption maximum of the rhodopsin in bovine rods, which is the focus of this article. Differences in this shift can result from different amino acids that interact with the chromophore; they explain the different absorption maxima in the pigments in the different cones, which are tuned to 430, 530 and 560nm, i.e. "blue", "green" and "red", respectively.
The protonation of the Schiff base in the protein environment is stabilized by a counterion, provided by the negatively charged carboxylic group of Glu^{113}. This interaction is part of a network of hydrogen bonds, the dominant interaction being a salt bridge between the

carboxylate anion and the protonated Schiff base. It is neutralized in mutants of rhodopsin, which are constitutively active in the dark. Although it is not clear whether this interaction is direct or *via* water molecules (which are not resolved in currently available structural models), a direct interaction is at least possible due to the extended conformation of the side chain of the Lys^{296}.

By site-directed mutagenesis other constitutively active opsin apoproteins can be produced, which are however all deactivated when 11-*cis*-retinal is recombined with the opsin. This shows that numerous interactions contribute to the stability of the ground state, including hydrogen bonded networks and hydrophobic clusters. The tripeptide Glu^{134}-Arg^{135}-Tyr^{136}, for example, participates in several hydrogen bonds with surrounding residues (see [5],[9] for details). It is important for maintaining the inactive conformation and for the activation process. Another example is Met^{257} in H-VI, which is surrounded by a Leu cluster that includes Leu^{76}, Leu^{79} in H-II, and Leu^{128}, Leu^{131} in H-III. Helix VIII from Lys^{311} to Cys^{322} exhibits an amphiphilic pattern of its side chain distribution, which allows interactions with H-I (Leu^{57}) and H-VII (Ile^{307}) and a hydrophobic interaction between Tyr^{306} and Phe^{313}. Recent work has shown that this latter pair of residues is of great functional importance (see below).

As a result of these interactions, the retinal is tightly fixed in the helix bundle and the helices are fixed in their position. The stable energy minimum provided by these interactions is a precondition of the function of rhodopsin as a single photon detector. In spite of the enormous number of rhodopsin molecules in the rod cell (ca. 10^8), which is needed to provide a dense target for light absorption, spontaneous thermal excitation of a rhodopsin molecule remains a very rare event.

Fig.2C shows in its upper part the $\alpha\beta\gamma$-heterotrimer of the rod G protein, transducin (G_t), in its inactive, GDP-bound state. Not only for the receptor but also for the G protein, one finds remarkable stability against spontaneous activation. In the presence of the cofactor GTP but absence of active receptor, G_t remains stable for hours and does not release its GDP. Fig.2C illustrates how the retinal in rhodopsin (R) and the cofactor in G_t are arranged.

5 Activation by light

5.1 Retinal isomerization and early intermediates

Absorption of a photon provides rhodopsin with the energy for forming the active state. As shown in Fig.4, we may distinguish three phases of the activation process: i) light induced *cis/trans* isomerization of the retinal, ii) thermal relaxation of the retinal-protein complex, and iii) the late meta states, which remain in thermodynamic equilibrium and play a role in the interactions with the G protein. Deactivation and regeneration are much slower, and it takes minutes to complete the rhodopsin cycle. Two thirds of the energy of 238kJ/mol (57kcal/mol) taken up by light absorption are stored in the photoisomerized chromophore-(all-*trans*-retinylidene)-opsin complex. The first stable high-energy intermediate thus formed, is called bathorhodopsin (λ_{max} 543nm). It can be stabilized at low temperature (–140°C), and Raman and FTIR difference spectroscopy have revealed that the chromophore is in a twisted 11-*trans*, 15-anti conformation. The reason for this distorted all-*trans* conformation might be a steric restriction due to an approach of the polyene chain of the retinal and the side chain of Ser^{186} after isomerization around the C_{11}-C_{12} double bond, which is consistent with the tight fit of the retinal in its pocket discussed above. The chromophore movement causes slight alterations in the backbone structure as indicated by small infrared difference bands in the

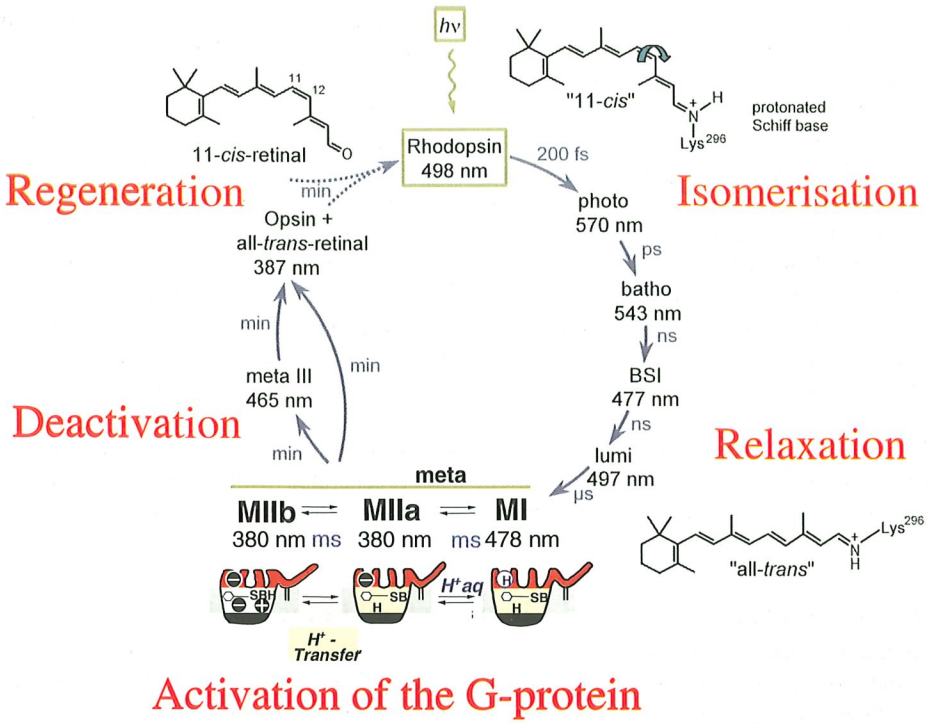


Fig. 4: *Functional cycle of rhodopsin*
Photoisomerization of the 11-cis double bond of the chromphore retinal occurs on a 200 femtoseconds time scale and yields the photoproduct photorhodopsin with a highly distorted 11-trans bond. The photolyzed pigment thermally relaxes through bathorhodopsin, blue shifted intermediate (BSI) and lumirhodopsin, each distinguished by its distinct λ_{max} of absorption, into the metarhodopsin intermediates. Identified meta species include meta I, meta IIa and meta IIb (M I, M IIa and M IIb, respectively). Up to meta I the retinylidene linkage (Schiff base) remains persistently protonated. Deprotonation of the Schiff base occurs only with the conversion of M I to M IIa and is reflected in a large blue-shift of λ_{max} to 380nm. The transition from M IIa into the M IIb state is spectrally silent and linked to proton uptake from the cytosol to Glu^{134} (Fig.1A). The M IIb species represents the active state of the rhodopsin receptor, capable to interact with G_t. From the meta equilibrium, rhodopsin decays either directly or via a species with a reprotonated retinal Schiff base, termed meta III (λ_{max} = 465 nm), into the apoprotein opsin and all-trans-retinal. The photolytic cycle is interrupted at this point due to irreversible hydrolysis of the Schiff base linkage. The ground state can only be regenerated with fresh 11-cis-retinal, provided by an extended retinal metabolism. Unlike invertebrate rhodopsin, the rhodopsin in rods cannot be regenerated in situ by reisomerization of retinal with a second photon. For details see text.

amide-I region (see [10] for details). FTIR difference bands evoked by protonated carboxylic acid group of the amino acid side chains of membrane-embedded Asp^{83} in H-II and Glu^{122} in H-III are weak, suggesting that the environment of these residues changes very little.
Thermal relaxation leads via a blue-shifted intermediate (BSI, Fig.4) that does not accumulate at low temperatures and can only be observed in time-resolved measurements, to lumirhodopsin. In lumirhodopsin, which can be stabilized at –40C, most of the twist in the polyene chain has relaxed as seen by low intensities of the hydrogen-out-of-plane (HOOP) vibrations of the chromophore. Retinal distortion could be released by relocation of the β-ionone ring of retinal. Photoaffinity labeling studies with a retinal analog with two photoactive moieties have indeed argued for this mechanism [11]. It was concluded that during the conversion from bathorhodopsin to lumirhodopsin the β-ionone moiety of the chromophore flips-over from the close-by Trp^{265} in H-VI into vicinity of Ala^{169} in H-IV. Such a mechanism, however, is not supported by recent NMR spectroscopic and structural data [12].
The β-ionone ring appears to remain in its ground state position, although it plays an important role in the activation process. Rhodopsins regenerated with retinal with a broken ring are impaired in the formation of the signaling state, metarhodopsin II (M IIa and M IIb in Fig.4).

5.2 Rhodopsin can be activated without retinal isomerization

Locked retinal analogs have a bridge between C_{10} and C_{13} which prevents isomerization around the C_{11}-C_{12} bond. Rhodopsins reconstituted with these analogs are stabilized in their inactive (or minimally active) conformation even under light conditions. Photoisomerization of these retinals, even if reconstituted with opsin, can still occur albeit not around the C_{11}-C_{12} double bond. Illumination of rhodopsin regenerated with an analog containing a 2 carbon bridge between C_{10} and C_{13} (6-membered ring) allows extraction of several isomers (9-*cis*- and/or 13-*cis*-forms of the ring-constrained 11-*cis*-retinal analog). Two of the four isomers were classified as 11-*cis*-like and the other two as all-*trans*-like by their affinities for 11-*cis*- and all-*trans*-retinol dehydrogenase, respectively [13].
At neutral pH, this photoisomerization does not result in significant activation of transducin or chromophore-induced activating structural changes in the opsin-moiety, as investigated by FTIR spectroscopy. However, later studies showed the surprising result that at appropriate pH the receptor is activated. Thus the central double bond isomerization is not necessary for the activation of the receptor. The possible movements of the locked chromophore are sufficient to overcome the trigger barrier to activation. Moreover, the presence of 11 - 12 locked *cis*-retinal in the binding site is compatible with a (pH-dependent) active conformation [14],[15].

5.3 The meta states

After having passed the earlier intermediates, an equilibrium between the meta I and meta II (M I and M II in Fig.4) states of rhodopsin is reached. M II can be stabilized at the expense of its predecessor M I by either low pH or high temperature.
M I is the first intermediate where larger changes in secondary structure of the protein occur as concluded from the intensity gain of the FTIR amide-I bands which are sensitive to structural changes. Spectroscopic evidence has also suggested that in M I, the distortion of the chromophore is still similar to Lumi and the Schiff base proton is weakly hydrogen-bonded. Transition to the signaling state meta II, which binds and activates the G protein is accompanied by a large shift of the absorption maximum from 480nm in M I to 380nm in M II. This is due to deprotonation of the Schiff base (Fig.4) and coupled to protonation of the counterion at Glu^{113}, suggesting a proton translocation in the hydrophobic core of the protein.

It is not clear yet, whether this occurs by direct proton transfer or *via* water molecules. FTIR difference spectra indicate that by far the largest light-induced alterations of the apoprotein structure take place upon M II formation.
Schiff base deprotonation and protonation of the counterion open the stabilizing salt bridge. This has been shown to be a necessary condition for forming the signaling state for the G-protein (see Ref. [9]). A second necessary step is proton uptake to Glu^{134}, the above mentioned highly conserved residue in the ERY motif. This second step can only occur when the salt bridge in the protein interior is broken first [2]. It can be assumed that the hydrogen bonded network linked to the ERY motif is broken in concert with proton uptake. Although the underlying mechanism is not understood, one may generally assume that proton transfer will only occur when intramolecular factors such as the relative orientation and distance between donor and acceptor, and the strength and orientation of the electrical field at the locus of the reaction are correctly adjusted. This may explain why it takes milliseconds after photoexcitation – much longer than the picosecond reaction time of the actual proton transfer reaction – until the proton translocations linked to M II formation occur.
Oliver Ernst and coworkers have developed a model which explains the role of the retinal in these processes [16]. It claims that the retinal itself helps to provide the scaffold for adjusting the proton donor and acceptor groups for proton transfer to one another. This model has provided an explanation for the experimental fact that a rhodopsin harboring the 9-desmethyl analog of retinal shows a specific defect in forming the M II intermediate [16],[17].
From these data, an analogy between the activation scheme of rhodopsin and of ligand activated GPCRs (Fig.1) can be deduced. It emphasizes the similarity between the M I and M II states and the low and high affinity binding conformations (R´ and R*, respectively) of ligand (L) -activated GPCRs, respectively:

Rhodopsin	***light-induced pathway***	$\rightarrow$	**M I + G**	$\rightleftharpoons$	**M II_a + G**	$\rightleftharpoons$	**M II_b•G**
GPCRs	**L + R + G**	$\rightleftharpoons$	**LR' + G**	$\rightleftharpoons$	**LR* + G**	$\rightleftharpoons$	***LR*•G***

5.4 Microstructural changes during rhodopsin activation

To explore the structural alterations which accompany the formation of rhodopsin´s signaling state, Khorana, Hubbell and coworkers have employed protein engineering to replace amino acids on rhodopsin´s cytoplasmic face with cysteins, which made it possible to label these locations for detection with electron paramagnetic resonance techniques.
Mobility and proximity relationships between the respective residues could thus be determined [18]. The structural changes detected occur mainly near Helix VI. Kinetics, pH dependence and the effect of mutants that break the salt bridge, argue that the structural changes occur during M II formation. It was postulated that Helix VI rotates as a solid body around its axis and tilts out of the helix bundle, with the hinge likely located at a Gly-Gly tandem in H-VI [19].
That helix motion is indeed a necessary condition for receptor activity could be confirmed by experiments, which showed that rhodopsin was reversibly inactivated in mutants with reversible chemical locking of helices [20].

5.5 Probing chromophore-protein interaction in the active state

A strategy to probe the active state was developed from the idea to photolyze meta II with blue light ($\lambda < 420$nm) and to observe the changes that are thereby induced. The same techniques as were used for the forward pathway (starting from the ground state) could be applied. In such an approach, one measures the constraints that come into play when one *leaves* the active state, and compares it to the changes when it is *formed.* It was generally assumed that blue light photolysis can revert the forward pathway (as it does in bacteriorhodopsin or halorhodopsin). At neutral pH, it was indeed measured [21] that ca. 50% of the photolyzed meta II goes through a shift of the absorption maximum indicating reprotonation of the retinal Schiff base and proton release, i.e. just the inverse of the proton transfer reactions observed in the forward pathway. However, the application of FTIR difference spectroscopy and retinal extraction led to a different, rather surprising result. It turned out that photolysis of meta II does not lead back to the ground state.

By FTIR vibrational spectroscopy, which measures absorption in the infrared range, information about changes of carboxyl groups, the protein backbone and chromophore of rhodopsin can be obtained [3],[22],[23]. Fig.5 shows the FTIR difference spectra of the rhodopsin $\rightarrow$ meta II conversion (*green line*; photolysis of rhodopsin with green light) and the meta II $\rightarrow$ P-product conversion (*blue line*; P arising form photolysis of meta II with blue light). When comparing the two spectra it is obvious that spectral features in a certain range are reversed, so that the two difference spectra are perfect mirror images of one another within this range. The reversible bands reflect structural alterations in the protein and comprise changes in hydrogen bonding and protonation of carboxyl groups (e.g. Glu^{113} at 1712cm^{-1}) and in the peptide backbone (amide-I and amide-II bands at 1700-1620 and 1570-1500cm^{-1}, respectively). However, no such reversal is seen for the bands in retinal "fingerprint" region which arise from changes in retinal geometry and retinal-protein interaction, reflected in C-C stretching vibrations (1238cm^{-1} band and its satellites), C=C stretching vibrations (1550-1570cm^{-1}) and hydrogen out of plane vibrations (HOOP region; 960/970cm^{-1}). Specific for meta II photolysis is a new band at 1349cm^{-1}. This indicates that the geometry of the retinal and its interactions with the chromophore in the P-product are different from the ground state [24]. It was concluded that meta II photolysis leads to deactivation without back-isomerization to the 11-cis conformation. The most recent data indicate that the product P is in all its measurable properties identical to meta III, the product that is formed when meta II decays by thermal activation (Ritter, E., and Bartl. F.J., unpublished). It has recently been shown [25] that meta III harbors the retinal in a 15-*syn* retinal configuration (in which the Schiff base bond and not the central C_{11}-C_{12} double bond is isomerized). Taken together, the results show that the chromophore constraints in meta II are such that changes in the geometry of the central part of the polyene chain are blocked. This is reminiscent of the locked analogs of retinal, where central isomerisation is prohibited by synthetic hydrocarbon ring structures.

As in these retinals [14],[15] the protein finds ways to change its state of activation via a "second switch" [24] although *cis/trans* isomerisation around the double bond is prohibited. Physiologically, the result underscores the uniqueness of the rhodopsin ground state, which cannot be regenerated by light but needs an extended metabolic network to provide opsin with its 11-*cis*-retinal chromophore [6].

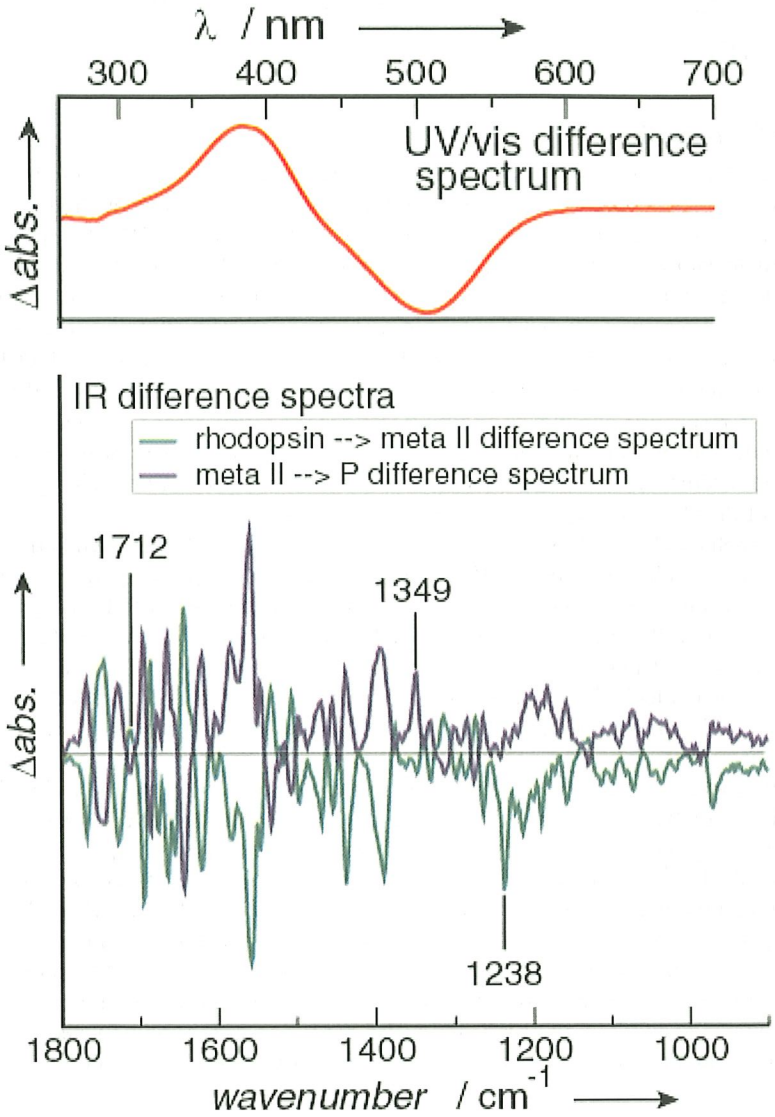


Fig. 5: *Probing the active state by photolysis of meta II*
The characteristic absorption spectrum of the rhodopsin ground state (λ_{max} = 500nm) originates from the interaction between the chromophoric ligand 11-cis-retinal (λ_{max} = 380nm) and its binding site in the apoprotein opsin. The UV/Vis difference spectrum (meta II with λ_{max} = 380nm minus rhodopsin) is obtained by subtracting the rhodopsin spectrum from pure meta II) light spectrum. Analogously a FTIR difference spectrum (meta II – rhodopsin, green) is obtained. Negative bands are due to vibrations of rhodopsin and positive bands are due to vibrations in meta II. Blue shows the difference spectrum of the products (P) formed by meta II photolysis minus meta II. Negative bands correspond to meta II and positive bands to the P-products. Between the two difference spectra, protein bands (ca. 1800 – 1400cm^{-1}) are reversed but not retinal fingerprint bands (e.g. 1238cm^{-1}). See text for details.

6 Catalytic interaction between rhodopsin and transducin

6.1 Loci of interaction

The formation of the *„interface"* between activated rhodopsin and the G protein evidently occurs via the cytoplasmic domain of rhodopsin. By making use of retinal as rhodopsin's own intrinsic reporter group for the transition to the active conformation, UV/Vis spectroscopy could be used to show that bound G-protein stabilizes meta II at the expense of inactive meta I, producing "extra meta II". In this regard the G_t holoprotein can be replaced by synthetic peptides from the $G_t\alpha$ and $G_t\gamma$ C-terminal sequences (Fig.2C), suggesting that the respective structures on transducin are interaction sites for rhodopsin. The interactions of activated rhodopsins with G_t and $G_t\alpha$-and $G_t\gamma$-derived peptides were also detected by flash-photolysis of the respective complexes and by FTIR spectroscopy. Both $G_t\alpha$- and $G_t\gamma$-derived peptides force the receptor into the same meta II state as is spontaneously formed at acidic pH [23]. Therefore, according to the spectroscopic criteria, the conformational changes induced upon binding of the $G_t\alpha$- and $G_t\gamma$-derived peptides are similar.
A similar approach can be used to explore the active surface of rhodopsin. Synthetic peptides derived form the cytoplasmic loops C-II, C-III und C-IV were found to compete with the active rhodopsin itself against binding to the G-protein and mutations of C-II and C-III affect the interaction in a well-defined manner (see Ref. [2]). Mutant rhodopsins with defects in either loop C-II or C-III interacted with G_t but failed to activate G_t. Charge reversal in the ERY motif however, blocked any interaction between the proteins, as concluded from the lack of extra meta II formation [26]. This work identified ERY as an important determinant of rhodopsin-G_t interaction.
In a recent investigation, it could be shown that signal transmission from the retinal binding site to G_t and eventually to the nucleotide binding site in G_t not only involves ERY but also the $NPxxY(x)_{5,6}F$ region, where helix VII and VIII are connected via a sharp rectangular kink (Fig.3). When Y (Tyr^{306}) or F (Phe^{313}) were replaced with Ala the respective mutant was entirely inactive, although meta II formed perfectly well, even in 9-desmethyl-rhodopsin and acyclic-rhodopsin, which contain only weak agonists after photoisomerization [27]. Mutation E134Q failed to rescue G_t activation, suggesting that the activating E134Q mutation cannot overcome the inhibition of an active receptor surface in the H-VII/H-VIII region.
It remains to be elucidated how the ERY and $NPxxY(x)_{5,6}F$ regions are activated, in parallel or sequentially, and how this relates to the sequence of events in nucleotide exchange catalysis.

6.2 Time ordered sequence of interactions

The problem of sequentiality in the activation mechanism becomes even more obvious by simple visual inspection of Fig.2C. Although the C-terminal regions of $G_t\alpha$ and $G_t\gamma$ have been localized to a common surface of G_t by X-ray studies, the distance between these two sites in a GDP-bound state of G_t appears too large (estimated 50Å) to interact with rhodopsin at the same time (the long diameter of rhodopsin's transmembrane bundle is about 40Å). The problem was recognized as soon as structures of G_t holoprotein was available and electron microscopic analysis of 2-dimensional crystals of rhodopsin showed the size of the helix bundle. Approaches to solve the discrepancy include the following:

1. Receptor and/ or G-protein go through a substantial conformational change in the course of catalytic interaction.
2. The receptor forms dimers or dimerises prior to or during catalysis.
3. The interaction of the G_t sites occurs sequentially rather than simultaneously.

The first explanation would be consistent with the fact that large structural rearrangements in certain GTPases (e.g. the elongation factor EF-Tu) do indeed occur during nucleotide exchange. The second explanation was favored in recent work by Engel, Palczewski and coworkers, based on highly resolution force microscopy (Ref. [28], see the contribution by Andreas Engel in this issue). Another approach was made in recent work of this laboratory (Herrmann,R., Heck,M., Kleuss,C., Hofmann,K.P. and Ernst,O.P., unpublished) employing "mass-tagged" peptides (in which the key binding sites were fused to functionally neutral maltose-binding-protein) and modified G proteins (wild type and modified in their attached lipids or C-terminal amino acid sequence) were used. Together with a kinetic near-infrared light scattering assay to monitor protein-protein interactions in real time, it was possible to show that nucleotide exchange requires a sequential two-step interaction of the G-protein with R*. The data suggest specifically that the encounter of the farnesylated C-terminus of the $G_t\gamma$-subunit with R* comes first and initiates the interaction with the C-terminus of the $G_t\alpha$-subunit, which is probably masked in the holo G-protein when not bound to R*. In the second step, R* shifts interaction from the $G_t\gamma$ towards the $G_t\alpha$ C-terminus; the nucleotide binding site now has low affinity for GDP and is prepared for the uptake of GTP. Such a sequential interaction does not exclude receptor dimerisation but it shows that it is not required to make rhodopsin and G_t interact productively.

6.3 Role of lipid

It has long been known that the formation of meta II depends on fatty acid chain length and saturation, and on the head group of the membrane lipids. In fluid detergents, formation of the intermediate is faster more complete than in the membrane (see [2]). Besides these global effects, there are interactions with specific phospholipids, which depend on the activation of the receptor. ESR spectroscopy together with lipid extraction assays has shown that one molecule of spin-labeled phosphatidylserine per Meta II formed becomes accessible to BSA extraction.

The effect occurs only during the lifetime of Meta II, while the extraction patterns of labelled phosphatidylethanolamine and phosphatidylcholine indicate an inverse effect, namely a tighter interaction during the decay of Meta II [29]. Since mutation of Glu^{134} to a neutral amino acid affects the FTIR difference bands assigned to receptor-bound phospholipid [30], one may suspect that this residue is involved in the control of phospholipid interaction.
Using similar techniques as were applied to rhodopsin, the anchoring of G_t and PDE with the membrane can be studied. We have recently shown that the signaling proteins interact with specific membrane lipids. This is a small cluster of (ca. 3 molecules per rhodopsin) phosphatidylserine in the case of the G_t and with a larger "cushion" of ca. 60 phosphatidylcholine and –ethanolamine molecules in the case of the PDE. These lipid interactions change with the state of activation of the respective protein, illustrating the complexity of time-dependent membrane bound protein-protein interaction [31].

7 General considerations

Rods detect single photons with high fidelity. To achieve the necessary signal to noise ratio rhodopsin has exceedingly low dark activity and thus stability of its ground state. In the time domain of the cellular response, no spontaneous signal must be generated and interfere with the receptor current. On the other hand, light is efficiently absorbed when passing through the cell. The resulting activation of the rhodopsin receptor and its interaction with the G-protein are just the first step in a cascade of protein-protein interactions and biochemical reactions, which are all activated by the absorption of a light quantum in one single rhodopsin molecule. All interactions are located on the disc membranes of the rods and occur only during certain phases of activity of the respective protein. Like with the receptor, all these later stages must be kept sufficiently inactive in the absence of photoexcitation. It is known, for example, that G_t is silent for hours in the absence of the light-activated rhodopsin. The presence of large signaling complexes [32] in which proteins are embedded, so-called transducisomes, and their defined interaction with membranes may generally provide specificity and sensitivity in signaling pathways. Key to the formation of the transducisome in rods appears to be an unusual set of glutamic-acid-rich proteins (GARPs) of unknown function.

Since the first discovery of the „Sehpurpur" more than 120 years ago, primary processes of vision were the subject of intense investigation. For one example, we have seen that, in spite of this detailed - and ever increasing - knowledge, many important questions could not yet be resolved. However, we have learned enough to see that the rod cell, as a masterpiece of evolution, will provide us with the material to ask the right questions.

Acknowledgments

The authors thank Drs. Oliver O. Ernst and Franz J. Bartl for helpful discussions.

References

[1] T. Okada, et al., Trends in Biochemical Sciences 26, 318 (2001)
[2] K. P. Hofmann. (2000) in Molecular Mechanism in visual transduction (Stavenga, D. G., DeGrip, W. J., and Pugh, E. N. J., eds) Vol. 3, pp. 91, Elsevier Science Publishers B.V., Amsterdam
[3] T.S. Shimizu, N. Le Novere, M.D. Levin, A.J. Beavil, B.J. Sutton, D. Bray, Nature Cell Biol. 2, 792 (2000)
[4] T. P. Sakmar, Current Opinion in Cell Biology 14, 189 (2002)
[5] S. Filipek, et al., Annu. Rev. Biophys. Biomol. Struct. 110601.142520 (2003)
[6] J. K. McBee, et al., Progress in Retinal and Eye Research 20, 469 (2001)
[7] J. M. Baldwin, et al., Journal of Molecular Biology 272, 144 (1997)
[8] K. Palczewski, et al., Science 289, 739 (2000)
[9] S. T. Menon, et al., Physiological Reviews 81, 1659 (2001)
[10] J. Herbst, et al., Science 297, 822 (2002)
[11] B. Borhan, et al., Science 288, 2209 (2000)
[12] P. J. Spooner, et al., Biochemistry 42, 13371 (2003)
[13] G.-F. Jang, et al., Journal of Biological Chemistry 276, 26148 (2001)
[14] V. Kuksa, et al., Journal of Biological Chemistry 277, 42315 (2002)
[15] G. Fan, et al., Journal of Biological Chemistry 277, 40229 (2002)
[16] C. K. Meyer, et al., Journal of Biological Chemistry 275, 19713 (2000)
[17] R. Vogel, et al., Biochemistry 39, 8895 (2000)
[18] W. L. Hubbell, et al., Nature Structural Biology 7, 735739 (2000)
[19] D. L. Farrens, et al., Science 274, 768 (1996)
[20] S. P. Sheikh, et al., Nature 383, 347 (1996) S. P. Sheikh, et al., Nature 383, 347 (1996)
[21] S. Arnis, and K. P. Hofmann, Biochemistry 34, 9333 (1995)
[22] R. Vogel, and F. Siebert, Biopolymers 72, 133 (2003)
[23] F. Bartl, et al., FEBS Letters 473, 259 (2000)
[24] F. J. Bartl, et al., Journal of Biological Chemistry 276, 30161 (2001)
[25] R. Vogel, et al., Biochemistry 42, 9863 (2003)
[26] R. R. Franke, et al., Science 250, 123 (1990)
[27] O. Fritze, et al., Proc Natl Acad Sci U S A 100, 2290 (2003)
[28] D. Fotiadis, et al., Nature 421, 127 (2003)
[29] E. Hessel, et al., Journal of Biological Chemistry 276, 2538 (2001)
[30] J. Isele, et al., Biophysical Journal 79, 3063 (2000)
[31] E. Hessel, et al., J Biol Chem 278, 22853 (2003)
[32] H. G. Korschen, et al., Nature 400, 761 (1999)
[33] Y. A. Ovchinnikov, FEBS Letters 148, 179 (1982)
[34] P. A. Hargrave, et al., Biophysics of Structure and Mechanism 9, 235 (1983)

INDEX

E

Index

A

α-actinin, C9.5
α-helix, D1.4
ab initio, C6.3
ABP, C9.5
absorption edge, B2.10
acetylcholine, D7.19
acetylcholine receptor, D7.2
ACh, D7.19
ACh receptor, D7.19, D7.22
actin, C8.4, C9.2
actin binding protein, C9.5
actin filament, C3.3, C3.15, C8.2, C8.6
actin network, C9.5, I1.20
actin solution, C9.10
action potential, D7.11, D7.14–D7.18
active microrheology, C9.8
adenosine triphosphat (ATP), A4.15
adhesion energy, D6.11
adhesion potential, D3.10
ADP, C8.16
AFM, D5.9
aggregate, C3.12
Airy stress function, D2.17
amino acid, A4.13, D1.4
amphipathic molecule, D1.5
amphiphile, A3.2, I1.14
amphiphilic molecule, D2.9
anomalous difference, B2.11
apical, D5.2
area difference elasticity, D3.7
asymmetric simple exclusion process, A1.30
atomic force microscopy (AFM), D5.9
ATP, A4.15, C8.2, C8.6, C8.9, C8.16, D1.2
ATP synthase, A4.17, C10.3
autocorrelation function, B6.13
averaging, B2.13

B

bacteriophage ϕ29, I1.17
bacteriophage HK97, D2.32
bacteriophage T4, D2.3
bacteriorhodopsin, B4.8, I1.17
base vector, B2.6
basolateral, D5.2
Bdellovibrio bacteriovorans, B8.4
bending energy, C3.16, D4.8
bending mode, C3.23
bending rigidity, D2.8, D3.3
bicontinuous phase, I1.17
bilayer, D1.4, D5.2, D5.14
bilayer couple model, D3.7
bilayer fluidity, A2.7
bilayer membrane, I1.19
binodal, C11.11, C11.14, C11.36
biological function, I1.9
biological membrane, A4.5
biological NMR, B3.2, B3.4, B3.13, B3.14
biological photoreceptors, B7.10
Biological Physics, I1.14
Biologically Inspired Physics, I1.14
biomolecular motor, I1.17
biopolymer, B2.2, C2.3
Bjerrum length, C3.6, C3.11
blob concept, C1.27
blobs, C2.34
Boltzmann weight, B5.11, I1.3
Boltzmann's law, C8.9
bond length, B2.14
bond-angle field, D2.19, D2.20
boundary condition, C3.21, C3.29
Bragg equation, B1.5
Brownian dynamics, C2.3
Brownian force, A2.35
Brownian motion, A1.23, A1.25, A2.2, A2.24, C2.28
Brownian time scale, A2.7
buckling radius, D2.26
bud, D1.13
budding, D1.11, D2.33, D4.9, D5.12
budding transition, D2.35
Burgers vector, D2.18, D2.19

C

Ca^{2+} channel, B7.20
Ca^{2+}-ATPase, D7.2, D7.6–D7.9
cadherin, D6.7
cantilever, B4.2

capping protein, C9.5
carbonyl oxygen atom, D7.25
CCD camera, B6.9, B8.2
cell, A2.24, B8.2, D1.2, D1.8
cell adhesion, D6.5, D6.7
cell matrix adhesion, D6.5
cell motility, C2.25, C8.3
cellular membrane, A4.5
central limit theorem, C1.14, C2.3
cGMP, D8.2
chain conformation, C1.2
chain rigidity, C2.35
chaperone, C6.2
characteristic ratio, C2.4, C7.18
chemical force, C8.9
chemical shift, B3.5, B3.7, B3.10, B3.11, B3.13
chemo-attractant, I1.9
chemotaxis, I1.9
chirality, C11.18
chloroplast, D1.10
cholesteric, C11.2, C11.5, C11.18, C11.19, C11.22, C11.24, C11.30, C11.32, C11.34
cholesterol, A4.5, D1.3, D5.2, D5.4, D5.17, D5.18
choline, D1.8
chromatin, A4.7, C3.2, C3.5, C3.13, D1.9
chromosome, A4.7
clathrin, D1.12, D1.13
clathrin triskelion, D2.33
clathrin-coated pit, D2.33
cmc, A2.12
coarse graining, A1.26
codon, A4.14
cofilin, C9.5
coherence length, B1.5
coherent scattering, C2.7
collective diffusion, A2.13
collective dynamics, C5.13
colloid, I1.14
Coloumb interaction, B2.3
columnar phase, C11.2, C11.19, C11.30, C11.32
completeness, B2.11, B2.13
complex plane, B2.7, B2.11
computer simulation, B5.2
confocal microscopy, D4.5
conformal diffusion, D3.8
conformational entropy, C2.3, C2.4
conformational exchange, B3.10
conformational space, B4.7, C6.6
constraint, B2.14
constraint release (CR), C2.23
constructive interference, B2.7
contact interaction, C6.6
contact mode, B4.2
contact site, D6.9
contour length fluctuation, C2.21, C2.23
contrast variation, B1.15
contrast variation technique, B1.8
correctness, B2.13
correlation function, C2.6
cosurfactant, A2.14
Coulomb potential, C3.4, C3.10
counterion condensation, C3.2, C3.3, C3.8
creeping flow equation, A2.29
critical exponent, I1.5, I1.6
critical micelle concentration, A2.12
critical point, I1.4
cross linking protein, C9.5
cross vectors, B2.9
cross-interaction parameter, C7.8
crowding, C11.5, C11.33
cryo-sectioning, B8.2
crystal structure, C10.4, D7.6, D7.7, D7.24, D7.26, D7.28
CuK_α-radiation, B2.3
curvature energy, D2.8
curvature model, D3.3
curvature tensor, D2.7, D2.39
cyclic guanosine monophosphate (cGMP), D8.2
cyclodextrin, D5.17
cytoplasm, A4.3, B8.2
cytoskeleton, A4.3, C1.21, C3.2, C3.3, C3.15, C8.3, C8.4, C9.2, I1.20
cytosolic protein, D1.8

D

de Broglie wavelength, B1.6
Debye function, C1.23, C2.8
Debye screening length, C3.6, C3.7
Debye-Hückel potential, C3.6

Debye-Waller factor, C2.19
deflection signal, B4.2
Delbrück, Max, I1.7, I1.12
depletion, C4.3
depolarization, D7.11, D7.15, D7.16, D7.19
designability, C6.11
desmin, C9.2
detailed balance, A1.13, B1.13, B5.12
detector, B2.4, B2.5
detergent, D5.16
difference density, B2.13
differential geometry, D2.5
diffraction, B2.8, B2.13
diffusion, A1.23, C5.11
diffusion constant, A2.6, C2.7, C3.26
diffusion equation, A1.36
diffusion towards receptors and membrane channels, A2.15
diffusive length scale, A2.7
diffusive time scale, A2.7
dilute solution, C2.3
dipolar coupling, B3.7, B3.10
disclination, D2.14
disclination, buckled, D2.23, D2.24, D2.40
disclination, flat, D2.15, D2.21, D2.40
dislocation, D2.14
dislocation, buckled, D2.24
dislocation, flat, D2.15, D2.21
disorder, B2.2, B2.13
displacement field, D2.14, D2.16
dissipation, D3.15
DLVO theory, B4.6
DMPC, D5.4
DNA, A4.10, A4.12, B1.25, B6.6, C3.2, C3.12, C3.18, C3.28, C8.6, C11.4, C11.5, C11.15, C11.21, C11.30, C11.34, D1.9
DPPC, D5.4
Drosophila melanogaster tripeptidyl peptidase, B8.3
dynamic force spectroscopy, B4.9
dynamic light scattering, C3.26
dynamic scattering function, B1.13
dynamic structure factor, C2.2, C2.16, C2.30, C3.26
dynamical exponent, A1.29
dynamical scaling, A1.29
dynamical structure function, A1.21
dynamical transition, C5.2
dynamically triangulated surfaces, D2.31
dystrophin, C9.5

E

edge-acquisition mechanism, D2.36
effective contact angle, D3.11
effective tension, D3.13
effector, D8.2
eigenfunction, C3.21
eigenvalue equation, C3.21
Einstein relation, A1.23
elasticity theory, D2.14
electron diffraction, C10.3
electron tomography, B8.2
end-to-end diffusion, C7.14
end-to-end distance, C1.3, C2.4
endocytic, D1.11
endocytosis, D1.12, D2.33
endoplasmic reticulum, A4.3, A4.4, D1.2, D1.9
endosome, D1.9, D1.10
endothermic, D5.4
energy density, B2.4
energy landscape, I1.14
entanglement, C2.2, C2.19
entanglement time, C9.11
entropic force, C2.28
entropy, A1.16
enzyme, D1.2, D1.4
epithelial cells, D5.2
equation of motion, C3.10, C3.20, D3.15
equilibrium potential, D7.12, D7.13
equipartition theorem, D2.12
eukaryote, A4.2, D1.15
eukaryotic cell, B8.2, C3.2, C3.13, C8.2, D1.2, D1.9
Euler characteristic, D2.9
Euler's formula, D2.9, D2.35
Ewald construction, B2.5, B2.8
excluded volume interaction, C1.25
exocytosis, D2.33
extracellular matrix, C9.2
Eyring rate theory, C8.13

F

F-actin, C2.25, C9.3, C11.4–C11.6, C11.15,

C11.30, C11.32–C11.34
Förster resonance energy transfer (FRET), B7.17
fd-virus, C11.6, C11.21, C11.22, C11.24, C11.27
Fick's diffusion coefficient, A2.20
Fick's law, A2.15
FIDA, B6.15
filament, C1.10
filamentous actin, C9.3
fimbrin, C9.5
first Born approximation, B1.12
flexibility, C11.4–C11.6, C11.8, C11.14
flexible polymer, C2.3
flippase, A4.7
Flory exponent, C1.24
Flory free energy, C3.7
fluctuation-dissipation theorem, A1.22, A1.33, D3.18
fluid mosaic model, A2.4
fluorescence correlation spectroscopy, B6.12
fluorescence intensity, B6.12
fluorescence lifetime, B6.12
fluorescence lifetime imaging (FLIM), B7.14
fluorescence microscopy, C8.6
fluorescent Ca^{2+} indicator, B7.12
fluorescent Cl^- indicator, B7.15
fluorescent dye, B6.4, B7.5
Fokker-Planck time scale, A2.7
Foppl-van Kármán number, D2.27, D2.33
force constant, B4.9
force induced dissociation of single bonds, D6.21
force spectroscopy, B4.9
force-extension relation, B4.9, C3.18
form factor, B1.19, B2.11, B2.15
Fourier transformation, B3.4, B3.7
Fourier-transform infrared (FT-IR), C10.6
fragmin, C9.5
free electron laser, B1.9
free energy, C8.10
free energy landscape, C6.3
freely jointed chain, C1.3
freely rotating chain, C1.6
frequency domain, B3.4, B3.8
frequency window, C5.7
friction coefficient, C2.13
Friedel's law, B2.10
funnel, C6.12

G

G-actin, C9.3
G-protein, D8.2, D8.3
G-protein coupled receptor, D8.2, D8.4
gating charge, D7.28, D7.30
Gauss-Bonnet theorem, D2.9, D3.4
Gaussian chain, C1.14, C2.3, C2.4, C2.12, C3.20
gelsolin, C9.5
gene regulation, I1.7
genetic algorithm, C6.6
genome, A4.10
GHK Equation, D7.14
globular actin, C9.3
globular protein, I1.16
glycine gating hinge, D7.27
glycocalix, C9.2
glycolipid, A4.5, D1.3
Goldman-Hodgkin-Katz (GHK) equation, D7.14
Golgi apparatus, D1.2, D1.9, D1.10
Golgi complex, A4.4
GPCR, D8.2
green fluorescent protein (GFP), B7.7
Green function, C1.16
ground state effect, C7.9
growth algorithm, C6.5
GTPases, D1.14
guanine nucleotide binding regulatory protein, D8.2

H

hairpin, C7.21
Halobacterium salinarum, C10.2, I1.17
Hamilton operator, I1.3
Hammond behavior, C7.10
hard sphere, B5.4
Harker construction, B2.11
harmonic dynamics, C5.7
harmonic mode, C5.14
heat capacity, C5.2
heavy atom, B2.9, B2.11
Helfrich free energy, A2.22
helix, C7.21
hexagonal phase, D4.12

hexatic phase, D2.22, D2.31
high-resolution NMR, B3.2, B3.6, B3.7, B3.10, B3.11, B3.14
histone, A4.8
homologous protein, C6.9
Hopfield model, I1.6
HSQC, B3.11–B3.13
hybrid Monte Carlo, C6.4
hydration, D7.2, D7.4, D7.23, D7.26
hydrocarbon, D1.3
hydrodynamic interaction, A2.27, C2.3, C2.28, C2.32, C2.34, C2.44, C3.20, C3.23
hydrodynamic screening, C2.44
hydrodynamic tensor, C3.24
hydrophobic, A2.13, C6.7, D1.5, D2.9
hydrostatic pressure, C8.10

I

importance sampling, B5.12
incoherent quasielastic scattering, C2.20
inextensional limit, D2.23
influenza virus, D5.12
insoluble monolayer, A2.14
integrin, C9.2, D6.8
intensity, B2.4
interaction, B5.4
interior-acquisition mechanism, D2.36
intermediate filament, C9.2
intermediate scattering function, B1.13
internal relaxation, C2.7
internal viscosity, C2.36
intracellular membrane, A2.5
ion channel, D7.9, D7.28
ion coordination, D7.26
ion gradient, D7.9
ion permeability, D7.2
ion selectivity, D7.11
Ising model, I1.4
isomorphous replacement, B2.11
isotope labelling, B3.13
isotropic, C11.2

J

joint probability distribution, C3.18

K

K^+ channel, D7.2, D7.12–D7.17, D7.19, D7.21, D7.23–D7.28
K^+ selectivity, D7.24
KcsA , D7.24, D7.26
keratin, C9.2
kinematic approximation, B1.12
kinesin, C8.2, C8.5–C8.7, C8.16
Kramers rate theory, C8.13
Kratky-Porod model, C3.16
Kuhn segment, C1.8

L

Lagrange theorem, C1.5
lamellar phase, D5.19
Lamé coefficient, D2.16
Langevin equation, A2.4, C2.5, C3.10, C3.20, C3.24
Laplace–Beltrami operator, D3.17
lattice model, C6.6
Leffler equation, C7.7
Lennard-Jones potential, B5.4, C4.7
Levinthal's paradox, C6.12
lift force, D3.20
ligand binding, B3.11
light-harvesting complex, B4.7
line energy, D4.7
linear response, A1.19
lipid, B8.2, D1.2, D1.8
lipid bilayer, A2.3, A4.6, D5.2, D5.8, D5.11, D5.14, D5.19, D7.2
lipid inhomogeneity, D4.4
lipid membrane, I1.19
lipid raft, A4.7, D5.2, D5.14, D5.18
lipidic cubic phase, C10.4, I1.17
liquid, D5.3
Lorentz force, I1.2
loss modulus, C9.6
lysosome, A4.4, D1.2, D1.9, D1.10
lytic enzyme, B8.4

M

m-RNA, A1.29
Mössbauer spectroscopy, C5.5
macrophage, C9.13
MAD, B2.11
magnetic tweezers, C9.6, C9.8
master equation, A1.25
Maxwell equations, I1.2

mean square displacement, A2.11, C2.7, C2.10, C2.15, C2.16, C2.20, C2.40, C3.23, C3.25, C5.2
mean square end-to-end distance, C3.17
melting, two-dimensional, D2.21
membrane, D1.2, D1.4, D4.2, D5.2, D5.19, D7.2
membrane adhesion, D6.9
membrane asymmetry, A2.7
membrane curvature, D4.6
membrane flickering, D2.12
membrane fluctuation, D2.11
membrane lipid, A2.6
membrane potential, D7.11–D7.16
membrane protein, B2.2, D7.5, I1.16
membrane receptor, D8.2
membrane tension, C8.10
membrane, self-assembling, D2.9
membrane, two-component, D2.33
mesoscopic method, B5.6
methyl-ß-cyclodextrin, D5.2
methylesterase, C10.12
methyltransferase, C10.12
metric tensor, D2.5, D3.22
Metropolis algorithm, B5.12, C6.4
micelle, A2.12
microemulsion, B1.19
micromechanics, I1.19
micropipette aspiration technique, D6.9
microrheological method, C9.6
microscope, B2.3
microstate, A1.4
microtubule, C8.6, C9.2, C11.5, C11.6
MIR, B2.11
misfolded protein, C6.12
mitochondrion, A4.4, A4.15, D1.2, D1.9, D1.10
mixing entropy, D4.16
modulated phase, D4.11
molecular biology, I1.8
molecular complex, B3.10
molecular dynamics, B3.9, B3.10, C5.5, C6.4
molecular dynamics simulations, B5.2, C3.10
molecular motor, C8.7, I1.17
molecular recognition, D6.4
molecular replacement, B2.10
Monge representation, D2.11, D4.12
Monte Carlo, C6.4
Monte Carlo simulation, A1.14, B5.2
motility assay, C8.4, C8.6
motor protein, C8.2, C8.4
MthK, D7.27
multi-parameter fluorescence detection, B6.16
multicanonical, C6.5
multicomponent membrane, D4.2
multidimensional NMR, B3.2, B3.9
multiminimum mode, C5.14
muscle, C8.2
myelin membrane, D1.8
myoglobin, C2.3, C2.41
myosin, C8.2, C8.4, C8.5

N

N-CAM, D6.7
Na^+ channel, D7.13, D7.15, D7.16, D7.19
Na^+-K^+-ATPase, D7.5–D7.7, D7.11
Na^+-K^+-pump, D7.5
native state, C6.2
Navier-Stokes equation, C2.28
nematic, C11.2, C11.3, C11.8
Nernst, D7.12, D7.14
neural network, I1.6
neurofilament, C9.2
neurotransmitter receptor, B7.13
neutral evolution, C6.11
neutron diffraction, C10.3
neutron research reactor, B1.10
neutron scattering, C2.2, C5.4
neutron spallation source, B1.10
neutron spin echo, C2.9
neutron spin echo (NSE), B1.12
neutron spin echo spectroscopy, C2.2
new PERM, C6.7
NOE, B3.7, B3.9
nonspecific adhesion, D6.2
normal coordinates, C2.6
normal mode approach, C3.28
NP-complete problem, C6.7
NSE, C2.2, C2.9
nuclear Overhauser enhancement (NOE), B3.7
nucleic acid, A4.10
nucleosome, C3.5

nucleotide, A4.11
nucleus, A4.4, D1.9

O

oblate, D3.6
olfactory sensory neuron, B7.14
Onsager, C11.2, C11.4, C11.8, C11.9
optical tweezers, B4.2, C3.28, C8.6, I1.19
order parameter, C11.2, C11.10, C11.13–C11.15, C11.19, C11.24, C11.27, C11.33
order-disorder transition, A2.7
organelle, B8.2, D1.10
Oseen tensor, A2.28, C2.28, C3.23, D3.16
osmotic pressure, A1.14, A1.16, C8.10

P

packing parameter (surfactant), A2.16
pair correlation function, B1.13, B5.8, C2.7
parallel tempering, C6.5
partition function, B5.11, C1.16, C3.16, C3.17
passive microrheology, C9.8, C9.9
Patterson function, B2.9
pawl, C8.15
PDMS, C2.9, C2.30, C2.37
peptidoglycan, B8.4
periodic boundary conditions, B5.6
permeability, D7.4
peroxisome, D1.9
persistence length, C1.10, C3.3, C3.16, C9.5, D2.13
phase, B2.5
phase behavior, C4.4
phase problem, B2.6, B2.10
phase transition, D5.3, I1.4
phenomenological potential, C6.3
phonon, B1.25
phosphatidylcholine, D5.10
phosphatidylethanolamine, D1.3
phosphodiesterase, D8.4, D8.6
phosphoglyceride, D1.2
phospholipid, A4.5, D1.2, D5.4, D5.18, D7.2
photoactive yellow protein (PYP), B7.10
photoisomerization, D8.10
photolytic cycle, D8.9
photon, B2.3
photon-counting histogramming, B6.12
photoreceptor, D8.2
photosynthetic reaction center, I1.16
phylogenetic branch, B8.4
phytochrome (phy), B7.16
PIB, C2.37
piconewton force transducer, D6.18
plane wave, B2.3
plasma membrane, A2.3, D2.2
plasmalogen, D5.17
pleiomorphic structure, B8.2
Poisson ratio, C9.9, D2.18
Poisson-Boltzmann equation, C3.6
polydispersity, C11.3–C11.5, C11.8, C11.17, C11.21, C11.29, C11.30, C11.32
polyelectrolyte, C3.2, C3.14
polyethylene, C2.18
polyethylethylene, C2.11
polymer, C4.6, I1.14
polymer depletion, C1.19
polymer dynamics, C2.3, C3.19
polymer melt, C2.3, C3.23
polymer network, C1.19, C1.27, C9.5
polymerized membrane, D2.29
polysaccharide, C4.3
POPC, D5.4
potential of mean force, C3.5
preaveraging approximation, C3.24, C3.28
principal component analysis, C5.14
PRISM, C3.8, C3.12
profilin, C9.5
prokaryote, A4.2, D1.15
prokaryotic cell, B8.2
prolate, D3.6
propagation of an action potential, D7.15
protein, C2.3, C2.40, C4.2, C8.8, D1.2, D1.4, D5.2, D5.14, D5.16, D5.19
protein crystal, B8.4
protein crystallization, I1.15
protein design, C6.2
protein dynamics, B3.10
protein fluorescence, B7.3
protein fold prediction, C6.2
protein function, C5.2
protein labelling, B7.5
protein structure, B3.10
protein synthesis, A1.29, D1.9

protein-lipid interaction, D8.15
proton gradient, C10.3
proton translocation, C10.2
pruned-enriched Rosenbluth Method, C6.5

Q

quasi-spherical vesicle, D3.9
quasielastic neutron scattering, C2.2
quasiharmonic mode, C5.14

R

R_{free}, B2.14
R_{work}, B2.14
Rabs, D1.14
radius of gyration, C1.5, C3.17
rafts, D5.10
Raman scattering, B6.2
random heteropolymer, C6.10
random walk, C1.3, C2.3
ratchet, C8.15
rate-equilibrium free energy relationship (REFER), C7.4
Rayleigh, B6.2
RDCs, B3.7, B3.9, B3.11, B3.13
receptor, D5.14
receptor aggregation, D6.14
reciprocal space, B2.5, B2.8, B2.10
red blood cell, D1.8, D3.2
refinement, B2.13
refraction, B2.3
relaxation, B3.4, B3.7, B3.9, B3.10
relaxation spectrum, C2.13
relaxation time, C2.6, C2.29, C2.35, C3.22, C3.25, C3.30, C9.11
reptation, C2.3, C2.14, C2.15, C2.17, C2.19, C2.25, I1.21
resonance assignment, B3.9
restraint, B2.14
retinal, C10.2, D8.6, D8.7
retinal isomerization, D8.8
retinal rod cell, D8.2
reversibility, A1.21, A1.24
Reynolds number, C8.7
rheology, C9.10
rhodopsin, D8.2, D8.4, D8.6, D8.9
ribosome, A1.30, A4.3, A4.15, D1.9
rigid valence bond, C6.3
RNA, A4.10, C8.6
rod cell, D8.4, D8.6
rotation, B2.5, B2.8, B2.10
Rotne-Prager tensor, C3.23
Rouse diffusion, C2.14
Rouse dynamics, C2.2, C2.34, C3.20
Rouse model, C2.4, C2.12, C2.18, C2.44, C3.20
Rouse rate, C2.20
Rouse relaxation time, C2.6, C2.15, C2.44, C3.29
rubber, C2.14
rubber elasticity, C1.19, C2.2
rubber plateau, I1.20

S

S-layer, B4.9, B8.4
S4 segment, D7.28
saddle-splay modulus, D2.8
scalar coupling, B3.6
scaling argument, C1.25, C1.27
scattering, B2.3, B2.10, B2.15
Schrödinger equation, I1.2
secondary structure, C6.2
secretion, D1.12
secretory vesicle, D1.9
sedimentation, A2.38
segment displacement, C2.10
segment probability distribution, C2.16
Sehpurpur, D8.16
selectin, D6.7
selectivity, D7.23
selectivity filter, D7.23, D7.25–D7.27
Selenium, B2.11
self diffusion, A2.11
self dynamic structure factor, A2.11
self-avoiding surfaces, D2.29
self-diffusion coefficient, A2.12, B5.9
self-interaction parameter, C7.8
self-similar structure, I1.5
self-similarity, C1.23
semidilute solution, C2.3
semiflexible, C11.6, C11.14
semiflexible polymer, C1.9, C2.25, C3.3, C3.15, C3.17, C9.5
semiflexible polymer network, C9.6, C9.7
sequential sampling, C6.4
servo, B4.2

shear flow, C11.36, D3.18
shear modulus, D2.17
signal transduction, D8.3
signal-to-noise ratio, B6.3
signaling complex, I1.13
signaling pathway, I1.9, I1.13
single molecule detection, B6.2
single molecule spectroscopy, B6.2
single molecule techniques, I1.13
small angle scattering, B1.11
smectic, C11.2, C11.8, C11.19
Smoluchowski equation, A2.34
SMS, B6.2
SNAREs, D1.14
Soft Condensed Matter, I1.14
Soft Matter, I1.14
solvent, B2.2, B2.13
solvent flattening, B2.13
solvent-accessible area, C6.4
specific adhesion, D6.2
specific bond, D6.4
specific heat, A1.8
specificity, D6.4
spectrin, C9.5
sperm, I1.9
sphingolipid, D5.2, D5.18
sphingomyelin, D1.2, D5.10
spinodal, C11.11, C11.36
SpM, D5.4
spontaneous curvature, D2.8, D3.8
Sporosarcina ureae, B8.4
starfish vesicle, D3.4
stationary shape, D3.5
statistical mechanics, A1.2
statistical physics, I1.3
stereo-chemistry, B2.14
steric membrane interaction, D2.12
steric potential, D3.12
stochastic tunnelling, C6.6
Stokes flow, D3.16
Stokes shift, B6.2
Stokes-Einstein relation, C2.43, C9.10
stomatocyte, D3.6
storage modulus, C9.6
strain tensor, D2.16
stress tensor, D2.17, D2.39
stretching energy, D2.16
stretching mode, C3.23
stretching ridge, D2.26
striped phase, D4.12
strong segregation limit, D4.7
structural change, C10.8
structure amplitude, B2.14
structure factor, B2.6, B2.10, B5.8, C2.20
super paramagnetic bead, C9.8
supramolecular assembly, B8.2
surface activity, A2.12
surface tension, D2.8, D2.9
surfactant, D5.15
symmetry, B2.10
synchrotron, B1.9, B2.3

T

tangent vector, D2.5
tank-treading motion, D3.19
tapping mode, B4.3
temperature factor, B2.12, B2.15
ternary mixture, D5.8
tertiary structure, C6.2
tethering potential, D2.29
thermal fluctuation, D3.8
thermodynamics, I1.3
theta solvent, C1.17, C1.27
threading, C6.9
thymosin, C9.5
tilt, D3.19
time domain, B3.4, B3.8
time scales, A2.6
TIRF, B6.10
TMV, C11.2, C11.5, C11.6, C11.21, C11.22, C11.24
Tobacco Mosaic Virus, C11.2
tomogram, B8.2
topographs, B4.2
topological constraint, C2.3
topological interaction, C2.14
topological invariant, D2.9
total-internal reflection fluorescence microscopy, B6.10
trajectory, C5.12
transcription, A4.13
transducin, D8.4, D8.14
transduction, D8.2
transition state, C7.2

transition temperature, C5.2
translation, B2.6
translocation, A4.15
transmembrane asymmetry, D4.4
transmission electron microscope, B8.2
treadmilling, C9.3, C9.4
triplet-triplet energy transfer (TTET), C7.14
triskelions, D1.13
TROSY, B3.7, B3.13
tryptophan (Trp) fluorescence, B7.3
two-photon microscopy, D4.11
two-state folding, C7.2

U

unbinding transition, D3.13
undulation mode, D2.12
undulator, B1.9
united-atom model, C6.6
universality, A1.29, I1.5
utrophin, C9.5

V

van der Waals, B4.6
velocity Verlet algorithm, B5.10
Verlet algorithm, B5.10
vesicle, B8.2, D1.12, D1.14, D3.2, D4.9, D5.7
vimentin, C9.2
viral capsid, D2.3, D2.31, I1.18
virus shape, D2.31
viscoelastic property, C9.7, C9.12
visible fluorescent protein, B7.7
voltage sensing, D7.27, D7.28
voltage sensor, D7.28, D7.29
voltage sensor paddles, D7.28
voltage-gated channel, D7.10, D7.19, D7.28

W

wave vector, A2.13
weak segregation limit, D4.13
white noise, C3.10, C3.21
Wigner crystal, C3.12
wormlike chain model, B4.12, C3.16

X

X-ray crystallography, C5.2, I1.15
X-ray diffraction, C10.2
X-ray photon correlation spectroscopy, B1.22

Y

Young modulus, D2.18

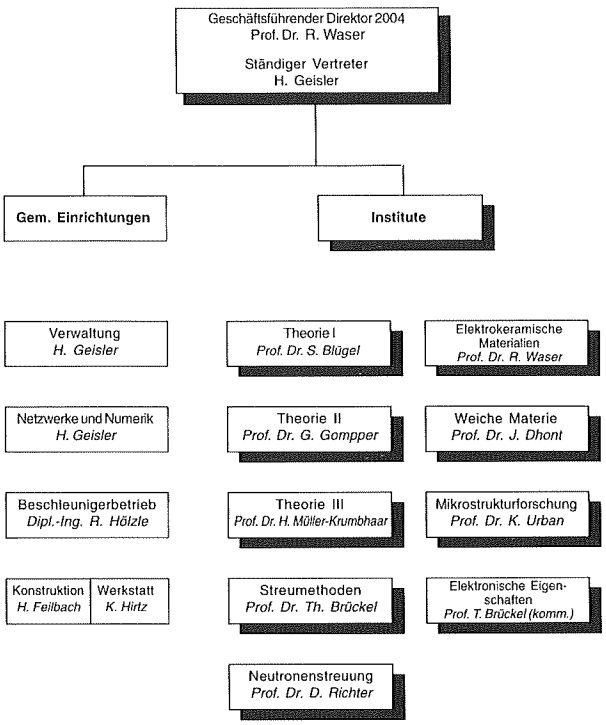
Z

Zimm model, C2.28, C2.30, C2.44, C3.20

Institute for Solid State Research (IFF)

D – 52425 Jülich
Tel. (0 24 61) 61 – 44 65 Fax (0 24 61) 61 – 24 10
http://www.fz-juelich.de/iff

Organisation of the IFF



1. The Mission of the Institute of Solid State Research

The Institute of Solid State Research (IFF) pursues research on condensed materials in the solid or liquid state. Based on its studies of bulk properties of condensed phases, the IFF is concerned with inhomogeneous systems and with the consequences of reduced dimensionality. This refers in particular to the physics boundaries, interfaces, thin films and membranes. With regard to both theory and experiment, the institute's research can be understood in terms of three strategically distinct categories:

- **Phenomena oriented research:**
The search for the discovery and explanation of general phenomena and behaviour in condensed matter systems, including the mathematical and physical concepts and structures underlying this behaviour.

- **Materials oriented research:**
The investigation of specific materials or classes of materials with a view of gaining an understanding of their special properties and, where appropriate, exploring their potential for practical application. Production of novel materials, and preparation as well as characterisation of well-defined samples.

- **Methods oriented research:**
Development of new methods and the improvement of existing methods, both in the experimental and the theoretical/numerical sectors.

The general physical basis for these research areas is provided by statistical physics and quantum mechanics. Together they describe on a microscopic scale the behaviour and the reaction to external influences of electrons, atoms and molecules, the building blocks whose aggregation and cooperation is responsible for the formation of condensed phases. The issues to which these connections give rise in the context of current research form the basis of the work of theoretical and experimental groups within the institute.

Based on new techniques of preparation and characterization on a microscopic atomic scale, the experimental groups carry out research projects to develop devices and taylor materials systems with new properties. This, in turn, has triggered the theoretical groups, employing modern techniques, to describe and come to an understanding of the extremely complex behaviour of these new systems.

The following specific foci of work in the IFF are typical for the mission of a Helmholtz-Research Centre:

- Contributions to the construction and operation of new experimental equipment and instruments for international sources of neutron and synchrotron radiation. In addition to the Jülich Research Reactor "DIDO", the neutron sources at Argonne, Berlin, Gaithersburg, Grenoble, München, Oak Ridge, Saclay and Appleton are used. Experiments with synchrotron radiation are carried out at BESSY, DELTA and HASYLAB in Germany, and internationally at Argonne (APS), Berkeley

(ALS), Brookhaven (NSLS), Grenoble (ESRF) and Trieste (ELETTRA).

- Use of the Jülich Computer Centre, mainly for theoretical investigations and large-scale numerical simulations.

- Projects that require high levels of investment and/or the continuity of a highly qualified team of staff (in most cases in the context of national and international cooperative projects and services). Examples are the Centre for High-Resolution Microscopy, the operation of accelerators, the Electroceramics laboratory, superconductor technology and the development of sophisticated numerical algorithms and programs.

The research results of the IFF have led to a broad and internationally recognized status. Special technical and scientific excellence was achieved in the following areas: application of synchrotron-radiation and neutron scattering, high-performance, electron-microscopy, electron-spectroscopy, high-temperature superconductivity, magnetoelectronics, the physics of clusters, electron theory, dynamical features of phasetransitions, quasicrystals, colloids, membranes and polymers.

A prerequisite for the focal areas of research in the IFF mentioned above and its position in the international community as a partner for cooperation is its strong position in basic research. The IFF carries out basic research in areas with potential for the future of condensed matter science, also as a spingboard for applied research projects.

Genuine innovation (as distinct from continuous development) occurs typically as a by-product of curiosity-driven research. This is exemplified particularly well by the discovery of the "giant-magnetoresistance" effect, which has become a very active area in solid state research. The effect has attractive applications for magnetic sensors e.g. as read-out heads for data storage and for control of moving parts. The license revenue from the patents on this effect represents the largest contribution to the total patent revenue of the Jülich Research Centre devolving from a single application! With this activity the IFF is also participating in the BMBF (Bundesministerium für Bildung und Forschung, i.e. Ministry of Education and Research) project "Magnetoelectronics".

2. Scientific and Technical Infrastructure

Research "at the cutting-edge" requires the capability to build and maintain experimental facilities in high-tech areas. This holds as well for the basic research performed by the IFF at national and international sources as for its applied research. Equipment must meet high international standards, and especially so when the work performed is to be "relevant for industry". A basic requirement for the work of the Department is an infrastructure, within the IFF and the Research Centre itself, which supplies the necessary expertise, effectiveness and facilities.

The group Networks and Numerics maintains the workstation cluster and computer networks, and supports scientific staff in the design and performance of experiments. It supplies expert assistance in hard- and software, maintains the system-software of the IFF-Workstation-Cluster and trains mathematical-technical assistants.

The Accelerator Group is responsible for the operation of the Compact-Cyclotron and the Tandetron.

Construction and Workshop develop and build equipment for the IFF, from simple mechanical components to sophisticated modern technology. The facility works closely together with the scientists and engineers of the Institute to achieve the ambitious quality and tolerance goals set by the experimentalists.

3. Other Scientific Activities: Collaborations with Universities, Institutions and Industry

The productivity of a Research Institute depends strongly on its capacity for scientific exchange and for scientific and technical cooperation. In this regard, the Jülich Research Center offers excellent possibilities. The Centre's top class infrastructure offers efficient assistance, also for visiting scientists. Modern neutron scattering instrumentation appropriate for cutting-edge studies are available to a large number IFF guest scientists in the DIDO reactor hall and in the neutron guide hall ELLA. The wide variety of collaborations in which the institutes for Neutron Scattering and Scattering Methods of the IFF are involved leads to optimal usage of these large scale instruments. On the other hand, the availability of complementary facilities at other institutions is essential for the IFF. With the construction of instruments the IFF contributed to the development and improved usage of external large scale facilities such as ILL and CEN (Grenoble), APS (Argonne, USA), DELTA (Dortmund), HASYLAB (Hamburg), BESSY (Berlin), NSLS (Brookhaven), and NIST (Washington).

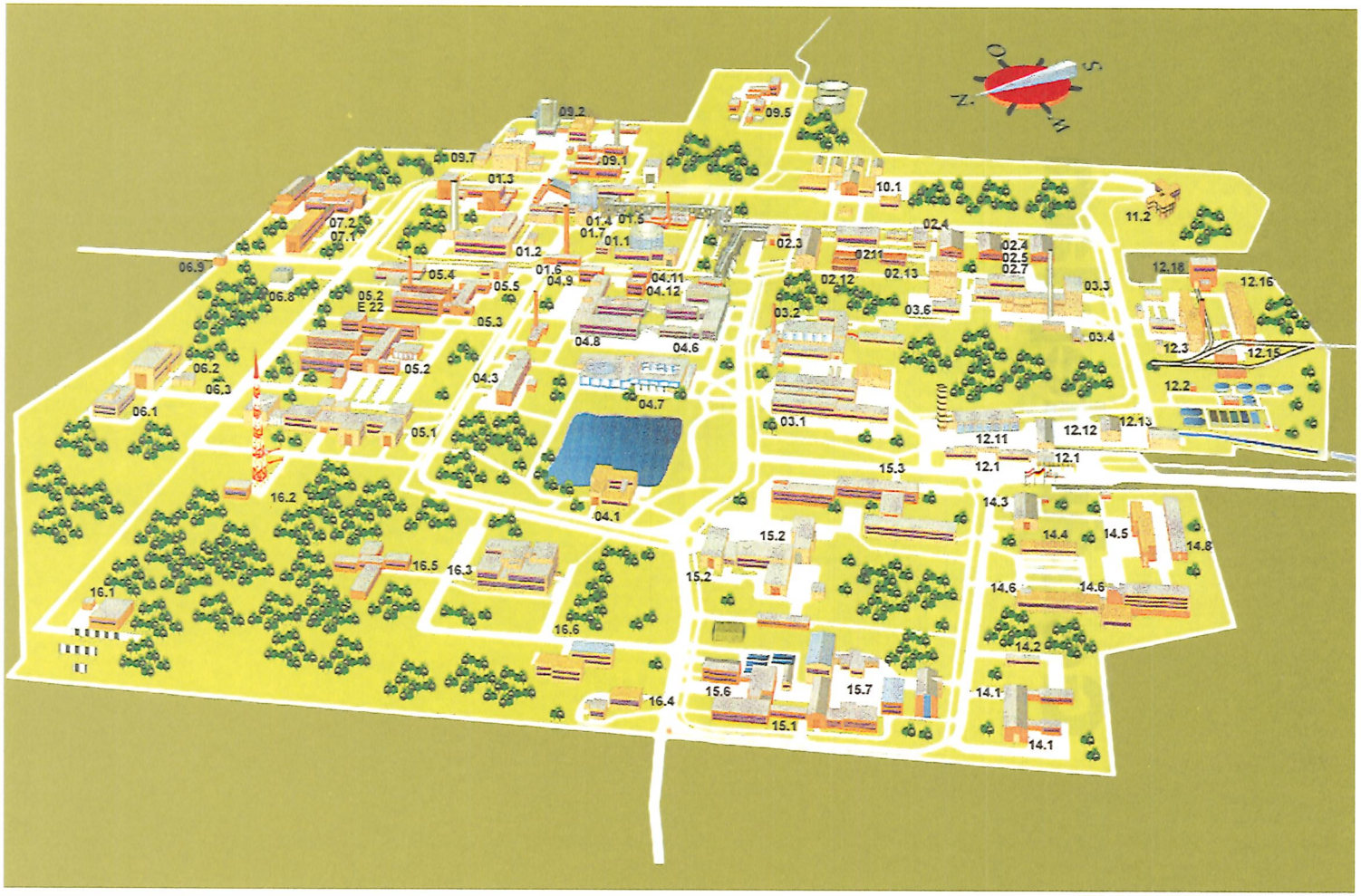
Bilateral agreements have been concluded with the Paul-Drude Institute, Berlin, RWTH Aachen, and the Universities Bonn, Göttingen and Kiel for the use of the Jülich centre for high-resolution microscopy for specialized studies in the area of quantitative and high-resolution microscopy, and the use of the Jülich microscopes with correction of spherical abberation.

In the past years, the IFF has expanded its activities as a sub-contractor to large companies (e.g. Bosch, Siemens, Daimler-Benz, Thompson, Philips) and contract work has become the main focus of the scientific work of several groups, especially in the context of large national (German) and international projects. Of special importance is the IFF's participation in the BMBF-Project "Magnetoelectronics". Under the auspices of the Technology Transfer Bureau of the Jülich Research Centre, the IFF leases equipment to companies that have established themselves in the Jülich Technology Centre. In addition, the IFF has designed and built pilot equipment for industrial companies.

One of the most important duties of a Research Centre is to make its large scale instruments available to universities and other research organisations for research projects that need not necessarily have a direct connection with active research programs of the Centre. In this context it is of paramount importance for the IFF to keep the 16 neutron scattering instruments at the research reactor DIDO operating and to improve them further. In the year 2002, neutrons were available for users at 185 days altogether. During this time 82 experiments were performed by IFF scientists and 116 experiments by external groups (38 groups from each national and international universities, 7 groups from national and 15 from international research institutes).

Many staff members of the IFF serve in scientific committees (e.g. advisory boards for institutes and scientific societies, program committees for international meetings, editors of proceedings, journals and databases, members or presidents of the boards of directors of scientific societies) as well as referees, e.g. for the German funding agencies, special research areas and prize committees.

Gebäudeplan


O
S
N.
N.
09.2
09.5
09.7
09.1
01.3
10.1
11.2
07.2
07.1
01.4
01.7
01.5
01.1
02.4
02.3
02.4
02.5
02.7
06.9
01.2
01.6
04.9
04.11
04.12
02.11
02.12
02.13
12.18
12.16
06.8
05.4
05.5
05.2
E 22
05.3
03.3
03.2
03.6
03.4
04.8
04.6
12.3
12.15
06.2
06.3
05.2
04.3
04.7
12.2
06.1
05.1
03.1
12.11
12.12
12.13
12.1
12.1
15.3
16.2
04.1
14.3
15.2
14.4
14.5
14.8
16.5
16.3
15.2
16.1
14.6
14.6
16.6
14.2
15.6
15.7
14.1
16.4
15.1
14.1

Forschungszentrum Jülich GmbH

12.15	Abfallager (ZFK-DE)
15.3	TIA, Technische Infrastruktur und Anlagenplanung
04.3	ASS, Abteilung Sicherheit und Strahlenschutz
04.3	ASS-AS, Arbeitsschutz
04.3	ASS-BS, Betrieblicher Strahlenschutz
04.3	ASS-GE, Genehmigungen
05.5	ASS-OS, Objektsicherung
04.3	ASS-ST, Sicherheitstechnik
04.3	ASS-UW, Umweltschutz
09.2	AVR, Arbeitsgemeinschaft Versuchs-Reaktor GmbH
14.1	Bauaufsicht
14.6	BEO, Projektträgerschaft Biologie, Energie, Ökologie
15.2	Betriebsärztlicher Dienst (PVA-MED)
14.5	Bildstelle (TIA-G)
14.4	BR, Betriebsrat
01.5	Brennstoffzellenlabor (IWE-BZ)
04.1	Casino I (See) (PVA-SAV)
11.2	Casino II (Wald) (PVA-SAV)
05.4	Chemiezellen (ISR-CZ)
07.1	COSY Jülich (IKP)
09.7	COSY-Bürogebäude
09.6	COSY-Testhalle
12.3	Dekontamination (ZFK-DE)
01.4	DIDO (ZFK-FRJ-2)
09.1	DLR, Deutsche Forschungsanstalt für Luft- und Raumfahrt (MAVIS)
14.5	Druckerei (TIA-G)
15.3	DSB, Datenschutzbeauftragter
15.3	EM, Einkauf und Materialwirtschaft
01.7	ELLA (IFF)
15.3	FA, Finanzabteilung
14.1	FBA, Finanzbauamt
12.11	Feuerwehr (TIA-F)
14.3	Fernmeldezentrale (TIA-N)
01.1	FRJ-1 (MERLIN ZFK)
01.4	FRJ-2 (DIDO ZFK)
06.9	Hambacher Tor (ASS-OS)
12.1	Hauptwache (ASS-OS)
03.3	Heizwerk (TIA-V)
04.11	HLRZ, Höchstleistungsrechenzentrum
01.2	HZ, Heiße Zellen (IWE-HZ)
01.3	IWE-2, Institut für Werkstoffe der Energietechnik
15.1	IBI, Institut für Biologische Informationsverarbeitung
15.7	IBT, Institut für Biotechnologie (IBT-1 und -2)
16.6	IBT, Institut für Biotechnologie (IBT-3)
05.2	ICG 1, Institut für Stratosphärische Chemie
05.2	ICG 2, Institut für Chemie der belasteten Atmosphäre
05.2	ICG 3, Institut für Atmosphärische Chemie
06.1	ICG 4, Institut für Erdöl- und organische Geochemie
06.2	IPC, Institut für Angewandte Physikalische Chemie
05.3	INC, Institut für Nuklearchemie
05.2	IET, Institut für Enzymtechnologie (Univ. Düsseldorf)
03.2	IEV, Institut für Energieverfahrenstechnik
04.6	IFF, Institut für Festkörperforschung
04.8	IFF, Institut für Festkörperforschung
05.2	IFF, Institut für Festkörperforschung
02.4	IGV, Institut für Grenzflächenforschung und Vakuumphysik
07.1	IKP, Institut für Kernphysik
15.2	IME, Institut für Medizin
14.6	INT, Internationales Büro
10.1	IPP, Institut für Plasmaphysik
09.1	IPP, Institut für Plasmaphysik
16.4	IRA, Institut für Radioagronomie
15.6	IRA, Institut für Radioagronomie
05.1	IWE-1, Institut für Werkstoffe der Energietechnik
02.11	ISI, Institut für Schicht- und Ionentechnik
02.13	ISI, Institut für Schicht- und Ionentechnik
04.9	ISI, Institut für Schicht- und Ionentechnik
16.5	ISR, Institut für Sicherheitsforschung und Reaktortechnik
09.1	ISR, Institut für Sicherheitsforschung und Reaktortechnik
06.8	Kaltwasserversorgung (TIA-V) Kälteanlage (TIA-V)
10.1	KFS, Projekt Kernfusion
12.13	Kfz-Halle (TIA-K)
12.12	Kfz-Werkstatt (TIA-K)
12.2	Kläranlage (ZFK-DE)
01.1	MERLIN (ZFK, FRJ-1)
16.2	Meteorologischer Turm (ASS-UW)
14.6	MUT, Programmgruppe Mensch, Umwelt, Technik
04.7	ÖA, Stabsstelle Öffentlichkeitsarbeit
14.2	OIV, Organisation und Informationsverarbeitung
16.5	OIV, Organisation und Informationsverarbeitung
04.12	PFR, Projektträgerschaft für Erforschung kondensierter Materie und Neue Technologien in den Geisteswissenschaften
14.6	PLR, Projektträgerschaft Material- und Rohstofforschung
15.3	PVA, Personal- und Verwaltungsabteilung
15.3	RPA, Rechts- und Patentabteilung
14.2	REV, Revision
14.3	Sicherheitszentrale
14.6	STE, Programmgruppe Systemforschung und Technologische Entwicklung
12.1	Tankstelle (KV)
01.4	Technikumshalle DIDO (ZFK)
10.1	TEXTOR (IPP)
14.5	TIA-E, Elektrotechnik
14.3	TIA-N, Nachrichtentechnik
14.5	TIA-G, Graphische Betriebe
02.7	TIA-H, Haustechnik und Instandhaltung TIA-V, Versorgungswerke TIA-A, Allgemeine Dienste
14.6	TFF, Programmgruppe Technologiefolgenforschung
14.1	TTB, Technologie-Transfer-Büro
04.5	Umspannstation Mitte (TIA-E)
16.1	Umspannstation Nord (TIA-E)
15.5	Umspannstation West (TIA-E)
06.3	Umweltprobenbank (IPC)
15.3	VS, Vorstand
01.6	Wache 01 (ASS-OS)
03.4	Wäscherei (TIA-A)
09.5	Wasserversorgung (TIA-V)
15.3	WTA, Wissenschaftlich-Technische Abteilung
14.6	WTR, Wissenschaftlich-Technischer Rat
16.3	ZAM, Zentralinstitut für Angewandte Mathematik
02.3	ZAT, Zentralabteilung Technologie
03.1	ZAT-MF, Mechanische Fertigung
02.12	ZAT-NPT, Neue Physikalische Technologien
04.7	ZB, Zentralbibliothek
04.8	ZCH, Zentralabteilung Chemische Analysen
02.5	ZEL, Zentrallabor für Elektronik
01.4	ZFK, Zentralabteilung Forschungsreaktoren und Kerntechnische Betriebe
12.3	ZFK-DE, Dekontamination
12.18	ZFK-DE, Abfallager-Halle IV
07.2	Zyklotron (IKP)

Vorlesungsmanuskripte des
17. IFF-Ferienkurses 1986

Dünne Schichten und Schichtsysteme

Einführung in die modernen Methoden der Herstellung, Modifikation, Mikroanalyse und technische Anwendungen dünner Schichten und Schichtsysteme.
Hauptthemen: Präparation · Haftung und innere Spannungen · Grenzflächen · Kohärente Vielfachschichten · Mikroanalytik · Schichten für Informationstechnik
744 Seiten, DM 25,00

nicht mehr verfügbar

Vorlesungsmanuskripte des
18. IFF-Ferienkurses 1987

Synchrotronstrahlung in der Festkörperforschung

Einführung in die Anwendung der Synchrotronstrahlung zu Strukturuntersuchungen in der Festkörperforschung.
Hauptthemen: Quellen · Elektronentheorie · Strukturuntersuchungen · Absorptionsspektroskopie · Photoelektronenspektroskopie · Litho-graphie · Mikroanalytik
812 Seiten, DM 25,00

nicht mehr verfügbar

Vorlesungsmanuskripte des
19. IFF-Ferienkurses 1988

Supraleitung und verwandte Quantenphänomene

Einführung in die Supraleitung und verwandte Quantenphänomene.
Hauptthemen: Supraleitung und Anwendungen · Supraflüssigkeit · Stark korrelierte Elektronen · Lokalisierung
786 Seiten, DM 25,00

nicht mehr verfügbar

Vorlesungsmanuskripte des
20. IFF-Ferienkurses 1989

Computersimulation in der Physik

Einführung in numerische Methoden zur Simulation komplexer Systeme in der Physik.
Hauptthemen: Molekular-Dynamik-Methoden · Monte-Carlo-Methoden · Partielle Differentialgleichungen
1011 Seiten, DM 35,00
ISBN 3-89336-013-1

Vorlesungsmanuskripte des
21. IFF-Ferienkurses 1990

Festkörperforschung für die Informationstechnik

Einführung in die materialphysikalischen Grundlagen der modernen Informationstechnik. Schwerpunkte: Physik der Halbleiter · Physik der Magnetspeichermaterialien · Funktion typischer Bauelemente · Industrielle Fertigung aktueller Bauelemente · Physikalische Aspekte zukünftiger Technik
1028 Seiten, DM 35,00
ISBN 3-89336-033-6

Vorlesungsmanuskripte des
22. IFF-Ferienkurses 1991

Physik der Polymere

Einführung in die physikalischen Stoffeigenschaften von Polymeren und Kunststoffen.
Schwerpunkte: Struktur, Dynamik und Thermodynamik · Konformation · Kristallisation und Glasumwandlung · Gummi- und Viskoelastizität Molekulare Dynamik · Computersimulation · Polymermischungen · Werkstoffe
900 Seiten, DM 45,00
ISBN 3-89336-055-7

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 02461/61-5368 · Fax 02461/61-6103 · e-mail: zb-publikation@fz-juelich.de

Vorlesungsmanuskripte des
23. IFF-Ferienkurses 1992

Synchrotronstrahlung zur Erforschung kondensierter Materie

Einführung in die Anwendung von Synchrotronstrahlung zur Bestimmung der geometrischen und elektronischen Struktur von Festkörpern und Grenzflächen.

Schwerpunkte: Erzeugung und Charakteristika der Synchrotronstrahlung · Instrumentierung · Wechselwirkung von Licht und Materie · Elektronische Struktur von Festkörpern, Oberflächen und Grenzflächen; Theorie und Experiment · Spinauflösende Messungen und Magnetismus · Halbleiter-Oberflächen und Schichtsysteme · Adsorbate auf Oberflächen · Streuung und Beugung von Röntgenstrahlung · Strukturanalyse mit EXAFS · Mikroskopie und Lithographie

ca. 900 Seiten, DM 55,00

ISBN 3-89336-088-3

Vorlesungsmanuskripte des
24. IFF-Ferienkurses 1993

Magnetismus von Festkörpern und Grenzflächen

Einführung in den Magnetismus von metallischen Festkörpern und Grenzflächen.

Schwerpunkte: Konventionelle und moderne Messmethoden · Theorie und Experimente zum Band-magnetismus · Magnetische Domänen und Abbildungsverfahren · Volumen- und Grenzflächenanisotropie · Spinwellen · Phasenübergänge und kritische Exponenten · KKY-Wechselwirkung · Kopplungsphänomene und elektrischer Transport in Schichtsystemen · Werkstoffe

ca. 900 Seiten, DM 55,00

ISBN 3-89336-110-3

Vorlesungsmanuskripte des
25. IFF-Ferienkurses 1994

Komplexe Systeme zwischen Atom und Festkörper

Einführung in die bei komplexen Systemen auftretende Phänomene und die zur Beschreibung verwendeten Konzepte und Theoretischen Methoden.

Schwerpunkte: Moleküle, Cluster, Fullerene (Dichtefunktional, Molekulardynamik, komplexe Energieflächen, Clusterspektroskopie) · Defekte in Quasikristallen, Fluiden, Mustern (Topologische Methoden, Energiefunktional) · Grenzflächen, Schichtsysteme (Adsorption, Benetzung, Aufrauhung, Reibung, STM, Kraftmikroskop, Heterostrukturen) · Komplexe Fluide (Hydrodynamik, Polymere, Polyelektrolyte, Kolloide, Membranen, Gläser) · Musterbildung, Chaos und Turbulenz (in Flüssigkeiten, chemischen Reaktionen und Biologie) · Poröse und granulare Medien (Transport, Nichtmischbare Flüssigkeiten)

1022 Seiten, DM 60,00

ISBN 3-89336-128-6

Vorlesungsmanuskripte des
26. IFF-Ferienkurses 1995



Elektrokeramische Materialien
Grundlagen und Anwendungen

Einführung in die physikalischen und physikalisch-chemischen Grundlagen elektrochemischer Materialien, vor allem mit Perowskit- oder Perowskit-ähnlicher Struktur, die sich durch außergewöhnliche elektrische und dielektrische Eigenschaften auszeichnen.

Schwerpunkte: Kristallstruktur · Elektronenstruktur · Defekt-Physik und Chemie · Elektronische und ionische Transporteigenschaften · Mechanische Eigenschaften

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 02461/61-5368 · Fax 02461/61-6103 · e-mail: zb-publikation@fz-juelich.de

Untersuchungs- und Präparationsmethoden: Impedanz-, dielektrische und mechanische Relaxationsspektroskopie · Neutronenstreuung · Dünne Schichten · Sintern · Kristallzüchtung
Anwendungen: Dielektrische Schichten · Varistoren · PTC · Piezo- und pyroelektrische Keramik · Ferroelektrika · Optische Bauelemente · Ionenleiter und Brennstoffzellen · Oxidische Supraleiter
ca. 900 Seiten, DM 65,00
ISBN 3-89336-146-4

Vorlesungsmanuskripte des
27. IFF-Ferienkurses 1996



Streumethoden zur Untersuchung kondensierter Materie

Einführung in die Wechselwirkung und Streuung von Strahlung zur Untersuchung der atomaren Eigenschaften von kondensierter Materie, insbesondere in die Streutheorie zur Interpretation der Meßdaten; Experimentelle Methoden der Diffraktometrie zur Strukturanalyse, der Kleinwinkelstreuung, der Reflektometrie und der Spektroskopie von Neutronen und Röntgenstrahlen. Strahlungsquellen und Instrumentierung. Relaxationsspektrometrie mit Licht.
Die Untersuchungsmethoden werden an den wichtigsten Stoffklassen demonstriert, beispielsweise Metalle, Legierungen, magnetische Verbindungen, Oxide, Gläser, Kunststoffe, Membranen und Proteine sowie Schichtsysteme von Polymeren und Halbleitern.
ca. 850 Seiten, DM 70,00
ISBN 3-89336-180-4

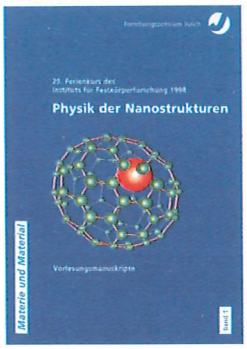
Vorlesungsmanuskripte des
28. IFF-Ferienkurses 1997



Dynamik und Strukturbildung in kondensierter Materie

Einführung in dynamische Vorgänge und Strukturbildung in verschiedenen Formen der kondensierten Materie auf atomaren bis hin zu makroskopischen Längen- und Zeitskalen.
ca. 900 Seiten, DM 75,00
ISBN 3-89336-204-5

Vorlesungsmanuskripte des
29. IFF-Ferienkurses 1998



Physik der Nanostrukturen

Vorstellung der Methoden, die es erlauben, die Nanostrukturen mit höchster Präzision herzustellen und zu charakterisieren. Vorstellung von Beispielen nanostrukturierter Festkörper und Oberflächen aus der aktuellen Forschung, in

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 02461/61-5368 · Fax 02461/61-6103 · e-mail: zb-publikation@fz-juelich.de

denen es gelungen ist, neue Materialeigenschaften zu beobachten und auch teilweise schon technologisch nutzbar zu machen. Erforschung von Clustern, die entweder isoliert im Molekularstrahl oder als Cluster-Materialien vorgestellt und diskutiert werden.
ca. 1000 Seiten, DM 80,00
ISBN 3-89336-217-7

Vorlesungsmanuskripte des
30. IFF-Ferienkurses 1999



Magnetische Schichtsysteme

Einführung in die Grundlagen der Wechselwirkungen, die zu kollektivem Magnetismus führen. Dazu gehört eine Einführung in die Dichtefunktionaltheorie und den Bandmagnetismus. Weiterhin wird eine Übersicht über verschiedene magnetische Materialklassen sowie relevante Messmethoden gegeben. Die Behandlung des Dünnfilmmagnetismus beginnt mit der Einführung der wichtigsten Präparationsmethoden und strukturellen Charakterisierungsmöglichkeiten. Daran schließt sich die Beschreibung der Eigenschaften des Magnetismus in Einzelschichten an, sowie deren Kopplungsphänomene in Multischichten. Besondere Beachtung erhalten Transportphänomene, die durch spinabhängige Elektronenstreuung und Reflektivität an den Grenzflächen oder einen spinabhängigen Tunneleffekt über isolierende Zwischenschichten verursacht werden. Auch die bereits in Einzelschichten auftretenden Anisotropien, sowie die Magnetostriktion werden theoretisch und experimentell behandelt, ebenso die Magnetooptik und der Röntgendichroismus.

Die wichtigsten Themen sind:
- Grundlagen von Magnetismus und magnetischen Materialien
- Präparation magnetischer Schichtsysteme
- Spinstrukturen in Schichtsystemen
- Zwischenschichtaustauschkopplung
- Magnetische Anisotropie und Magnetooptik
- Spinabhängiger Transport

ca. 1300 Seiten, zahlreiche z.T. farb. Abb., DM 110,00
ISBN 3-89336-235-5

Vorlesungsmanuskripte des
31. IFF-Ferienkurses 2000



Femtosekunden und Nano-eV

Dynamik in kondensierter Materie

Die charakteristischen Längen- und Zeitskalen der Dynamik kondensierter Materie erstrecken sich über viele Größenordnungen: von ultraschnellen elektronischen Prozessen im Subfemtosekundenbereich über die Bewegung von Makromolekülen im Nano- bis Mikrosekundenbereich bis hin zu langsamen kollektiven Relaxationsprozessen im Stundenbereich, wie sie etwa bei Spingläsern auftreten.
Diese extreme Bandbreite, die in anderen Disziplinen ihresgleichen sucht, erfordert einen ganzen Zoo von experimentellen Untersuchungsmethoden, die auf der Wechselwirkung von elektromagnetischer Strahlung und von Teilchenstrahlung mit Materie beruhen. Durch revolutionäre Entwicklungen bei den Strahlungsquellen und neue raffinierte experimentelle Verfahren wurden in den letzten Jahren die Grenzen der Festkörperspektroskopie zu extremen Werten der Orts-, Zeit- und Energieauflösung verschoben.

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 02461/61-5368 · Fax 02461/61-6103 · e-mail: zb-publikation@fz-juelich.de

So können heute Nanopartikel mit einer Zeitauflösung im fs-Bereich bzw. Energieauflösungen im neV-Bereich untersucht werden.

In dem IFF-Ferienkurs 2000 werden diese modernen experimentellen Verfahren vorgestellt, wobei als Sonden elektromagnetische Strahlung (von Infrarot- bis harter Röntgenstrahlung), Neutronen und Elektronen diskutiert werden. Viele Experimente wurden erst durch neue Strahlungsquellen realisierbar, wie Synchrotronstrahlungsquellen der dritten Generation oder Spallationsneutronenquellen.

Es werden die Eigenschaften dieser Quellen diskutiert, wobei Begriffe wie Kohärenz und Zeitstruktur eine zentrale Rolle spielen, und ein Ausblick auf Großprojekte im neuen Jahrtausend gegeben. Moderne experimentelle Verfahren wie fsec-Laserspektroskopie, neV-Neutronen-spinechospektroskopie, Röntgenkorellations-spektroskopie oder mehrdimensionale Kern-spinresonanz werden zusammen mit den wissen-schaftlichen Anwendungsfeldern vorgestellt.

Die Dynamik der kondensierten Materie wird in ihrer ganzen Breite behandelt: von elektronischen Anregungen und Transport über die klassische Elementaranregungen wie Phono-nen und Magnonen hin zu diffusiven Bewe-gungen oder die wesentlich entropisch getriebene Dynamik weicher Materie z.B. von Poly-meren, Membranen und komplexen Flüssig-keiten. Weitere Schwerpunkte bilden ultraschnelle Prozesse in Molekülen, der zeitliche Ablauf chemischer Reaktionen, metallische und magnetische Systeme und schließlich die dynamischen Prozesse an Oberflächen. Anwen-dungsnahe Themen wie etwa die Magneto-elektronik werden ebenfalls angesprochen.

Die wichtigsten Themen sind:

- fsec-Spektroskopie
- inelastische Steuung von elektromagnetischer Strahlung, Neutronen
- Elektronenelektronische Anregungen und Transport
- Magnetisierungsdynamik und Magnetoelektronik
- Oberflächendynamik
- chemische Reaktionen und transiente Zustände
- Diffusion, Relaxationsprozesse, Elementaranregungen.

ca. 1000 Seiten, zahlreiche z.T. farb. Abb.,
DM 95,00 (Hardcover)

Vorlesungsmanuskripte des
32. IFF-Ferienkurses 2001



Neue Materialien für die Informationstechnik

Die Möglichkeiten der heutigen Informations- und Kommunikationstechnik werden zunehmend bestimmt durch den Einsatz neuer elektronischer Materialien - jenseits der klassischen Siliziumtechnik. Dies betrifft sowohl die Erweiterung der konventionellen Mikroelektronik durch neuartige Funktionen als auch den Einsatz in integrierten Hybridsystemen. Der IFF-Ferienkurs 2001 behandelt alle Materialklassen, die in modernen Systemen der Informationstechnik heute bereits eingesetzt werden oder ein hinreichend großes Marktpotential für zukünftige Anwendungen aufweisen.

Der Ferienkurs hat einen betont interdisziplinären Charakter. Er verbindet Aspekte der Festkörperphysik, der Kristall- und Molekülchemie sowie der Biologie mit der Mikroelektronik. Der Schwerpunkt liegt auf der Verknüpfung des mikroskopischen Verständnisses der Materialien und der Nutzung ihrer Eigenschaften in Bauelementefunktionen.

Es werden materialübergreifende Konzepte der Informationstechnik diskutiert und aufgezeigt, in wie großer Breite neue Materialien in Bauelemente und Systeme vordringen. Das Spektrum der Vorlesungen umfasst u. a. folgende Themen:

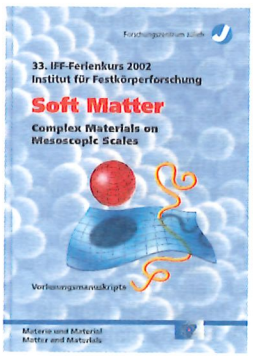
- dielektrische und ferroelektrische Oxide für künftige Generationen schneller, nicht flüchtiger Arbeitsspeicher und hochintegrierter Logik-ICs,

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 0 24 61 / 61 - 53 68 · Fax 0 24 61 / 61-61 03 · e-mail: zb-publikation@fz-juelich.de

- ultradünne Metallschichten mit spin-abhängiger Leitfähigkeit für magnetische Speicher und Leseköpfe heutiger Festplattengenerationen,
- Phasenwechsel-Materialien für wiederbeschreibbare DVDs,
- photoadressierbare Polymere für holographische 3-D-Speicher,
- akusto-optische und elektro-optische Oxide für die optische Datenübertragung,
- abstimmbare keramische Dielektrika für die Mikrowellentechnik, komplexe organische Verbindungen für hochauflösende Flüssigkristall-Displays und die künftige Molekularelektronik,
- Komponenten für Plasma- und Feldemissionsdisplays,
- Materialien für die Verknüpfung der biologischen und der mikroelektronischen Informationsverarbeitung.

ca. 800 Seiten, zahlreiche z. T. farb. Abb.,
EUR 61,36 (Hardcover)

Lecture manuscripts of the
33rd IFF Spring School



Soft matter:

complex materials on mesoscopic scale

In the last years, the traditional research areas of polymers, biological macromolecules, colloids, amphiphilic systems and membranes, as well as liquid crystals have merged into a new research field - *Soft Matter*. This field combines all materials, which are characterized by structures on typical length scales between nanometers and micrometers. Due to the large structural length scale, the number density of their translational degrees of freedom is many orders of magnitude smaller than for an ordinary, molecular material. This and the weak interactions between the structural units, which is typically on the order of the thermal energy k_BT, implies that these materials are easily deformable by external forces - they are soft.

The growing together of the different, previously disjoint areas of *Soft Matter* arises on one hand from the recognition of the same under-lying mechanisms in the structure and the properties of these systems, and on the other hand from the combination of many of these components in a single material. Examples are polymer-colloid mixtures such as ink, or the cell membranes of biological cells, in which a large number of different, cooperative components are involved.

Research in the field of *Soft Matter* is an interdisciplinary enterprise. This includes chemistry, which, with the synthesis of increasingly complex molecules, provides the building blocks for new materials. Physics provides the methods to investigate the properties of soft materials, and aims at reaching a detailed understanding of the connection between the molecular units and their interaction, and the observed macroscopic properties. Some of the investigated questions are derived from, or have implications, for example, for pharmacology or cell and molecular biology. As far as applications are concerned, material science is interested in finding materials with properties, which can be custom-tailored and tuned over a wide range.

The book consists of four parts:

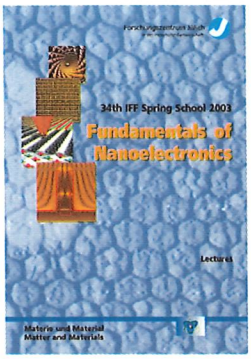
- Techniques, Methods and Synthesis. Scattering, NMR, dielectric spectroscopy, computer simulations and the synthesis of polymers and colloids, including templating methods, are discussed.
- Structure and Phase Behavior. After an introductory lecture on statistical mechanics, the microstructural properties and phases of many different systems are dealt with, like polymers, polyelectrolytes, films and surfaces, amphiphilic systems, membranes and colloids.
- Equilibrium and Non-equilibrium Dynamics. This part discusses dynamics and kinetics of soft matter in- and out- of equilibrium,

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 0 24 61 / 61 - 53 68 · Fax 0 24 61 / 61-61 03 · e-mail: zb-publikation@fz-juelich.de

like polymer and colloid dynamics, glassy polymers and colloids, membrane and microemulsion dynamics, phase separation kinetics and rheology.
- Industrial Applications. This part discusses applications of soft matter systems in industrial processes.

c. 1000 pages, many partly coloured illustrations, EUR 70,00 (Hardcover)

Lecture manuscripts of the
34th IFF Spring School



Fundamentals of Nanoelectronics

Information technology stands today at the edge of a revolutionary change: from micro-electronics to nano-electronics. The characteristic structure sizes will soon fall below the 100 nm mark and will continue to shrink to approximately 20 nm around the year 2015. On this length scale quantum effects are decisive and lead to completely new possibilities and challenges.
As an alternative to the continually increasing cost of conventional semiconductor technologies for manufacturing integrated circuits, new strategies are examined in research, which are based on fundamental principles of physics and chemistry. For example, molecular self organization mechanisms are developed in order to manufacture well defined nanostructures with desired properties.
The IFF Spring School treats the basic principles of physics, chemistry and information technology that prepare the way into the new and fascinating world of nanoelectronics.
The spectrum of the lectures covers, among other subjects, the following topics:

- Fundamentals
  Theory of electron scattering
  Tunnelling processes
  Coulomb blockade effects
  Spin dependent transport
  Quantum communication/computing
- Analyses
  Scanning probe methods
  Spectroscopy with atomic resolution
  High resolution electron microscopy
- Technology
  New lithography techniques
  Atomic layer by layer deposition
  Self organization techniques
  Scanning probe manipulation
- Concepts for nanoelectronic devices
  Limits of Si technology
  Metallic nanowires
  Spintronics
  Electronics based on carbon nanotubes
  Single electron devices
  Concepts for QC devices
  Molecular electronics

c. 500 pages, many partly coloured illustrations, EUR 55,00 (Hardcover)

Bezugsadresse:
Forschungszentrum Jülich GmbH · ZENTRALBIBLIOTHEK · 52425 Jülich
Tel. 0 24 61 / 61 - 53 68 · Fax 0 24 61 / 61-61 03 · e-mail: zb-publikation@fz-juelich.de

Schriften des Forschungszentrums Jülich
Reihe Materie und Material / Matter and Materials

1. **Physik der Nanostrukturen**
   Vorlesungsmanuskripte des 29. IFF-Ferienkurses vom 9. bis 20.3.1998
   (1998), ca. 1000 Seiten
   ISBN: 3-89336-217-7

2. **Magnetische Schichtsysteme**
   Vorlesungsmanuskripte des 30. IFF-Ferienkurses vom 1. bis 12.3.1999
   (1999), ca. 1300 Seiten, zahlreiche z.T. farb. Abb.
   ISBN: 3-89336-235-5

3. **Femtosekunden und Nano-eV: Dynamik in kondensierter Materie**
   Vorlesungsmanuskripte des 31. IFF-Ferienkurses vom 13. bis 24.3.2000
   (2000), ca. 1000 Seiten, zahlreiche z. T. farb. Abb.
   ISBN: 3-89336-205-3 (Hardcover)

4. **Massenspektrometrische Verfahren der Elementspurenanalyse**
   Kurzfassungen der Vorträge und Poster des 5. Symposiums und des
   12. ICP-MS-Anwendertreffens vom 18. bis 21. September 2000 in Jülich
   zusammengestellt von H.-J. Dietze, J. S. Becker (2000), xvi, 109 Seiten
   ISBN: 3-89336-267-3

5. **Neutron Scattering**
   Lectures of the Laboratory Course
   held at the Forschungszentrum Jülich from 26 September to 6 October 2000
   herausgegeben von Th. Brückel, G. Heger, D. Richter (2000), 382 pages
   ISBN: 3-89336-269-X

6. **Baryon Excitations**
   Lectures of the COSY Workshop
   held at the Forschungszentrum Jülich from 2 to3 May 2000
   edited by T. Barnes and H.-P. Morsch (2000), 203 pages
   ISBN: 3-89336-273-8

7. **Neue Materialien für die Informationstechnik**
   Vorlesungsmanuskripte des 32. IFF-Ferienkurses vom 05. bis 16.03.2001
   herausgegeben von R. Waser (2001), ca. 1200 Seiten, zahlreiche z. T. farb. Abb.
   ISBN: 3-89336-279-7

8. **Deutsche Neutronenstreutagung 2001**
   Kurzfassungen der Vorträge und Poster der
   Deutschen Neutronenstreutagung 2001 vom 19. bis 21. Februar 2001 in Jülich
   zusammengestellt von Th. Brückel (2001), xvi, 186 Seiten
   ISBN: 3-89336-280-0

9. **Neutron Scattering**
   Lectures of the 5th Laboratory Course
   held at the Forschungszentrum Jülich from 18 September to 28 September 2001
   edited by Th. Brückel, G. Heger, D. Richter, R. Zorn (2001), 418 pages
   ISBN: 3-89336-288-6

10. **Soft Matter. Complex Materials on Mesoscopic Scales**
    Lecture manuscripts of the 33rd IFF Spring School on March 04 - 14, 2002
    edited by J. K. G. Dhont, G. Gompper, D. Richter (2002),
    c. 1000 pages, many partly coloured illustrations
    ISBN: 3-89336-297-5 (Hardcover)

11. **Proceedings of the symposium on threshold meson production in pp and pd interaction**
    Extended COSY-11 Collaboration Meeting Cracow, 20 - 24 June 2001
    edited by P. Moskal, M. Wolke (2002), 274 pages
    ISBN: 3-89336-307-6

12. **Polarized Neutron Scattering**
    Lectures of the 1st Summer School held at the Forschungszentrum Jülich from 10 to 14 September 2002
    edited by Th. Brückel, W. Schweika (2002), 360 pages
    ISBN: 3-89336-314-9

13. **Beam Cooling and Related Topics**
    255th International WE-Heraeus-Seminar, Bad Honnef, May 13 - 18, 2001
    Proceedings
    edited by D. Prasuhn (2002), PDF-Datei auf CD
    ISBN: 3-89336-316-5

14. **Fundamentals of Nanoelectronics**
    34th Spring School of the Department of Solid State Research,
    10 - 21 March 2003, Jülich, Germany
    edited by S. Bluegel, M. Luysberg, K. Urban, R. Waser (2003), c. 1000 pages
    ISBN: 3-89336-319-X (Hardcover)

15. **Neutron Scattering**
Lectures of the Laboratory course held at the Forschungszentrum Jülich
edited by Th. Brückel; G. Heger; D. Richter; R. Zorn (2003), 375 pages
New, revised edition of Materie und Material/Matter and Materials Volume 9
ISBN: 3-89336-324-X

16. **Proceedings of the 1st Summer School 2002 and Workshop on COSY Physics**
28. August - 4. September 2002
Organizing Committee: K. Brinkmann, M. Büscher, et al. (2003), CD-ROM
ISBN: 3-89336-331-9

17. **Resource-orientated Analysis of Metallic Raw Materials: Findings of CRC 525 for Aluminium**
edited by W. Kuckshinrichs, P. N. Martens (2003), CD-ROM
ISBN: 3-89336-342-4

18. **The Physics of Spallation Processes: Theory, Experiments and Applications**
F. Goldenbaum (2004), V, 174 pages
ISBN: 3-89336-346-7

19. **Physics meets Biology: From Soft Matter to Cell Biology**
Lecture manuscripts of the 35th IFF Spring School from March 22 – April 2, 2004, Jülich, Germany
edited by G. Gompper, U. B. Kaupp, J. K. G. Dhont, D. Richter, R. G. Winkler (2004), c. 900 pages
ISBN: 3-89336-348-3 (Hardcover)