

John von Neumann Institute for Computing (NIC)

NIC Workshop 2006 From Computational Biophysics to Systems Biology 2006

edited by

Jan Meinke

Olav Zimmermann

Sandipan Mohanty

Ulrich H.E. Hansmann

Symposium, 06. - 09. June 2006

Forschungszentrum Jülich

Proceedings

organized by

John von Neumann Institute for Computing

NIC Series

Volume 34

ISBN-10: 3-9810843-0-6

ISBN-13: 978-3-9810843-0-6

Die Deutsche Bibliothek – CIP-Cataloguing-in-Publication-Data
A catalogue record for this publication is available from Die Deutsche
Bibliothek.

Publisher: NIC-Directors

Technical Editor: Fabian Fredel

Distributor: NIC-Secretariat
Research Centre Jülich
52425 Jülich
Germany

Internet: www.fz-juelich.de/nic/

Printer: Graphische Betriebe, Forschungszentrum Jülich

© 2006 by John von Neumann Institute for Computing

Permission to make digital or hard copies of portions of this work
for personal or classroom use is granted provided that the copies
are not made or distributed for profit or commercial advantage and
that copies bear this notice and the full citation on the first page. To
copy otherwise requires prior specific permission by the publisher
mentioned above.

NIC Series Volume 34

ISBN-10: 3-9810843-0-6

ISBN-13: 978-3-9810843-0-6

Published in full on the internet

Persistent Identifier: [urn:nbn:de:0001-00367](http://nbn-resolving.org/urn:nbn:de:0001-00367)

<http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0001-00367>

Preface

Over the last few decades the scope of biology has changed from categorizing experimental observations to understanding the underlying mechanisms. This is at least in part due to the increasing use of computers. Computers allow us to study biological processes and systems *in silico*. Computer simulations are already used to study single macromolecules, the interaction between various biomolecules, or their regulation in the cell. For many medical and pharmaceutical applications such as the understanding of disease pathways or the working of drugs at the level of an individual cell one would like an integration of these simulations — situated at a molecular level — into a functioning model of cells. This is one of the big challenges in computational biology requiring both new hardware and new techniques.

The many benefits of such an extension of computational biophysics into systems biology motivated the workshop that the research group "Computational Biology and Biophysics" held June 6-9, 2006 at the John von Neumann Institute for Computing in Jülich, Germany. Specifically, the workshop targeted the following general areas at the interface between (computational) biophysics and systems biology:

- Protein folding (both structure prediction and mechanics)
- Mis-folding and aggregation
- Interaction between proteins and other molecules
- Assembly of nano-structures, multi-protein, protein-DNA/RNA complexes
- Cellular systems at the molecular level

In scientific presentations and numerous informal discussions the participants explored a wide range of topics "From Computational Biophysics to Systems Biology" therefore justifying the name of the workshop. More than 70 researchers from Europe, Asia and the USA discussed in lectures and posters recent developments in both hardware and algorithms. Physicists, Chemists, Biologists and Computer Scientists described successful applications of high-performance computing to biology and biochemistry. This proceeding volume collects selected presentations from the 4-day long workshop that may serve as starting point for further discussions. It is divided into articles by invited speakers, contributed speakers, and such originally presented as posters, as the interdisciplinary nature of the articles often defies a simple classification according to subject areas.

Besides the editors Thomas Neuhaus, Jaroslav Skrivánek and Longhui Wang helped organize the workshop. For their most valuable help with the local arrangements we are greatly indebted to Helga Frank, Erika Wittig and Fabian Fredel. We also wish to thank the Forschungszentrum Jülich for financial support and IBM for the very special price in our raffle.

Jülich, July 2006

U. H. E. Hansmann

Contents

Invited Talks

Cluster Distance Geometry and Protein Folding <i>G.M. Crippen</i>	1
From Simulation to Therapy: A Systems Biology Approach to Oncogene Detection <i>A. Ghosh, D. Pant, A. Kumar, R. Zou, D. Miller</i>	7
Protein Aggregation and Unfolding Studied Using an All-Atom Model with a Simplified Energy Function <i>A. Irbäck, S. Mitternacht</i>	15
Modeling Protein Structure, Dynamics and Thermodynamics with Reduced Representation of Conformational Space <i>A. Kolinski, D. Gront, S. Kmiecik, M. Kurcinski, D. Latek</i>	21
A High-Speed Special-Purpose Computer for Molecular Dynamics Simulations: MDGRAPE-3 <i>T. Narumi, Y. Ohno, N. Futatsugi, N. Okimoto, A. Suenaga, R. Yanai, M. Taiji</i>	29
The Two Aspects of the Protein Folding Problem <i>H.A. Scheraga, A. Liwo, St. Ołdziej, C. Czaplewski, M. Khalili, J.A. Villa, D.R. Ripoll</i>	37
De novo all atom folding of helical proteins <i>A. Verma, S. Murthy, K. H. Lee, E. Starikov, W. Wenzel</i>	45
Dewetting in Nanoscale Hydrophobic Plates Collapse and Protein Complex Folding <i>R. Zhou</i>	53

Contributed Talks

Reduction to the Simplest – The Complexity of Minimalistic Heteropolymer Models <i>M. Bachmann, W. Janke</i>	59
Multiplexed-Replica Exchange Molecular Dynamics with the UNRES Force-Field as an Effective Method for Exploring the Conformational Energy Landscape of Proteins. <i>C. Czaplewski, S. Kalinowski, S. Ołdziej, A. Liwo, H. A. Scheraga</i>	63

Single Amino Acid Contributions to Binding Affinity in Enzyme-Inhibitor Interactions: a Docking-Based Screening of BPTI-Beta Trypsin interaction <i>D. Dell’Orco, P. G. De Benedetti, F. Fanelli</i>	67
Sub-type Specific Sites for SMAD Receptor Binding Identified by Sequence Comparison Using “Sequence Harmony” <i>K. A. Feenstra, W. Pirovano, J. Heringa</i>	73
High Throughput Method for Protein Structure Prediction <i>D. Gront, S. Kmiecik, A. Kolinski</i>	79
From Microscopic Monte-Carlo Simulations to Macroscopic Solvation Models <i>M. Gruzziel, W. R. Rudnicki, B. Lesyng</i>	83
A Very General Solvation Model for BioMolecular Simulation <i>S. Höfinger</i>	87
Prediction of Domain Interactions in <i>C. elegans</i> <i>F. Morcos, M. Boxem, N. Klitgord, M. Vidal, J. A. Izaguirre</i>	91
Adsorption Phenomena at Organic-Inorganic Interfaces <i>M. Bachmann, W. Janke</i>	95
Modelling protein-DNA interactions in chromatin <i>G. La Penna, S. Furlan, L. Arcesi, A. Perico</i>	99
Molecular Dynamics Study of Amyloid Formation of Two Abl-SH3 Domain Peptides <i>I. Liepina, S. Ventura, C. Czaplewski, A. Liwo</i>	103
Optimization of the United-Residue UNRES Force Field for Langevin Dynamics Simulations <i>A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, H. A. Scheraga</i>	107
Simple Physics-Based Analytical Formulas for the Potentials of Mean Force for the Interaction for Amino-Acid Side Chains in Water. Tests with Simple Spherical Systems <i>M. Makowski, A. Liwo, K. Maksimiak, J. Makowska, L. Chmurzyński, H. A. Scheraga</i>	111
The Role of Flexible Stem Geometries in Protein Loop Structure Prediction <i>M. Mönnigmann, C. A. Floudas</i>	115
Modeling of Water Explicitly in the Replica-Exchange Simulation Method for Protein Folding <i>M. Yuguang, Y. Ye</i>	119
Molecular Transport Through Channels and Pores <i>W. Nadler, W. R. Bauer</i>	125

Integrating Structural and Kinetic Enzymatic Information in Systems Biology <i>M. Stein, R. R. Gabdouliline, R. C. Wade</i>	129
Secondary Structure of Compstatin Analogues: Insights from Molecular Dynamics Simulations in Explicit Water <i>Ph. Tamamis, G. Archontis</i>	133
Using Simple Fluid Wetting as a Model for Cell Spreading <i>M. Reuter, C. Taylor</i>	137
Posters	
Parallel Cost-Sensitive Support Vector Machine Software for Classification <i>T. Eitrich, B. Lang</i>	141
A Study on the Structure and Dynamics of Spin Labeled Proteins by MD Simulations and EPR Spectroscopy. <i>P. Gajula, I. V. Borovykh, Ch. Beier, P. Gast, M. Huber, H.-J. Steinhoff</i>	145
An Improved Evolutionary Strategy for Protein Structure Prediction <i>S. M. Gopal, K. H. Lee, W. Wenzel</i>	149
Ab initio Simulations of Cu Binding Sites in the N-Terminal Region of PrP <i>S. Furlan, F. Guerrieri, G. La Penna, S. Morante, G.C. Rossi</i>	153
Cross-Check Methods in Protein Simulations <i>Ch. Junghans, U.H.E. Hansmann</i>	157
The Influence of Molecular Surface Composition on the Outcome of Poisson Boltzmann Calculations Performed to Obtain Solvation Free Energies <i>P. Kar, Y. Wei, U.H.E. Hansmann, S. Höfinger</i>	161
Generalized Born and Nonpolar Models in Solvation Free Energies. <i>P. Kmiec, J. Kalinowski, B. Lesyng</i>	165
Determination of Short-Range Potentials for Physics-Based Protein-Structure Prediction <i>U. Kozłowska, K. Wachucik, A. Liwo, H. A. Scheraga</i>	169
Large Scale Molecular Dynamics Simulations for the Derivation of Solvation Free Energies of Strongly Hydrophobic Molecules <i>R. Mahajan, D. Kranzlmüller, J. Volkert, S. Höfinger</i>	173
Simulation of Thick Polymers <i>T. Neuhaus, O. Zimmerman, U.H.E. Hansmann</i>	177
Symplectic Algorithms of Molecular Dynamics. An Application to the United-Residue Model. <i>F. Rakowski, P. Grochowski, B. Lesyng, A. Liwo, H. A. Scheraga</i>	181

A Fast Wavelet Based Evaluation of Coulomb Potentials in Molecular Systems <i>G. Sutmann, S. Wädow</i>	185
A Highly Accurate and Optimal Method to Calculate Long Range Interactions <i>M. Bolten, G. Sutmann</i>	189
Evolutionary Analysis in Pathogenesis-Related Proteins <i>N. M. Scherer, C. E. Thompson, L. B. Freitas, S. L. Bonatto, F. M. Salzano</i>	193
A Parameterized Procedure for Consensus Sequences - Validation of the Method <i>S. Walkowiak, B. Lesyng</i>	197
Prediction of Parallel and Antiparallel Beta Sheets Based on Sequence Profiles Using Support Vector Machines <i>L. Wang, O. Zimmerman, U.H.E. Hansmann</i>	201
Side-Chain Ordering in PolyGlutamic Acid <i>Y. Wei, W. Nadler, U. H. E. Hansmann</i>	205

Cluster Distance Geometry and Protein Folding

G.M. Crippen

College of Pharmacy,
University of Michigan, Ann Arbor, MI 48109-1065, USA
E-mail: gcrippen@umich.edu

Cluster distance geometry is a recent generalization of distance geometry whereby protein structures can be described at even lower levels of detail than one point per residue. Protein conformations can be summarized in terms of alternative contact patterns between clusters, where each cluster contains four sequentially adjacent residues. A very simple potential function involving 210 adjustable parameters can be determined that favors the native contacts of 31 small monomeric proteins over their respective sets of nonnative contacts. This potential then favors the native contacts for 698 small proteins, even though they have low sequence identity with the training set.

1 Introduction

The ultimate aim of many research groups is to develop predictive models for the thermodynamics and kinetics of protein folding. The standard general approach is to first devise a Hamiltonian or at least a potential energy function that includes terms for all the significant physical effects, such as excluded volume, solvation, and electrostatics. The protein molecule is usually described either in terms of atomic Cartesian coordinates or dihedral angles, as in the rigid valence geometry approximation. The solvent may be treated either as explicit water molecules having atomic coordinates, or by different implicit models, such as Poisson-Boltzmann, generalized Born, or solvent accessible surface area. Only then does the search begin for energetically important protein conformations, folding pathways, or statistical mechanical averages over an appropriate ensemble of states. Unfortunately, the search/sampling problem is very difficult because there are many degrees of freedom, and the energy function depends in a complicated way on the different conformational parameters.

Recently we have been exploring a radically different approach to the problem where the choice of conformational parameters and energy function guarantee that the search problem and the evaluation of the partition function are computationally feasible. Only then does one adjust the energy function to give results in agreement with experiment. We call this paradigm SMEUSE = Statistical Mechanics Enabled Using Separable Energies, where the energy-like function consists of a sum of terms, each one of which depends on a few conformational parameters that are not used by any other term. Thus we have

$$E_{tot}(x_1, \dots, x_n) = \sum_i E_i(x_i) \quad (1)$$

and the potential energy part of the canonical partition function can be simplified from a multidimensional integral to a product of simple integrals.

$$Z = \int \cdots \int \exp(-E_{tot}/kT) dx_1 \cdots dx_n = \prod_i \int \exp(-E_i(x_i)/kT) dx_i \quad (2)$$

One version of SMEUSE for protein thermodynamics was to represent the polypeptide chain as a sequence of C^α points and then use the Haar wavelet transform on the x , y , and z coordinates as a function of sequence number, much as one might transform the amplitude of a signal as a function of discrete time steps.⁴ So that the conformational parameters would be invariant under rotation, the transformed coordinates were combined into the corresponding distance transform.

$$\hat{d}_{w,j} = (\hat{x}_{w,j}^2 + \hat{y}_{w,j}^2 + \hat{z}_{w,j}^2)^{1/2} \quad (3)$$

The physical interpretation of the different $\hat{d}_{w,j}$ s is the distance from the centroid of the first half of the chain to the second half, the distance from the first quarter to the second quarter, the distance from the third quarter to the fourth quarter, the distance from the first eighth to the second eighth, etc. For n residues, there are n distance transforms, $\hat{d}_{w,j}$, and these are taken to be the conformational parameters. The energy was taken to be a simple sum over all the wavelet distances

$$E = \sum_{w,j} (\hat{d}_{w,j} - s_{w,j} \cdot a_w)^2 \quad (4)$$

where the vector $s_{w,j}$ encodes the amino acid composition of the part of the chain corresponding to the wavelet, and the adjustable parameters are the components of the vector a_w . Skipping over some details, for short polypeptide chains involving the standard 20 residue types, there were only 26 adjustable parameters, and these were adjusted so that the free energy for a flexible region around the PDB structure should be more favorable than the free energy for all other conformations. Using only seven proteins of a variety of SCOP fold types (α , β , $\alpha + \beta$, and α/β), the 26 parameters could be adjusted to favor their native states. Then among a set of 1822 PDB entries having single polypeptide chains with lengths from 30 to 128 residues, we found 497 of them to have predicted stability of their native states at some sufficiently low temperature, even though almost all of them had negligible sequence similarity to the training set of proteins. The good news is that such an incredibly simple model has any predictive power. The bad news is that with only n conformational parameters for an n residue chain, the most favorable conformations are not precisely specified. In any case, there was no conformational sampling problem by construction.

An alternative set of conformational parameters that are translation and rotation invariant would be the C^α - C^α interresidue distances, d_{ij} . These are particularly appealing since they have a more direct connection to interresidue interactions and residue desolvation than the Haar transform parameters, which only reflected the degree of extension of different parts of the chain. However, for n residues there are $(n-1)(n-2)/2$ variable d_{ij} s, taking the virtual bond lengths $d_{i,i+1}$ to be invariant. There are additional nonlinear relations among these d_{ij} s such that they represent conformations in three-dimensional space, reducing the degrees of freedom to the $2n-5$ permitted for a (more or less) freely jointed chain.¹ Another way to reduce the number of conformational parameters is cluster distance geometry,¹ where the chain is represented by n/b blocks of b sequential residues each. Then the parameters become the sum of the squared distances between the residues

of block I and block J .

$$D_{IJ} = \sum_{i \in I} \sum_{j \in J} d_{ij}^2 \quad (5)$$

Many of the same relations hold among the D_{IJ} as among the d_{ij}^2 , such as the triangle inequality, but they are algebraically more complicated because the D_{II} reflect the (nonzero) radius of gyration of block I , whereas always $d_{ii} = 0$. The good news is that there are bounds on the D_{II} reflecting the limits of extension and compression of the segment of polypeptide chain, and otherwise the value represents its secondary structure. Similarly, the D_{IJ} between different blocks corresponds to the squared distance between the two centroids of the chain segments, thus indicating the degree of contact between them.

All protein chain conformations for $n = 128$ residues viewed at the $b = 8$ cluster level can be represented as points in a 120 dimensional space.² Shorter chains can be represented in the same space by adding on $128 - n$ imaginary non-interacting residues, much as one zero-fills short signals in discrete fourier transforms. Due to chain connectivity, geometric relations, and steric effects, not all points within the hyperrectangle between the upper and lower bounds in all 120 dimensions correspond to real, three-dimensional structures. The allowed regions can be better approximated by the union of 71 smaller, non-overlapping hyperrectangles ("ranges") chosen so that extended conformations fall into broad ranges while a training set of 32 globular proteins tended to fall into mostly different, smaller, better defined ranges. The energy is taken to be a sum over all 120 variable D_{IJ} of a simple square well, where all residues of block I are in contact with all residues of block J if a chosen contact cutoff $\kappa > D_{IJ}$. When two blocks are in contact, the energy contribution is taken to be a simple linear combination of all residue type pairs, $\mathbf{a} \cdot \mathbf{t}_{IJ}$, where $\mathbf{a}$ is the vector of 210 adjustable energy parameters, and the corresponding vector $\mathbf{t}_{IJ}$ shows how many of each of the 210 (unordered) residue type pairs there are when block I comes in contact with block J . Fitting the model amounts to adjusting κ and $\mathbf{a}$ so that the free energy calculated from evaluating the partition function Eq. 2 over the 71 ranges favors the range that the PDB structure falls in (the native state) over that of the 70 nonnative ranges. Then there were 146 short proteins having $n \leq 128$ and negligible sequence similarity to the training set that had favorable free energies of their respective native states compared to the denatured state. The good news is that the model can fit many proteins and has some predictive power. The bad news is that even though the ranges containing the compact native structures are relatively small, they are still large enough that a wide range of conformations are contained within one of them. In one test case, alternative conformations within the native range differed by as much as 11.5 Å from the PDB structure. Whatever the crudities of the model are, at least all conformations were covered so that free energies are more than wishful thinking.

2 Methods

Recently we have developed an alternative cluster distance geometry approach that is more knowledge based, but it has better predictive power.³ The key improvements were a better treatment of contacts between blocks of residues as a function of D_{IJ} and a better

treatment of different chain lengths n , especially when they are not a multiple of the block size b .

Instead of having a simple square well interaction function, better results can be achieved using a piecewise linear continuous contact function $c(D_{IJ})$ defined relative to a midpoint C_{IJ} and transition halfwidth $0 < w < 1$ so that $c(D_{IJ}) = 1$ when $D_{IJ} < (1 - w)C_{IJ}$, $c(D_{IJ}) = 0$ when $D_{IJ} > (1 + w)C_{IJ}$, and otherwise linearly interpolating in between so that $c(D_{IJ}) = 1/2$ when $D_{IJ} = C_{IJ}$. A value of $w = 0.2$ works well, and then C_{IJ} was chosen for a given choice of b so that 1860 PDB chains had relatively diverse matrices of interblock contacts/noncontacts. For diagonal elements, C_{II} was considerably smaller and was less than the value of D_{II} observed for helical segments of chain.

A second key improvement was to consider all b possible segmentations of n residues into blocks of b residues, rather than appending imaginary noninteracting residues. For example, for $b = 3$ and $n = 7$, the chain 1234567 can be subdivided into 123—456 or 234—567 or 345—, where leftover end pieces having fewer than b residues are simply deleted. That way, each block has a full set of b real residues, and all possible phase shifts of grouping residues are considered. Then the energy is take to be a sum over all segmentations s

$$E = \sum_{s=1}^b \sum_{I \leq J} (\mathbf{a} \cdot \mathbf{t}_{sIJ}) c(D_{sIJ}) \quad (6)$$

using the sigmoidal contact function defined above.

The coverage of conformation space is essentially done in terms of Voronoi polyhedra in the sense that the ranges of the previous study are replaced by the sets of all conformations closer to a given representative conformation than to any other representative. The representative conformations are taken to be contiguous parts of PDB structures of the appropriate chain length that differ adequately from each other in a measure that is harmonious with the chosen energy function. Thus the distance between conformations A and B is

$$S(A, B) = \left\langle \sum_s \sum_{I \leq J} |c(D_{A,s,IJ}) - c(D_{B,s,IJ})| \right\rangle \quad (7)$$

and $S(A, B) < 0.1$ means that conformation B belongs to the region of conformation space associated with representative conformation A . For every chain length n in the training and test sets, PDB was searched until 50 or 100 different representative conformations were found. For short chains, there were often fewer than 50 different representatives, corresponding to an exhaustive coverage of conformation space. For long chains, sometimes fewer than 50 representatives could be found because of insufficient variety available in PDB.

The 210 energy parameters corresponding to the interaction of all residue type pairs were adjusted so that the representative conformation closest to the native (PDB) conformation of each protein in the training set has a lower energy than that of any other

representative. Since the energy is a linear function of the parameters (Eq. 6), this amounts to solving a linear program, much as other energy functions have been trained to favor the native fold over a set of decoys. The main difference is that the region of conformation space associated with one representative can be quite large, especially when there are few contacts. Then this trained energy function discriminates between whole regions, rather than between narrowly defined single conformations, as is the case with most decoy studies using rapidly varying energy functions. For a training set, we selected 211 PDB entries involving a single polypeptide chain with $n \leq 128$, without disulfide groups or substantial ligands, all apparently stable as monomers due to intrachain interactions. There are other PDB entries matching these criteria for which we are unable to adjust the parameters satisfactorily, for reasons that remain unclear. By the nature of linear programming, at the solution that satisfies all inequalities generated by all 211 natives vs. all nonnative representatives, not all inequalities are "active" (i.e., satisfied with very small margins). Indeed, often many of the training proteins could have been deleted, and the same solution would have been found.

3 Results

Not surprisingly, large values of $b > 15$ give no satisfactory energy function, nor do discontinuous ($w = 0$) or very broadly varying ($w \approx 1$) contact functions. As can be seen in Table 1, smaller values of b require fewer active proteins in training and give more successful predictions. Increasing the number of representative conformations makes it harder to fit the training set and somewhat harder to make predictions. In any case, the predicted proteins are sequentially dissimilar to the training set.

b	Representatives	Actives	Predicted
8	50	69	407
4	50	31	698
4	100	56	556

Table 1. Training and prediction results.

While the number of predictions is certainly gratifying, there are nagging doubts about the thoroughness of coverage of conformations. Since the set of representatives implicitly includes realistic features of chain connectivity, excluded volume, three dimensions, peptide flexibility, and secondary structure, these features need not be enforced by the energy function. However, since the representatives are derived by surveys over PDB structures, there is no guarantee that all possible folds are covered. Some representatives cover large ranges of conformations, particularly with regard to parts of the chain that are not in contact with the rest. On the other hand, sometimes the choice of sufficient difference with regard to contacts corresponds to 6.5 Å rmsd for one example of a 76 residue protein. In other words, two different representatives may represent only moderate conformational differences. Beyond that, the current model does not weight the volumes of conformation space covered by the different representatives, so at present we do not attempt to calculate the corresponding free energies of the native and denatured states.

Acknowledgments

This work was supported in part by the Vahlteich Research Award Fund of the College of Pharmacy, University of Michigan.

References

1. G. M. Crippen, *Cluster distance geometry of polypeptide chains*, J. Comput. Chem. **25**, 1305–1312 (2004).
2. G. M. Crippen, *Statistical mechanics of protein folding by cluster distance geometry*, Biopolymers **75**, 278–289 (2004).
3. G. M. Crippen, *Recognizing protein folds by cluster distance geometry*, Proteins **60**, 82–89 (2005).
4. J. Wang and G. M. Crippen, *Statistical mechanics of protein folding with separable Energy Functions*, Biopolymers **74**, 214–220 (2004).

From Simulation to Therapy: A Systems Biology Approach to Oncogene Detection

Avijit Ghosh¹, Dhruv Pant², Aparna Kumar³, Rui Zou⁴, and Dave Miller¹

¹ Department of Physics, Drexel University
3141 Chestnut St, Philadelphia, PA 19104

² Watson School of Biological Sciences, Cold Spring Harbor Laboratory
One Bungtown Road, Cold Spring Harbor, NY 11724

³ Merrimack Pharmaceuticals
101 Binney St, Cambridge, MA 02142

⁴ Department of Biomedical Engineering, Drexel University
3141 Chestnut St, Philadelphia, PA 19104

In silico models of signal transduction pathways have been successful both qualitatively as well as quantitatively in describing how complex protein networks control cell function. Moreover, the study of networks has been used to elucidate not only how these pathways control the complex regulation and response mechanism of cells, but also provide insight into how a breakdown in the biological circuitry can lead to particular disease states.

We have recently examined the circuitry within the MAPK signal transduction pathway to understand how changes within this canonical network may lead to malfunction, notably the rise of proto-oncogenic cells. In addition we have developed a new complementary technique that provides insight into which key players within the pathway are most likely to be most conducive to selective inhibition within this transformed line of cells. These tools have been made freely available to the public, as part of a software suite developed by our group, *Cellsim*¹. I will give an overview on how *Cellsim* may be used to quantitate cell function and moreover malfunction.

1 Introduction

Computational Biophysics has always played a complementary role to the experimental biological sciences. The role of a computational biophysicist, as such, is not to develop tools that simply reassure the experimentalists that well-regarded experiments may, in fact, be duplicated *in silico*, but instead must also provide new predictive and quantitative tools that provide new insight into biological mechanism or function. New tools from the development of new experimentally designed unified force fields such as UNRES² to new special purpose hardware techniques such as MDGRAPE³ have given rise to new predictive mathematical and computational methods that can probe behavior of individual proteins on a femtosecond scale. It is in this sense that new tools such as systems biology have been developed to address the complementary issue of how these proteins can act *en masse* to dictate not only form, but intra-cellular behavior⁴⁻⁸. The scale is perhaps different (from angstrom to micrometer, from individual proteins to micromolar concentrations, from picosecond to millisecond and so forth), but the philosophy is the same. The method of systems biology is to take experimental data that is relatively simple to reproduce, and to use it to provide insight into phenomena that can not be easily elucidated by an existing set of experiments.

Systems biology uses these experimentally derived, computational techniques to elucidate both how cells have the ability to respond to external stimuli and how intra-cellular signaling circuitry that is essential for cell function is controlled. The mechanism by which this occurs depends fundamentally on the way in which cells use protein networks as the mechanism for translating extra-cellular signals into intra-cellular behavior. Hence, the complexity of signal transduction networks is based on the interplay between different aspects of the signaling process, any of which may change with subtle external or internal changes to the cell⁹. The focus of our group is on one particularly important signaling cascade: the MAPK signal transduction pathway.

2 The Canonical MAPK Signal Transduction Pathway

The canonical MAPK signaling cascade (Figure 1) is perhaps one of the most well studied signal cascades, both experimentally and computationally¹⁰. This central cascade is critical for governing cell growth and proliferation as well as actin cytoskeleton rearrangement^{11–14}. Stimulation of the cascade activates many downstream effectors including PI3K¹⁵, Bcl-2¹⁶, and PKC,^{17,18} among others¹⁰. The central cascade is activated via the following mechanism: The epidermal growth factor (EGF) signaling begins with the epidermal growth factor receptor (EGFR) and traverses a series of signaling proteins to the Ras protein. The Ras protein works in part by activating a series of kinases starting from Raf (a Mitogen Activated Protein Kinase Kinase Kinase (MAPKKK)) which activates the Mitogen Activated Protein ERK Kinase (MEK), This in turn activates the extracellular-signal regulation kinase (ERK), which subsequently translocates to the nucleus and stimulates a series of growth promoting transcription factors¹⁹. The pathways described represent a simplified description of the full process of cell signaling, as this cascade is but a single member of a complex set of parallel, interacting pathways⁹.

2.1 Oncogenic Transformation of the MAPK Pathway

The central member of the MAPK pathway, Ras, illustrates the importance of transformations within the MAPK signal transduction cascade. The Ras protein is a GTP-binding signaling protein, activating downstream effectors by binding to GTP, while inactive in the GDP bound form. The Ras oncogene was first found experimentally by its ability to induce tumor-like growth in fibroblasts^{20–22}. The major oncogenic transformation involves specific mutations in Ras that prevent hydrolysis of GTP-bound Ras by GTPase Activating Proteins (GAP), leaving Ras continuously activated (and thus persistently signaling downstream effectors). These particular mutations have been implicated in approximately 30% of all human cancers, predominantly in lung, colon and pancreatic cancers²³. This is likely an under-estimation of oncogenic transformation of Ras related cascades as mutations in other effectors downstream may cause a similar transformation in absence of mutations in the Ras gene itself. The transformation to tumor cells does not occur by a mutational event in Ras alone, but through a series of malignant transformations along or between several distinct pathway species²⁴.

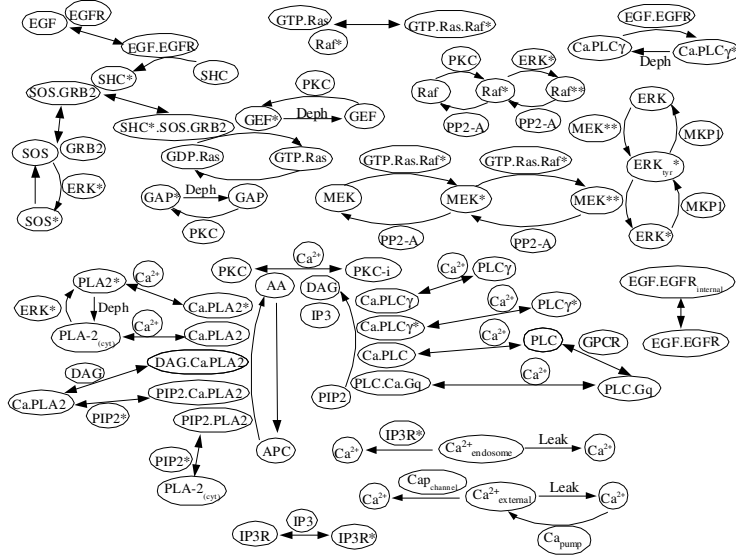


Figure 1. The canonical MAPK pathway

3 Methods

Systems biology methods developed by our group use either ordinary differential equations (ODEs) or partial differential equations (PDEs) to describe the overall temporal or spatio-temporal behavior of the protein network within (and between) compartments of a cell^{7,8}. Enzymatic reactions and other chemical interactions are represented as simply a system of ODEs which couple to active and passive transport. Passive transport includes processes such as simple diffusive processes. Active transport includes explicit advective terms, modeling for instance transport along actin filaments and other ATP driven processes.

Elementary chemical reactions describe the enzymatic and non-enzymatic reactions within each compartment. These reactions may be written as:



where a set of reactant species R_i with stoichiometric coefficients n_i inter-converts into a set of product species P_j with stoichiometric coefficients n_j with rate constants k_f and k_b . As collisions that are greater than bimolecular are rare, the order of an elementary chemical reaction is not typically greater than 2. Characteristic of signaling pathways are enzymatic reactions such as phosphorylation or dephosphorylation events. These reactions may be expressed as a combination of a reversible and an irreversible chemical reaction as follows:



where E represents an enzyme which catalyzes the substrate S . The intermediate species $E.S$ first forms reversibly with rate constants k_1 and k_2 followed by an irreversible catalytic step with rate constant k_3 which releases the activated substrate S^* and the enzyme for further catalysis.

These reactions lead to a set of ordinary differential equations such that one may express the time rate of change in concentration of all species as a system of unimolecular and bimolecular reactions such that:

$$\frac{d[C_i]}{dt} = \sum_j k_{ij}[C_j] + \sum_{l>m} k_{ilm}[C_l][C_m] + \sum_j T(C_i, C_j) \quad (3)$$

where k_{ij} is the rate constant for a unimolecular reaction involving species C_i and C_j at concentrations $[C_i]$ and $[C_j]$ respectively. If k_{ilm} is positive then k_{ilm} represents the rate constant of formation of species C_i from a bimolecular reaction between species C_l and C_m with concentrations $[C_l]$ and $[C_m]$. Conversely, if k_{ilm} is negative then k_{ilm} represents the rate constant of disassociation of species C_i into two species C_l and C_m . $T(C_i, C_j)$ represents a function governing the passive transport of a species C_i into a different compartment at which time it is labeled with a subscript j as C_j via passive channels.

3.1 Mutations

A mutation in a particular gene in a signaling pathway manifests itself in one of two ways. In the first case, the mutation may directly affect the *interaction* between two species. If two species A and B reversibly associate/disassociate with rate constants k_f/k_b :



then a mutation of this type will perturb k_f or k_b by some amount. For instance, lowering k_f by some amount represents a mutation that hinders the ability of species A or B to associate into $A.B$. We define a mutation of this type as an “interaction” mutation. This mutation may occur in either species A or B , as the effect is the same. A simple analysis using Arrhenius theory may be used to connect the free energy change from a mutation with the corresponding kinetic parameters k_f and k_b ^{6,5}.

The rate of interconversion from reactant to product may be given as:

$$k_f = Ae^{-E_a/k_B T} \quad (5)$$

where A is a constant prefactor, E_a is the barrier energy of activation and k_B is Boltzmann’s constant. Mutations in the enzyme may affect a transition rate by either increasing the barrier height or changing the free energy of the initial state of the system by some amount ΔE . The new mutated system may therefore be considered to be a perturbed system with a new barrier of height E'_a and a forward transition rate of k'_f (Figure 2). The ratio k'_f/k_f is :

$$k'_f/k_f = \frac{e^{-(E_a+\Delta E)/k_B T}}{e^{-E_a/k_B T}} \quad (6)$$

and simplifying

$$k'_f/k_f = e^{-\Delta E/k_B T} \quad (7)$$

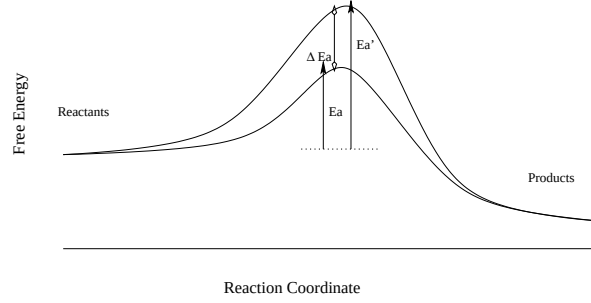


Figure 2. Graphical representation of the energy barriers involved in a reaction. An increase in the barrier height lowers the rate constant. The lower rate constant k'_f corresponds to the greater barrier height E'_a .

The right side of eq. (7) is a function of the *change* in the barrier height and not of the barrier height E_a itself.

Eq. 7 can be used to simulate the *effect* of any single mutation on the normal MAPK signal transduction pathway without having to explicitly delineate the underlying cause. The key effect governing the transformation of the normal MAPK signal transduction pathway, is the ability to activate downstream ERK *without* EGF stimulation. The results from this analysis will be published in a forthcoming manuscript²⁵.

Moreover, such mutations may be rank ordered in terms of ΔE . The top 10 mutations (those that need the smallest ΔE increase to activate ERK) are shown in table 1.

Rank	k_{mod}	Reaction	Dir
1	0.75	$X^* + PP2A \rightleftharpoons X^*-PP2A$	f
2	0.75	$X^*-PP2A \rightarrow PP2A + X$	f
3	0.6	$Raf^* + GTP-Ras \rightleftharpoons Raf-GTP-Ras^*$	b
4	0.55	$Raf^* + PP2A \rightleftharpoons Raf^*-PP2A$	f
5	0.55	$GTP-Ras + GAP \rightleftharpoons GTP-Ras-GAP$	f
6	0.55	$GTP-Ras-GAP \rightarrow GAP + GDP-Ras$	f
7	0.55	$AA \rightarrow APC$	f
8	0.55	$Ca-Ca_{pump} \rightarrow Ca_{pump} + Ca_{ext}$	f
9	0.5	$X + Raf-GTP-Ras^* \rightleftharpoons X-Raf-GTP-Ras^*$	b
10	0.5	$Ca + Ca_{pump} \rightleftharpoons Ca-Ca_{pump}$	f

Table 1. Ranking of the ‘interaction’ mutations based on the highest value of k_{mod} below which ERK activation occurs. These reactions represent the ‘inhibiting’ set of mutations. Reactions with X (or X^*) or (g) represent group reactions.

4 Drug Targeting

With drug development costs now reaching 500 million dollars and more, development strategies represent a significant hurdle in bringing new therapeutics to the marketplace²⁶.

The experimental development cycle can be optimized in a manner that minimizes the number of false positives during costly clinical trials using systems biology techniques similar to those described previously in this manuscript to detect proto-oncogenes. Rather than modeling the effect of a mutation, modeling of an inhibitor may be performed with the addition of a single chemical reaction representing simple competitive binding between the substrate and the target protein:



The binding free energy can be calculated from the equilibrium constant k_{eq} of the reversible binding reaction. A particular inhibitor may bind to any substrate within the MAPK pathway. The efficacy of the inhibitor against a particular target is gauged by its ability to stop auto-activation of the pathway of the entire set of transformed cells described in the previous section. Furthermore, targets that successfully inhibit all transformed cell lines are further ranked by the minimum binding affinity and concentration needed to deactivate all cell lines.

5 Concluding Remarks

Computational biophysics has been successful in underscoring how quantitation and simulation can be used to address difficult problems of interest in biology, as evidenced by the many articles within this book. Systems biology continues this tradition with an emphasis on the *macroscopic* rather than the *microscopic*, focusing on not single molecules or proteins but rather how entire systems interact. As illustrated, systems biology can be used to not only quantitate how behavior is governed, but moreover how malfunctions in the signaling process can give rise to aberrant signaling processes. Finally, an analysis of these aberrant networks can be used to suggest novel treatment strategies.

References

1. A. Ghosh, D. Pant, A. Kumar, D. Miller, R. Zou, H. Lin, Y. Wang, and T. Hoppe. <http://silicon.physics.drexel.edu/cellsim>, 2004. CellSIM Home Page.
2. A. Liwo, P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy, and H.A. Scheraga. A method for optimizing potential-energy functions by hierarchical design of the potential-energy landscape: application to the unres force field. *PNAS*, 99:1937–42, 2002.
3. Y. Komeiji, H. Yokoyama, M. Uebayasi, M. Taiji, T. Fukushige, D. Sugimoto, R. Takata, A. Shimizu, and K. Itsukashi. A high performance system for molecular dynamics simulation of biomolecules using a special-purpose computer. *Pac Symp. Biocomput.*, pages 472–87, 1996.
4. R. Zou and A. Ghosh. Automated sensitivity analysis of stiff biochemical systems using a fourth-order adaptive step size rosenbrock integration method. *IEE Proc. Systems Biology*, 153:79–90, 2006.
5. D.K. Pant and A. Ghosh. A systems biology approach for the study of cumulative oncogenes, with applications to the mapk signal transduction pathway. *Biophys. Chem.*, 119:49–60, 2006.

6. D.K. Pant and A. Ghosh. Automated oncogene detection in complex protein networks, with applications to the mapk signal transduction pathway. *Biophys. Chem.*, 113:275–288, 2005.
7. A. Ghosh, D. Miller, R. Zou, B. Sokhansanj, and A. Kriete. *Spatio-Temporal Systems Biology*, pages 327–382. Elsevier Press, 2006.
8. A. Ghosh, D. Miller, R. Zou, H. Pais, B. Sokhansanj, and A. Kriete. Integrated spatio-temporal model of cell signaling. In *FOSBE 2005 Proceedings*. FOSBE, 2005.
9. U.S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283:381–387, 1999.
10. T. Joneson and D. Bar-Sagi. Ras effectors and their role in mitogenesis and oncogenesis. *J. Mol. Med.*, 75:587–93, 1997.
11. C.J. Marshall. Ras effectors. *Curr. Opin. Cell. Biol.*, 8:197–204, 1996.
12. A.F. Chambers and A.B. Tuck. Ras-responsive genes and tumor metastasis. *Crit Rev Oncog*, 4:95–114, 1993.
13. A.B. Vojtek and C.J. Der. Increasing complexity of the ras signaling pathway. *J. Biol. Chem.*, 273:19925–19928, 1998.
14. D.T. Denhardt. Signal-transducing protein phosphorylation cascades mediated by ras/rho proteins in the mammalian cell: the potential for multiplex signalling. *Biochem J*, 318:729–47, 1996.
15. B.M. Marte and J. Downward. Pkb/akt: connecting phosphoinositide 3-kinase to cell survival and beyond. *Trends Biochem Sci*, 22:355–8, 1997.
16. T. Tsuchida, H. Kijima, S. Hori, Y. Oshika, T. Tokunaga, K. Kawai, H. Yamazaki, Y. Ueyama, K.J. Scanlon, N. Tamaoki, and M. Nakamura. Adenovirus-mediated anti-k-ras ribozyme induces apoptosis and growth suppression of human pancreatic carcinoma. *Cancer Gene Ther*, 7:373–83, 2000.
17. F. Uberall, K. Hellbert, S. Kampfer, K. Maly, A. Villunger, M. Spitaler, J. Mwanjewe, G. Baier-Bitterlich, G. Baier, and H. H. Grunicke. Evidence that atypical protein kinase c-lambda and atypical protein kinase c-zeta participate in ras-mediated reorganization of the f-actin cytoskeleton. *J. Cell. Biol.*, 144:413–425, 1999.
18. S. Kampfer, F. Uberall, S. Giselbrecht, K. Hellbert, G. Baier, and H. H. Grunicke. Characterization of pkc isozyme specific functions in cellular signaling. *Adv Enzyme Regul.*, 38:35–48, 1998.
19. S. J. Mansour, W. T. Matten, A. S. Hermann, J. M. Candia, S. Rong, K. Fukasawa, G. F. Vande Woude, and N. G. Ahn. Transformation of mammalian cells by constitutively active map kinase kinase. *Science*, 265:966–70, 1994.
20. L. F. Parada, C. J. Tabin, C. Shih, and R. A. Weinberg. Human ej bladder carcinoma oncogene is homologue of harvey sarcoma virus ras gene. *Nature*, 297:474–8, 1982.
21. C. J. Tabin, S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, E. M. Scolnick, R. Dhar, D. R. Lowy, and E. H. Chang. Mechanism of activation of a human oncogene. *Nature*, 300:143–9, 1982.
22. R. W. Ellis, D. DeFeo, J. M. Maryak, H. A. Young, T. Y. Shih, E. H. Chang, D. R. Lowy, and E. M. Scolnick. Dual evolutionary origin for the rat genetic sequences of harvey murine sarcoma virus. *J Virol*, 36:408–20, 1980.
23. J.L. Bos. Ras oncogenes in human cancer: a review. *Cancer Res.*, 49:4682–4689, 1989.
24. R.A. Weinberg. Oncogenes, antioncogenes, and the molecular bases of multistep

- carcinogenesis. *Cancer Res*, 49:3713–21, 1989.
25. A. Kumar, R. Zou, and A. Ghosh. A systems biology approach for lead drug target prediction with applications to the mapk signal transduction pathway. In Prep.
 26. R.G. Frank. New estimates of drug development costs. *J. Health Econ.*, 22:325–30, 2003.

Protein Aggregation and Unfolding Studied Using an All-Atom Model with a Simplified Energy Function

Anders Irbäck and Simon Mitternacht

Computational Biology and Biological Physics Group,
Department of Theoretical Physics, Lund University,
Sölvegatan 14A, SE-223 62 Lund, Sweden
E-mail: {anders, simon}@thep.lu.se

We discuss two Monte Carlo studies based on a simple all-atom model for proteins. The first study explores the early aggregation steps of $A\beta_{16-22}$, an amyloid fibril-forming 7-residue fragment of Alzheimer's $A\beta$ peptide. The other study deals with the mechanical and thermal unfolding of ubiquitin, a 76-residue protein that was studied in recent single-molecule constant-force experiments.

1 Introduction

Protein aggregation into amyloid fibrils is a recurrent theme in several human disorders, including Alzheimer's and Parkinson's diseases,¹ and there is evidence that amyloid structures can have a functional role, too.² The mechanisms of amyloid formation are currently being intensely investigated, both experimentally and by computer simulations. These studies are not limited to fibrillar aggregates; small assemblies get more and more attention, because of findings that link soluble oligomers to pathology.³ A broad set of sequences is studied, from disease-associated proteins like the Alzheimer's $A\beta$ peptide to designed amyloid sequences like the hexapeptide STVIIIE.⁴

Here we discuss a study of small assemblies of the $A\beta$ fragment $A\beta_{16-22}$,⁵ which was performed using an all-atom model with a simplified energy function. In addition, we discuss a study of the mechanical and thermal unfolding of ubiquitin,^{6,7} based on exactly the same model. This model was developed through folding studies of a set of well characterized peptides,^{8,9} including α -helical as well as β -sheet peptides. For these peptides, the model was found to give a good description of both structure and thermodynamics.⁹

Mechanical unfolding has been studied experimentally at the single-molecule level for several proteins. These studies have provided valuable insights into the elastic properties of, e.g., the muscle protein titin.^{10,11} Typically, these experiments focus on the extension-versus-force behavior. Computer simulations have the potential to provide information not captured by the experiments, and thereby give a more complete picture of the unfolding process.

2 Model and Methods

The model we use contains all atoms of the protein chains, including hydrogen atoms, but no explicit water molecules. It assumes fixed bond lengths, bond angles and peptide torsion angles (180°), so that each amino acid only has the Ramachandran torsion angles ϕ , ψ and a number of side-chain torsion angles as its degrees of freedom.

The energy function

$$E = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{hp}} \quad (1)$$

is composed of four terms. The term E_{loc} is local in sequence and represents interactions between adjacent backbone dipoles along the chain. The other three terms are non-local in sequence. The excluded volume term E_{ev} is a $1/r^{12}$ repulsion between pairs of atoms. E_{hb} represents two kinds of hydrogen bonds: backbone-backbone bonds and bonds between charged side chains and the backbone. The last term E_{hp} represents an effective hydrophobic attraction between non-polar side chains. It is a simple pairwise additive potential based on the degree of contact between two non-polar side chains. A detailed description of all the different terms can be found elsewhere.^{8,9}

Despite its simplicity, this energy function is able to fold several α -helical and β -sheet peptides with about 20 amino acids.⁹ One type of interaction that the model neglects is the Coulomb interaction between charged side chains. For a small peptide, these charges tend to be exposed to, and therefore screened by the solvent. To be able to study larger proteins and protein aggregates, we expect that it will be necessary to refine the energy function to take into account, e.g., the interactions between side-chain charges.

All our studies were carried out using PROFASI,¹² which is a Monte Carlo software package for simulations of this model.

Let us stress that the amino acid sequence is the only input to the model. All model parameters were thus kept the same in our different studies.

3 $A\beta_{16-22}$ Aggregation

A characteristic “cross- β ” X-ray fiber diffraction pattern reveals that the core structure of amyloid fibrils is composed of β -sheets whose strands run perpendicular to the fibril axis.¹³ For $A\beta_{16-22}$ fibrils, it has been found, by solid-state NMR, that the β -strands have an antiparallel organization.^{14,15} The $A\beta_{16-22}$ sequence (Ac-KLVFFAE-NH₂) consists of five inner residues that are hydrophobic, and two end residues that are oppositely charged. Coulomb interactions between such charges provide a possible explanation for the antiparallel β -strand organization in $A\beta_{16-22}$ fibrils.

Computer simulations of $A\beta_{16-22}$ aggregation have been reported by several groups¹⁶⁻¹⁹. In our calculations, we studied systems of one, three and six $A\beta_{16-22}$ peptides. The simulations were started from random configurations, and the energy function was exactly the same as in our folding studies.⁹

The isolated $A\beta_{16-22}$ peptide turned out to be disordered in our simulations. The three- and six-chain systems, on the other hand, self-assembled into ordered, β -strand-rich aggregates. There was no single dominating free-energy minimum, but rather a number of more or less degenerate minima. Fig. 1 shows two snapshots of such minima.

The β -strand organization in these structures is interesting since our model neglects the interactions between charged side chains, which might be responsible for the antiparallel organization seen in $A\beta_{16-22}$ fibrils. A detailed analysis showed that mixed configurations with both parallel and antiparallel β -strand pairs were common in our simulations, but nevertheless there was a clear statistical preference for the antiparallel organization over the parallel one. Our model thus favors the antiparallel organization despite that these Coulomb interactions are neglected. This, of course, does not mean that these interactions

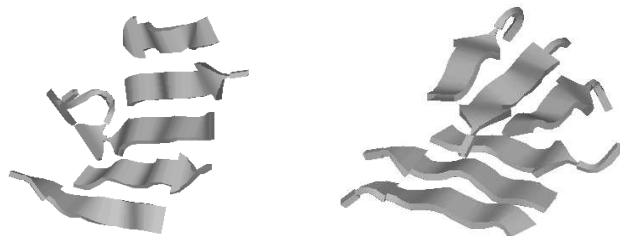


Figure 1. Two typical low-energy structures from our simulations of six $A\beta_{16-22}$ peptides: a five-stranded β -sheet (left), and two three-stranded β -sheets “sandwiching” several of their hydrophobic side chains between them (right). Drawn with RasMol.²⁰

are unimportant, but it suggests that other factors (e.g., hydrogen bond geometry) play a significant role, too. We also did some simulations with attraction/repulsion between unlike/like side-chain charges. Not unexpectedly, this led to an increased preference for the antiparallel organization.

A configuration type that did not occur at all in these simulations was closed barrel-like structures. For this sequence, it seems that six chains are not enough to permit the formation of such structures. By contrast, we have seen the formation of closed barrel-like structures in simulations for nine $A\beta_{16-22}$ peptides.

4 Mechanical and Thermal Unfolding of Ubiquitin

Ubiquitin is a 76-residue α/β protein, whose unfolding and refolding properties have been extensively studied experimentally.²¹ Its native structure contains an α -helix and a five-stranded β -sheet (see Fig. 2).

The mechanical unfolding of ubiquitin was recently investigated by the Fernandez group by single-molecule methods.^{23–25} One study examined the unfolding behavior under a constant stretching force, using end-to-end linked polyubiquitin.²⁵ Here, the time evolution of the end-to-end distance r was followed, and a total of about 800 unfolding traces were collected. In most cases, unfolding occurred in one step, but several examples of unfolding through intermediate states were also observed. The size of the unfolding step to the typical intermediate state was consistent with what one would expect if the α -helix and the N-terminal β -hairpin (A and B in Fig. 2) remain folded in this state, whereas the rest of the molecule is unfolded. Interestingly, these two structures have been found to be the most stable ones in several different experiments at zero force.²¹

We studied the unfolding of ubiquitin under a constant stretching force using the same strengths of the applied force as in the experiments (100 pN, 140 pN and 200 pN).²⁵ The energy function was $E_f = E - \vec{f} \cdot \vec{r}$, where the internal energy E is the same as before (see Eq. 1), $\vec{f}$ is the applied force, and $\vec{r}$ is the end-to-end vector. Because of the existence of multiple unfolding pathways, we performed a set of 500 runs for each force.

As in the experiments, we saw both one-step unfolding and unfolding through intermediate states in our simulations. Furthermore, properties such as the size of the unfolding step, the frequency of occurrence of intermediate states, and the position of the typical

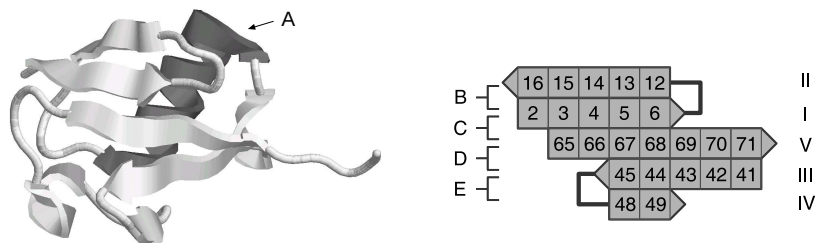


Figure 2. The native structure of ubiquitin with our labels for secondary-structure elements, A–E. Left: A 3D model (PDB code 1d3z, first model²²) drawn with RasMol.²⁰ Right: The organization of the β -sheet.

intermediate state were all found to be in reasonable agreement with experimental data. Having verified these properties, we performed more detailed measurements which, in particular, aimed at characterizing the typical intermediate state.

For this purpose, we investigated the order of breaking of five secondary-structure elements, labeled A–E (see Fig. 2). The structure A is the α -helix, whereas B–E are the four pairs of adjacent strands in the five-stranded β -sheet. To determine the order of breaking, the native hydrogen bonds in these structures were monitored as a function of the end-to-end distance r (r increased essentially monotonically with time). Fig. 3 summarizes the results of this analysis, at 100 pN. From this figure it is immediately clear that the structures A–E do not break in a random order but instead in a statistically preferred order, namely CBDEA. C and B tend to break below the typical r for intermediate states, $r_{\text{int}} \approx 12$ nm, whereas D, E and A tend to break above r_{int} . Our results thus suggest that the typical intermediate is composed of D, E and A rather than A and B.

Fig. 3 shows averages over all events, and therefore does not tell how strong the statistical preference is for the unfolding order CBDEA. To investigate this, we also did an event-by-event analysis, and found that 61 % of the events followed the unfolding pathway CBDEA. Another 23 % of the events had the order of B and D apparently interchanged, the path being CDBEA. In these events, B does unfold before D, but partially reforms after D is gone. The partial refolding of B is not a step back toward the native state, because when B reforms, D is gone, and r is larger. Hence, 84 % of the events followed the same basic pathway.

Several aspects of this calculated unfolding order can be understood in terms of native topology and pulling geometry. That C breaks first is inevitable; the other parts cannot sense the force until C is broken. The native state is mechanically resistant because C is pulled longitudinally, so that several hydrogen bonds must break at the same time. Once C is gone, nothing keeps B from unzipping, one bond at a time. Unzipping requires less force than separation by longitudinal pulling. Therefore, it seems reasonable that B breaks soon after C, as it did in our simulations.

That B breaks early implies that ubiquitin shows a different behavior in these simulations than in various experiments at zero force. In particular, there are experiments suggesting that B along with A are the thermally most stable parts of ubiquitin.^{26,27} Therefore, one must ask whether one actually sees a difference between mechanical and thermal unfolding in our model. To address this question, we performed a set of 800 thermal un-

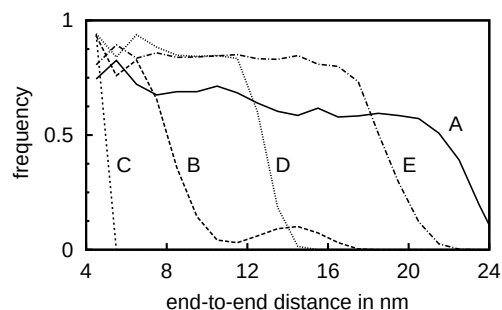


Figure 3. Frequency of occurrence of native hydrogen bonds in the structures A–E against end-to-end distance. Each curve represents an average over all the native hydrogen bonds in a given structure and over all events.

folding simulations at a fixed temperature.⁷ In a majority of these events, A and B unfolded after C and D (E was not analyzed due to noisy data), which indeed is in agreement with the experiments. The agreement with experimental data in the thermal case strengthens our proposed mechanical unfolding order, which remains to be verified experimentally.

Acknowledgments

We thank Sandipan Mohanty for useful discussions. This article is based on work supported in part by the Swedish Research Council.

References

1. For a review, see C. M. Dobson, *The structural basis of protein folding and its links with human disease*, Phil. Trans. R. Soc. Lond. **B 356**, 133–145 (2001).
2. D. M. Fowler, A. V. Koulov, C. Alory-Jost, M. S. Marks, W. E. Balch, and J. W. Kelly, *Functional amyloid formation within mammalian tissue*, PLoS Biol. **4**, e6 (2006).
3. S. Lesné, M. T. Koh, L. Kotilinek, R. Kaye, C. G. Glabe, A. Yang, M. Gallagher, and K. H. Ashe, *A specific amyloid- β assembly in the brain impairs memory*, Nature **440**, 352–357 (2006).
4. M. López de la Paz and L. Serrano, *Sequence determinants of amyloid fibril formation* Proc. Natl. Acad. Sci. USA **101**, 87–92 (2004).
5. G. Favrin, A. Irbäck, and S. Mohanty, *Oligomerization of amyloid A β_{16-22} peptides using hydrogen bonds and hydrophobicity forces*, Biophys. J. **87**, 3657–3664 (2004).
6. A. Irbäck, S. Mitternacht, and S. Mohanty, *Dissecting the mechanical unfolding of ubiquitin*, Proc. Natl. Acad. Sci. USA **102**, 13427–13432 (2005).
7. A. Irbäck and S. Mitternacht, *Thermal versus mechanical unfolding of ubiquitin*, to appear in Proteins.
8. A. Irbäck, B. Samuelsson, F. Sjunnesson, and S. Wallin, *Thermodynamics of α - and β -structure formation in proteins*, Biophys. J. **85**, 1466–1473 (2003).
9. A. Irbäck and S. Mohanty, *Folding thermodynamics of peptides*, Biophys. J. **88**, 1560–1569 (2005).

10. M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, *Reversible unfolding of individual titin immunoglobulin domains by AFM*, Science **276**, 1109–1112 (1997).
11. M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamente, *Folding-unfolding transitions in single titin molecules characterized by laser tweezers*, Science **276**, 1112–1116 (1997).
12. A. Irbäck and S. Mohanty, *PROFASI: a Monte Carlo simulation package for protein folding and aggregation*, to appear in J. Comp. Chem.
13. M. Sunde and C. Blake, *The structure of amyloid fibrils by electron microscopy and X-ray diffraction*, Adv. Protein Chem. **50**, 123-159 (1997).
14. J. J. Balbach, Y. Ishii, O. N. Antzutkin, R. D. Leapman, N. W. Rizzo, F. Dyda, J. Reed, and R. Tycko, *Amyloid fibril formation by A β_{16-22} , a seven-residue fragment of the Alzheimer's β -amyloid peptide, and structural characterization by solid state NMR*, Biochemistry **39**, 13748-13759 (2000).
15. D. J. Gordon, J. J. Balbach, R. Tycko, and S. C. Meredith, *Increasing the amphiphilicity of an amyloidogenic peptide changes the β -sheet structure in the fibrils from antiparallel to parallel*, Biophys. J. **86**, 428-434 (2004).
16. B. Ma and R. Nussinov, *Stabilities and conformations of Alzheimer's β -amyloid peptide oligomers (A β_{16-22} , A β_{16-35} , and A β_{10-35}): sequence effects*, Proc. Natl. Acad. Sci. USA **99**, 14126-14131 (2002).
17. D. K Klimov and D. Thirumalai, *Dissecting the assembly of A β_{16-22} amyloid peptides into antiparallel β sheets*, Structure **11**, 295-307 (2003).
18. S. Santini, N. Mousseau, and P. Derreumaux, *In silico assembly of Alzheimer's A β_{16-22} peptide into β -sheets*, J. Am. Chem. Soc **126**, 11509–11516 (2004).
19. S. Gnanakaran, R. Nussinov, and A.E. García, *Atomic-level description of amyloid β -dimer formation*, J. Am. Chem. Soc. **128**, 2158–2159 (2006).
20. R. A. Sayle and E. J. Milner-White, *RasMol: biomolecular graphics for all*, Trends Biochem. Sci. **20**, 374–376 (1995).
21. For a review, see S. E. Jackson, *Ubiquitin: a small protein folding paradigm*, Org. Biomol. Chem. **4**, 1845–1853 (2006).
22. G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax, *Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute crystalline phase*, J. Am. Chem. Soc. **120**, 6836–6837 (1998).
23. M. Carrion-Vazquez, H. Li H, H. Lu, P. E. Marszalek, A. F. Oberhauser, and J. M. Fernandez, *The mechanical stability of ubiquitin is linkage dependent*, Nat. Struct. Biol. **10**, 738–743 (2003).
24. J. M. Fernandez and H. Li, *Force-clamp spectroscopy monitors the folding trajectory*, Science **303**, 1674–1678 (2004).
25. M. Schlierf, H. Li, and J. M. Fernandez, *The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques*, Proc. Natl. Acad. Sci. USA **101**, 7299–7304 (2004).
26. F. Cordier and S. Grzesiek, *Temperature-dependence of protein hydrogen bond properties as studied by high-resolution NMR*, J. Mol. Biol. **317**, 739–752 (2002).
27. H. S. Chung, M. Khalil, A. W. Smith, Z. Ganim, and A. Tokmakoff, *Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin*, Proc. Natl. Acad. Sci. USA **102**, 612–617 (2005).

Modeling Protein Structure, Dynamics and Thermodynamics with Reduced Representation of Conformational Space

Andrzej Kolinski, Dominik Gront, Sebastian Kmiecik,
Mateusz Kurcinski, and Dorota Latek

Laboratory of Theory of Biopolymers, Faculty of Chemistry
Warsaw University, Pasteura 1, 02-093 Warsaw, Poland
E-mail: kolinski@chem.uw.edu.pl

In this contribution we describe a successful approach to protein modeling which is based on reduced representation of protein conformational space, all-atom-refinement, evaluation and selection of the best molecular models. During the sixth CASP (Critical Assessment of protein Structure Prediction) community-wide experiment our methodology (referred further as CABS) proven to be one of the best performing methods for protein structure prediction, applied both for comparative modeling and to *de novo* folding. The newest applications of the CABS modeling technology include: study of protein folding thermodynamic, dynamics in the denatured state and folding pathways, structure prediction based on sparse and inaccurate experimental data and prediction of protein-protein interactions or flexible ligand docking. The CABS reduced model could be easily integrated with the all-atom approaches providing solid starting point for reliable multiscale simulations of large biomolecular systems.

1 Introduction

Due to the systematic sequencing of numerous genomes¹ the number of recognizable gene products, and therefore protein sequences, grows exponentially in the recent years. The number of known protein sequences exceeds 30 millions. Understanding of protein biological function, from enzymatic activity, through transport and signaling, to mechanisms and thermodynamics of complex macromolecular assemblies, requires knowledge of proteins' three dimensional structures. For many purposes it is also important to know protein folding mechanism and the dynamics near the native state, as well as the dynamics at denatured state. Understanding protein dynamics may be even more challenging than theoretical prediction of protein structure. Nevertheless, at the moment protein structure prediction seems to be the most urgent and the most advanced general goal of theoretical structural biology. It has an increasing impact on rational drug design, development of new biotechnologies, and on many other areas of biomedical sciences and technology.

Theoretical approaches are important regardless the great progress in experimental methods of protein structure determination. In spite of this progress the X-ray^{2,3} or NMR based methods for structure determination remain very costly and time-consuming⁴. As a result, the number of good-quality experimental structures, although impressive, is still very small in comparison with the number of known sequences, and is now in the range of 30 thousands. It is not quite clear if we know already the majority of possible protein folds, i.e. loosely defined distinct three-dimensional structures. This is an important issue, since the most advanced and the most accurate methods of theoretical

prediction of protein structures employ already determined structures as templates in the framework of comparative modeling techniques. Reasonable comparative modeling can be now done for some 50-60% of newly identified proteins, depending on a genome⁵. For the remaining proteins it is difficult, or even impossible, to find structural templates and different approaches need to be applied. In such cases various *de novo* modeling techniques could be helpful. A number of distinct approaches to the *de novo* structure prediction have been recently proposed and evaluated. In general, the most advanced techniques are quite successful when applied to relatively simple and small proteins, although dependable structure prediction of larger proteins remains elusive.

In principle, the simplest and the most straightforward approach to the *de novo* folding would be to use a detailed all-atom-representation of polypeptide chain immersed in a proper number of water molecules with a proper number of ions around. Such systems could be sampled using Molecular Dynamics or Monte Carlo techniques. This is unfortunately still limited to very small proteins or peptides. The conformational space of proteins is enormous, and in reality a typical folding time of globular proteins (in vivo as well as in vitro) is range of milliseconds to minutes. This time frame is few orders of magnitude too large for the contemporary computers. Thus, the geometrical representation of proteins, their interactions and sampling techniques need to be simplified in order to make the task feasible. Up to the date the two distinct, reduced-space, approaches to the *de novo* folding simulations of proteins seem to be the most successful. The first one, adopted in the Rosetta model of Baker and collaborators^{6,7} employs short structural fragments excised from the known protein structures. These fragments are 3-9 residues long and their representation in the algorithm is limited to the main chain and single united atoms for the side groups. During the sampling process the model chains undergoes a long series of deletion/insertions of such building blocks. The acceptance criteria for these modifications include short range geometrical restraints, a simplified model of side chain interactions and a model of main chain hydrogen bonds. The final models are subject to a clustering procedure, rebuilding of the atomic details and final selection of the best models. A different class of models assumes even a more simplified representation of the main chain backbone, where only the alpha carbons are treated in an explicit fashion, although the sampling process allows for a broader range of conformations, controlled by properly designed force fields. Good example of such model of protein folding is UNRES, designed by Scheraga and coworkers^{8,9}. The UNRES force field is carefully designed and optimized, and is derived from basic physical principles. The conformational space of UNRES is continuous. In order to further speed-up the sampling process discretized space models have been developed. Below, we briefly describe the CABS model¹⁰, which besides the grid-type representation employs knowledge-based force field derived from statistical analysis of structural regularities observed in known proteins. Rosetta and CABS based methods proven to be among the most successful^{11,12} during the last round of the CASP (Critical Assessment of Protein Structure Prediction) community-wide experiment. The summary of the CASP results can be found in our homepage <http://biocomp.chem.uw.edu.pl> or at the CASP homepage.

The reduced representation of the CABS model¹⁰ employs up to four interaction centers per residue (alpha carbon, beta carbon, the center of mass of the side group

and the center of peptide bond). Carefully designed, tuned and tested force field¹³ of CABS consists of several potentials of mean-force derived from statistical analysis of structural regularities seen in known protein structures. The sampling of conformational space of the model proteins employs various variants of the Monte Carlo method, including very efficient muticopy simulated tempering algorithms^{14–16}. A number of supplementary bioinformatics tools have been developed to handle data processing and analysis of the large scale simulations of protein systems using the CABS modeling system. CABS methodology proven to be one of the best performing methods for protein structure prediction, from comparative modeling to de novo folding. It has been clearly demonstrated during the sixth CASP (Critical Assessment of protein Structure Prediction) community-wide experiment. The groups employing the CABS-based methodology scored among the best. The design of the CABS model has been recently described in great detail¹⁰. In the next section, the most typical recent applications in protein structure prediction, refinement of NMR data and study of protein dynamics and interactions are presented.

2 Structure Prediction Based on Sparse NMR Data

NMR based protein structure determination is a time consuming and costly process. However, some sparse NMR data as Chemical Shifts (CS), Residual Dipolar Coupling (RDC) and some sparse NOE's are relatively easy to obtain at early stages of structure determination process. Such data are usually insufficient for molecular model building using the standard computational procedures. The restraints derived from sparse experimental data (NMR in this case) are easy to implement in the CABS modeling tool. The first applications focused on the CS and RDC data¹⁷. The CS data provide loose restrictions on a fraction of Φ , Ψ angles of the protein backbone. Due to the reduced $C\alpha$ representation of the main chain in the CABS algorithm it was necessary to develop a translating procedure, where the Φ , Ψ angles are transformed into θ , γ angles of the $C\alpha$ -trace. θ is the planar angle of the trace, while γ is the dihedral angle defined by three consecutive $C\alpha$ pseudobonds (Figure 1a). The ranges of experimental inaccuracies need also to be properly translated. The θ , γ restraints are easy to implement in the CABS simulations controlled by the simulated tempering MC scheme. It has been shown that for not too-complex structures CS experimental data are often sufficient for a moderate resolution structure prediction. It has been also demonstrated that a proper combination of the CS data with the artificial intelligence based secondary structure predictions significantly increases the accuracy of the generated molecular models and increase the range of applicability of the method - larger proteins could be modeled. Using simulated (extracted from known PDB structures) CS data it was demonstrated that increase of the fraction of assigned phi-psi angles from around 60% to some 80% improves qualitatively the accuracy of the obtained molecular models. It points onto a possible way of improving and speeding-up the NMR based structure prediction procedures. Increasing the accuracy of the CS measurements or/and supplementing the CS data with the local NOE's (or other measurements of the short range angular correlations) would be extremely beneficial for a fast computational model building with the CABS algorithm.

While the CS data are of the local nature the RDC data provide global, although very

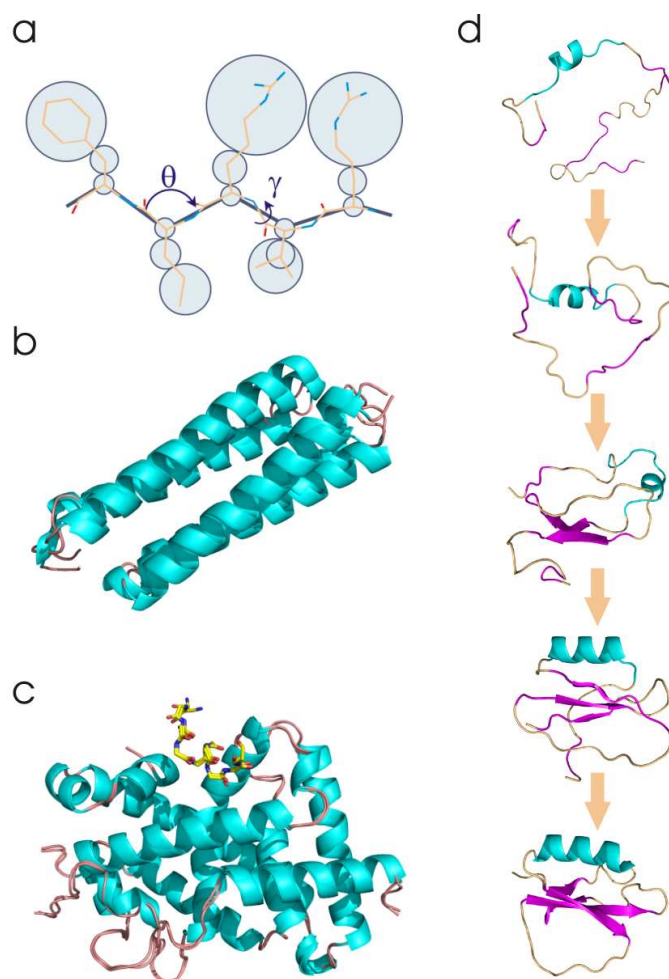


Figure 1. **(a)** Translation from Φ , Ψ angles (obtained from NMR chemical shifts) to the CABS representation (θ , γ angles). **(b)** De-novo assembly of ROP homodimer: theoretical model superimposed on NMR structure (pdb code: 1RPR) with cRMSD equal to 2.97 Å. **(c)** Predicted model of vitamin D receptor bound with short peptide co-activator fragment, superimposed on crystallographic structure (pdb code: 1RJK). Accuracy of obtained model is 0.61 Å. **(d)** Example snapshots from folding simulation, illustrating the folding pathway of chymotrypsin inhibitor 2. The portions of the protein that form β -strands in the native state are highlighted in magenta, the α helix is colored in cyan.

inaccurate, restraints. The RDC data need to be also properly translated onto the CABS geometry. Preliminary computational studies show that the RDC based restraints are very useful in guiding the simulations for more complex structures - they provide a bias towards the right fold topology. The local details are tuned by the CABS force field, eventually supplemented by the CS-based restraints. Finally, even very small number of the long range restraints from NOE's (range of $N/12$, where N is the number residues in the chain) allows for the proper fold assembly even for quite complex structures.

In summary, CABS based modeling could be a powerful tool for structure determination from sparse (and of low accuracy) experimental data. It could be also used as a part of the full process of structure determination using the NMR techniques. Namely, the initial data (CS, RDC, etc.) could be used for building an initial model, or a number of alternative models. These models could be subsequently used as a guide in the assignments of the NOE signals.

3 Flexible Docking with CABS

At present, the CABS force field is applicable only to proteins and peptides. This allows studies of protein complexes and docking of peptide ligands (co-factors, etc) to proteins. The docking procedure can be performed in various regimes; from completely unrestrained simulations of two or more polypeptide (peptide) chains to simulations with partially restrained internal coordinates of the interacting macromolecules. It has been shown that unrestrained simulations of protein dimers (GCN4 leucine zippers, ROP-dimer, etc) lead to a rapid assembly of the correct dimeric structures (Figure 1b). Weak restraints superimposed onto internal coordinates of the interacting molecules lead to a higher resolution of the obtained assemblies.

A number of test docking experiments have been performed assuming a limited flexibility of a receptor and the full conformational mobility of the interacting peptides²¹. No knowledge of the binding sites was assumed. At the beginning of simulations peptides were placed at random position at a large distance from the receptors. In all studied cases the peptides attached correctly to various nuclear receptors (Figure 1c). In about half of cases the accuracy was range of 1 Å RMSD, as measured for the alpha carbons of the peptide after the best superposition of the model receptor structure onto its crystallographic structure. In other cases the resolution was range of 2-2.5 Å. This implies, that always the pattern of the side chain interactions was correctly predicted. Thus, it has been demonstrated that the new modeling tool is already capable of producing correct high-resolution structures of protein - peptide complexes and of proper assembly of protein mutimers (in the last case the process is computationally expensive for larger structures). This should be very important for better understanding of protein interactions, signaling pathways and for computer aided design of new drugs.

4 Modeling of Protein Dynamics and Folding Pathways

The force field of CABS has been designed basing on statistical analysis of structural regularities observed in known protein structures^{10,13}. Thus, it may appear that the model is applicable only to structure prediction, but not to study of protein dynamics in the denatured state, or modeling the folding mechanisms. This is however not the case. Apparently, the nature of the mean field interactions in the denatured state is very similar to these in the compact native structures. For a number of small proteins a detailed data describing the nature of the denatured state and the folding mechanism are available. Using the CABS model isothermal simulations just above the folding temperature were performed and the degree of exposure to the solvent for all residues computed from the resulting trajectories¹⁹. These correlated extremely well with the measurements of the protection factors for

the burst intermediates. The unfolded model proteins exhibited the features of the molten globule state²⁰, believed to be a common folding intermediate for globular proteins. Interestingly, the most frequently observed long range contacts in the simulated denatured conditions overlapped with the folding nuclei observed in experiments. Thus, it seems to be safe to conclude, that the CABS model could be used not only in studies of protein folded structures, but also in semiquantitative modeling of protein dynamics, folding pathways (Figure 1d) and mechanism of macromolecular assembly.

5 Conclusions

CABS is a high resolution, lattice based model of protein structure and protein stochastic dynamics. The force field of the model consists of several statistical potentials of mean force, derived from the regularities seen in the known protein structures. The solvent in this force field is treated in an implicit fashion. Due to computational speed CABS can be used in a large scale protein modeling (large scale in respect to the size of the modeled system, as well as in respect to the number of proteins that could be structurally annotated in a reasonable time). Recent applications include protein structure prediction (*de novo*, as well as supported by sparse experimental data), modeling of protein interactions in macromolecular assemblies and study of protein dynamics and folding mechanisms. Finally, it should be noted that CABS is compatible with the classical all-atom modeling tools. The spatial accuracy of the CABS models is sufficient for a meaningful reconstruction of the atomic details. Thus multiscale simulations at various levels of resolution become feasible. Future extensions of the model will include interactions with non-peptide ligands, membranes and nucleic acids.

Acknowledgments

This work was partially supported by the grant # PBZ-KBN-088/P04/2003

References

1. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *haemophilus influenzae* rd. *Science*, 269(5223):496–512, July 1995.
2. W. A. Hendrickson. Synchrotron crystallography. *Trends Biochem Sci*, 25(12):637–643, December 2000.
3. A. Schmidt and V. S. Lamzin. Veni, vidi, vici - atomic resolution unravelling the mysteries of protein function. *Curr Opin Struct Biol*, 12(6):698–703, December 2002.
4. D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nat Struct Biol*, 8(6):559–566, June 2001.
5. M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

6. C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3):565–577, August 1998.
7. C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods Enzymol*, 383:66–93, 2004.
8. A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry*, 18(7):849–873, 1997.
9. A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *The Journal of Chemical Physics*, 115(5):2323–2347, 2001.
10. A. Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*, 51(2):349–371, 2004.
11. J. Moult. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol*, June 2005.
12. A. Kolinski and Janusz M. M. Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, September 2005.
13. D. Gront and A. Kolinski. A new approach to prediction of short-range conformational propensities in proteins. *Bioinformatics*, 21(7):981–987, April 2005.
14. C. J. Geyer. Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*, pages 156–163. Fairfax Station, 1991.
15. U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules, *Chem. Phys. Lett*, 281:140 1997.
16. D. Gront, A. Kolinski, and J. Skolnick. Comparison of three monte carlo conformational search strategies for a proteinlike homopolymer model: Folding thermodynamics and identification of low-energy structures. *The Journal of Chemical Physics*, 113(12):5065–5071, 2000.
17. D. Plewczynska and A. Kolinski. Protein folding with a reduced model and inaccurate short-range restraints. *Macromolecular Theory and Simulations*, 14(7):444–451, 2005.
18. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, January 2000.
19. S. Kmiecik, M. Kurcinski, A. Rutkowska, D. Gront, and A. Kolinski. Denatured proteins and early folding intermediates simulated in a reduced conformational space. *Acta Biochim Pol*, December 2005.
20. D. Ekonomiuk, M. Kielbasinski, and A. Kolinski. Protein modeling with reduced representation: statistical potentials and protein folding mechanism. *Acta Biochim Pol.*, 52(4):741–748, 2005.
21. M. Kurcinski and A. Kolinski. Steps towards flexible docking: Modeling of three-dimensional structures of the nuclear receptors bound with peptide ligands mimicking co-activators’ sequences. *Journal of Steroid Biochemistry and Molecular Biology*, (submitted)

A High-Speed Special-Purpose Computer for Molecular Dynamics Simulations: MDGRAPE-3

**Tetsu Narumi, Yousuke Ohno, Noriyuki Futatsugi, Noriaki Okimoto,
Atsushi Suenaga, Ryoko Yanai, and Makoto Taiji**

High Performance Molecular Simulation Team,
Computational and Experimental Systems Biology Group,
Genomic Sciences Center, RIKEN
61-1-1F, Ono-cho, Tsurumi, Yokohama, Kanagawa, 230-0046, JAPAN
E-mail: narumi@gsc.riken.jp

We developed a high-speed computer, MDGRAPE-3, which is dedicated to perform molecular dynamics simulations. It is a special-purpose computer for force calculation between particles and has become the first Pflops machine. Coulomb and van der Waals force calculations are accelerated by the MDGRAPE-3, and the other calculations are performed on a usual general-purpose computer. By using this architecture, we can get higher performance and 10 times lower cost, size, and power consumption per performance than those of general-purpose computers. MDGRAPE enables us to perform precise molecular dynamics simulations, because explicit water molecules and non-cutoff Coulomb interaction can be efficiently accelerated by this hardware.

1 Introduction

One of the main target of the recent supercomputers is molecular dynamics (MD) simulations, which require huge computational power. To satisfy this demand, cluster of computers is often used because its cost per performance is very low. Even the high-end supercomputers are made of commodity based PCs. For example, 20-25% of the top 50 supercomputers are PC clusters¹. However, PC clusters have many disadvantages; too big and power consuming. One of the reasons why only one PC cluster is in the top 10 supercomputers is because of its size and power consumption level.

One approach is to develop a computer which is specially packed in a small form factor. IBM has developed the Blue Gene/L², which is the fastest in the world for the general-purpose computer with relatively small and low power consumption.

We took another approach to reach the highest performance: Special-purpose computer for particle simulations. GRAPE³⁻⁵ is a series of special-purpose computers for astrophysical N -body and MD simulations, etc. MDGRAPE-3 is the latest computer for MD simulations⁶. MDGRAPE-3 becomes the first Pflops machine, and its cost, size and power consumption per performance is 10 times smaller than those of any other computers.

In this paper, hardware, software, and performance of MDGRAPE-3 system are presented as well as the short report of MD simulations with our hardware. Hardware and software of MDGRAPE-3 system are shown in Secs. 2 and 3 respectively. Performance is presented in Sec. 4. MD simulation results and conclusion are described in Secs. 5 and 6 respectively.

2 Hardware of MDGRAPE-3 System

The GRAPE is a series of application specific processor designs, which is specially built to accelerate the inter-particle force calculation. The GRAPE was originally developed for accelerating the gravitational interactions, while MD simulations are also supported in the case of GRAPE-2A⁷, MD-GRAPE^{8,9} and MDM (MDGRAPE-2)^{10–12}. The main difference in hardware which supports MD and which does not is that the arbitrary central force can be calculated with MD-supported machines instead of the gravitational force.

Machines other than GRAPE have been developed with similar idea: such as Delft Molecular Dynamics Processor¹³, ATOMS¹⁴, FASTRUN¹⁵ and so on. However, these machines were not so successful compared with GRAPE because their hardware were complicated. MDGRAPE is simple because it only accelerates the long range interactions (Coulomb and van der Waals). Other calculations are performed in the host computer which is connected to MDGRAPE. In MD simulations, long range interactions dominate the total calculation. Therefore, the simple hardware can accelerate MD simulations more effectively.

To reach high performance, we developed a processor, MDGRAPE-3 chip, which is dedicated to MD simulations. It has 180 or 216 Gflops equivalent performance with the clock frequency of 250MHz or 300MHz when the effective number of floating-point operations of a pairwise force calculation is assumed to be 36. In this assumption, we counted division and square root operations as ten floating-point operations respectively. There are two main reasons why the MDGRAPE-3 chip has high performance. One reason is that MDGRAPE-3 chip has 20 special pipelines, each of which calculates pairwise interaction between atoms. Many hundreds of arithmetic units and their data path in the pipelines are hard wired. Therefore, all the arithmetic units work every clock cycle basically, and it makes highly efficient calculation possible. On the other hand, conventional CPU, such as Pentium, can perform only several operations per clock cycle. Another reason is that all the pipelines can share one memory due to the parallelism of particle interaction. This lower bandwidth of memory enables simpler hardware, which means it is easy to parallelize to get higher performance.

Figure 1(left) shows the MDGRAPE-3 board which has 12 MDGRAPE-3 chips. The memory which stores particle positions etc. is integrated in the MDGRAPE-3 chip, and

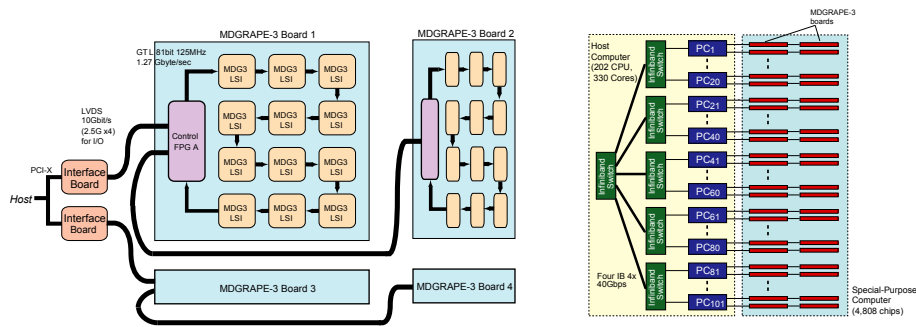


Figure 1. Block diagram of the MDGRAPE-3 board, which has 12 MDGRAPE-3 chips connected in a daisy-chain and an FPGA (left). Block diagram of the MDGRAPE-3 system, which is consisted of the PC cluster with 101 nodes and 402 MDGRAPE-3 boards (right).

System	Peak speed (Gflops)	Cost/speed (\$/Gflops)	Power/speed (Watt/Gflops)	Size/speed (liter/Gflops)
MDGRAPE-3	1,000,000	<u>9</u>	<u>0.2</u>	<u>0.03</u>
Blue Gene/L	360,000	280	4	0.3
PC (Pentium D 3.0GHz)	12	80	25	3

Table 1. Advantages of MDGRAPE-3 against general-purpose computers. The MDGRAPE-3 is superior to general-purpose computers in all of the three points which are cost/speed, power/speed, and size/speed as underlined.

the board is also simple. It is connected to the host computer via interface board. Figure 1(right) shows the total system. It is composed of a PC cluster with 101 nodes, each of them has four MDGRAPE-3 boards. Total number of CPU cores is 330, because we used 64 nodes of dual Xeon 5150 and 37 nodes of dual Xeon 3.2D GHz CPUs.

MDGRAPE-3 has many advantages compared to a conventional computer. Table 1 compares special- and general-purpose computers in four points: Peak speed, cost per speed, power consumption per speed, and size per speed. We chose Blue Gene/L and a PC for general-purpose computers, because Blue Gene/L is the fastest supercomputer, one of the lowest power, and smallest size computer per performance, and PC is one of the most cost effective computer. MDGRAPE-3 is superior to general-purpose computers in all four categories above. It has 10 times lower cost, size, and power consumption per performance than those of general-purpose computers.

3 Software of MDGRAPE-3 System

To get the benefit of MDGRAPE-3, the application program must be modified to fit our hardware. We have already ported AMBER¹⁶ and CHARMM¹⁷ programs, though some limitations exist. Users just need to add keyword to use MDGRAPE-3 in input files.

Other application can be accelerated with our APIs (Application Programming Interfaces). Table 2 is one of the APIs to calculate the Coulomb force between atoms. Software for the MDGRAPE-3 can be downloaded from the web site¹⁸. We support C and Fortran APIs and Linux operating system.

```
void MR3calccoulomb(double x[], int n, double q[],double rscale, int tblno,
                   double size, int pflag, int flag, double force[])
x ----- Positions of atoms.   n ----- Number of atoms.
q ----- Charges of atoms.     tblno - Potential type between atoms.
xmax --- Simulation box size.   flag -- Special flag to control calculation.
rscale - Alpha parameter for real-space of Coulomb of the Ewald method.
pflag -- Flag to specify periodic boundary condition or not.
force -- Calculated forces (output).
```

Table 2. Example of APIs in C to calculate the Coulomb force between atoms

Number of atoms	3,339	4,953	13,101
Without MDGRAPE-3 card (sec)	346.73	844.52	9231.70
With MDGRAPE-3 card (sec)	6.38	7.33	34.31
Acceleration	54	115	269

Table 3. Performance of the MDGRAPE-3 card

4 Performance of MDGRAPE-3 System

Table 3 summarizes the performance of MDGRAPE-3. We used 2-chip version of MDGRAPE-3 card, which is commercially available from the venture company¹⁹. We compared the calculation time of the 1,000 steps of 'sander' program in AMBER8 with and without MDGRAPE-3 card. The calculated system has a Scytalone dehydratase (2,715 atoms) and water molecules. Coulomb and van der Waals interactions are calculated without cutoff under free boundary condition. Pentium D 3.0 GHz (only single core is used) and Pentium 3.2 GHz CPUs are used with and without MDGRAPE-3 calculation respectively. We got more than 100 times acceleration.

For PME (Particle Mesh Ewald) method simulation, the acceleration is not so high because the wavenumber-space part of the Coulomb calculation becomes the bottleneck. FFT (Fast Fourier Transform), which is used for wavenumber-space, cannot be accelerated with MDGRAPE-3. With newer method such as IPS²⁰ (Isotropic Periodic Sum), we will be able to accelerate more efficiently under periodic boundary condition in the near future.

5 MD Simulation Results

In this section, we briefly present MD simulation results with previous MDGRAPE-2 hardware. First example is RNA polymerase II²² which is the fundamental and important enzyme in the transcription process. As pre-translocation state has not yet been determined by the X-ray crystallographic study, we first constructed a model structure of the Pol II elongation complex with the 50 base pairs of DNA-24 bases of RNA, including the unwound bubble of DNA. Then we got its reliable structure by 2 ns of MD simulation [Fig.

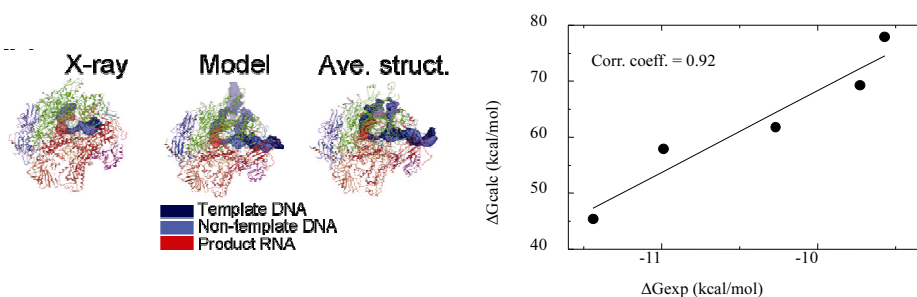


Figure 2. X-ray²¹, initial model, and average structures of the Pol II elongation complex are shown from left to right (left). Correlation between calculated and experimental binding free energy between SH2 domain of growth factor receptor binding protein 2 and ErbB receptor-derived phosphotyrosyl peptides (right).

2(left)]. It was a big system with 450 k atoms in total, and MDGRAPE hardware enabled us to perform accurate simulation with explicit water molecules and non-cutoff of Coulomb and van der Waals interaction. We also performed a short simulation with a targeted MD simulation approach²³ and found that the conformational change of a loop in the Pol II, fork loop 1, couples with the unidirectional movement of the Pol II along DNA.

Second example is free energy calculation between the SH2 domain of growth factor receptor binding protein 2, Grb2, and ErbB receptor-derived phosphotyrosyl peptides²⁴. Binding free energies for nine phosphotyrosyl peptides were calculated by using the MM-PBSA (Molecular Mechanics Poisson-Boltzmann Surface Area) continuum solvent method²⁵, and we got a good correlation coefficient of 0.92 between simulated and experimental data of surface plasmon resonance analysis [Fig. 2(right)]. Such precise binding free energy calculation would be useful for *in silico* drug screening because conventional docking programs tend to be inaccurate because they often treat proteins as rigid and water molecules as implicit models. With full MDGRAPE-3 system, over 2,000 of simulations, each of which is 2 ns long and has 13 k atoms, can be performed per day.

6 Concluding Remarks

We have described the MDGRAPE-3, special-purpose computer for MD simulations. It is the first Pflops computer, and has 10 times lower cost, size, and power consumption per performance than those of general-purpose computers. Our architecture of special-purpose computer would be promising approach in the near future because power consumption and size are becoming more and more important in conventional processors. We can perform precise MD simulations with our hardware because it efficiently accelerates the simulation with explicit water molecules and non-cutoff interactions.

Acknowledgments

This work was partially supported by Research Revolution 2002 “Protein 3000 Project”, operated by the Ministry of Education, Culture, Sports, Science and Technology, Japan. Authors thank Drs Yoshiyuki Sakaki, Tomoya Ogawa, Shigeyuki Yokoyama, and Toshikazu Ebisuzaki for their support to this project. Authors also thank Japan SGI Inc. for the system support. In particular, we would like to thank Masahiro Yamaguchi, Hiroshige Tatsu, and Junya Shinohara for their help with the system installation. Authors thank Intel Corporation for providing their early processors and system support for them.

References

1. <http://www.top500.org/list/2005/11/100>
2. <http://www.research.ibm.com/bluegene/index.html>
3. D. Sugimoto, Y. Chikada, J. Makino, T. Ito, T. Ebisuzaki, and M. Umemura, *A Special Purpose Computer for Gravitational Many-body Problems*, Nature **345**, 33-35 (1990).
4. J. Makino and M. Taiji, *Scientific Simulations with Special-Purpose Computers: The Grape Systems*, John Wiley & Sons, (1998).

5. P. Hut, and J. Makino, *Computational Physics - Astrophysics on the GRAPE Family of Special-Purpose Computers*, Science **283**, 5401 (501-505)1999.
6. M. Taiji, T. Narumi, Y. Ohno, N. Futatsugi, A. Suenaga, N. Takada, and A. Kona-gaya, *Protein Explorer: A Petaflops Special-Purpose Computer System for Molecular Dynamics Simulations*, in Proceedings of SC2003, CDROM, Phoenix, (2003).
7. T. Ito, J. Makino, T. Fukushige, T. Ebisuzaki, S. K. Okumura, and D. Sugimoto, *A Special-Purpose Computer for N-Body Simulations: GRAPE-2A*, Publications of the Astronomical Society of Japan **45**, 339-348 (1993).
8. M. Taiji, T. Fukushige, J. Makino, T. Ebisuzaki, and D. Sugimoto, *MD-GRAPE: A Parallel Special-Purpose Computer System for Classical Molecular Dynamics Simulations*, in Proceedings of the 6th Joint EPS-APS international conference on Physics Computing, Geneva, 200-203 (1994).
9. T. Fukushige, M. Taiji, J. Makino, T. Ebisuzaki, and D. Sugimoto, *A Highly-Parallelized Special-Purpose Computer for Many-body Simulations with an Arbitrary Central Force: MD-GRAPE*, Astrophys. J. **468**, 51-61 (1996).
10. T. Narumi, R. Susukita, T. Ebisuzaki, G. McNiven, and B. Elmegreen, *Molecular Dynamics Machine: Special-purpose Computer for Molecular Dynamics Simulations*, Mol. Sim. **21**, 401-415 (1999).
11. T. Narumi, A. Kawai, and T. Koishi, *An 8.61 Tflop/s Molecular Dynamics Simulation for NaCl with a Special-Purpose Computer: MDM*, in Proceedings of SC2001, CDROM, Denver, (2001).
12. R. Susukita, T. Ebisuzaki, B. G. Elmegreen, H. Furusawa, K. Kato, A. Kawai, Y. Kobayashi, T. Koishi, G. D. McNiven, T. Narumi, and K. Yasuoka, *Hardware Accelerator for Molecular Dynamics: MDGRAPE-2*, Computer Physics Communications **155**, 115-131 (2003).
13. A. F. Bakker and C. Bruin, in *Special Purpose Computers*, edited by B. J. Alder (Academic Press, San Diego), 183-232, (1988).
14. A. F. Bakker, G. H. Gilmer, M. H. Grabow, and K. Thompson, *A Special Purpose Computer for Molecular Dynamics Calculations*, J. Comput. Phys. **90**, 313-335 (1990).
15. R. Fine, G. Dimmler, and C. Levinthal, *FASTRUN: A Special Purpose, Hardwired Computer for Molecular Simulation*, Proteins **11**, 242-253 (1991).
16. D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, B. Wang, D. A. Pearlman., M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. W. Caldwell, W. S. Ross, P. A. Kollman, Amber 8.0, University of California San Francisco 2004.
17. B. Brooks, R. Brucoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus, *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*, Comp. Chem. **4**, 87-217 (1983).
18. <http://mdgrape.gsc.riken.jp>
19. <http://www.peta.co.jp>
20. Xiongwu Wu and Bernard R. Brooks, *Isotropic periodic sum: A method for the calculation of long-range interactions*, J. Chem. Phys. **122**, 044107 (2005).
21. H. Kettenberger, K.-J. Armache, and P. Cramer, *Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS*, Mol. Cell **22**, 955-965 (2004).

22. A. Suenaga, N. Okimoto, N. Futatsugi, Y. Hirano, T. Narumi, Y. Ohno, R. Yanai, T. Hirokawa, T. Ebisuzaki, A. Konagaya, and M. Taiji, *Structure and dynamics of RNA polymerase II elongation complex*, Biochemical and Biophysical Research Communications **343**, 90-98 (2006).
23. C. Guilbert, D. Perahia, L.A. Mouawad, *A method to explore transition paths in macromolecules applications to hemoglobin and phosphoglycerate kinase*, Comput. Phys. Comm. **91**, 263-273 (1995).
24. A. Suenaga, M. Hatakeyama, M. Ichikawa, X. M. Yu, N. Futatsugi, T. Narumi, K. Fukui, T. Terada, M. Taiji, M. Shirouzu, S. Yokoyama, and A. Konagaya, *Molecular Dynamics, Free Energy, and SPR Analyses of the Interactions Between the SH2 Domain of Grb2 and ErbB Phosphotyrosyl Peptides*, Biochemistry **42**, 5195-5200 (2003).
25. P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Hou, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham III, *Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models*, Acc. Chem. Res. **33**, 889-897 (2000).

The Two Aspects of the Protein Folding Problem

Harold A. Scheraga¹, Adam Liwo^{1,2}, Stanislaw Ołdziej^{1,2}, Cezary Czaplewski^{1,2},
Mey Khalili¹, Jorge A. Vila^{1,3}, and Daniel R. Ripoll⁴

¹ Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca, New York 14853-1301
E-mail: has5@cornell.edu

² Faculty of Chemistry, University of Gdansk,
80-952 Gdansk, Poland

³ Universidad Nacional de San Luis, IMASL-CONICET,
Ejército de los Andes 950, (5700) San Luis, Argentina

⁴ Computational Biology Service Unit-Cornell Theory Center,
Cornell University, Ithaca, New York 14853

A physics-based approach to the protein folding problem is presented. It is concerned with the computation of folding pathways and final native structures, given the amino acid sequences, an empirical all-atom potential energy function, and a procedure to identify the global minimum of the potential energy. Whereas the all-atom approach has provided three-dimensional structures of relatively small molecules and for helical proteins containing up to 46 residues, it has been necessary to develop a hierarchical approach to treat larger proteins. In the hierarchical approach, global optimization was originally carried out with a simplified united residue (UNRES) description of a polypeptide chain to locate the *region* in which the global minimum lies. Conversion of the UNRES structures in this region to all-atom structures is followed by a local search in this region. The performance of this physics-based approach in successive CASP blind tests for predicting protein structure is described. More recently, a molecular dynamics treatment with UNRES has been introduced to compute not only native structures but also folding pathways.

1 Introduction

Ever since Anfinsen¹ demonstrated that a polypeptide chain can fold spontaneously into the three-dimensional structure of a native protein, experimental and theoretical chemists have tried to determine the interactions that govern the folding process (the protein-folding problem). Actually, there are two protein-folding problems: (a) determination of the folding pathways, and (b) identification of the folded native structure corresponding to the global minimum of the potential energy according to Anfinsen's thermodynamic hypothesis¹ that the native structure is the thermodynamically most stable one.

Early experimental efforts² were devoted to identifying inter-residue interactions in the native structure, and subsequent experiments identified multiple folding pathways^{3,4}. Early theoretical approaches were based on the use of empirical interatomic potential energy functions⁵ together with a menu of procedures⁶ for global optimization of the potential energy. Some of these early efforts are summarized in reference 7. This article is devoted to recent theoretical work on the two protein-folding problems.

While the use of all-atom potential energy functions led to success for relatively small proteins, containing up to 20 residues⁷, it was only after the development of large-scale

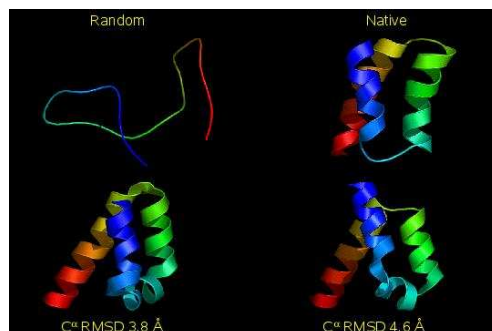


Figure 1. Results of all-atom calculations on protein A⁹, starting from a randomly-generated structure. The native structure and the lowest-energy structures obtained with two different hydration models^{12,13} are also illustrated.

parallel-processor computers⁸ that it has been possible to carry out computations with all-atom potentials for larger proteins, the largest one treated thus far being the 46-residue staphylococcal protein A⁹.

2 All-Atom Calculations on Protein A

Calculations on protein A⁹ were carried out with the empirical potential function ECEPP/3¹⁰ and the electrostatically-driven Monte Carlo (EDMC) procedure¹¹, together with two implicit hydration models, OONS¹² and SRFOPT¹³, starting from four different random conformations, one of which is illustrated schematically⁹ in Figure 1. Three of the four runs converged to the same native-like fold illustrated in Figure 1 for each of the two hydration models; the fourth converged to the mirror-image conformation.

Protein A is larger than the 36-residue α -helical protein from the villin headpiece, for which all-atom simulations, starting from an extended structure, were previously carried out by other groups^{14,15}. Those simulations were carried out with explicit solvent, which increases the computing time considerably compared to the time required for the implicit solvent models used in our simulations⁹.

It is not yet clear what the largest size protein is that can be treated by our all-atom EDMC procedure. Nevertheless, it is encouraging that an all-atom representation of the chain, and global optimization of the corresponding potential energy, can identify the native-like fold without resorting to knowledge-based information in the search procedure.

3 Hierarchical Procedure to Predict Protein Structure

Without waiting for further extensions of the EDMC procedure, we have developed a hierarchical procedure to treat larger proteins containing both α and β folded portions. In this procedure, global optimization is carried out by using a Conformational Space Annealing (CSA) method^{16,17} with a united-residue (UNRES) representation of the protein chain^{18–20}. This is the key stage of the hierarchical algorithm. It is designed to locate the *region* of the global minimum rapidly and efficiently. The lowest-energy structures

obtained from the UNRES representation in this stage are then converted to the all-atom representation^{21,22}, and a local search is carried out in the restricted region located with the UNRES/CSA approach. This is accomplished with the EDMC method and the ECEPP/3 force field¹⁰, together with the SRFOPT hydration model¹³. Initially, the backbone of the chain is constrained to the structures obtained by UNRES and CSA, but the constraints are gradually reduced as the calculations proceed.

The UNRES model^{18–20} consists of a virtual-bond chain, i.e., a sequence of α -carbons, united peptide groups, and united side chains represented by ellipsoids whose size depends on the nature of the amino acid residue. The α -carbons are not centers of interaction, but merely serve to locate the backbone. The centers of interaction are the united peptide groups and united side chains, with a united-residue potential given as the sum of interactions involving side chain-side chain, side chain-peptide, peptide-peptide, virtual-bond torsional, virtual-bond double torsional, virtual-bond-angle bending, internal side-chain motional, and multi-body (correlation) energies. The variables to change conformation are the angles between virtual bonds, the torsional angle for rotation about the virtual bonds, and the position angle and rotational angle of the side chains.

The CSA method^{16,17}, used to search conformational space, starts with an initial set of widely-spaced UNRES minima. CSA is based essentially on a build-up and genetic algorithm to force these minima to coalesce to the *region* of the global minimum. All UNRES minimum-energy conformations in the final coalesced clusters are converted to the all-atom representation^{21,22}, and the global optimization search is continued from these starting conformations with the EDMC procedure¹¹.

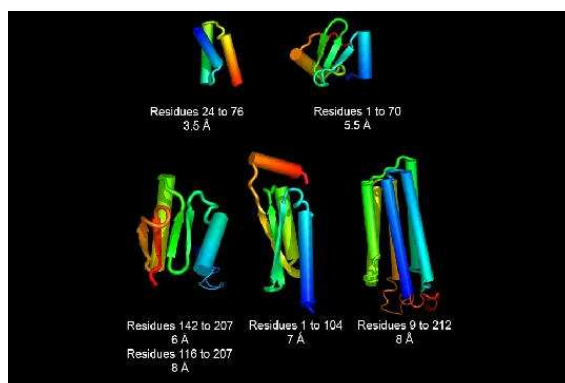


Figure 2. Results of blind CASP6 predictions.

4 CASP Results with Hierarchical Procedure

After carrying out initial tests of the hierarchical procedure on proteins of known structure, we participated in successive blind tests (CASP, Critical Assessment of Protein Structure Prediction), beginning with CASP3 in 1998. Various improvements of the procedure and

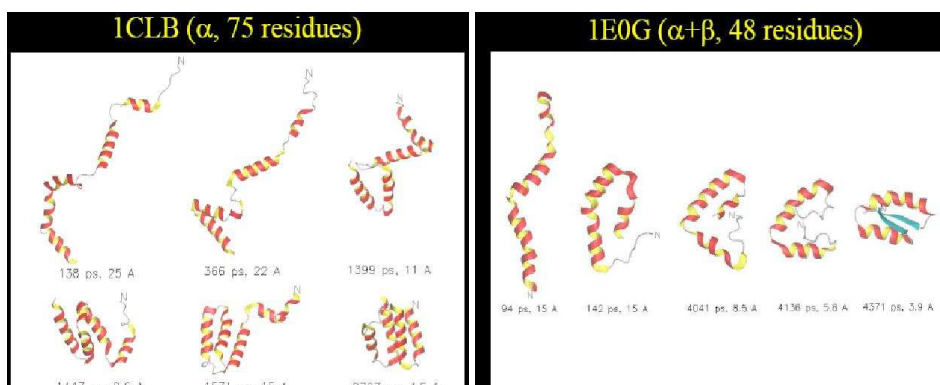


Figure 3. Right: Example of a fast-folding pathway of 1CLB obtained in Langevin dynamics simulations. The N-terminus of the chain is marked for tracing purposes. Left: Examples of a folding pathway of 1E0G obtained in Langevin dynamics simulations. The N-terminus of the chain is marked for tracing purposes.

its associated physics were implemented in successive tests. Some results from the most recent test (CASP6 in 2004) are illustrate in Figure 2.

5 Calculations of Folding Pathways

As pointed out in the Introduction (section 1), a second type of protein folding problem is the computation of the structural pathways by which the completely unfolded polypeptide chain proceeds to the folded native conformation, i.e., the progression from a given unfolded state (with no native contacts or native hydrogen bonds) to the final folded structure. One approach assumes that both the initial and final structures are known, and computes the folding pathways by used of the stochastic difference equation method of Elber *et al*²³, and has been applied with a full-atom treatment to protein A²⁴. A second approach, applied to protein A and larger proteins, makes use of the UNRES force field, in which the fast degrees of freedom are averaged out, and carries out Langevin molecular dynamics^{25–28}. The structure of protein A was obtained within an RMSD of 3.0Å. The largest protein to which th is second approach was applied is the 75-residue, α -helical protein 1CLB, and the structure was obtained within an RMSD of 4.5Å (see Figure 3(left)).

The structure of a 48-residue $\alpha + \beta$ protein, 1E0G, was obtained within an RMSD of 3.9Å (see Figure 3(right)). This molecular dynamics technique provides not only the final folded structure, but also the intermediate structures along the folding pathways. Further, by computing 400 trajectories to obtain good statistics, it was possible to compute the folding kinetics of protein A. As illustrated in Figure 4, the kinetics is either two-state (single exponential) or three-state, depending on the quantity that is computed (or measured experimentally), i.e., two-state for helix content (without considering interhelical interactions), or three-state (bi-exponential) if the RMSD is computed (which includes interhelical interactions).

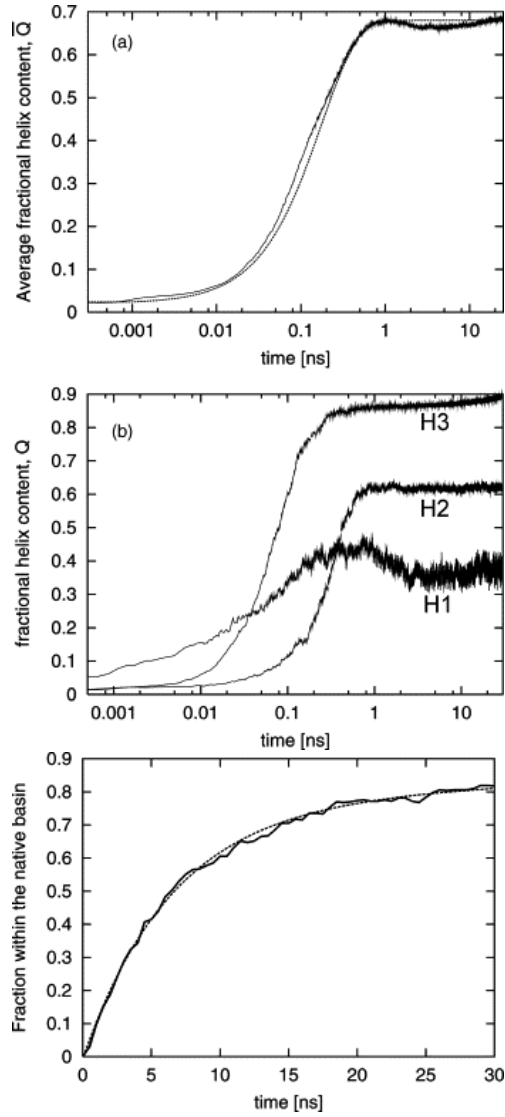


Figure 4. (a) Plot of the helix content of protein A ($\bar{Q}$), averaged over 400 trajectories (continuous line) and the exponential fit to $\bar{Q}$ (broken line). (b) Plots of the helix content of segments corresponding to helices H1, H2, and H3, from the N- to the C-terminus. (c) Plot of the fraction of native structure (secondary + tertiary) as a function of time (continuous line), and the fit of a bi-exponential function to these data (broken line).

6 Conclusions

The evolution of computational methodology has led from an all-atom treatment illustrated in Figure 1 to a hierarchical treatment of larger proteins (Figure 2) and to a longer-time molecular dynamics (MD) approach to treat not only the final folded structures but also

the intermediate structures and the folding kinetics illustrated in Figures 3, 4, respectively. Current work is focused on use of the UNRES/MD approach with improvement of the force field and inclusion of entropic effects.

Acknowledgements

This research was supported by grants from NIH (GM-14312, TW-7193, TW-6335) and NSF (MCB00-03722), and was carried out with resources of (a) our 392-double processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, and (d) the Cornell Theory Center which receives funding from Cornell University, New York State, Federal agencies foundations, and corporate partners.

References

1. C. B. Anfinsen *Principles that Govern Folding of Protein Chains*, Science **181**, 223–230 (1973).
2. H. A. Scheraga *Structural studies of pancreatic ribonuclease*, Fed. Proc. **26**, 1380-1387 (1967).
3. T. E. Creighton and D. P. Goldberg *Kinetic role of a meta-stable native-like two-disulfide species in the folding transition of bovine pancreatic trypsin inhibitor*, J. Mol. Biol. **179**, 497-526 (1984).
4. D. M. Rothwarf, Y. J. Li and H. A. Scheraga *Regeneration of bovine pancreatic ribonuclease A. Detailed kinetic analysis of two independent folding pathways*, Biochemistry **37**, 3767-3776 (1998).
5. H. A. Scheraga *Calculations of conformations of polypeptides*, Adv. Phys. Org. Chem. **6**, 103-184 (1968).
6. J. A. Vila, H. A. Baldoni and H. A. Scheraga *Position dependence of the ^{13}C chemical shifts of α -helical model peptides. Fingerprint of the 20 naturally occurring amino acids*, Protein Science **13**, 2939-2948 (2004).
7. H. A. Scheraga, A. Liwo, S. Ołdziej, C. Czaplewski, J. Pillardy, D.R. Ripoll, J.A. Vila, R. Kazmierkiewicz, J.A. Saunders, Y.A. Arnautova, A. Jagielski, M. Chinchio and M. Nanas *The protein folding problem: Global optimization of force fields*, Frontiers in Bioscience **9**, 3296-3323 (2004).
8. J. Lee, J. Pillardy, C. Czaplewski, Y. Arnautova, D. R. Ripoll, A. Liwo, K.D. Gibson, R.J. Wawak, and H. A. Scheraga *Efficient parallel algorithms in global optimization of potential energy functions*, Comput. Physics Commun. **128**, 399-411 (2000).
9. J. A. Vila, D. R. Ripoll and H. A. Scheraga *Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures*, Proc. Natl. Acad. Sci. U.S.A. **100**, 14812-14816 (2003).
10. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey and H. A. Scheraga *Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm*,

- with application to proline - containing peptides,
J. Phys. Chem. **96**, 6472-6484 (1992).
11. D.R. Ripoll, A. Liwo, and H. A. Scheraga *New Developments of the electrostatically driven Monte Carlo method: Test on the membrane-bound portion of melittin*, Biopolymers **46**, 117-126 (1998).
 12. T. Ooi, M. Oobatake, G. Némethy and H. A. Scheraga *Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides*, Proc. Natl. Acad. Sci., U.S.A. **84**, 3086-3090 (1987). Erratum: *ibid.* **84**, 6015 (1987).
 13. J. Vila, R. L. Williams, M. Vasquez and H. A. Scheraga *Empirical solvation models can be used to differentiate native from near - native conformations of bovine pancreatic trypsin inhibitor*, Proteins: Structure, Function, and Genetics **10**, 199-218 (1991).
 14. Y. Duan and P.A. Kollman *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution*, Science **282**, 740-744 (1998).
 15. B. Zagrovic, C.D. Snow, M.R. Shirts and V.J. Pande *Simulation of folding for a small α -helical protein in atomistic detail using worldwide-distributed computing* J. Mol. Biol. **323**, 927-937 (2002).
 16. J. Lee, H. A. Scheraga and S. Rackovsky *New optimization method for conformational energy calculations on polypeptides: Conformational space annealing*, J. Comput. Chem. **18**, 1222-1232 (1997).
 17. J. Lee and H.A. Scheraga *Conformational space annealing by parallel computations: Extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin*, Intl. J. of Quantum Chem. **75**, 255-265 (1999).
 18. A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky and H.A. Scheraga *A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data*, J. Comput. Chem. **18**, 849-873 (1997).
 19. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej and H.A. Scheraga *A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization*, J. Comput. Chem. **18**, 874-887 (1997).
 20. A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga *United-residue force field for off-lattice protein-structure simulations; III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials*, J. Comput. Chem. **19**, 259-276 (1998).
 21. R. Kazmierkiewicz, A. Liwo and H.A. Scheraga *Energy-based reconstruction of a protein backbone from its α -carbon trace by a Monte-Carlo method*, J. Comput. Chem. **23**, 715-723 (2002).
 22. R. Kazmierkiewicz, A. Liwo and H.A. Scheraga *Addition of side chains to a known backbone with defined side-chain centroids*, Biophys. Chem. **100**, 261-280 (2003). Erratum: Biophys. Chem. **106**, 91 (2003).

23. R. Elber, A. Ghosh and A. Càdenas *Long-time dynamics of complex systems*,
Acc. Chem. Res. **35**, 396-403 (2002).
24. A. Ghosh, R. Elber and H.A. Scheraga *An atomically detailed study of the folding pathways of protein A with the stochastic difference equation*,
Proc. Natl. Acad. Sci., U.S.A. **99**, 10394-10398 (2002).
25. M. Khalili, A. Liwo, F. Rakowski, P. Grochowski and H.A. Scheraga *Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode*,
J. Phys. Chem. B. **109**, 13785-13797 (2005).
26. M. Khalili, A. Liwo, A. Jagielska and H.A. Scheraga *Molecular dynamics with the united-residue model of polypeptide chains. II. Langevin and Berendsen-bath dynamics and tests on model α -helical systems*,
J. Phys. Chem. B **109**, 13798-13810 (2005).
27. A. Liwo, M. Khalili and H.A. Scheraga *Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains*,
Proc. Natl. Acad. Sci., U.S.A. **102**, 2362-2367 (2005).
28. M. Khalili, A. Liwo and H.A. Scheraga *Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains*,
J.Mol. Biol. **355**, 536-547 (2006).

De novo all atom folding of helical proteins

A. Verma¹, S. Murthy², K. H. Lee³, E. Starikov², and Wolfgang Wenzel²

¹ Research Center Karlsruhe
Institute for Scientific Computing
PO Box 3640, D-76021 Karlsruhe, Germany

² Research Center Karlsruhe
Institute for Nanotechnology
PO Box 3640, D-76021 Karlsruhe, Germany

³ Supercomputational Materials Laboratory
Korean Institute of Science and Technology
39-1 Hawolgok-dong, Seonbuk-gu, Seoul 136-791, Korea
E-mail: wenzel@int.fzk.de

We recently developed an all-atom free energy forcefield (PFF01) for protein structure prediction with stochastic optimization methods. We demonstrated that PFF01 correctly predicts the native conformation of several proteins as the global optimum of the free energy surface. Here we review recent folding studies, which permitted the reproducible all-atom folding of the 20 amino-acid trp-cage protein, the 40-amino acid three-helix HIV accessory protein and the sixty amino acid bacterial ribosomal protein L20 with a variety of stochastic optimization methods. These results demonstrate that all-atom protein folding can be achieved with present day computational resources for proteins of moderate size.

1 Introduction

De novo protein tertiary structure prediction (PSP) and the elucidation of the mechanism of the folding process are among the most important outstanding problems of biophysical chemistry^{1,2}. The many complementary proposals for PSP span a wide range of representations of the protein conformation, ranging from coarse grained models to atomic resolution. The choice of representation often correlates with the methodology employed in structure prediction, ranging from empirical potentials for coarse grained models^{3,4} to complex atom-based potentials that directly approximate the physical interactions in the system. The latter offer insights into the mechanism of protein structure formation and promise better transferability, but their use incurs large computational costs that has confined all-atom protein structure prediction to all but the smallest peptides^{5,6}.

It has been one of the central paradigms of protein folding that proteins in their native conformation are in thermodynamic equilibrium with their environment⁷. Exploiting this characteristic the structure of the protein can be predicted by locating the global minimum of its free energy surface without recourse to the folding dynamics, a process which is potentially much more efficient than the direct simulation of the folding process. PSP based on global optimization of the free energy may offer a viable alternative approach, provided that suitable parameterization of the free energy of the protein in its environment exists and that global optimum of this free energy surface can be found with sufficient accuracy⁸.

We have recently demonstrated a feasible strategy for all-atom protein structure prediction^{9–11} in a minimal thermodynamic approach. We developed an all-atom free-energy forcefield for proteins (PFF01), which is primarily based on physical interactions with important empirical, though sequence independent, corrections¹¹. We already demonstrated the reproducible and predictive folding of four proteins, the 20 amino acid trp-cage protein (1L2Y)^{9,12}, the structurally conserved headpiece of the 40 amino acid HIV accessory protein (1F4I)^{10,13} and the sixty amino acid bacterial ribosomal protein L20¹⁴. In addition we showed that PFF01 stabilizes the native conformations of other proteins, e.g. the 52 amino-acid protein A^{5,15}, and the engrailed homeodomain (1ENH) from *Drosophila melanogaster*¹⁶.

1.1 Forcefield

We have recently developed an all-atom (with the exception of apolar CH_n groups) free-energy protein forcefield (PFF01) that models the low-energy conformations of proteins with minimal computational demand^{17,10,11}. In the folding process at physiological conditions the degrees of freedom of a peptide are confined to rotations about single bonds. The forcefield is parameterized with the following non-bonded interactions:

$$V(\{\vec{r}_i\}) = \sum_{ij} V_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - \left(\frac{2R_{ij}}{r_{ij}} \right)^6 \right] + \sum_{ij} \frac{q_i q_j}{\epsilon_{g(i)g(j)} r_{ij}} + \sum_i \sigma_i A_i + \sum_{\text{hbonds}} V_{hb}. \quad (1)$$

Here r_{ij} denotes the distance between atoms i and j and $g(i)$ the type of the amino acid i . The Lennard Jones parameters (V_{ij} , R_{ij} for potential depths and equilibrium distance) depend on the type of the atom pair and were adjusted to satisfy constraints derived from as a set of 138 proteins of the PDB database^{18,17,19}. The non-trivial electrostatic interactions in proteins are represented via group-specific dielectric constants ($\epsilon_{g(i),g(j)}$ depending on the amino-acid to which atom i belongs). The partial charges q_i and the dielectric constants were derived in a potential-of-mean-force approach²⁰. Interactions with the solvent were first fit in a minimal solvent accessible surface model²¹ parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides²². A_i corresponds to the area of atom i that is in contact with a fictitious solvent. Hydrogen bonds are described via dipole-dipole interactions included in the electrostatic terms and an additional short range term for backbone-backbone hydrogen bonding (CO to NH) which depends on the OH distance, the angle between N,H and O along the bond and the angle between the CO and NH axis¹¹.

1.2 Optimization Methods

The low-energy free energy landscape of proteins is extremely rugged due to the comparatively close packing of the atoms in the native structure. Suitable optimization methods must therefore be able speed the simulation by avoiding high energy transition states, adapt large scale move or accept unphysical intermediates. Here we report on four different optimization methods, the stochastic tunneling method²³, the basin hopping technique^{24,25}, the parallel tempering method^{26,27} and a recently employed evolutionary technique.

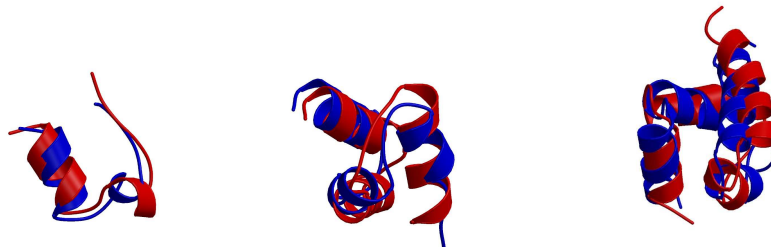


Figure 1. Overlay of the native(red) and folded (blue) structures of trp-cage protein²⁸, the HIV accessory protein¹³ and the bacterial ribosomal protein L20¹⁴.

2 Results

2.1 The trp-Cage Protein

Using the PFF01 forcefield we simulated 20 independent replicas of the 20 amino acid trp-cage protein^{29,6} (pdb code 1L2Y) with a modified versions of the stochastic tunneling method^{23,9}. Six of 25 simulations reached an energy within 1 kcal/mol of the best energy, all of which correctly predicted the native experimental structure of the protein(see Fig 1 (left)). We find a strong correlation between energy and RMSD deviation to the native structure for all simulations. The conformation with the lowest energy had a backbone root mean square deviation of 2.83 Å.

We also folded this protein with the parallel tempering method¹². We found that the standard approach, which preserves the thermodynamic equilibrium of the simulated populations, did not reach very low energies even for the low-temperature replicas. and introduced the adaptive temperature control. The best final structure associated with the lowest temperature in the simulation with 30 replicas had a RMSB deviation of 2.01 Å. We found convergence of the method using eight to thirty replicas. However, a minimal number of at least eight replicas appears to be required to fold the protein, for lower replica numbers it appears that even the adaptive temperature scheme fails to generate rapid replica exchange while spanning both high and low temperatures required for the speedy exploration of the free energy surface and the refinement of local minima respectively.

Finally we have folded the trp-cage protein protein with the basin hopping technique. with a starting temperature of $T_s = 800K$ and a final temperature of $T_f = 3K$ the lowest six of 20 simulations converged to the native structure. A total of 12 of these simulations approached the native conformation as its estimate of the optimum. While all methods correctly identify the folding funnel, the basin hopping approach results in the lowest energies. Note that the second best simulation has an RMSB of only 1.8Å to the native conformation and loses in energy with less the 0.5 kcal/mol.

2.2 The HIV Accessory Protein

Encouraged by this result, we applied a the modified basin hopping or Monte-Carlo with minimization (MCM) strategy^{8,25} to fold the structurally conserved 40-amino acid head-piece of the HIV accessory protein¹⁰. We performed twenty independent simulations and

Name	RMSB	Energy	Secondary Structure Content
N	0.00		cchhhhhhhhhhhc lcbhhhhhhhhhhc lccchhhhhhhhhc
D01	2.34	-119.54	chhhhhhhhhhh lcbchhhhhhhhhhhbhhhhhhhhhhc
D02	2.41	-117.52	chhhhhhhhhhh lcbhhhhhhhhhhhhbhhhhhhhhhhc
D03	2.76	-116.25	chhhhhhhhhhh lcbhhhhhhhhhhhhbhhhhhhhhhhc
D04	2.40	-115.85	chhhhhhhhhhh lbbhhhhhhhhhhhhbhhhhhhhhhhc
D05	2.43	-114.67	chhhhhhhhhhh lcbhhhhhhhhhhc bhhhhhhhhhhc
D06	6.48	-114.06	chhhhhhhhhhh ccc bhhhhhhhhhhhhbhhhhhhhhhhc
D07	2.57	-113.65	chhhhhhhhhhh lbbchhhhhhhhhhhbhhhhhhhhhhc
D08	4.61	-107.72	chhhhhhhhhc c lccchhhhhhhhhhh lcl hhhhhhhhhc
D09	4.14	-106.29	chhhhhhhhhhh c bcbhhhhhhhhhhbb lchhhhhhhhhc
D10	5.92	-103.88	chhhhhhhhhhh lchhhhhhhhhbcbcc lbhhhhhhhhhhc

Table 1. Energies (in kcal/mol) of the 10 lowest energy decoys obtained in the basin hopping simulations of the HIV accessory protein. The table shows the backbone RMS deviation to the NMR structure and secondary structure content. The first row designates the secondary structure content of the NMR structure.

found the lowest five to converge to the native structure (see Table (1))¹⁴. The first non-native decoy appears in position six, with an energy deviation of 5 kcal/mol and a significant RMSB deviation. The table demonstrates that all low-energy structures have essentially the same secondary structure, i.e. position and length of the helices are always correctly predicted, even if the protein did not fold correctly.

The good agreement between the folded and the experimental structure is also evident from Figure (1)(center), which shows the secondary structure alignment of the native and the folded conformations. The good physical alignment of the helices illustrates the importance of hydrophobic contacts to correctly fold this protein. An independent measure to assess the quality of these contacts is to compare the C_{β} - C_{β} distances (which correspond to the NOE constraints of the NMR experiments that determine tertiary structure) in the folded structure to those of the native structure. We found that 66 % (80 %) of the C_{β} - C_{β} distance distances agree to within one (1.5) standard deviations of the experimental resolution.

We also performed a simulation of the HIV accessory protein using the adapted parallel tempering method¹³. We used 20 processors of an INTEL XEON PC cluster and ran the simulation for a total of 30×10^6 energy evaluations for each configuration, which corresponds to approximately 500 CPU hours on an 2.4 GHz INTEL XEON processor. All simulations were started with random conformations at high temperatures to allow for rapid, unbiased relaxation of the structures and the temperature distribution. The final conformation with the lowest energy/temperature had converged to within 1.23 / 2.46 Å backbone root mean square (RMSB) deviation to the best known decoy / NMR structure of the HIV accessory protein. The overlay of the experimental and the converged structure (see Figure (1)) demonstrates the good agreement between the conformations, the difference in NOE constraints demonstrates that not only short range, but also long range distances are correctly predicted. Considering the ensemble of final conformations, we find many structures closely resembling the native conformation. The RMSB deviations of the next four lowest conformations (all within 1.5 kcal/mol of the minimal energy) have RMSB deviations of 3.14/2.23/3.78/3.00 Å respectively to the native decoy.

Energy	RMSB	3-state secondary structure
	0.01	cCHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCCCC
-167.87	4.64	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-166.15	8.25	cCHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-165.91	4.41	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHcHHHHHHHHHHHHHHHHHHCC
-164.11	5.54	cCHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-163.99	3.79	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-163.93	4.04	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-163.45	8.52	CCCCHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-163.20	4.37	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC
-162.67	5.55	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHcHHHHHHHHHHHHHHHHHHCC
-162.52	3.78	cHHHHHHHHHHHHCCCCCHHHHHHHHHHHCCCCCCCCCHHHHHHcHHHHHHHHHHHHHHHHCC

Table 2. Energies (in kcal/mol) of the 10 lowest energy decoys of the final population with backbone RMS deviation to the NMR structure and secondary structure content. The first row designates the secondary structure content of the NMR structure.. The letters H and c indicate amino acids in Helix and coil structure respectively. Green letters indicate correct, red incorrect secondary structure.

2.3 The Bacterial Ribosomal Protein L20

In the course of the simulations on the HIV accessory protein we explored methods to share information between the independent basin hopping simulations in order to improve the overall convergence. For the 60 amino acid bacterial ribosomal protein L20 (pdb-code 1GYZ) we thus experimented with the evolutionary technique described in the methods section. Starting from a seed population of random structures we performed the folding simulation in three phases: (1) generation of starting structures of the population, (2) evolutionary improvement of the population and (3) refinement of the best resulting structures to ensure convergence.

The energies and structural details of the best ten resulting conformations are summarized in Table (2). Again the best conformation had approached the native conformation to about 4.6 Å RMSB deviation. In total six of the lowest ten conformations approach the native structure, while four others misfolded. Note that the selection criterion for the active population (see methods section) precludes the occurrence of the same configuration to within 3 Å RMSB, this dominance of near native conformations of the total ensemble is particularly encouraging.

In order to quantify the overall improvement of native content during the simulation, we defined the native content of the simulated ensemble as a weighted average of the deviations of the population and the native conformation: For a population of size N we add $100(N-R+1)/N$ for each near-native decoy (RMSB less than 4Å) ranked at position R by energy to the total native score of this population. A score of 100 thus corresponds to a native decoy placed at the top position, while a near native decoy at the very bottom contributes just unity. Non-native conformations contribute nothing. Using this measure the final population contains in excess of 20% of near native conformations, its native score exceeds 800, increasing sixty-fold during the simulation phases (2) and (3).

3 Conclusion

Since the native structure dominates the low-energy conformations arising in all of these simulation, our results demonstrate the feasibility of all-atom protein tertiary structure prediction for three different proteins ranging from 20-60 amino acids in length with a variety of different optimization methods. The free energy approach thus emerges as viable trade-off between predictivity and computational feasibility. While sacrificing the folding dynamics, a reliable prediction of its terminus, the native conformation — which is central to most biological questions — can be achieved.

The computational advantage of the optimization approach stems from the possibility to visit unphysical intermediate conformations with high energy during the search. This goal is realized with different mechanism in all of the employed stochastic optimization methods. In the stochastic tunneling method the nonlinear transformation of the PES permits the dynamical process to traverse arbitrarily high energy barriers at low temperatures, in basin hopping and parallel tempering, simulation phases at very high temperatures accomplish the same objective.

Acknowledgments

We thank the BMBF, the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2, WE 1863/14-1) and the Kurt Eberhard Bode Stiftung for financial support. Part of the simulations were performed at the KIST teraflop cluster.

References

1. D. Baker and A. Sali. *Science*, 294:93–96, 2001.
2. J. Schonbrunn, W. J. Wedemeyer, and D. Baker. *Curr. Op. Struc. Biol.*, 12:348–352, 2002.
3. N. Go and H. A. Scheraga. *Macromolecules*, 9:535–542, 1976.
4. P. Ulrich, W. Scott, W. F. van Gunsteren, and A. E. Torda. *Proteins, SF&G*, 27:367–384, 1997.
5. C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele. *Nature*, 420:102–106, 2002.
6. C. Simmerling, B. Strockbine, and A. Roitberg. *J. Am. Chem. Soc.*, 124:11258–11259, 2002.
7. C. B. Anfinsen. *Science*, 181:223–230, 1973.
8. Z. Li and H.A. Scheraga. *Proc. Nat. Acad. Sci. U.S.A.*, 84:6611–6615, 1987.
9. A. Schug, T. Herges, and W. Wenzel. *Phys. Rev. Letters*, 91:158102, 2003.
10. T. Herges and W. Wenzel. *Phys. Rev. Letters*, 94:018101, 2004.
11. T. Herges and W. Wenzel. *Biophys. J.*, 87(5):3100–3109, 2004.
12. A. Schug, T. Herges, and W. Wenzel. *Europhysics Lett.*, 67:307–313, 2004.
13. A. Schug, T. Herges, and W. Wenzel. *Proteins*, 57:792–798, 2004.
14. A. Schug, T. Herges, and W. Wenzel. *J. Am. Chem. Soc.*, 126:16736–7, 2004.
15. H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Shimanda. *Biochemistry*, 40:9665–9672, 1992.
16. U. Mayor, et.al. *Nature*, 421:863–867, 2003.

17. T. Herges, H. Merlitz, and W. Wenzel. *J. Ass. Lab. Autom.*, 7:98–104, 2002.
18. R. Abagyan and M. Totrov. *J. Molec. Biol.*, 235:983–1002, 1994.
19. T. Herges, A. Schug, B. Burghardt, and W. Wenzel. *Intl. J. Quant. Chem.*, 99:854–893, 2004.
20. F. Avbelj and J. Moulton. *Biochemistry*, 34:755–764, 1995.
21. D. Eisenberg and A. D. McLachlan. *Nature*, 319:199–203, 1986.
22. K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig. *Biochemistry*, 30:9686–9697, 1991.
23. W. Wenzel and K. Hamacher. *Phys. Rev. Lett.*, 82:3003, 1999.
24. A. Nayeem, J. Vila, and H.A. Scheraga. *J. Comp. Chem.*, 12(5):594–605, 1991.
25. J. P.K. Doye and D. Wales. *J. Chem. Phys.*, 105:8428, 1996.
26. G. J. Geyer. *Stat. Sci.*, 7:437, 1992.
27. K. Hukushima and K. Nemoto. *Journal of the Physical Society of Japan*, 65:1604–1608, 1996.
28. A. Schug, A. Verma, T. Herges, and W. Wenzel. submitted to *Proteins*, 2005.
29. J. W. Neidigh, R. M. Fesinmeyer, and N. H. Anderson. *Nature Struct. Biol.*, 9:425–430, 2002.

Dewetting in Nanoscale Hydrophobic Plates Collapse and Protein Complex Folding

Ruhong Zhou^{1,2}

¹ IBM Thomas J. Watson Research Center,
Yorktown Heights, NY 10598
E-mail: ruhongz@us.ibm.com

² Department of Chemistry,
Columbia University, New York, NY 10027

We have performed molecular dynamics simulations of the collapse of graphite-like model plates, fluorinated carbon monolayers, and protein melittin tetramer, to examine how water molecules mediate hydrophobic collapse in nanoscale (size 3–4nm) hydrophobic plates as well as protein complexes (with similar sizes). We found that the normal graphite plates do not display any dewetting (water drying) transition during the double plate collapse, while a reduction in carbon-water van der Waals interaction will result in a dewetting transition, consistent with previous findings from carbon nanotubes. On the other hand, the fluorinated carbon plates display a strong dewetting transition with a critical distance up to 10Å and a very fast collapse speed (~100ps). In the melittin tetramer folding, we also observed an unexpected dewetting transition inside a nanoscale channel formed by the tetramer, with a channel size of up to 2–3 water diameters. These dewetting transitions, although occurring on a microscopic length scale, are reminiscent of the first order phase transition from liquid to vapor, which provide an enormous driving force towards the further collapse.

1 Introduction

Hydrophobic effects play a key role in many important physio-chemical processes, such as protein folding, micelle formation, water permeation in membrane channels (aquaporins), capillary evaporation, and superhydrophobic surface coating. Many biomolecules are characterized by surfaces containing extended nonpolar regions, and the aggregation and subsequent removal of water molecules between these hydrophobic surfaces is believed to be crucial. It is now widely known that small hydrophobic solutes hydrate differently than large ones. Small solutes such as methane can fit into the water hydrogen bond network without destroying much of the hydrogen bonds¹, whereas larger hydrophobic solutes induce reorganization of water molecules.^{2–4} More interestingly, there might exist a drying layer as large as several water molecules around strongly hydrophobic surfaces, as first suggested by Stillinger⁵ and then studied by many other groups^{6–8} (interested readers can refer to the review by Pratt⁸ for more details). Most of these early studies focus on single plates (solutes) using macroscopic theories or simplified computational models. One recent study on double plates also adopted such a simplified model where the solute-water interaction is purely repulsive.⁹ Water dramatically reorganizes between such plates as they are brought together, so much so, that for inter-plate distances smaller than a certain critical distance D_c , there is a spontaneous drying transition, which is reminiscent to the first order liquid-vapor phase transition.⁹ When the solute-water interaction enhances, the dewetting transition is expected to be weaker and weaker until completely disappears.⁹ In this mini-review, we go over some of our recent studies on the dewetting transition in more

realistic systems, from all-atom graphite-like plates, to fluorinated carbon monolayers, and protein complexes. Some interesting and striking results have been found.

2 Results and Discussion

2.1 Graphite-like Plates

We first use graphite-like plates to study how sensitive the dewetting transition is to the solute-solvent attractions and how water contact angle can be a useful guide in estimating the dewetting critical distance. Figure 1 shows the diagram of the double graphite-plate system, with each graphite plate consisting of 170 carbon atoms and a diameter of $\sim 28\text{\AA}$. The graphite plates are then solvated in a $60 \times 60 \times 60 \text{\AA}^3$ SPC water box. A series of 1.0 ns NPT molecular dynamics (MD) simulations are then performed after the standard equilibration (see Method section for details). It is found that the graphite plates with normal interaction ($\epsilon_{CC}=0.086 \text{ kcal/mol}$ and $\epsilon_{CO}=0.1156 \text{ kcal/mol}$) do not show any dewetting transition no matter how close we push the double plates (of course no water molecule can stay in-between the plates once the separation is so small, $\leq 6.2\text{\AA}$, that the steric effect takes control). This is consistent with previous studies with carbon nanotubes and graphite plates.^{10,11} More simulations are then followed with the reduced water-carbon interactions to study the drying transition sensitivity. The inter-plate distance is fixed at $6.8\text{\AA}$ but the strength of the water-plate interaction is decreased gradually. When the water-carbon interaction ϵ_{CO} is decreased to 0.0611 kcal/mol ($\epsilon_{CC}=0.024 \text{ kcal/mol}$), the dewetting transition is observed.¹² When the water-carbon interaction is further reduced, the dewetting becomes more profound. It is also interesting to note that when $\epsilon_{CO}=0.0647 \text{ kcal/mol}$, the system oscillates between the “dry” and “wet” states, indicating that $D = 6.8\text{\AA}$ is the critical distance for the graphite-like system with a $\epsilon_{CC}=0.027 \text{ kcal/mol}$.

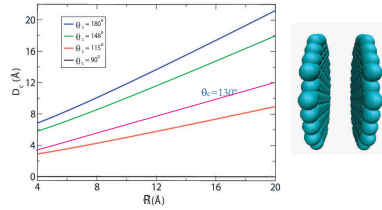


Figure 1. Schematic diagram for the two graphite-like plates system, and the dewetting critical distance versus the radius of the plates with different water contact angle (predicted from Eq. (1) of the macroscopic theory).

It is known from the macroscopic theory that the critical distance for drying between the plates will increase as the strength of the attractive interaction decreases⁹. In general, it can be shown that when the water contact angle with the plate, θ_c , is obtuse, the critical distance for dewetting $D_c \propto \cos \theta_c$.⁹ The critical distance also increases linearly with the plate size when the plate is small.⁹ Since the contact angle increases as the strength of the attraction between water and the plate decreases, the critical distance D_c should also increase. Here is the equation for the critical distance between two plates derived from a

simple model system⁹:

$$D_c = \frac{-2\gamma_{lv}\cos(\theta_c)}{(P - P_V) + b\frac{\gamma_{lv}}{R}} \quad (1)$$

where γ_{lv} is the liquid(water)-vapor surface tension, θ_c is the water contact angle on the surface, P is the external pressure and P_V vapor pressure, R is the radius of the plate, and b is a constant representing the geometrical factor (for cylindrical disks, $b=2$). The water contact angle for the graphite plate is only 86° ,¹³ so there is no dewetting transition for normal graphite plates (see Fig. 1). The water contact angle will increase when the carbon-water attraction is decreased, which might result in a dewetting transition, as shown in the above all-atom simulations. Similarly, the water contact angle for paraffin is estimated to be about 115° ,¹² which gives a critical distance of $6-8\text{\AA}$ according to Eq. (1) (see Fig. 1), agreeing well with molecular dynamics simulations.¹² For the following fluorinated plates, a larger contact angle up to 130° was found¹⁴, thus a dewetting transition with larger critical distance is expected. For even more hydrophobic model plates, such as those with a water contact angle of 148° , a critical distance up to $14\text{\AA}$ has been found.⁹

2.2 Superhydrophobic Fluorocarbon Plates

Recently, Genzer and Efimenko have discovered a new family of molecules which display superhydrophobic surfaces in their mechanically assembled monolayers¹⁴. The common characteristic shared by these molecules is that they are all chain-like molecules with semi-fluorinated carbon groups. It has been previously found that fluorinated chains might possess a higher hydrophobicity than their hydrogenated counterparts. Thus from the above analysis based on the simple macroscopic theory (Eq. (1)), we might expect a stronger dewetting transition for these superhydrophobic fluorinated carbon monolayers (plates).

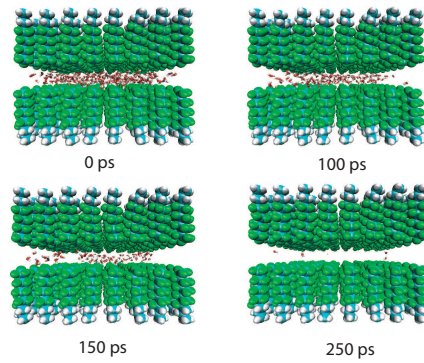


Figure 2. The snapshots of the double superhydrophobic fluorocarbon plates with $D = 8\text{\AA}$. Only water molecules within the inter-plate gap region are shown. It takes only about 100-150ps for the sharp dewetting transition to finish once it initiates.

We devised a double plate system, with each plate consisting of 8×8 such molecules ($\text{F}(\text{CF}_2)_8(\text{CH}_2)_2\text{SiH}_3$) and their terminal $-\text{CF}_3$ groups facing each other (see Fig. 2). We

then solvate the double plate system in SPC water, with water molecules at least $8\text{\AA}$ away from the solute surfaces. Figure 2 displays a few snapshots of the double plate system with separation distance $D = 8\text{\AA}$. Only water molecules within the inter-plate gap region are shown for clarity. Clearly, a sharp dewetting transition is observed and the transition takes place within $\sim 100\text{ps}$. More dewetting simulations reveal that a critical distance of $D_c = 10\text{\AA}$ exist for this system. These findings are also confirmed by rewetting simulations – starting from initial dry conditions and waiting for water molecules to rewet the region, indicating that the dry configurations are indeed thermodynamically more stable when the inter-plate distance D is less than $10\text{\AA}$. Another interesting point to notice is that although the time it takes to the initiation of the dewetting transition varies across different systems, the actual transition time after the initial diffusive motion is almost the same for all cases – about 100ps ! This is also consistent with findings from the above graphite-like plates with reduced carbon-water interactions.

2.3 Melittin Tetramer

It is of great current interest to see whether any protein complex also displays a similar dewetting transition when the complex forms during the final stages of folding. A recent study by ten Wolde and Chandler on the collapse of a hydrophobic polymer using a coarse grained model shows that the mechanism of collapse is much like that of a first-order phase transition – the water drying and hydrophobic core formation happens roughly at the same time. The authors¹⁵ thus speculate “Our findings would seem also pertinent to the mechanism of biological assembly, such as protein folding, but to demonstrate so with simulation will require an analogous simulation study of a protein-like chain.” In this part of work, we study the water drying transition inside the a protein complex - the collapse of the melittin tetramer.

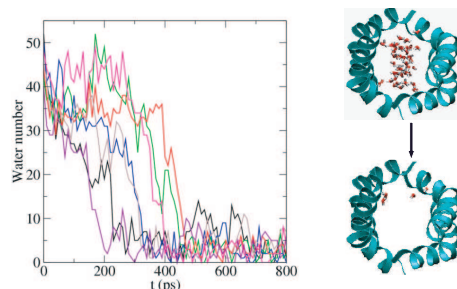


Figure 3. The number of water molecules inside the channel versus MD time and two representative snapshots of water molecules inside the channel of the melittin tetramer.

The two dimers of the melittin tetramer (initial crystal structure from PDB, 2mlt.pdb) were separated by a distance D to create a “nanoscale channel,” then solvated in a water box with water molecules extending at least $8\text{\AA}$ from the protein surface. Ten different initial configurations are prepared (same protein conformation in slightly different water boxes) for each distance D . Figure 3 shows the number of water molecules inside the channel versus the MD time for all trajectories with $D=6.0\text{\AA}$. All trajectories show that there

is a sharp dewetting transition and essentially all water molecules are expelled from the nanoscale channel after the transition. Figure 3 also shows two representative snapshots, one from before the transition, and one from after the transition. Again, the time it takes to the initiation of the dewetting transition varies with different initial configurations, but the actual transition time after the diffusive motion is almost the same for all trajectories – about 100ps! These results clearly indicate that there is a strong water drying (dewetting) transition inside this nanoscale melittin tetramer channel. Simulations with many other separation distances show that the drying transition critical distance D_c for this melittin tetramer system is approximately 5.5–7.0 Å, which is equivalent to 2-3 water molecule diameters. The reason why the melittin tetramer channel shows a drying transition while the previous two-domain protein BphC does not¹⁶ appears to be two-fold. First, the melittin channel is like a tube while the inter-domain region in the two domain protein is a slab. It is less costly with respect to free energy to disrupt the hydrogen bonds in a tube-like channel. Second, the unique surface topology springing from the isoleucine residues in melittin disrupts the water hydrogen bonds in the channel thus destabilizes the wet state.

3 Conclusion

The nanoscale dewetting transitions provide a unique way to study the hydrophobic interaction in complex systems. In this mini-review, we have gone over some of our recent studies on the dewetting transition in the collapse of nanoscale graphite-like plates, fluorinated carbon monolayers, and protein melittin tetramer using molecular dynamics simulations. It was found that the normal graphite plates do not display any dewetting transition during the double plate collapse, while a reduction in carbon-water attraction can cause a dewetting transition, consistent with previous findings from carbon nanotubes by Hummer and coworkers.¹¹ The fluorinated carbon plates, on the other hand, display a strong dewetting transition with a critical distance up to 10Å and a very fast collapse speed (~ 100 ps). In the melittin tetramer folding, we also observed an unexpected dewetting transition inside a nanoscale channel formed by the tetramer, with a channel size of up to 2-3 water diameters. These dewetting transitions, although occurring on a microscopic length scale, are reminiscent of the first order phase transition from liquid to vapor. The drying induced collapse provides an enormous driving force towards the further collapse. Finally, a simple macroscopic theory based on surface tensions can also provide useful guidelines for estimate of the dewetting critical distances from the water contact angle and plate size.

4 Simulation Methods

The GROMACS program is used for the molecular dynamics simulations of all the systems studied here. The OPLSAA force field is adopted for the fluorinated carbons and protein melittin tetramer (the graphite plate system is simple and the parameters are listed in the text above), and the SPC model for the explicit water solvent. The Particle Mesh Ewald (PME) method is used for the long-range electrostatic interactions, whereas a typical 10Å cutoff is applied to the van der Waals interactions. The timestep used for all molecular dynamics simulations is 1.0fs. A standard equilibration protocol is used, with a 1000 steps of conjugate gradient minimization, followed by a 100 ps position restrained

MD equilibration. The equilibrated systems are then used as the starting points for data collection (NPT, 1atm and 300K). In most of the “dewetting” simulations, the positions of the solute atoms are restrained in space with a harmonic potential ($k = 10 \text{ kJ/mol/\AA}^2$), unless otherwise explicitly stated.

Acknowledgments

I would like to thank Bruce Berne, Gerhard Hummer, Haiping Fang, Xin Li, Maria Eleftheriou, Xuhui Huang, Pu Liu, and Jingyuan Li for the help with the manuscript and many useful discussions.

References

1. C. Pangali, M. Rao, and B. J. Berne. Hydrophobic hydration around a pair of apolar species in water. *J. Chem. Phys.*, 71:2982–2990, 1979.
2. A. Wallqvist, E. Gallicchio, and R. M. Levy. A model for studying drying at hydrophobic interfaces: Structural and thermodynamic properties. *J. Phys. Chem. B*, 105:6745–6753, 2001.
3. M. E. Paulaitis and L. R. Pratt. Hydration theory for molecular biophysics. *Adv. Protein Chem.*, 62:283–310, 2002.
4. K. Lum and A. Luzar. Pathway to surface-induced phase transition of a confined fluid. *Phys. Rev. E.*, 56(6):6283–6286, 1997.
5. F. H. Stillinger. Structure in aqueous solutions of nonpolar solutes from the standpoint of scaled-particle theory. *J. Solution Chem.*, 2:141, 1973.
6. D. M. Huang and D. Chandler. The hydrophobic effect and the influence of solute-solvent attractions. *J. Phys. Chem. B*, 106:2047–2053, 2002.
7. K. Lum, D. Chandler, and J. D. Weeks. Hydrophobicity at small and large length scales. *J. Phys. Chem. B*, 103:4570–4577, 1999.
8. L. R. Pratt. Molecular theory of hydrophobic effects: She is too mean to have her name repeated. *Annu. Rev. Phys. Chem.*, 53:409–436, 2002.
9. X. Huang, C. J. Margulis, and B. J. Berne. The dewetting and collapse of nanoscale hydrophobic objects. *Proc. Nat. Acad. Sci. U.S.A.*, 100:11953–11958, 2003.
10. N. Choudhury and B. M. Pettitt. *J. Phys. Chem. B*, 109:6422–6429, 2005.
11. G. Hummer, J. R. Rasaiah, and J. P. Noworyta. Water conduction through the hydrophobic channel of a carbon nanotube. *Nature*, 414:188–190, 2001.
12. X. Huang, R. Zhou, and B. J. Berne. Drying and hydrophobic collapse of paraffin plates. *J. Phys. Chem. B*, 109:3546–3552, 2005.
13. M. Lundgren, N. L. Allen, and T. Cosgrove. Wetting of water and water/ethanol droplets on a non-polar surface: A molecular dynamics study. *Langmuir*, 18:10462–10466, 2002.
14. Jan Genzer and Kirill Efimenko. Creating long-lived superhydrophobic polymer surfaces through mechanically assembled monolayers. *Science*, 290:2130–2133, 2000.
15. P. R. T. Wolde and D. Chandler. Drying-induced hydrophobic polymer collapse. *Proc. Natl. Acad. Sci. U.S.A.*, 99:6539–6543, 2002.
16. R. Zhou, X. Huang, C. J. Margulius, and B. J. Berne. Hydrophobic collapse in multidomain protein folding. *Science*, 305:1605–1609, 2004.

Reduction to the Simplest – The Complexity of Minimalistic Heteropolymer Models

Michael Bachmann and Wolfhard Janke

Institut für Theoretische Physik and Centre for Theoretical Sciences (NTZ),
Universität Leipzig, 04109 Leipzig, Germany
E-mail: {bachmann, janke}@itp.uni-leipzig.de

Simple coarse-grained hydrophobic-polar models for heteropolymers as the lattice HP and the off-lattice AB model allow a general classification of characteristic behaviors for hydrophobic-core based tertiary folding. The strongly reduced computational efforts enable one to reveal systematically the thermodynamic properties of comparatively long sequences in a wide temperature range of conformational activity. Based on a suitable cooperativity parameter, characteristic folding channels and free-energy landscapes, which have strong similarities with realistic folding paths, can be analysed.

1 Introduction

The understanding of the sequence-dependent stability of native protein structures, the interplay of sequence and conformation, and the classification of conformational phases are essential aspects in protein folding studies. By means of computer simulations of detailed microscopic models with modern algorithms, great progress has been achieved in the past few years in predicting native structures for exemplified amino acid sequences. Also, the relation between conformational cooperativity and thermodynamic activity in the folding process has been subject of numerous selective studies. Nonetheless, a general framework, which would allow the classification of characteristic global folding events on the basis of an appropriate “order” or cooperativity parameter, is still lacking. Another difficult problem is the systematic analysis of the influence of the amino acid order in the sequence on folding kinetics and native fold. Since these questions cannot or only partially be answered with all-atom protein models, it is necessary to increase the level of abstraction in appropriate modelling. It is known that most of these problems are related to the so-called tertiary folding process which concerns almost all amino acid residues in a cooperative manner, in contrast to the much more localised secondary structure segments. The main idea is that tertiary folding is strongly related to the hydrophobic effect. Many proteins fold spontaneously in an aqueous solution and the hydrophobicity of the distinguished side chains is mainly responsible for the position of the side chains in the native fold. Polar residues favor contact with the polar environment and screen more hydrophobic monomers, which typically form a compact core in the interior of the protein structure. In connection with volume exclusion, it is the hydrophobic force that enters into minimalistic hydrophobic-polar models.

2 The Hydrophobic-Polar (HP) Lattice Model

The simplest hydrophobic-core based approach is the HP lattice model¹, where in the original form only nearest-neighbor hydrophobic monomers, being nonadjacent along the self-

avoiding chain, attract each other. A remarkable result obtained from exactly enumerating the entire sequence and conformation space of HP lattice proteins in this parameterization²⁻⁴ is that only a small fraction of possible sequences is nondegenerate, i.e., for these only one conformation representing the “native” state exists. This is interpreted as indication of stability in the funnel-like free-energy landscape: There is an energy gap towards higher-energetic states. Such sequences are called *designing*. On the other hand, it is interesting that the number of designable conformations, i.e., structures being native folds of designing sequences, is also very small, typically even smaller than the number of designing sequences.⁴ This means that several designing sequences fold into the same native conformation. In studies of thermodynamic activity accompanying the folding process, it turned out that, depending on the sequence, the folding trajectory can pass highly compact intermediate globular states. Favoring hydrophobic contacts energetically, however, hydrophobic-core folds are not necessarily maximally compact.^{4,5} Identifying low-lying energetic states and analyzing thermodynamic properties of HP lattice proteins is not straightforward and requires sophisticated computational methods.^{5,6} Although the tendency of the HP model to favor folding paths with intermediary conformational phases is seen as sign of too little cooperativity, its power lies in the fact that sequences with far more than 100 monomers can be analyzed. These sequences are much longer than what can be studied by means of other models which are not knowledge-based, i.e., not biased towards a certain target structure.

3 Getting Rid of the Lattice: A Minimal Continuum Model

A straightforward off-lattice generalization of the HP model is the AB model (*A*: hydrophobic, *B*: polar), which was originally proposed for heteropolymers in two dimensions.⁷ Volume exclusion and effective inter-residue interactions are modeled by Lennard-Jones potentials, and a “ferromagnetic” bending energy models effectively stiffness and torsional barriers in the three-dimensional representation.^{8,9} For studies of folding channels, we define an angular overlap parameter of torsional and bond angles. Comparing a conformation $\mathbf{X}$ with the native conformation $\mathbf{X}_0$, this similarity parameter is defined by $Q(\mathbf{X}) = 1 - d(\mathbf{X})$, where $d(\mathbf{X})$ is a measure of the angular deviation of $\mathbf{X}$ from $\mathbf{X}_0$.^{8,9} Restricting the canonical partition function at temperature T to the ensemble of conformations with overlap Q , $Z(Q) = \int \mathcal{D}\mathbf{X} \delta(Q - Q(\mathbf{X})) \exp\{-E(\mathbf{X})/k_B T\}$, the overlap free energy is given by $F(Q) = -k_B T \ln Z(Q)$. Figure 1 shows the overlap free-energy landscapes for the 20-mer $BA_6BA_4BA_2BA_2B_2$, which behaves like a typical two-state folder. For permuted sequences, we also find characteristic folding through intermediate conformations and metastability.⁹

4 Concluding Remarks

Protein folding is a highly complex process where microscopic details are of essential importance. Nonetheless, the effective hydrophobic force which is responsible for the tertiary fold is of rather mesoscopic nature, since it is due to the interaction with the bath of surrounding solvent molecules. For this reason, appropriate effective coarse-grained models should allow a qualitative description of characteristic features accompanying conformational folding transitions. By means of minimalistic standard hydrophobic-polar models

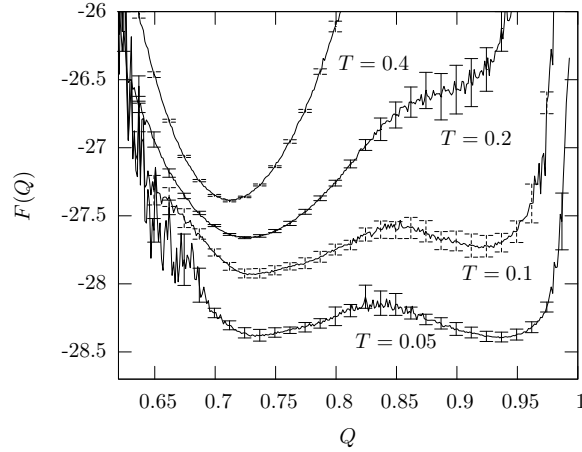


Figure 1. Free-energy landscapes $F(Q)$ at different temperatures for the sequence $BA_6BA_4BA_2BA_2B_2$, exhibiting typical two-state folding behavior.⁹

we have shown that even at this level of abstraction typical properties of native folds and cooperative effects in the folding process can be described qualitatively.

Acknowledgments

We thank H. Arkin, R. Schiemann, S. Schnabel, T. Vogel, A. Kallias, J. Schluttig, and C. Junghans for cooperation. This work is partially supported by the DFG (German Science Foundation) under Grant No. JA 483/24-1 and the computer time Grant No. hlz11 of the John von Neumann Institute for Computing (NIC), Forschungszentrum Jülich.

References

1. K. A. Dill, *Biochemistry* **24**, 1501 (1985).
2. C. Tang, *Physica A* **288**, 31 (2000).
3. H. Li, R. Helling, C. Tang, and N. Wingreen, *Science* **273**, 666 (1996); H. Cejtin, J. Edler, A. Gottlieb, R. Helling, H. Li, J. Philbin, C. Tang, and N. Wingreen, *J. Chem. Phys.* **116**, 352 (2002).
4. R. Schiemann, M. Bachmann, and W. Janke, *J. Chem. Phys.* **122**, 114705 (2005); *Comp. Phys. Commun.* **166**, 8 (2005).
5. M. Bachmann and W. Janke, *Phys. Rev. Lett.* **91**, 208105 (2003); *J. Chem. Phys.* **120**, 6779 (2004), and references therein.
6. H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, *J. Chem. Phys.* **118**, 444 (2003).
7. F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, *Phys. Rev. E* **48**, 1469 (1993).
8. M. Bachmann, H. Arkin, and W. Janke, *Phys. Rev. E* **71**, 031906 (2005).
9. S. Schnabel, M. Bachmann, and W. Janke, to be published (2006).

Multiplexed-Replica Exchange Molecular Dynamics with the UNRES Force-Field as an Effective Method for Exploring the Conformational Energy Landscape of Proteins.

Cezary Czaplewski^{1,2}, Sebastian Kalinowski², Stanisław Ołdziej^{1,2}, Adam Liwo^{1,2}, and Harold A. Scheraga¹

¹ Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca, NY 14853-1301, U.S.A.,
E-mail: has5@cornell.edu

² Faculty of Chemistry, University of Gdańsk,
Sobieskiego 18, 80-952 Gdańsk, Poland

The theoretical prediction of the three-dimensional structures of proteins purely from an amino-acid sequence is a grand challenge of computational biophysics. Here, we present simulations using an effective search algorithm combined with the united residue force-field. In the replica exchange method, the kinetic trapping problem is overcome by exchanging noninteracting replicas simulated at different temperatures. To enhance sampling, the replicas are multiplexed with a number of independent molecular dynamics runs at each temperature. The simulated free-energy landscape is presented for 1CLB protein. Application of the multiplexed-replica exchange algorithm to large-scale parallel computing is discussed.

1 Introduction

Efficient conformational sampling algorithms are an essential component of methods for studying protein structure and dynamics. Methods such as canonical Molecular Dynamics (MD) or Metropolis Monte Carlo (MC) can be used for estimating thermodynamic properties as well as for a global search, but in practice they easily become trapped, and thus are not effective methods for studying rough free-energy landscapes of proteins.

One of the most effective sampling methods, the Replica Exchange method (also known as Exchange MC¹, or Parallel Tempering²), was initially developed to improve sampling in glassy systems in statistical physics. However, following Hansmann's use of the method in Monte Carlo simulations of a simple peptide, Met-enkephalin² and Okamoto and Sugita's formulation of an MD version of the algorithm,³ the Replica Exchange method has been applied extensively in protein-folding simulations.

2 Methods

In the united-residue physics-based UNRES model,⁴⁻⁶ a polypeptide chain is represented by a sequence of α -carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p) positioned in the middle between two consecutive C^α atoms. The UNRES force field has been derived as a Restricted Free Energy (RFE) function decomposed into factors arising from interactions within and between a

given number of united interaction sites⁵. The selected parameters of the UNRES force-field were refined by hierarchical optimization⁶ of the potential-energy function.

The Replica Exchange MD (REMD) method combines the idea of simulated annealing and Monte Carlo methods, and is one of the generalized-ensemble algorithms that performs a random walk in energy space due to a free random walk in temperature space. In the REMD method, n replica systems, each in the canonical ensemble, and each at a different temperature, are simulated. At given intervals, swaps, or exchanges, of the configurational variables between systems are accepted with the Metropolis criteria. In this paper, we are using multiplex REMD (MREMD) introduced by Pande⁷. In MREMD, to enhance sampling, the replicas are multiplexed with a number of independent molecular dynamics runs at each temperature. Exchanges of configurations only between random replicas with neighboring temperatures are tried.

3 Results and Discussion

The free energy is the most important quantity for the description of equilibrium properties of proteins. An example of the simulated free-energy landscape for the 1CLB protein is shown in Figure 1.

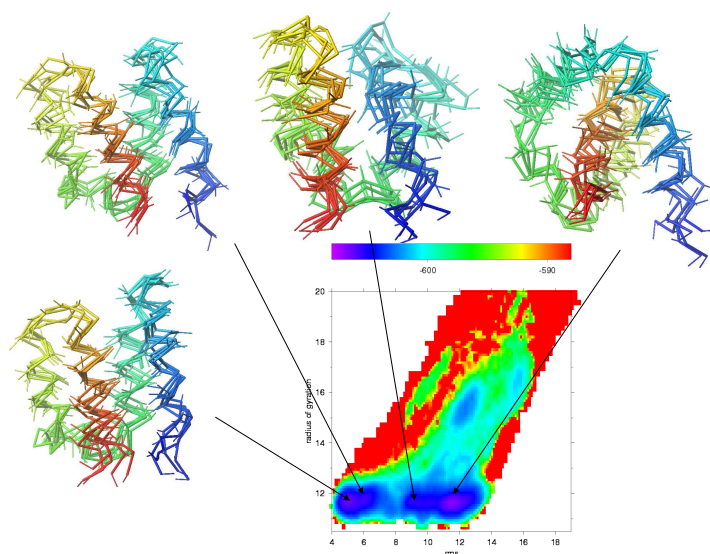


Figure 1. Free energy landscape of 1CLB protein as a function of rmsd and radius of gyration, and representative structures from low free-energy regions.

To calculate free energy profiles, we used the densities of states obtained from the multi-histogram analysis of MREMD trajectories.⁸ From the densities of states, we calculated the microcanonical entropy, $S(E_i) = k_B \ln[n(E_i)]$, for all conformations collected from the simulations, and used it to compute the microcanonical free energies with the

following expression: $F(E_i, T) = E_i - TS(E_i)$. To plot the free energy as a function of rmsd (r) and radius of gyration (ρ), we calculated the free energy by evaluating the following expression for each grid point:

$$F(r, \rho, T) = -k_B T \ln \sum_{E_i \in N(r, \rho)} \exp\left(\frac{-F(E_i, T)}{k_B T}\right) \quad (1)$$

where the index i enumerates conformations within the histogram bins, $N(r, \rho)$, for given ranges of rmsd and radius of gyration.

Parallelization of the MREMD method has been enhanced by removing the synchronization step. The exchanges of configurations between random replicas with neighboring temperatures are tried, not at the same number of MD steps for each replica but are forced by the replica with the lowest temperature independent of the number of steps performed by the other replicas. The improved algorithm scales almost linearly up to 1,000 processors with over 80% average performance.

Acknowledgments

Support from the National Institutes of Health (grant GM-14312), the National Science Foundation (grant MCB00-03722), the NIH Fogarty International Center (grant TW7193), the Polish Ministry of Education and Science (grant 3 T09A 032 26) as well as use of supercomputing resources at the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center is gratefully acknowledged.

References

1. K. Hukushima and K. Nemoto. Exchange monte carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.*, 65:1604–1608, 1996.
2. U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–150, 1997.
3. Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.
4. A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18:849–873, 1997.
5. A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.*, 115:2323–2347, 2001.
6. S. Ołdziej, J. Łągiewka, A. Liwo, C. Czaplewski, M. Chinchio, M. Nancias, and H. A. Scheraga. Optimization of the unres force field by hierarchical design of the potential-energy landscape. 3. use of many proteins in optimization. *J. Phys. Chem. B*, 108:16950–16959, 2004.
7. Y. M. Rhee and V. S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.*, 84:775–786, 2003.

8. M. Nianias, C. Czaplewski, and H. A. Scheraga. Replica exchange and multicanonical algorithms with the coarse-grained united-residue (UNRES) force field. *J. Chem. Theor. Comput.*, 2:513–528, 2006.

Single Amino Acid Contributions to Binding Affinity in Enzyme-Inhibitor Interactions: a Docking-Based Screening of BPTI-Beta Trypsin interaction

Daniele Dell’Orco¹, Pier G. De Benedetti², and Francesca Fanelli¹

¹ Dulbecco Telethon Institute,
Department of Chemistry, University of Modena and Reggio Emilia, Italy
via Campi 183, 41100 Modena, Italy
E-mail: {dellorco,fanelli}@unimo.it

² Department of Chemistry, University of Modena and Reggio Emilia, Italy
via Campi 183, 41100 Modena, Italy
E-mail: deben@unimo.it

Enzyme-inhibitor interactions are crucial for normal functioning of many biological pathways. Point mutations in either the enzyme or the inhibitor molecule often lead to a modulation of the binding affinity (ΔG^o) with no major alterations in the 3D structure of the complex. The possibility to screen *in silico* the effects of point mutations on ΔG^o is of high interest, especially for protein design purposes. We have recently developed a computational protocol based on an existent rigid-body docking algorithm, which has shown a good capability to predict ΔG^o changes upon mutations in protein-protein interactions. Here, we present the results obtained for the bovine pancreatic trypsin inhibitor (BPTI) and beta-Trypsin (β -Tryp) interaction. In this system, the BPTI^{Lys15} residue was replaced by eight different amino acids, hence varying the physico-chemical nature of the interface. The X-ray structure of each variant is available, as well as the effect of each substitution on the binding energetics. We compare our approach with data arising from both *in vitro* measurements and from another structure-based empirical approach, based on changes in solvent accessible surface areas (ΔASA) upon binding. The performance and the limitations of our approach are discussed.

1 Introduction

Enzymes are the unquestioned protagonists of catalysis for virtually all the biochemical reactions occurring within cells. Inhibitors are usually proteins or other molecules specifically competing with enzyme substrates, hence exerting a regulatory mechanism of the biochemical pathway they belong to. Understanding the physico-chemical features of enzyme-inhibitor binding is fundamental, especially for designing specific molecules for therapeutic and biomedical purposes. Moreover, enzyme-inhibitor represent an interesting class of high-affinity protein-protein complexes.

Naturally occurring or artificially induced point mutations may significantly affect both the structural and the kinetic/thermodynamic features of the association between an enzyme and its protein inhibitor. Here, we report a computational study concerning the high affinity interaction between bovine pancreatic trypsin inhibitor (BPTI) and beta-Trypsin (β -Tryp), see Fig. 1. In line with *in vitro* experiments, we have used the X-ray structures of eight mutants of the residue Lys 15 in BPTI to perform a rigid-body docking simulation of BPTI/ β -Tryp association. We have followed a computational protocol previously developed in our laboratory¹, which is based on the recent fast Fourier-transform (FFT) protein-protein docking algorithm ZDOCK2.3³. Besides the present system, we investigated other

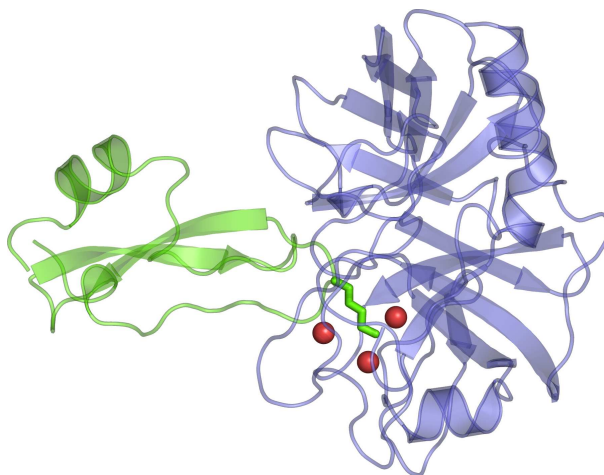


Figure 1. Three-dimensional structure of the BPTI- β -Tryp complex (Protein Data Bank (PDB) entry: 3BTK⁴. BPTI, i.e. the molecule used as a probe to sample the roto-translational space, is colored in green. β -Tryp, i.e. the molecule used as a target and hence kept fixed, is colored in slate. Structural water molecules are represented by red spheres, whereas BPTI^{Lys15}, i.e. the residue target of in silico mutagenesis, is represented in sticks.

enzyme-inhibitor complexations, namely human ribonuclease inhibitor-angiogenin interaction and barnase-barstar interaction. Our results are satisfactory for every investigated system, both in structural and in thermodynamic terms. Fairly accurate correlative predictions concerning the effects of point mutations on ΔG^o were achieved for each tested enzyme-inhibitor complex². Moreover, we could gain some insights into the interpretation of kinetic data².

In this brief paper, we focus on the important case of a single amino acid replacement at the binding interface of an enzyme-inhibitor complex, indeed BPTI/ β -Tryp. We show that accurate molecular modeling can be successfully employed in the absence of high resolution structural data, such as X-ray diffraction maps, to estimate the thermodynamic effects of specific amino acid substitutions. Finally, we compare our results with predictions arising from another broadly used empirical approach for thermodynamic analysis of protein-protein interactions starting from high resolution structures.

2 Solvent-Accessible Surface Area (ASA)-Based Empirical Approaches to Protein-Protein Interactions

The parameterization of the heat capacity, enthalpy, and entropy of protein-protein association developed by Murphy⁵ and Freire⁶ has been widely used in the literature to empirically calculate the free energy change from structural information. The basic assumption is that enthalpy and entropy changes, and hence ΔG^o itself, can be both linearly related to changes in the solvent accessible surface area (ΔASA) of polar and nonpolar protein atoms through empirical relationships. The well-known relationships:

$$\Delta G^o = -RT \ln K_A \quad (1)$$

and:

$$\Delta G^o(T) = \Delta H^o(T) - T\Delta S^o(T) \quad (2)$$

are used to determine empirically K_A . The first assumption is that at a constant pressure, ΔH^o can be expressed in terms of the heat capacity change upon association of the protein-protein complex, by the relationship:

$$\Delta H^o(T) = \Delta H^o(T^o) + \Delta C_p(T - T^o) \quad (3)$$

where $\Delta H^o(T^o)$ is the standard reference enthalpy of association at some reference temperature T^o ⁵. On the other hand, the assumption concerning the entropic term is that in ordinary conditions (e.g. no proton transfers, etc.) it may be dissected in a solvation entropy term ($\Delta S_{solv}(T)$), a conformational entropy term (ΔS_{conf}) and an overall rotational/translational entropy term (ΔS_{rt}):

$$\Delta S^o(T) = \Delta S_{solv}(T) + \Delta S_{conf} + \Delta S_{rt} \quad (4)$$

Starting from the known 3D structure of a protein-protein complex, equations 1-4 thus allow for an estimate of ΔG^o , and hence, through equation 1, for a prediction of the equilibrium constant K_A .

2.1 Structure-Thermodynamics Correlations Analysis of BPTI- β -Tryp Interaction

We used the empirical method by Murphy and Freire^{5,6} to predict K_A 's for BPTI- β -Tryp interaction. This was possible as the X-ray structure of eight variants of BPTI^{lys15} are available⁴. We will refer to this method as Structure-Thermodynamics Correlation (STC) approach throughout the paper.

Results from the STC analysis are reported in Fig.2A. Although for some substitutions (i.e. Gly, Thr, Glu, Met and Phe) the affinity change follows a roughly correct trend (Fig.2A), the method apparently fails when considering the overall set of mutants. Indeed, the STC method correctly predicts that the Gly substitution has the most detrimental effect on the K_A , whereas it clearly underestimates the effect of the Phe substitution. Moreover, the overall range of K_A variation is predicted to be about one pK unit, whereas it experimentally exceeds four pK units.

3 Docking Score-Based Approaches to Protein-Protein Interaction

We recently demonstrated that rigid-body docking simulations of protein-protein association may lead to quantitative models based on tested correlative equations, which can be used to predict thermodynamic and kinetic properties¹. The method works best with high-affinity (i.e. $\Delta G^o \leq -9$ kcal/mol) protein-protein complexes that do not drastically change their conformation upon binding¹. The FFT-based algorithm ZDOCK2.3³ performs a search for optimization of intermolecular complementarity and provides a score

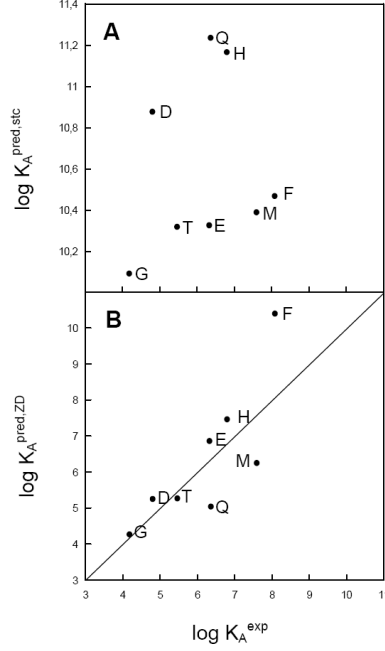


Figure 2. Comparison between the STC (A) and the docking-based (B) predictions for the equilibrium constants K_A (M^{-1}) in BPTI- β -Tryp interaction. Each individual amino acid substituting for BPTI^{lys15} is represented by one letter code.

(ZD) for each docking solution, which can be summarized as a linear combination of three components:

$$ZD = S_{sc} + S_{el} + S_{des} \quad (5)$$

where *sc*, *el* and *des* indicate respectively the shape complementarity, electrostatics and desolvation terms, each implicitly including a scaling factor for the relative energy term proposed by the ZD developers³. In detail, we previously found a convincing correlation between the ZD score and ΔG^o for a number of protein-protein complexes that do not share structure similarity¹. Our results, and particularly the correlative models obtained, are consistent with the interpretation of the ZD score as an empirically determined free energy function, which does not require the decomposition in enthalpic and entropic contributions. In other words, ZD can be considered as an index linearly related to ΔG^o , if we assume that this latter can be dissected as follows:

$$\Delta G^o = \Delta E_{vdw} + \Delta G_{des} + \Delta E_{el} + \Delta G_{const} \quad (6)$$

where ΔE_{vdw} represents the energy term due to van der Waals interactions, accounted for by ZD through the *sc* component, ΔG_{des} represents the free energy of desolvation

of atoms at the binding interface, accounted for by ZD through the *des* component, and ΔE_{el} represents the electrostatic interaction energy, hence accounted for by ZD through the *el* term. ΔG_{const} is the only term, which is not accounted for by ZD due to the rigid-body approximation. Indeed, this term includes translational, rotational and vibrational free energy changes upon binding, i.e. the conformational and the cratic entropies. The correlative context in which ZD is used, however, evidently allows such a neglect.

3.1 Docking-Based Correlative Analysis of Mutational Effects of BPTI^{Lys15} substitutions

Similarly to what we did in our previous work¹, we run docking simulations of BPTI- β -Tryp binding considering both the wild-type and mutated forms². A 128x 128 x 128 point grid with a 1.2 Å spacing was used for digitalizing the interacting molecules. A rotational sampling interval of 6° was employed and the best 4000 solutions from each run were retained and ranked according to their ZD score. Three independent sets of docking runs were performed for each complex, i.e. one starting from the X-ray coordinates and the other two randomizing the initial positions. All computations were run on a Linux cluster with 2.6 GHz Opteron CPUs.

Among the solutions generated by the algorithm, we selected as native-like structures all the docked complexes characterized by a C_α -RMSD lower than 1.0 Å from the native structure. For each docking simulation, the ZD score was averaged over all the native-like complexes resulting from the three independent runs, and employed for the correlation analysis with the thermodynamic data². Figure 2B reports the predicted K_A for each BPTI^{Lys15} substitute.

The docking based-approach leads to a fair prediction of the effects of amino acid replacements in the BPTI- β -Tryp interaction, in a correlative context. Indeed, we found an acceptable correlation between ZD score and $\log K_A$, with a coefficient $R=0.86$ and a probability $p=0.006$, when the docking was done starting from the X-ray structure of each mutant². Moreover, when the mutants were modeled by introducing the substitutions in the structure of the wild type, the correlation slightly worsened ($R=0.79$, $p=0.019$) but the trend remained substantially the same², indicative of a good potential of the method even in the absence of experimental 3D structure of each mutant.

Figure 2 reports a comparison of the results obtained in this study with the two different approaches presented above. A leave-one-out correlative analysis performed on ZD and $\log K_A$ lead to a good prediction of the effects of BPTI^{Lys15} substitutions on the equilibrium constant ($R=0.80$, slope= 1.2 for the least mean square fitting line, Fig.2B). Our approach is thus able to provide a reasonable estimate of K_A , which is consistent with the experimental values, independently of the physico-chemical nature of the substitutions.

4 Concluding Remarks

We have described a docking-based approach with an effective capability to predict the contributions of single amino acids to the binding energetics in a high-affinity enzyme-inhibitor interaction. The simplicity of the protocol, its tested capability^{1,2} and the modest computational cost makes it an interesting tool also for protein design purposes. The protocol is applicable in cases of high affinity enzyme-inhibitor interactions, where there is

evidence of no major conformational changes of the partners upon binding. Moreover, an assumption is made that the point mutations investigated may alter the affinity but not drastically the conformation of either the proteins involved.

Acknowledgments

Helpful technical suggestions by Emiliano Specchia and Michele Seeber are gratefully acknowledged. This work was supported by a Telethon-Italy grant (Grant S00068TELA to F.F.).

References

1. D. Dell’Orco, M. Seeber, P. G. De Benedetti, and F. Fanelli *Probing fragment complementation by rigid-body docking: in silico reconstitution of calbindin D9k*, J. Chem. Inf. Model. **45**(5), 1429–38 (2005).
2. D. Dell’Orco, P. G. De Benedetti, and F. Fanelli *In Silico screening of Mutational effects on enzyme-inhibitor affinity: a docking-based approach*. Submitted for publication (2006)
3. R. Chen, L. Li, , and Z. Weng *ZDOCK: an initial-stage protein-docking algorithm*, Proteins **52**(1), 80–7 (2003).
4. R. Helland, J. Otlewski, O. Sundheim, M. Dadlez, and A. O. Smalas *The crystal structures of the complexes between bovine beta-trypsin and ten P1 variants of BPTI*, J. Mol. Biol. **287**(5), 923–42 (1999).
5. B. M. Baker, and K .P Murphy *Dissecting the energetics of a protein-protein interaction: the binding of ovomucoid third domain to elastase.*, J. Mol. Biol. **268**(2), 557–69 (1997).
6. I. Luque, and E . Freire *Structure-based prediction of binding affinities and molecular design of peptide ligands.*, Methods Enzymol. **295**, 100–27 (1998).
7. V. J. Hilser, J. Gomez, and E . Freire *The enthalpy change in protein folding and binding: refinement of parameters for structure-based calculations.*, Proteins **26**(2), 123–33 (1996).

Sub-type Specific Sites for SMAD Receptor Binding Identified by Sequence Comparison Using “Sequence Harmony”

K. Anton Feenstra*, Walter Pirovano*, and Jaap Heringa

Centre for Integrative Bioinformatics VU (IBIVU)
Free University, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
E-mail: heringa@few.vu.nl

We have examined the sub-type specific functions within the MH2 domain of the SMAD-family of transcription factors and found that our novel algorithm, called “Sequence Harmony”, has a high specificity for identification of sites important for the functional differences. For the SMAD MH2 domain, 40 sub-type specific functional sites are predicted, which in the structure form clusters of similar function, i.e. for receptor binding, co-repressor binding and binding to transcription factors. From these clusters, putative functions were assigned to eleven out of fourteen predicted functional sites with unknown function. We propose these fourteen sites of unknown function as interesting candidates for further (experimental) investigation.

1 Introduction

Protein families and sub-families separated on the basis of functional properties.^{1,2} It is therefore not surprising that a fair number of methods are in use for the comparison of amino-acid composition at different positions between groups of proteins from different families and/or sub-types.^{3,4} It is surprising, however, that apparently in the vast majority of those studies, relatively little thought has been given to the underlying formalism of sequence comparison. Starting from a multiple sequence alignment (MSA) of the proteins of interest, the aim is to identify sites that are possibly conserved within a group, but certainly different between the groups.

Current practice seems to focus on sites that are conserved in both groups, but still different between them,³ thereby excluding sites that are not totally conserved but nevertheless different between groups. This may not seem a serious problem at first hand, but let us consider an example of proteins that bind a certain receptor (the ‘binders’) and those that don’t (the ‘non-binders’). Certainly, one can expect sites that are crucial for binding to be conserved in the group of ‘binders’. To exclude binding, on the other hand, several of many reasons may suffice and it seems imprudent to expect those sites to be conserved as well in the group of ‘non-binders’. If the binders also interact with different, related, receptors, also the restriction to sites conserved in the group of ‘binders’ may not be a sensible one. Relative entropy is a measure for the difference in information content between both distributions of amino acid types.⁵ Unfortunately, for those sites that interest us the relative entropy is degenerate. We will introduce an alternative similarity measure named *Sequence Harmony* for comparison of groups of sequences.

We have applied our method to the interactions of the SMAD proteins with the cell-membrane associated receptors TB β RI and BMPRI. The SMAD-TB β RI and SMAD-BMPRI interactions are relatively well-studied⁶ and provide a good background of experimental data on which to validate our method.

* KAF and WP acknowledge financial support from Biorange/Netherlands Bioinformatics Centre.

2 Theory

Commonly used methods for sequence comparison from multiple sequence alignments share several drawbacks. As an alternative solution, we propose:

$$SH_i^{AB} = \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B} \quad (1)$$

where the ‘relative entropy’ of group A is calculated relative to the sum of the probabilities of both groups ($p^A + p^B$). This function operates opposite to the relative entropy; zero for maximally different sites and one for sites with identical distributions. It is therefore that we coin this measure *Sequence Harmony* (*SH*) as it indicates the amount of correspondence of amino acid composition between two groups of sequences.

3 Results

Protein sequences for the R-SMADs were collected using the NCBI query for sequence retrieval (www.ncbi.nih.gov), yielding 32 sequences: 17 BR-SMADs and 15 AR-SMADs, including that of the SMAD2-MH2 structure (PDB-ID: 1KH⁷), and aligned using PSI-Praline (www.ibivu.cs.vu.nl/programs/pralinewww).⁸ From the alignment obtained, the MH2 domain was selected for further analysis, and divided according into two subgroups, AR-SMADs (binding TB/ β RI) and BR-SMADs (binding BMPRI).

Sequence Harmony was calculated using Eq. 1 for AR-SMADs and BR-SMADs at all positions in the alignment, see Figure 1. Relatively few sites are completely non-harmonious (32 $SH=0$) whereas the vast majority are overall conserved (137 $SH=1$), and the remaining are intermediate (44 $0 < SH < 1$). Furthermore, the low harmony sites are not spread completely randomly along the sequence, but clusters can be seen of up to about five low harmony sites spread over a sequence stretch of up to about ten residues.

Out of the 40 low harmony sites, 26 have a known function (65%), and for the 32 non-harmonious sites, 22 have a known function (69%). Of the 171 remaining high-harmony

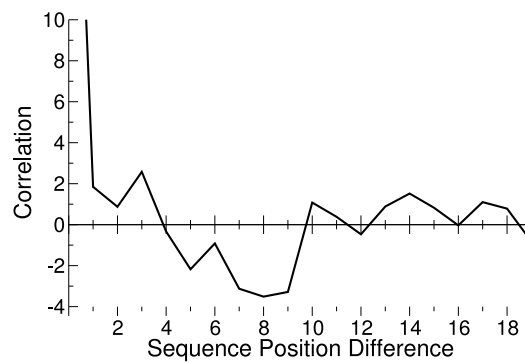


Figure 1. Sequence harmony for AR vs. BR-SMADs plotted along the sequence of the MH2 domain of SMADs.

sites, on the other hand, to the best of our knowledge no other specific sites have been identified as important for receptor binding specificity.

Figure 2 shows the SMAD2-MH2 domain colored by Sequence Harmony. The distribution of non-harmonious sites (red) is far from random, and in addition, the low harmony sites (orange) cluster along with them. In fact, we can identify a limited number of clustered regions high in low harmony sites, which are summarized in Table 1. The sites of known function in most clusters, allow us to assign putative functions to the remaining sites of unknown function.

Three of these clusters (#4, 5 and 7) are associated with receptor binding. The second-largest cluster (#2) is associated with c-Ski/SnoN interactions, and consists of a group of six largely non-sequentially low harmony sites. In addition, all sidechains point in the same general direction, forming a putative interaction surface. Three clusters (#1, 3 and 6) are associated with FAST1, Mixer and/or SARA binding. FAST1, Mixer share a SMAD interaction motif which is similar to that of SARA and have been shown to compete for binding to AR-SMADs.¹⁰ Here as well, the sidechains form a putative interaction surface. The surface-patch formed by the sidechains of the largest cluster, #1, is less regular. The two smallest clusters (#8 and 9) are formed of sites of unknown function. In the structure of the functional trimeric form of the SMAD2-MH2 domain residues of cluster #8 are close to several sites of known function *across the protein interface*.

4 Discussion & Conclusion

Sites of low Sequence Harmony correspond very specifically to functionally relevant sites in the SMAD-MH2 domain, with a very sharp separation between conserved positions

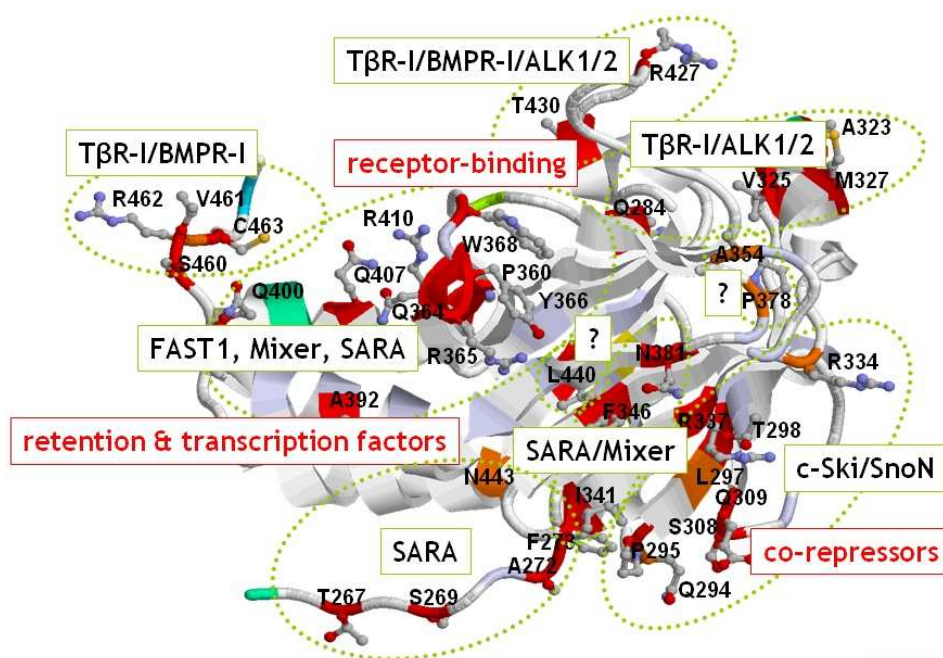


Figure 2. Sequence Harmony for AR-SMADs versus BR-SMADs colour-coded onto a SMAD2-MH2 structure (1KHx), cf. the values in Figure 1. Non-harmonious (SH zero) is red and labeled with residue names and numbers, maximal harmony (SH one) is white, intermediate values are rainbow colours from red to light blue.

Cluster	Sites	Ref.	Function
1	360, 364-366, 368, 392, 400, 407, 410	7,11,10,12	FAST1, Mixer, SARA, ?
2	294, 295, 297-298, 308, 309, 334, 337	7,13	c-Ski/SnoN, ?
3	267, 269, 272, 273, 443	7	SARA, ?
4	284, 323-325, 327	6,14	T β RI/BMPRI/ALK1/2
5	460-463	12	T β RI/BMPRI/ALK1/2
6	341, 346, 381	7,10	SARA/Mixer
7	427,430	12	T β RI/BMPRI/ALK1/2
8	354,378	-	?
9	440	-	?

Table 1. Structural clusters of low harmony sites (*cf.* Figure 2) and their respective known functions.

(which are the majority) and those that show a clear difference between the TGF- β and the BMP-binding sub-types. The Sequence Harmony scale corresponds to an intuitive interpretation of the differences in sequence composition and can be interpreted as sites that are more or less likely to be of functional importance. From the available experimental evidence in literature, it is difficult to identify false positive predictions, and almost no direct evidence is present to discriminate true from false negatives.

We have identified 13 sites of low Sequence Harmony in the SMAD-MH2 domain of unverified function and hereby suggest these as promising candidates for further elucidation of their function in determining the specificity of the TGF- β and BMP signalling pathways. Specifically, it would be very interesting to confirm (or rebuke) the putative functional roles we assigned to these sites of unknown function based on their proximity to low harmony sites of known function.

References

1. K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington. Homstrad: a database of protein structure alignments for homologous families. *Protein Sci.*, 7(11):2469–71, 1998.
2. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Res.*, 32(Database issue):D138–41, 2004.
3. S. S. Hannenhalli and R. B. Russell. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Molec. Biol.*, 303(1):61–76, 2000.
4. F. Pazos and M. J. E. Sternberg. Automated prediction of protein function and detection of functional sites from structure. *Proc. Nat. Acad. Sci. USA*, 101(41):14754–14759, 2004.
5. I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Molec. Biol.*, 336(5):1265–1282, 2004.
6. M. Huse, T. W. Muir, L. Xu, Y. G. Chen, J. Kuriyan, and J. Massague. The tgf beta receptor activation process: an inhibitor- to substrate-binding switch. *Mol. Cell*, 8(3):671–82, 2001.
7. G. Wu, Y. G. Chen, B. Ozdamar, C. A. Gyuricza, P. A. Chong, J. L. Wrana, J. Massague, and Y. Shi. Structural basis of smad2 recognition by the smad anchor for receptor activation. *Science*, 287(5450):92–7, 2000.

8. V. A. Simossis and J. Heringa. Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic. Acids. Res.*, 33(Web Server issue):W289–94, 2005.
9. J. W. Wu, A. R. Krawitz, J. Chai, W. Li, F. Zhang, K. Luo, and Y. Shi. Structural mechanism of smad4 recognition by the nuclear oncoprotein ski: insights on ski-mediated repression of tgf-beta signaling. *Cell*, 111(3):357–67, 2002.
10. R. A. Randall, S. Germain, G. J. Inman, P. A. Bates, and C. S. Hill. Different smad2 partners bind a common hydrophobic pocket in smad2 via a defined proline-rich motif. *Embo J.*, 21(1-2):145–56, 2002.
11. Y. G. Chen, A. Hata, R. S. Lo, D. Wotton, Y. Shi, N. Pavletich, and J. Massague. Determinants of specificity in tgf-beta signal transduction. *Genes Dev.*, 12(14):2144–52, 1998.
12. R. S. Lo, Y. G. Chen, Y. Shi, N. P. Pavletich, and J. Massague. The l3 loop: a structural motif determining specific interactions between smad proteins and tgf-beta receptors. *Embo J.*, 17(4):996–1005, 1998.
13. M. Mizuide, T. Hara, T. Furuya, M. Takeda, K. Kusanagi, Y. Inada, M. Mori, T. Imamura, K. Miyazawa, and K. Miyazono. Two short segments of smad3 are important for specific interaction of smad3 with c-ski and snon. *J. Biol. Chem.*, 278(1):531–6, 2003.
14. Y. G. Chen and J. Massague. Smad1 recognition and activation by the alk1 group of transforming growth factor-beta family receptors. *J. Biol. Chem.*, 274(6):3672–7, 1999.

High Throughput Method for Protein Structure Prediction

Dominik Gront, Sebastian Kmiecik, and Andrzej Kolinski

Laboratory of Theory of Biopolymers,
Faculty of Chemistry, Warsaw University
Pasteura 1, 02-093 Warsaw, Poland
E-mail: dgront@chem.uw.edu.pl

Recently several successful methods for protein structure prediction have been proposed. Next step towards modeling on a genomic scale is to combine existing tools into a single automated protocol. Such methods are crucial to fill the gap between the number of currently known protein sequences and structures. Here we utilize a lattice based coarse-grained modeling algorithm together with several accompanying tools to build a generalized pipeline for protein structure prediction. Our strategy was successfully applied during the CASP6 experiment.

1 Introduction

Building accurate 3D structural models for protein sequences of unknown structure is a challenging problem in contemporary computational biology. Large-scale genomic sequencing efforts provide increasing number of sequences. On the contrary the number of experimentally determined structures remains relatively small. Several computational methods that have been recently proposed can help to narrow this gap. The gap however is still getting wider. Therefore, development of high throughput protocol that would be able to predict protein structures with no human intervention becomes an urgent task.

2 Generalized Approach to Structure Prediction in a Reduced Conformational Space

The CABS modeling tool has been designed in a way allowing easy implementation of various restraints. Such restraints could be derived theoretically using various bioinformatics tools and databases of known structures, or experimentally from sparse NMR data, site-directed mutagenesis, etc. The approach is called „generalized”, since essentially the same strategy was employed to all types of protein targets, from comparative modeling (CM), through the fold recognition category (FR), to the most difficult new fold (NF) cases. This strategy combines FRankenstein (FR - Fold Recognition) method of Bujnicki¹² used for derivation of structural restraints, CABS simulation employed in a consensus model building, clustering and evaluation of models. The „FRankenstein Monster” algorithm builds models from large molecular fragments, properly extracted from the databases of known protein structures. These models are often quite accurate in the regions of regular secondary structure, while the loop connections are usually poorly predicted. The general idea was to generate a number of FRankenstein models, extract a large number of distance restraints (often self-contradictory) from these models and to apply them as the set of soft constraining potentials in the CABS simulations. Only the $C\alpha$ - $C\alpha$ distance restraints were used. The entire prediction pipeline could be outlined as follows (see flowchart on Figure 1a):

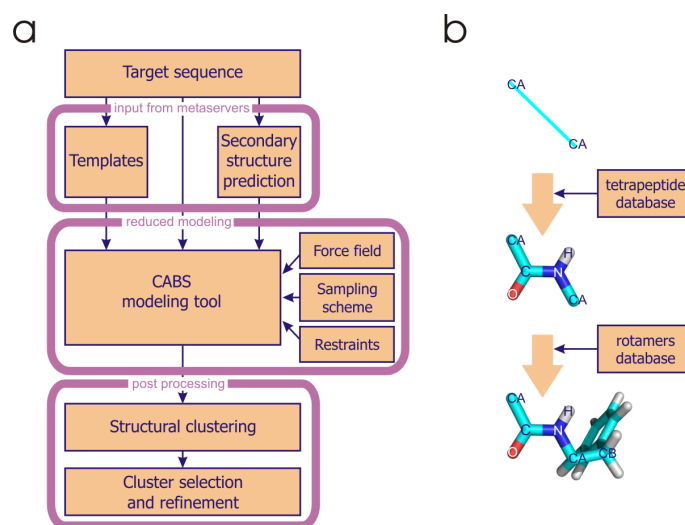


Figure 1. a) Flowchart representing main stages in protein structure prediction by the CABS modeling tool. b) Two-step procedure for building an all-atom model from its reduced representation

- Using GeneSilico metaserver¹² plausible templates (or their fragments) were identified, secondary structure of the target predicted, and FR alignments generated.
- Basing on the data from the metaserver the Frankenstein's all-atom models were built.
- The obtained models were evaluated by VERIFY3D method², and the high scoring fragments of each model were used as a source of distance restraints. The poorly scoring fragments of the initial models were left unrestrained. In the cases of the apparent New Folds very weak restraints were derived from fragments of FR models and from few alternative models generated by the Robetta server³.
- A number of copies of the target structure was generated using the CABS lattice discretization and the initial Frankenstein's models. These constituted the set of the starting replicas for a long simulated tempering Monte Carlo simulations. To obtain a sufficient number of replicas, not only different templates, but also different alignment variants were used in this stage.
- Thousands models in reduced representation from CABS simulations were clustered using HCPM (Hierarchical Clustering of Protein Models)^{4,5} algorithm, leading to 5-20 clusters, depending on the degree of convergence of the MC simulations. For a representative alpha-carbon backbone from each cluster a full-atom model was rebuilt.

This procedure has proven to be very efficient. Good predictions were achieved in all categories of the CASP targets and the group Kolinski-Bujnicki¹ using the outlined methodology has been classified as the second best, when averaging scores from all categories. Interestingly, for a number of targets the predicted molecular models were

closer to the true (released after the CASP meeting) structures of the targets than to any of the templates employed in modeling. In some cases of the NF targets reasonable models were obtained that were qualitatively different from all templates used. Apparently, the self-contradictory restraints became „diffused” by the CABS force field, while a small subset (for instance defining a plausible supersecondary element of the fold) of qualitatively correct restraints restricted sufficiently the conformational space during the simulations.

In the high-throughput modeling, especially when many proteins are modeled in a similar way, data processing in an automated fashion is especially important. As a result of our experience gained during the long process of multistage modeling of all CASP targets, we created a BioShell package⁶. BioShell is a way of performing and integrating computations employing UNIX shell as an interface platform. The package itself is composed of several programs, which extend functionality of shell or scripting languages, such as Python or Perl. Most of the examples given in the BioShell website (<http://biocomp.chem.uw.edu.pl/>) are related to the CABS modeling tool. Nevertheless, owing to standard file formats and algorithms, BioShell programs can be applied to virtually any modeling protocol.

3 Reconstructing the Full-Atom Representation

Simulations of reduced lattice models are usually a couple of orders of magnitude faster than simulations employing equivalent all-atom continuous models¹. One of the major drawbacks is the necessity to reconstruct all-atom representation from approximate coordinates of $C\alpha$ atoms. Here we describe a very efficient algorithm that can be applied on a genomic scale (Figure 1b)

For the non- $C\alpha$ backbone atoms average relative positions were derived in a local coordinate system by statistical analysis of PDB structures. Our method follows previous approaches by Purisima and Scheraga⁷ and by Milik⁸. Assuming a constant length of all $C\alpha$ - $C\alpha$ vectors, each backbone configuration for a tetrapeptide can be described by three $C\alpha$ distances: R_{13} (between first and third $C\alpha$ atoms in a tetrapeptide), R_{24} and R_{14} (defined similarly to R_{13}). The chirality of the backbone is taken into account by applying a sign to R_{14} . Negative values represent left-handed, and positive values represent right-handed conformations. These three distances form a three-dimensional grid in which average positions of C, O, and N atoms are accumulated from the PDB structures according to the local backbone configuration measured by R_{13} , R_{24} , and R_{14} after transformation into the local coordinate system. To make our implementation more accurate the grid spacing was chosen as 0.2 Å, rather than the 0.3 Å used by Milik et al.⁸. In order to reconstruct the C, O, and N backbone atoms, for each tetrapeptide average positions for C, O, and N are taken from the grid described above according to R_{13} , R_{24} , and R_{14} and transformed back into the original coordinate system. In the final step of the modeling, side chain atoms are reconstructed. For this purpose the SCWRL⁹ method can be used.

Our approach to backbone reconstruction turned out to be reliable, fast and very accurate. In a test on a set of native structures taken from PDB we compared our algorithm with several existing methods: bb¹⁰, MaxSprout¹¹, Pulchra¹² and Sybyl (Tripos) program which implements an algorithm by Claessens et al¹³. MaxSprout and Sybyl are the methods that

utilize fragment libraries derived from known structures. In many cases the two methods were not able to find a suitable fragment: MaxSprout succeeded only in 46% and Sybyl in 91% of the native C α traces in a test set. Among the other three methods our approach was found to be both the fastest and the most accurate. Average crmsd error for 70 chains measured on the fully reconstructed backbone is 0.42Å.

Acknowledgments

This work was partially supported by the grant # PBZ-KBN-088/P04/2003

References

1. A. Kolinski and J. M. M. Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, September 2005.
2. D. Eisenberg, R. Luthy, and J. U. Bowie. Verify3d: assessment of protein models with three-dimensional profiles. *Methods Enzymol*, 277:396–404, 1997.
3. D. Chivian, D. E. Kim, L. Malmstrom, P. Bradley, T. Robertson, P. Murphy, C. E. Strauss, R. Bonneau, C. A. Rohl, and D. Baker. Automated prediction of casp-5 structures using the robetta server. *Proteins*, 53 Suppl 6:524–533, 2003.
4. D. Gront and A. Kolinski. HCPM—program for hierarchical clustering of protein models. *Bioinformatics*, 21(14):3179–3180, July 2005.
5. D. Gront, U. H. E. Hansmann, and A. Kolinski. Exploring protein energy landscapes with hierarchical clustering. *Int J Quantum Chem*, 105(6):826–830, 2005.
6. D. Gront and A. Kolinski. Bioshell - a package of tools for structural biology computations. *Bioinformatics*, 22(5):621–622, March 2006.
7. E. O. Purisima and H. A. Scheraga. Conversion from a virtual-bond chain to a complete polypeptide backbone chain. *Biopolymers*, 23(7):1207–1224, July 1984.
8. M. Milik, A. Kolinski, and Jeffrey Skolnick. Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. *Journal of Computational Chemistry*, 18(1):80–85, 1996.
9. Jr Dunbrack and M Karplus. Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543–574, March 1993.
10. S. A. Adcock. Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield. *J Comput Chem*, 25(1):16–27, January 2004.
11. L. Holm and C. Sander. Database algorithm for generating protein backbone and side-chain co-ordinates from a c alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol*, 218(1):183–194, March 1991.
12. M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, and C. L. Brooks. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins: Structure, Function, and Genetics*, 41(1):86–97, 2000.
13. M. Claessens, E. van Cutsem, I. Lasters, and S. Wodak. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng.*, 2(5):335–345, January 1989.

From Microscopic Monte-Carlo Simulations to Macroscopic Solvation Models

Magdalena Gruziel^{1,2}, Witold R. Rudnicki², and Bogdan Lesyng¹

¹ Department of Biophysics,
Warsaw University, Poland
E-mail: {magd, lesyng}@icm.edu.pl

² Interdisciplinary Centre for Mathematical and Computational Modelling,
Warsaw University, Poland
E-mail: rudnicki@icm.edu.pl

The hydration free energy of model ionic and hydrophobic solutes are studied. Based on Metropolis Monte-Carlo simulations in the extended NpT ensemble, the Landau free energy surfaces, dependent on solute parameters, were determined. In particular, the solute charge as well as the diameter and well depth of the Lennard-Jones potential were chosen as the order parameters.

Regarding the charge and the Lennard-Jones potential diameter, simulations confirmed and/or completed other results obtained so far. The role of the third parameter, the potential well-depth, was studied in a more detail. It modifies slightly the solvation radius of the solute, and therefore influences the solvation free energy.

1 Introduction

Solvation/hydration phenomena take place in biomolecular systems and play very important role in technological processes. Therefore, theoretical description of these phenomena are very important. First of all, one has to estimate the free energy of solvation in such processes, which is a difficult problem by itself, and until now not solved sufficiently accurately. Also theoretical description of "subprocesses", such as hydrophobic and hydrophilic solvation, is still a challenging task.

In this work we study the solvation free energy of model ionic and hydrophobic solutes. From the results of exhaustive series of Metropolis Monte-Carlo simulations we determined the free energy functions, which are dependent on solute "control parameters". Our main goal is interpretation of the simulation results based on implicit solvent models and description of the solvation processes and binding free energies of more complex systems in the space of control parameters.

2 Methodology

Sampling of the configuration space was carried out using Metropolis Monte-Carlo method. Each simulation was carried out in the extended isothermic-isobaric ensemble with additional degree of freedom, either the solute charge q or one of the solute Lennard-Jones interaction potential parameters, ε or σ . In each simulation only one solute parameter was varied with the other two being fixed. The multi-window umbrella sampling approach was

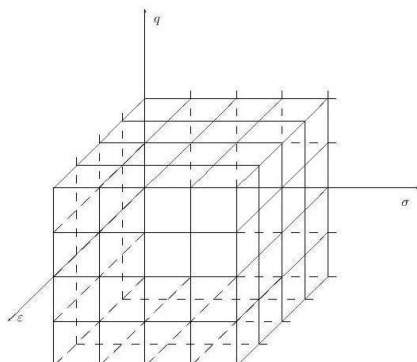


Figure 1. Simulation scheme. The axes represent 3 solute parameters under studies, ε , σ and q . Each line corresponds to a set of simulations (in consecutive windows) with one parameter being a simulation degree of freedom and the other two being fixed. For practical reasons, only one parameter was changed during one simulation. The ranges of parameters covered in this work: $\varepsilon \in [0.05; 12.8] kJ/mol$, $\sigma \in [2; 5] \text{\AA}$, $q \in [-1; 1]e$.

used to assure effective sampling of the parameter space. The results from consecutive simulation windows were merged together based on the following formula:

$$\Delta G(\xi) = -k_B T \ln \frac{\rho_{bias}(\xi_j)}{\rho_{bias}(\xi_i)} - \Delta V_{bias,ij}(\xi) \quad (1)$$

where ξ is either q , ε or σ , ρ is the probability density in the extended ensemble and V_{bias} is the biasing potential. The indices i, j indicate that the ΔG_{ij} is obtained from two consecutive and overlapping windows. Lynden-Bell et al.¹ applied this method earlier in the molecular dynamics simulations, sampling only q and σ .

3 Results

Results of our simulations with variable charge, are in good agreement with earlier work¹, namely, we have observed similar asymmetry of the free energy of ionic solvation for positive and negative charges. The dependence of the the solvation free energy on ε has not been studied before. The selected results are shown in Figure 2 (left plot).

One may notice the difference between curves corresponding to the neutral solute, in comparison to the charged solutes. Also clear asymmetry between positive and negative ions is visible. Using the results of one-dimensional simulations we were able to represent the free energy of solvation as a function of single parameter, with two other parameters fixed. We also were able to develop two-dimensional free energy functions. The example of the free energy of solvation as a function of Lennard-Jones parameters - $G(\varepsilon, \sigma)$ - is shown in Figure 2 (right plot). We failed to merge the $G(\varepsilon)$ and $G(\sigma)$ results of simulations with fixed negative solute charges.

The simulations yielded many structural data concerning the first solvation shell of water molecules. This information is very helpful in understanding the solvation process and also the shape of the free energy curves. These data are not presented here.

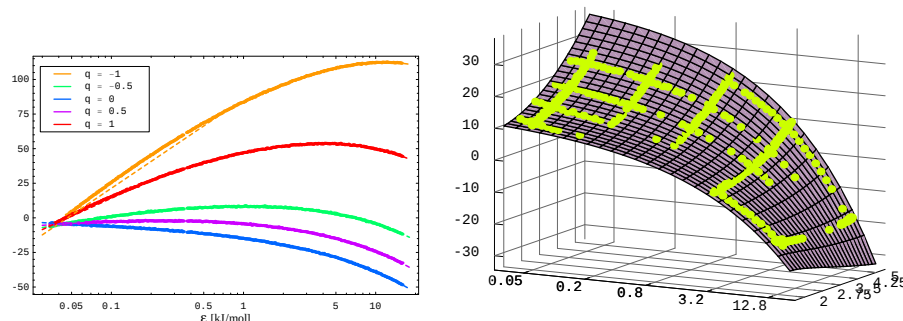


Figure 2. The solvation free energy curves with fitted functions. Left plot: $\Delta G(\varepsilon)$ for five fixed values of the solute charge, q (curves were shifted for better view). In all presented cases σ was fixed at $3.5\text{\AA}$. Right plot: $\Delta G(\varepsilon, \sigma)$, the solute charge was fixed at $q = 0$; vertical axis presents $\Delta G[\text{kJ/mol}]$, left horizontal axis shows the solute parameter $\varepsilon[\text{kJ/mol}]$, and right horizontal axis shows the solute parameter $\sigma[\text{\AA}]$. The bivariate free energy surface was calibrated with respect to experimental solvation free energies of methane².

4 Concluding Remarks

The results from the series of Monte Carlo simulations allowed us to propose solvation free energy functions dependent on solute parameters: $G(q)$, $G(\varepsilon)$, $G(\sigma)$ and some of $G(\varepsilon, \sigma)$. These solvation free energy functions may be further interpreted based on the mesoscopic solvation models, such as the one recently developed by Gallicchio et al.³.

The ε and σ sampling results from simulations with the fixed negative charge ($q = -1$, $q = -0.5$) did not merge, probably because the first solvation shell of water molecules did not maintain the tetrahedral structure around the small solute ($\sigma = 2\text{\AA}$) during the ε sampling. However, such structure is present during the σ sampling simulations (in the region of small σ and fixed small ε).

Acknowledgments

The simulations were carried in the Interdisciplinary Centre for Mathematical and Computational Modelling at Warsaw University. These studies were also partially supported by CoEs MAMBA and BioExploratorium.

References

1. R. M. Lynden-Bell, J. C. Rasaiah *From hydrophobic to hydrophilic behaviour: A simulation study of solvation entropy and free energy of simple solutes* J. Chem. Phys. **107**, 1981-1991 (1997)
2. A. Ben-Naim, Y. Marcus *Solvation thermodynamics of nonionic solutes*, J. Chem. Phys. **81**, 2016-2027 (1984).
3. E. Gallicchio and R. M. Levy *AGBNP: An analytic smplectic solvent model suitable for molecular dynamics simulations and high-resolution modeling* J. Comput. Chem. **25**, 479-499 (2004)

A Very General Solvation Model for BioMolecular Simulation

Siegfried Höfner

Michigan Technological University,
Department of Physics,
1400 Townsend Drive, 49931 Houghton, MI, USA
E-mail: shoefing@mtu.edu

A continuum description of solvation for macromolecular structure is presented in the framework of classical molecular mechanics. The essential kernel of the proposed solvation model is similar in spirit to the popular PCM approach frequently used in Quantum Chemistry. However, electronic degrees of freedom are not considered explicitly here and a partial charge discretization is employed instead. The major advantage with such a description is the clear-cut separation into individual physical terms that can all be parameterized independently. Three contributions will constitute the model, a polarization term derived from Poisson Boltzmann theory, a cavitation term obtained from free energy perturbation calculations and a dispersion term using modified Caillet-Claverie coefficients that can be determined from first principle calculations. The model description is very general and can be easily adapted for any kind of solvent considered.

1 Introduction

A proper theoretical description of the native behaviour of BioMolecules will always have to consider effects coming from the environment. Proteins, DNA, RNA, lipids, carbohydrates and all other classes of essential biochemical matter do usually exist and function only in their native environment, i.e. in water, membranes or some other kind of cellular matrix. If we therefore address complex biological behaviour with theoretical methods then we also have to somehow think of appropriate means to account for the environment.

The majority of cases in structural biology usually deals with aqueous systems. Water can be incorporated into theoretical models by either explicit solvation (addition of a shell of molecular water in full atomic detail) or by implicit solvation models (consideration of the environment as a structureless continuum having a specific dielectric constant, i.e. $\epsilon = 80$ in the case of water). In the following we want to focus on an example of implicit solvation. It is the classical analogon to the very powerful *Polarizable Continuum Method* widely used in the Quantum Chemistry community¹. Instead of quantum resolution of the electronic degrees of freedom we restrict ourselves to the approximation of atom-centered partial charges taken from the AMBER force field². Because the present description aims at covering non-electrostatic effects too, we would like to term this approach an *Enhanced Implicit Solvation Model*. It is according to the statement used by Dill et. al.³ a typical BIPSE model, which stands for *Break Into Pieces Sum-up the Energies*. The formal decomposition reads

$$\Delta G^{solv} = \Delta G^{pol} + \Delta G^{cav} + \Delta G^{disp} \quad (1)$$

and each of the individual terms together with their critical aspects shall be discussed next.

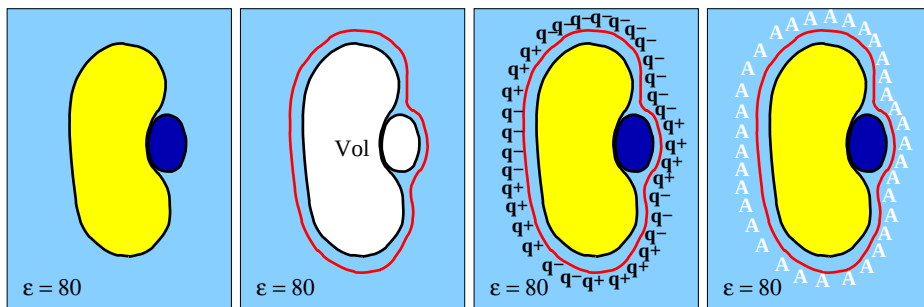


Figure 1. Elements of an Enhanced Implicit Solvation Model. a) Leftmost panel: A protein (yellow) in complex with some ligand (blue) resides in water considered as a dielectric continuum (light blue background). b) Second panel to the left: Formation of the boundary (red line) delivers the molecular surface and the molecular volume which in turn gives access to ΔG^{cav} . c) Second last panel to the right: Solution of the Poisson Boltzmann equation delivers a set of induced surface charges that allow calculation of ΔG^{pol} . d) Rightmost panel: Evaluation of response vector functions on the dielectric interface with appropriate Cailliet-Claverie dispersion coefficients give access to an estimate of ΔG^{dis} .

2 Methods

2.1 System Set Up

Consider a protein (with potentially complexed ligand) in aqueous solution (see leftmost panel in Figure 1). The first thing to accomplish within our enhanced implicit solvation model will be the definition of a dielectric interface where the dielectric constant will abruptly switch from $\epsilon = 1$ (protein interior) to $\epsilon = 80$ (solvent continuum). Due to the intended advantageous description restricted to the solute-solvent boundary⁴ the first step will comprise computation of the molecular surface and the molecular volume depicted as white area enclosed by a red line in Figure 1, second panel to the left. At first glance this might appear a trivial task, but in practice it turns out to represent a very crucial process with significant influence on overall performance and accuracy. A systematic study for a diverse set of structures using a variety of commonly employed molecular surface programs is currently underway in our group. The essential ingredients computed here are, boundary element location, size and corresponding normal vector, the net molecular surface area and the net molecular volume. All this implies prior assignment of AMBER charges and van der Waals radii.

2.2 Calculation of ΔG^{cav}

Having access to the molecular volume, or more specifically to the solvent excluded molecular volume (see second panel to the left in Figure 1), will enable us to derive an effective radius, $B_{eff} = \left(\frac{3}{4\pi} V^{slv.xcl}\right)^{\frac{1}{3}}$, that may be used in formulas of the *revised Pierotti Approach*⁵. Here a set of expansion coefficients, k_0 , k_1 , k_2 , was derived from Free Energy Perturbation Calculations based on Molecular Dynamics Simulations of molecular liquids (see Table 1 in reference⁵).

2.3 Calculation of ΔG^{pol}

The major computational effort will be devoted to calculation of the polarization term. This is in fact the only one term considered in many other solvation models. The problem may be efficiently solved within the *Boundary Element Method*⁶. Here a set of surface charges is derived from the assumption that the change in the normal component of the electric field at the dielectric boundary cannot become discontinuous. Taking into account these additional surface charges to the original set of partial charges of the protein will allow an estimation of the polarization term, ΔG^{pol} (see second last panel to the right in Figure 1). The problem may be efficiently solved from an iterative solution scheme using specialized computer hardware⁷.

2.4 Calculation of ΔG^{disp}

Similar to the electrostatic solute-solvent interaction an attractive van der Waals contribution can be obtained. Within the surface integral description the approach of Floris et. al. became well known⁸. Therefore we want to follow this strategy and calculate response vector functions A_i from re-parameterized Caillet Claverie parameters⁹ (see rightmost panel in Figure 1). These critical dispersion coefficients may nowadays be determined from ab-initio calculations as outlined by Amovilli et. al.¹⁰.

3 Concluding Remarks

Upon splitting the overall solvation free energy ΔG^{solv} into individual physical contributions and getting separate parameterizations for each of them independently, a rather solid description of environmental effects should arise. Work towards all of the described terms is currently in progress in our lab.

Acknowledgments

This work was supported in part by the National Institutes of Health Grant GM62838.

References

1. J. Tomasi, B. Mennucci, R. Cammi, *Quantum Mechanical Continuum Solvation Models*, Chem. Rev. **105**,8, 2999–3094 (2005).
2. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*, J. Am. Chem. Soc. **117**, 5179–5197 (1995).
3. N. T. Southall, K. A. Dill, A. D. J. Haymet, *A View of the Hydrophobic Effect*, J. Phys. Chem. B. **106**, 521–533 (2002).
4. A. H. Juffer, E. F. F. Botta, B. A. M. van Keulen, A. van der Ploeg, H. J. C. Berendsen, *The Electric Potential of a Macromolecule in a Solvent: A Fundamental Approach*, J. Comp. Phys. **97**, 144–171 (1991).

5. S. Höfinger, F. Zerbetto, *Simple models for hydrophobic hydration*, Chem. Soc. Rev. **34**, 1012–1020 (2005).
6. R. J. Zauhar, R. S. Morgan, *A new method for computing the macromolecular electric potential*, J. Mol. Biol. **186**, 815–820 (1985).
7. S. Höfinger, *Solving the Poisson-Boltzmann Equation with the Specialized Computer Chip MD-GRAPE-2*, J. Comp. Chem. **26**, 1148–1154 (2005).
8. F. M. Floris, J. Tomasi, J. L. Pascual Ahuir, *Dispersion and repulsion contributions to the solvation energy: Refinements to a simple computational model in the continuum approximation*, J. Comp. Chem. **12**, 784–791 (1991).
9. J. Caillet, P. Claverie, *Theoretical evaluations of the intermolecular interaction energy of a crystal: application to the analysis of crystal geometry* Acta Cryst. A **31**, 448–461 (1975).
10. C. Amovilli, B. Mennucci, *Self-Consistent-Field Calculation of Pauli Repulsion and Dispersion Contributions to the Solvation Free Energy in the Polarizable Continuum Model* J. Phys. Chem. B **101**, 1051–1057 (1997).

Prediction of Domain Interactions in *C. elegans*

Faruck Morcos¹, Mike Boxem², Niels Klitgord², Marc Vidal², and
Jesús A. Izaguirre^{1,3}

¹ Department of Computer Science and Engineering,
Notre Dame IN 46556 USA.

E-mail: {amorcosg, izaguirr}@nd.edu

² Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology,
Dana-Farber Cancer Institute, and Department of Genetics,
Harvard Medical School, Boston MA 02115 USA

E-mail: {mike_boxem, niels_klitgord, marc_vidal}@dfci.harvard.edu

³ Corresponding author.

High throughput experiments and computational methods allow the elucidation of networks of protein-protein interactions (PPI) in several organisms. Since about 80% of all proteins consist of multiple self-folding units called domains, it is desirable to determine which domains are responsible for a given interaction. We use an integrative computational and experimental approach to predict domain-domain interactions (DDI) from the known PPI for the metazoan *C. elegans*. We use Maximum Likelihood Estimation and Maximum Specificity Set Cover algorithms to predict the DDI. We validate the results using a small and accurate yeast 2 hybrid (Y2H) screen on 8 protein-protein interactions involving 13 proteins in *C. elegans*. The computational predictions are consistent with Y2H in 7 of the 13 interaction domains, including the prediction of domain fusions. These results illustrate the predictive power of a combined computational and experimental approach to DDI determination.

1 Introduction

Macromolecular interactions among proteins, nucleic acids, and lipids, form the basis of biological activity. High-throughput experiments have enabled the elucidation of a substantial number of protein-protein interactions (PPI) in several organisms. About 80% of these proteins consist of multiple domains, which are self-folding polypeptide units. Many of these domains are repeated across different proteins, and hence allow for a modularization of the interaction networks. The domain architecture of many proteins is available in Pfam¹. We use a network of PPI and Pfam to infer a network of domain-domain interactions (DDI). We validate this network on a set of 8 well-characterized protein interaction pairs involving 13 proteins from the roundworm *C. elegans*. Validation is done using an experimental yeast 2 hybrid (Y2H) approach. Our results show substantial agreement between the computational prediction methods and experiment (agreement for 7 out of 13 interaction domains). Several of the predicted interaction domains consist of domain fusions. Several of the domain interactions are predicted by two different computational methods, adding to the confidence level of the predictions. Failures of the computational predictions can be partly explained by the ability of the experiments to discover novel interaction domains, and in one case by the fusion of 3 domains, which was not considered in the computational model. This integrative approach has the potential to explain at a greater level of detail observed PPI.

Interaction	Description
SQV-4 – SQV-4	UDP-glucose 6-dehydrogenase, required for cytokinesis.
EMB-27 – B0511.9	Anaphase promoting complex subunit and novel gene.
MEI-1 – MEI-2	Together forms the microtubule severing katanin complex.
DYRB-1 – DYCI-1*	Homology to dynein subunits (light and intermediate chains).
Y65B4BR.5 – ICD-1	Predicted transcription factor and β NAC.
MEL-26 – MEI-1s	Ubiquitin ligase substrate adaptor targeting MEI-1.
MEL-26 – MEL-26	Ubiquitin ligase substrate adaptor targeting MEI-1.
ATN-1* – ATN-1*	α -actinin, actin bundling protein that homodimerizes.
GPR-1* – GOA-1	GPR-1 regulates the activity the G-protein subunit GOA-1.
* Interaction domain known	

Table 1. Y2H protein interaction set of high confidence.

2 Methods

The network of PPI examined consists of a set of 9,044 interactions among 4,625 proteins, collectively named WI7². From the 9,044 PPI, 4,736 are experimental, 3,359 based on homology with *Drosophila* interactions, and 949 based on homology with *S. cerevisiae* interactions. The domain architecture was obtained from Pfam-A and Pfam-B, yielding a total of 8,558 interactions with known domain architecture. This was used as training set for two different algorithms, which yield a network of DDI. The first is the maximum likelihood estimation (MLE) procedure which has been applied successfully to the problem of estimating the DDI from experimental PPI for yeast³. The other algorithm is a weighted set cover algorithm that we developed, called Maximum Specificity Set Cover (MSSC), which selects domain-domain interactions while minimizing the number of false positives in the training set⁴. Domain fusions of 2 domains are included in the set of potential interaction domains.

Previous work has validated the ability of using the network of DDI found by MLE and MSSC to predict PPI. In this work we attempt to directly evaluate the quality of a small portion of these DDI. Suppose that an interaction domain pair $d_i - d_j$ is part of the DDI produced by any of the above methods and can explain a protein interaction $p_i - p_j$. Y2H is capable of accurately identifying the interaction domain for a protein-protein interaction. First the domain d_i in p_i , when interacting with p_j , is identified; then d_j in p_j is identified when interacting with p_i . From the Worm Interactome⁵ version 5, we selected 8 interactions involving 13 proteins (Table 1). These interactions were selected because they are high-confidence Y2H interactions, and the proteins are involved in a diverse set of biological processes. This set of interactions is used to compare experimental and computational predictions.

3 Results

MLE and MSSC were run from the PPI server (<http://ppi.cse.nd.edu>). MLE selected 5,170 domains and 36,489 DDI. MSSC explained the same network with 3,138 domains and 12,870 DDI. MSSC predicted 189 domain fusions involved in 1,872 DDI. Predicted interaction domains for the small screen test are shown in Table 2. By inspection, 7 of the 13 interaction domains are consistent with Y2H. SQV4 homodimerization could not be predicted by only using fusion of 2 domains, since Y2H indicates that 3 domains are needed. MEI-1 requires the entire protein in the experiment, and the predicted interaction

Protein Interaction	Domain Interaction	Remarks
SQV-4 (3) – SQV-4 (3)	UDPG_MGDP_dh – UDPG_MGDP_dh	Y2H indicates 3 domains needed –
EMB-27 (2) – B0511.9 (1)	Pfam-B_41051 – Pfam-B_92146 [†]	Pfam-B_41051 not consistent with Y2H – Y2H consistent with Pfam-B_92146
MEI-1 (1) – MEI-2 (1)	AAA – Pfam-B_62049	Y2H consistent with AAA Pfam-B_62049 not consistent with Y2H –
DYRB-1 (1) – DYCI-1* (4)	Robl_LC7 – Pfam-B_4434	DYRB-1 had no hit in Y2H – Y2H consistent with Pfam-B_4434
Y65B4BR.5 (3) – ICD-1 (2)	UBA_NAC – NAC	Y2H consistent with fusion Y2H consistent with NAC
MEL-26 (3) – MEL-26 (3)	BTB_MATH – BTB_MATH	Y2H consistent with fusion
ATN-1* (4) – ATN-1* (4)	Pfam-B_843 – Pfam-B_843 [†]	Pfam-B_843 not consistent with Y2H –
GPR-1* (2) – GOA-1 (1)	Pfam-B_66388 – G-alpha [†]	Pfam-B_66388 not consistent with Y2H – Y2H consistent with G-alpha
* Interaction domain known, [†] Predicted by both MLE and MSSC		

Table 2. Set of DDI predicted by MLE and MSSC. The notation (#) besides the protein name represents the number of domains that conform such protein. A – at the end of a remark indicates this was a negative result.

domain AAA is contained within this area. The interaction domain in MEI-2 is consistent with experiment although Y2H used a smaller piece of the complete predicted interaction. Predicted networks of DDI can serve both as a starting point in designing Y2H screens, as well as the basis for predicting further PPI. Another direction that we are pursuing is to further validate the interaction domains through structure prediction and protein-protein docking.

Acknowledgments

JAI and FM were partially supported by NSF grants ACI-0135195 and DBI-0450067. FM is recipient of a Kellogg graduate fellowship.

References

1. A. Bateman et al. The Pfam Protein Families Database. *Nucl. Acids Res.*, 32, 2004.
2. K.C. Gunsalus et al. Predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861–865, 2005.
3. M. Deng, S. Mehta, F. Sun, and T. Cheng. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, 12:1540–1548, 2002.
4. C. Huang, F. Morcos, S. P. Kanaan, S. Wuchty, D. Z. Chen, and J. A. Izaguirre. Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Trans. on Comput. Biology Bioinformatics*, 2006. In press.
5. S. Li et al. A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543, 2004.

Adsorption Phenomena at Organic-Inorganic Interfaces

Michael Bachmann and Wolfhard Janke

Institut für Theoretische Physik and Centre for Theoretical Sciences (NTZ),
Universität Leipzig, 04109 Leipzig, Germany
E-mail: {bachmann, janke}@itp.uni-leipzig.de

The qualitative solvent- and temperature-dependent conformational behavior of a peptide in the proximity of solid substrates with different adsorption properties is investigated by means of a simple lattice model. The resulting pseudophase diagrams exhibit a complex structure, which can be understood by analysing the minima of the free-energy landscape in dependence of appropriate system parameters.

1 Introduction

The interest in understanding adsorption of polymers and peptides to inorganic material has grown recently with the development of nanotechnological equipment that enables present and future applications ranging from nanoelectronic devices [as, for example, the celebrated organic light-emitting diodes (OLEDs)] to pattern-recognising nanosensors in biomedicine. In recent experiments, evidence for substrate specificity of synthetic peptide adsorption to semiconductor substrates has been found.¹ The investigation of organic-inorganic interfaces is connected with the substantial problem of setting up an effective microscopic model for the adsorption of organic to inorganic matter in addition to the intrinsic interactions governing the folding properties of the organic substance. Van der Waals based interaction models² were employed, e.g., in studies of gold-binding peptides.³ Here, we study a minimalistic lattice model, where the folding part is governed by the hydrophobic-polar (HP) model,⁴ and the interaction with the substrate is restricted to monomer-type dependent surface-layer attraction.^{5,6}

2 Conformational Transitions in a Simple Hybrid Interface Model

Based on the assumption that peptide folding is mainly due to the hydrophobic effect, we distinguish only hydrophobic (H) and polar residues (P). According to the minimalistic HP model⁴, the intramolecular energy of a conformation is related to the number of hydrophobic nearest-neighbor contacts n_{HH} . For the study of specificity of surface-binding, we investigate three attractive substrate models: (a) type-independent attractive, (b) hydrophobic, and (c) polar substrate. In case (a), the energy of the system is proportional to the total number of monomer-surface contacts, n_s^{H+P} , while in the cases (b) and (c) the respective hydrophobic (n_s^H) and polar (n_s^P) surface contacts are energetically favored. The generic model can be defined by $E(n_s, n_{HH}) = -n_s - sn_{HH}$, where, depending on the substrate model, $n_s = n_s^{H+P}$, n_s^H , or n_s^P . The parameter s controls the solvent quality (the larger the value of s , the worse the solvent). The contact numbers n_s and n_{HH} are natural system parameters that allow the discrimination of conformational pseudophases. In our investigations of this system, we employed the contact-density chain-growth method which

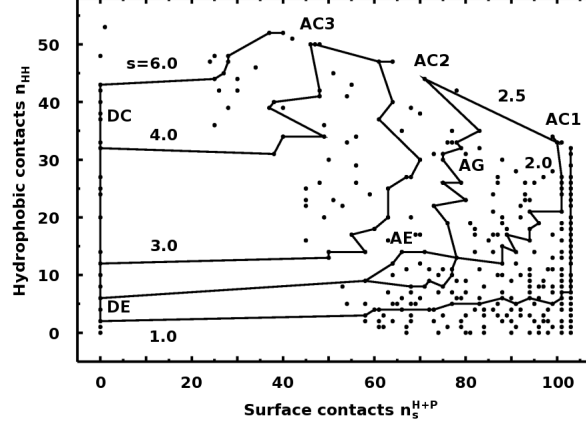


Figure 1. Contact-number map of all free-energy minima for an exemplified peptide with 103 monomers⁵ and a substrate which is equally attractive to all monomers. Circles correspond to minima of the contact free energy $F_{T,s}(n_s^{H+P}, n_{HH})$ in the parameter space $T \in [0, 10]$, $s \in [-2, 10]$. Lines illustrate how the contact free energy changes with the temperature at constant values of the solvent parameter s .

has successfully been applied in adsorption studies of polymers.⁸ This method, which is a suitably generalized variant of the multicanonical chain-growth algorithm,⁷ allows us to estimate the contact density $g(n_s, n_{HH})$, which simply counts the number of conformations for a given pair of contact numbers n_s and n_{HH} . For fixed temperature T and solubility s , we define the contact free energy $F_{T,s}(n_s, n_{HH}) \sim -T \ln g(n_s, n_{HH}) \exp(-E_s/T)$. Assuming that the minimum of the free-energy landscape $F_{T,s}(n_s^{(0)}, n_{HH}^{(0)})$ is related to the class of macrostates with $n_s^{(0)}$ surface and $n_{HH}^{(0)}$ hydrophobic contacts, this class dominates the pseudophase the system resides in. For this reason, it is instructive to calculate all minima of the contact free energy and to determine the associated contact numbers in a wide range of values for the external parameters. This can easily be done with the knowledge of $g(n_s, n_{HH})$ by simple reweighting.

In Fig. 1, the map of all possible free-energy minima (circles) in the range of external parameters $T \in [0, 10]$ and $s \in [-2, 10]$ is shown for a peptide with 103 monomers (37 hydrophobic, 66 polar) in the vicinity of a substrate that is equally attractive for both hydrophobic and polar monomers.^{5,6} There are a few regions of circles which can be associated with pseudophases. Clearly, where the number of surface contacts n_s^{H+P} is zero, the heteropolymer is desorbed and in the bulk regime two main pseudophases can be identified: desorbed extended (DE) and desorbed collapsed (DC) hydrophobic-core conformations. In case of adsorption, we distinguish also extended (AE) and compact (AC1-3) conformations, and an intermediate globular pseudophase (AG), where conformations are globally compact, but do not exhibit pronounced hydrophobic domains. The discrimination of the three AC subphases is due to layering: In dependence of the solvent quality, hydrophobic domains form layers: single-layer domains in AC1 (good solvent, substrate contact favored prior formation of intrinsic hydrophobic contacts), and respective hydrophobic double- and triple-layer cores in AC2 and AC3. Solid lines in Fig. 1 visualize “paths” through the

free-energy landscape when changing temperature, but leaving the solvent parameter unchanged ($s = \text{const}$). Recalling that larger s values correspond to worse quality of the solvent, it is obvious that the $s = 6$ path at low temperatures starts in the highly compact AC3 subphase. Increasing the temperature, the monomers reorder and the system enters for entropic reasons the globular pseudophase AG. Approaching $T \approx 2.3$ a noticeable first-order-like unbinding transition occurs towards pseudophase DC, before the compact structure decays (DE). Under very good solvent conditions as, e.g., for the trajectory of $s = 1$, the favored conformations at low temperatures are film-like. Increasing the temperature slightly to $T \approx 0.2$, the system crosses over to the AE pseudophase. This subphase is very stable under these solvent conditions, but close to $T \approx 2.3$ the heteropolymer desorbs into the dissolved bulk phase DE.

For the other two substrates, the pseudophase diagrams look quantitatively different but most of the main pseudophases are also present in these cases.⁵

3 Concluding Remarks

Beside the expected adsorbed and desorbed phases, hybrid organic-inorganic systems exhibit a rich subphase structure in the adsorbed phases with compact hydrophobic domains. In this region, the binding behavior is strongly substrate-specific and depends in detail on the quality of the solvent. Since current experimental equipment can reveal molecular structures at the nanometer scale, it should be possible to investigate the grafted structures in dependence of the solvent quality. Such experiments would be essential for studies of binding forces that are strong enough to refold peptides or proteins.

Acknowledgments

This work is partially supported by the DFG (German Science Foundation) under Grant No. JA 483/24-1 and the computer time Grant No. hlz11 of the John von Neumann Institute for Computing (NIC), Forschungszentrum Jülich.

References

1. S. R. Whaley, D. S. English, E. L. Hu, P. F. Barbara, and A. M. Belcher, *Nature* (London) **405**, 665 (2000); K. Goede, P. Busch, and M. Grundmann, *Nano Lett.* **4**, 2115 (2004).
2. R. Hentschke, *Macromol. Theory Simul.* **6**, 287 (1997).
3. R. Braun, M. Sarikaya, and K. Schulten, *J. Biomater. Sci. Polym. Ed.* **13**, 747 (2002).
4. K. A. Dill, *Biochemistry* **24**, 1501 (1985).
5. M. Bachmann and W. Janke, *Phys. Rev. E* **73**, 020901(R) (2006);
6. M. Bachmann and W. Janke, *Chain-growth simulations of lattice-peptide adsorption to attractive substrates*, in: *Proceedings of the NIC Symposium 2006*, John von Neumann Institute for Computing, Jülich, NIC Series vol. **32**, ed. by G. Münster, D. Wolf, M. Kremer (NIC, Jülich, 2006), p. 245.
7. M. Bachmann and W. Janke, *Phys. Rev. Lett.* **91**, 208105 (2003).
8. M. Bachmann and W. Janke, *Phys. Rev. Lett.* **95**, 058102 (2005); *Phys. Rev. E* **73**, 041802 (2006).

Modelling protein-DNA interactions in chromatin

Giovanni La Penna, Sara Furlan, Luca Arcesi, and Angelo Perico

Institute for Macromolecular Studies,
National Research Council, 16149 Genova, Italy
E-mail: {lapenna}@ge.ismac.cnr.it

Proteins with positive charges strongly interact with DNA and, in the important cases of histones and histone-like proteins, they pack and bend long segments of DNA into highly ordered structures. In this contribution we report the results of combining numerical simulations in generalized ensembles (modulated metastatistics) with polyelectrolyte theory. The example of the interaction between the N-terminal tail of the H3 histone with segments of DNA will be illustrated, to explain the role of histone charged tails in the compaction of the genetic code in eukaryotic cells.

1 Introduction

In the eukaryotic cells, DNA is localized in the nucleus. The localization and length of the DNA filaments, require a high degree of packing for the entire genetic material. This high degree of packing also allows silencing of DNA during the cell interphase and a limited well-regulated expression of the genes¹.

The storage of DNA, that carries 2 negative charges per bp at physiological pH, in the nuclei is performed by aggregates of histone proteins, carrying a net positive charge and, therefore, attempting to form a neutral complex with DNA. The resulting complex of DNA and histone proteins is called chromatin. Each particle formed by 8 histone aggregated proteins and a portion of DNA wrapped around it is the nucleosome core particle (NCP). NCPs are linked by the non-wrapped portions of DNA called “linkers”.

Only recently, the structure of chromatin has been investigated at an atomic resolution by X-ray diffraction² and by other techniques. The structure of highly dense chromatin is now hypothesized as a two-start helix, with the linker DNA significantly bent. The reasons of such bending has never been explored and they are the object of this study. The best candidates for the linker bending are the N-terminal tails of H3 histones: these tails are disordered and long enough to touch the linker DNA.

2 The N-terminal tail of H3 histones

The behaviour of these tails both in the absence of DNA and in the vicinity of DNA can be simulated using atomic models. First, a self-avoiding least-biased random walk³ for a model of 25 residues moving around a DNA segment of 10 bp was performed. Ten different initial orientations of DNA with respect to the protein were used as ten independent replica. The protein was anchored at its C-terminus (residue 26) to mimick the anchoring of the N-terminal tail to the H3 rigid core.

Configurations were studied in terms of the most relevant configurational variable related to protein-DNA interactions, *i.e.* the number of electrostatic contacts, C . This quantity is defined in terms of $C\zeta(\text{Arg})/N\zeta(\text{Lys}) - P(\text{DNA})$ distances: when this distance is lower

Z	l (bp)	R_{12} (nm)
0	0	23.5
5	4	15.2
10	8	10.2
15	13	6.8

Table 1. - Length l of bent DNA around a circle of $R = 1.5$ nm with different charge Z . R_{12} is the distance between two coplanar NCPs linked by a 65 bp linker DNA. Therefore, the first row is for a straight linker DNA.

than 0.5 nm a contact is counted. From the entire basin of configurations obtained through the random walk, 7 configurations with high ($C = 8$) and 7 with low ($C = 2$) numbers of protein-DNA contacts were immersed in a bath of water molecules and sodium counterions. These 14 systems were simulated via MD both in the presence and absence of DNA.

The results of these simulations can be summarized as in the following⁶: on average the protein increases its helical content when DNA is present in the nearby. and, consequently, the radius of gyration R is smaller ($R \sim 1$ nm) than when proteins are completely disordered or they are simulated in the absence of DNA ($R > 1.5$ nm). The MD simulations carried the protein towards configurations with about 6 protein-DNA contacts on average, irrespective of the initial number of contacts. The analysis of energetics shows that a larger number of contacts produces an increase of hydrophobic energy that may even balance the decrease of electrostatic energy. Finally, the 14 simulations of protein in the nearby of DNA did not show a specific bound configuration, rather the configurations are of different types: protein packed in the DNA major groove; protein stuck on the minor groove; protein bridging two sides of the major groove. These different configurations have similar energies.

3 The long-range bending of DNA

Once we know that the H3 N-terminal tail can be efficiently and non specifically packed by its nearby DNA, we ask if the resulting charge density (8 positive charges localized in a sphere of about 1 nm radius) can be responsible of a significant linker DNA bending.

The free energy of the protein-DNA complex can be written as a function of the length l of DNA wrapped on the protein, assuming the protein be a circle of radius R with positive charge Z . The free energy accounts for counterion decondensation that occurs when the protein comes in contact with DNA and replaces the shell of counterions condensed by the high negative charge density of DNA^{4,5}. The length of wrapped (and therefore bent) DNA is obtained by finding the minimum of the free energy. Results are summarized in Table 1, assuming the bulk salt concentration $C_s = 0.1$ M.

From these results, the answer to the question at the beginning of the section is: a charge of +10 confined in a region of 1.5 nm radius is able to bend 8 bp of DNA and to bring two NCPs at distances even lower than their contact distance ($R_{12} < 11$ nm).

4 Conclusions

The significant bending of linker DNA in chromatin can be performed by the N-terminal tails of H3 histones that are projected in the internucleosomal space. The extent of bending (~ 8 DNA bp) can be derived by placing a positive charge of $Z \sim 10$ within a sphere of radius $R \sim 1.5$ nm, that is similar in charge density with the collapsed H3 N-terminal tail of 25 residues as it is modelled using atoms. By using these parameters, the distance between two nucleosome core particles linked by about 65 bp of DNA (a value in the range of experimental values) are brought at contact distance and a high degree of chromatin packing is obtained only acting on the linker DNA bending. This confirms the possibility for the H3 N-terminal tail for acting as switches for keeping the chromatin dense when tails are charged, and for relaxing the structure when tails are neutralized. This latter event is actually performed by lysine acetylation due to specific enzymes and is limited to the N-terminal tails because of their larger accessibility compared to the histone cores.

Acknowledgments

This work has been done with the financial support of the FIRB 2001 project RBA01Y3SN of MIUR (Italy). The authors acknowledge prof. Eligio Patrone (CNR-ISMac) for stimulating discussions.

References

1. J. Widom, *Structure, dynamics and function of chromatin in vitro*, Ann. Rev. Biophys. Biomol. Struct. **27**, 285–327 (1998).
2. T. Schalch, S. Duda, D. F. Sargent, and T. J. Richmond, *X-ray structure of a tetranucleosome and its implications for the chromatin fibre*, Nature **436**, 138–141 (2005).
3. G. La Penna, S. Morante, G. C. Rossi, and A. Perico, *Designing generalized statistical ensembles for numerical simulations of biopolymers*, J. Chem. Phys. **121**, 10725–10741 (2004).
4. G. S. Manning, *Is a small number of charge neutralizations sufficient to bend nucleosome core DNA onto its superhelical ramp?* J. Am. Chem. Soc. **125**, 15087–15092 (2003).
5. A. Perico, G. La Penna, and L. Arcesi, *Electrostatic interactions with histone tails may bend linker DNA in chromatin*, Biopolymers **81**, 20–28 (2006).
6. G. La Penna, S. Furlan, and A. Perico, *Modelling H3 histone N-terminal tail and linker DNA interactions*, Biopolymers, to be published (2006).

Molecular Dynamics Study of Amyloid Formation of Two Abl-SH3 Domain Peptides

Inta Liepina¹, Salvador Ventura², Cezary Czaplewski³, and Adam Liwo³

¹ Latvian Institute of Organic Synthesis,
Riga, LV1006, Latvia
E-mail: inta@osi.lv

² Institut de Biotecnologia i de Biomedicina,
Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
E-mail: salvador.ventura@uab.es

³ Faculty of Chemistry, University of Gdańsk,
80-952 Gdańsk, Poland
E-mail: {czarek, adam}@chemik.chem.univ.gda.pl

Two and ten antiparallel strand β -sheets built from DLSFMKG (MK) and DLSFKKGE (KK) peptides, were immersed in periodic water boxes and subjected to molecular dynamics simulations (MD) at 30K, 170K/190K - for two and ten strand β -sheets respectively, and to 300 K. MD simulations showed that 2xMK β -sheet is more stable than 2xKK β -sheet, and 10xMK β -sheet is more stable than 10xKK β -sheet, suggesting that MK systems are fibril creating and the KK systems are not.

1 Introduction

Amyloid formation and deposit is connected with Alzheimer's disease, Parkinson's disease, and Finnish familial amyloidosis. After protein misfolding short peptide sequences act as "hot spots" providing the driving force for protein aggregation in amyloid fibrils.

2 Results and Discussion

Previously we have characterized in detail one of these "hot spots" in the diverging turn of PI3-SH3 domain, one of the most extensively studied amyloidogenic proteins not related to a disease^{1,2}. Based on an homology search we have identified an aggregation prone region in the same structural element of the related Abl-SH3 domain of *Drosophila* (DLSFMKGE) whereas the human homologous region (DLSFKKGE) is predicted to be less amyloidogenic. The possible reason for the difference of amyloid formation propensities of the two peptides was investigated by molecular dynamics (MD) of β -sheet structures. The antiparallel alanine β -sheets consisting of two and ten strands were constructed, minimized, and mutated to the sequences DLSFMKGE and DLSFKKGE. All four systems: 1) DLSFMKGE - two strands (2xMK) (fig 1 a-f), 2) DLSFKKGE - two strands (2xKK) (fig 1 g-k), 3) DLSFMKGE - ten strands (10xMK) (fig 2 a-d), 4) DLSFKKGE - ten strands (10xKK) (fig 2 e-h), were surrounded by 10 Å layer of water molecules over the solute and subjected to MD, Amber 8.0 force field, NTP protocol. The MD runs were started at the temperature of 10 K and the temperature was elevated stepwise by 10 degrees till 300 K. Longer MD runs were done at 30 K (78 ns), 170 K (140 ns) and 300

K for the two strand systems, and at 30 K (47 ns), 190 K (55 ns) and 300 K for the ten strand systems. The results (fig 2) show considerably higher hydrogen bond percentage for DLSFMKGE than that one for DLSFKKGE during the course of the simulation, thus suggesting that DLSFMKGE is a potential fibril- maker, but DLSFKKGE is not. Two strand β -sheet systems were stable until 170 K (fig 1). The ten strand β -sheets appears to be clearly more stable (fig 2). The fact that with the course of the simulation a β -sheet partly or completely melts at higher temperatures suggests, the need for a complex of several face to face oriented β -sheets to provide stability to the amyloid fibril.

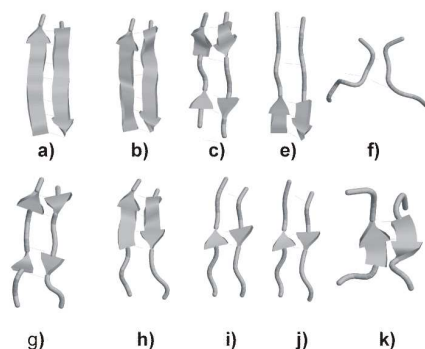


Figure 1. MD snapshots of DLSFMKGE two strand system : a) at 30K, 23159 ps, b) at 170K, 42359 ps, c) at 170 K, 91467 ps, e) at 170 K, 99126 ps, f) at 300 K, 64231 ps; MD snapshots of DLSFKKGE two strand system: g) at 30K, 19941 ps, h) at 170K, 42880 ps, i) at 170K, 92658 ps , j) at 170K, 98969 ps, k) at 300K, 94276 ps.

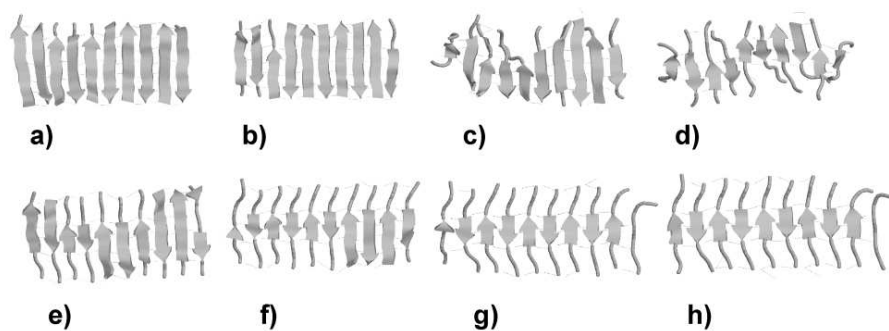


Figure 2. MD snapshots of DLSFMKGE ten strand system: a) at 30 K, 15343 ps, b) at 190 K, 58023 ps, c) at 300K, 58489 ps, d) at 300 K, 64419 ps; MD snapshots of DLSFKKGE ten strand system: e) at 30 K, 15950 ps, f) at 190K, 58588 ps, g) at 300 K, 58466 ps, h) at 300 K, 64038 ps.

Acknowledgments

This work was supported by NATO Collaborative Linkage Grant LST.CLG.979807, by stipend for I.L. from the "Kasa im. Jozefa Mianowskiego" Fund, Poland, by Latvian Science Council Grant 05.1768, by Gdansk Academic Computer Center TASK.

References

1. S. Ventura, E. Lacroix, L. Serrano, *Insights into the origin of the tendency of the PI3-SH3 domain to form amyloid fibrils*, J.Mol.Biol. **332**, 1147-58 (2002).
2. S. Ventura, J. Zurdo, S. Narayanan, M. Parreno, R. Mangues, B. Reif, F. Chiti, E. Giannoni, C.M. Dobson, F.X. Aviles, L. Serrano, *Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case*, Proc. Natl. Acad. Sci. USA **101**, 7258-63 (2004).

Optimization of the United-Residue UNRES Force Field for Langevin Dynamics Simulations

Adam Liwo^{1,2}, Mey Khalili¹, Cezary Czaplewski^{1,2}, Sebastian Kalinowski²,
Stanisław Oldziej^{1,2}, and Harold A. Scheraga¹

¹ Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca, NY 14853-1301, U.S.A.,
E-mail: has5@cornell.edu

² Faculty of Chemistry, University of Gdańsk,
Sobieskiego 18, 80-952 Gdańsk, Poland

Recently, with success, we applied the united-residue UNRES force field developed in our laboratory to carry out molecular dynamics simulations. This communication is a preliminary report of the development of a procedure for optimizing the UNRES force field for canonical simulations.

1 Introduction

In the last decade we have been developing a united-residue physics-based force field termed UNRES¹⁻³ for energy-based prediction of protein structures from amino-acid sequences. Recently⁴, we extended the application of UNRES to mesoscopic molecular dynamics simulations and found that, with this approach, we can simulate protein folding pathways in real time. This new application requires reparameterization of the force field, which is the subject of our current work. We describe preliminary results in this communication.

2 Methods

In the UNRES model, a polypeptide chain is represented as a sequence of α -carbon atoms (C^α) with attached united side chains (SC) and united peptide groups (p), each of which is positioned in the middle between two consecutive C^α atoms. The effective energy function is a sum of different terms corresponding to interactions between the SC, SC and p, and p sites, as well as local and correlation terms, each of which is multiplied by an appropriate weight³, w . The expressions for these terms had been derived² based on a Kubo cluster cumulant⁵ expansion of a polypeptide chain in water, where the degrees of freedom not present in the model had been integrated out. In the present work we introduced explicit dependence of the cumulant-based terms on temperature.

The hierarchical method of force field optimization developed in our laboratory³ aims at obtaining energy landscapes of selected training proteins such that the free energy of each of the training proteins decreases with increasing native likeness. In the present study, we computed the free energies below, at, and above the folding-transition temperatures and extended the approach by the requirements that the free-energy relations be inverted above the folding-transition temperature. We used the multiplexing replica-exchange molecular dynamics (MREMD)⁶ to generate decoy sets with given energy-function parameters

and the weighted-histogram analysis method (WHAM)⁷ to process the results in order to compute the free energies of the ensembles.

3 Results and Discussion

The structures of the three training proteins used in this work and the representatives of the most probable conformations obtained with optimized force fields (a separate optimization was carried out for each protein) are shown in Figure 1.

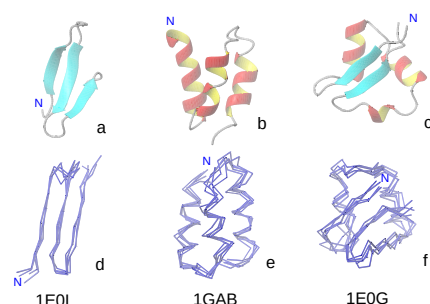


Figure 1. Experimental structures of 1E0L (a), 1GAB (b), and 1E0G (c) and some of the most probable conformations at temperatures below the folding-transition temperature of these proteins (d–f). The experimental structures are shown in the ribbon representation and the simulated structure are shown as C $^{\alpha}$ -traces. The N-termini are marked for tracing purposes.

The results of test *ab initio* prediction runs (using the multiplexing replica-exchange molecular dynamics and the force fields parameterized using 1GAB on a number of α -helical proteins are summarized in Table 1. It can be seen that the force field predicts the structures of proteins both with simple three- or four-helix bundle folds (1BDD, 1CLB, 1LQ7, 1E68, 1P68) and those with more complex topology (1POU, 1KOY, 1PRU), although its performance on the former is better. Therefore the force field parameterized using 1GAB appears transferable. We are currently working on including proteins with more complex α -helix topology as well as more complex $\alpha + \beta$ -proteins in parameterization.

Acknowledgments

Support from the National Institutes of Health (grant GM-14312), the National Science Foundation (grant MCB00-03722), the NIH Fogarty International Center (grant TW7193), the Polish Ministry of Education and Science (grant 3 T09A 032 26) as well as use of supercomputing resources at the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw is gratefully acknowledged.

PDB ID	No of residues	lowest RMSD	index(es) of native cluster(s)	probability
1BDD	46	2.2	1–4	0.84
1LQ7	67	1.6	5	0.10
1E68	70	3.9	4	0.08
1CLB	75	3.9	3	0.15
1P68	102	2.7	2	0.23
1POU	71	5.3	7	0.05
1PRU	56	5.3	— ^a	— ^a
1KOY	62	4.4	9	0.003

Table 1. Results of tests on α -helical proteins of the force field parameterized on 1GAB.

^aNo clear cluster of native-like structures was located.

References

1. A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18:849–873, 1997.
2. A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.*, 115:2323–2347, 2001.
3. S. Ołdziej, A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Optimization of the unres force field by hierarchical design of the potential-energy landscape: 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B*, 108:16934–16949, 2004.
4. A. Liwo, M. Khalili, and H. A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2362–2367, 2005.
5. R. Kubo. Generalized cumulant expansion method. *J. Phys. Soc. Japan*, 17:1100–1120, 1962.
6. U. H. E. Hansmann and Y. Okamoto. Comparative study of multicanonical and simulated annealing algorithms in the protein folding problem. *Physica A*, 212:415–437, 1994.
7. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13:1011–1021, 1992.

Simple Physics-Based Analytical Formulas for the Potentials of Mean Force for the Interaction for Amino-Acid Side Chains in Water. Tests with Simple Spherical Systems

Mariusz Makowski^{1,2}, Adam Liwo^{1,2}, Katarzyna Maksimiak³, Joanna Makowska^{1,2}, Lech Chmurzyński², and Harold A. Scheraga¹

¹ Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca, NY 14853-1301, U.S.A.,
E-mail: has5@cornell.edu

² Faculty of Chemistry, University of Gdańsk,
Sobieskiego 18, 80-952 Gdańsk, Poland

³ School of Computational Sciences, Korea Institute for Advanced Study,
207-43 Cheongnyangni 2dong, Dongdaemun-gu, Seoul 130-722, Korea

Simple analytical functions consisting of electrostatic, polarization, Lennard-Jones, and cavity terms were proposed to express the potential of mean force of particles interacting in water analytically. A simple analytical approximation based on the integral over the surface-hydration energy density of the interacting sites was derived for the cavity term; the solvation energy density of each site is represented as a difference of two Gaussian terms. This expression can easily be generalized to interaction sites with non-spherical symmetry. The analytical expressions were fitted to the potential of mean force of model systems determined by umbrella-sampling molecular dynamics. The analytical formulas fitted the potentials of mean force very well, and the resulting parameters of the expressions were physically reasonable.

1 Introduction

Effective potentials of amino-acid side chain interactions are a key part of united-residue force fields, because they encode most of sequence specificity. In the UNRES force field developed in our laboratory^{1,2} the side chain–side chain interaction potentials are presently assigned arbitrary Gay-Berne³ functional forms that take into account anisotropy of interactions, and their parameters were determined based on fitting to the correlation functions and side-chain-contact energies determined from the PDB¹. These potentials have the symmetry of an ellipsoid of revolution¹, which does not capture the amphiphilic feature of side chains such as that of aspartic acid or asparagine, which are composed of a charged/polar “head” and a non-polar “tail” and possess, therefore, a lower symmetry of interactions.

In this communication, we introduce a new model of charged and polar side chains consisting of two centers: non-polar ones and charged/polar ones. The interaction potential consists of three components: a Gay-Berne component to reproduce steric interactions, a solvent-accessible-surface area component to reproduce hydrophobic association and a Generalized Born⁴ component to reproduce electrostatic interactions between polar/charged parts. We also introduce an analytical expression for the free energy of hydrophobic solvation, which is based on a Gaussian model of the solvent density around a

solute molecule. We fitted the resulting analytical expressions to the potentials of mean force of model systems determined by umbrella-sampling molecular dynamics simulations.

2 Methods

Molecular dynamics simulations for the systems studied were performed with the AMBER 7.0 force field⁵. The TIP3P⁶ model of water was used. The potentials of mean force (PMFs) were determined by using the WHAM method⁷. The detailed computational procedure was described previously⁸. Fitting of analytical formulas to the PMFs was accomplished by minimizing the sum of squares of the differences between the PMF values computed from analytical formulas and determined from MD simulations by using the Marquardt method⁹. We used two expressions for the cavity potential, one derived in this work based on the Gaussian model of solvent energy density and one based on molecular surface area.

3 Results and Discussion

An example of a PMF of the pair of two methane molecules together with curves computed with fitted analytical expressions are shown in Figure 1.

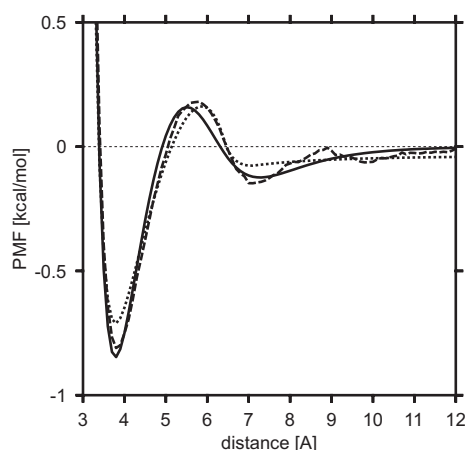


Figure 1. PMF curve of a pair of methane molecules in water determined by simulation (dashed line), the fitted curve with the cavity term expressed by the Gaussian model (solid line) and the fitted curve with the cavity term expressed as molecular surface area (dotted line).

It can be seen from Figure 1 that the Gaussian model of solvent density is able to reproduce both the desolvation maximum and the solvent-separated minimum in the PMF, while the molecular-surface area model can reproduce only the desolvation maximum. Unlike the model based on molecular-surface area, the Gaussian model of hydrophobic solvation can be easily generalized to non-spherical sites and does not produce discontinuities in the energy gradient.

Acknowledgments

This work was supported by grants from the Polish Ministry of Education and Science (1 T09A 099 30 and BW/8000-5-0015-6), the U.S. National Institutes of Health (GM-14312), and the U.S. National Science Foundation (MCB00-03722), National Institutes of Health Fogarty International Center (TW7193). This research was conducted by using the resources of (a) our 392-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, (c) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdansk, (d) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdansk, and (e) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

References

1. A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18:849–873, 1997.
2. S. Ołdziej, A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Optimization of the UNRES force field by hierarchical design of the potential-energy landscape: 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B*, 108:16934–16949, 2004.
3. J. G. Gay, B. J. Berne. Modification of the overlap potential to mimic a linear site–site potential. *J. Chem. Phys.*, 74:3316–3319, 1981.
4. D. Bashford, and D. A. Case. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.
5. D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Chaetham III, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Straton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman. AMBER 7. *University of California, San Francisco*, 2002.
6. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.
7. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J. Comput. Chem.*, 13:1011–1021, 1992.
8. K. Maksimiak, S. Rodziewicz–Motowidło, C. Czaplewski, A. Liwo, and H. A. Scheraga. Molecular simulation study of the potentials of mean force for the interactions between models of like-charged and between charged and nonpolar amino acid side chains in water. *J. Phys. Chem. B*, 107:13496–13504, 2003.
9. D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, 11:431–441, 1963.

The Role of Flexible Stem Geometries in Protein Loop Structure Prediction

M. Mönnigmann¹ and C. A. Floudas²

¹ Process Systems Engineering
RWTH Aachen University, Templergraben 55, 52056 Aachen
E-mail: moennigmann@lpt.rwth-aachen.de

² Department of Chemical Engineering,
Princeton University, Princeton, NJ 08544-5263
E-mail: floudas@titan.princeton.edu

Previous works on loop structure prediction treat the loop reconstruction problem, that is, the geometry of the protein into which the loop must fit is assumed to be given. In *ab initio* protein structure prediction, however, this information is not available, but loop structure and the anchoring stem region structure have to be predicted simultaneously. The resulting loop structure prediction problem with flexible stems treated here is considerably more difficult than the reconstruction problem.

The proposed approach is based on (i) dihedral angle sampling, (ii) structure optimization by energy minimization with a physically based energy function, and (iii) clustering. In contrast to previous works, clustering is not only used to identify conformers that are likely to be close to the native structure, but also to identify far-from-native decoys.

The method is tested on a large test set. Surprisingly, the average of the smallest rmsd found in ensembles of 2000 conformers depends linearly on the length of the sequence. Since our test set is large, different methods for selecting close-to-native conformers from ensembles can be compared in a meaningful way.

1 Introduction

Without the great structural flexibility of loops, many proteins could not fold into compact structures. Since loops possess great geometric flexibility, loop structure is difficult to predict. Recent approaches assume information on the location of the anchoring residues is available, that is, the loop reconstruction problem is treated¹. While using information on the geometry of the anchoring residues may be justified in homology modeling, approaches of this type are ruled out in *ab initio* prediction. In order to benchmark the prediction precision that can be achieved in *ab initio* loop structure prediction, we apply a methodology that predicts both, the geometry of the anchoring residues, and the loop structure itself, to a large test set.

2 Methods

Conformers are generated with probability functions in a discretized (ϕ, ψ) -space similar to those used by DePristo et al.². We distinguish between three types of residues, (i) α -helical amino acids as defined by DSSP code H, (ii) β -strand (DSSP code E), and (iii) loop³. To qualify as a loop an amino acid sequence must not be of DSSP type E or H, not be at either terminal, and must be located between strands or helices. Probability distributions

n_r [-]	10	12	14	16	18	20
rmsd [Å]	.9842	1.412	1.825	2.159	2.451	2.772

Table 1. Average of smallest rmsd in ensembles of given loop length. The symbol n_r denotes the number of residues in a loop including the stem residues. Rmsds depend linearly on the loop length¹. Results for odd loop lengths are omitted for brevity.

are determined by investigating a reference set of known structures. This set comprises all proteins with experimental resolution of 2.2Å or better in the PdbSelect25 set³.

In order to take the influence of the anchor regions into account, we add three stem residues at both ends and subject these additional six residues to the same prediction procedure as the loop residues. We optimize side chain angles using the Dunbrack rotamer library⁴. Energies are calculated with the ECEPP/3 force field⁵. After the side chain optimization the entire structure is subjected to an energy minimization with NPSOL⁶.

Clustering methods have been used before to identify groups of conformers that are *likely* to be similar to each other and, ultimately, *likely* to be similar the native structure¹. In contrast, our clustering approach identifies conformers that are likely *not* to be similar to the unknown native structure, and discards those structures. As a result the overall quality of the ensemble is improved, and any strategy for picking out the best conformer in the remaining ensemble is more likely to identify a conformer that is close to native. The clustering method used here is based on the observation that cluster size and rmsd to the native structure are correlated¹.

3 Results

We applied our method to 3215 loops extracted from the PdbSelect25 set of proteins and to a set of 65 loops from the CASP6 targets¹. Due to space restrictions, we can only report a fraction of our results¹.

Table 1 assesses the quality of the conformers generated by dihedral angle sampling. These results are obtained by determining the minimum rmsd to native for all conformers generated by dihedral angle sampling, and subsequently averaging over the minimum rmsds of all ensembles of loops of a given length. From the data in table 1 we infer that dihedral angle sampling does not restrict the prediction accuracy¹.

Figure 1 summarizes some of our prediction results¹. Results labeled *average rmsds by energy* are determined by identifying the lowest energy conformer in each ensemble, recording its rmsd to native, and subsequently averaging over the rmsds for all ensembles of loops of a given length. The average rmsds by colony energy are determined correspondingly, where the candidate conformer in each ensemble is picked by identifying the lowest colony energy⁷. The remaining data shown in Figure 1 is obtained by applying our clustering algorithm¹.

From Figure 1 it is evident that the conformer that generates the largest cluster after any clustering step $k = 0, 1, 2$ is on average a better prediction for the native structure than the lowest energy or lowest colony energy conformer. Cluster size is therefore a better criterion for the identification of good conformers than energy or colony energy. Moreover, the largest cluster conformer found after step $k + 1$ of the clustering algorithm is as good

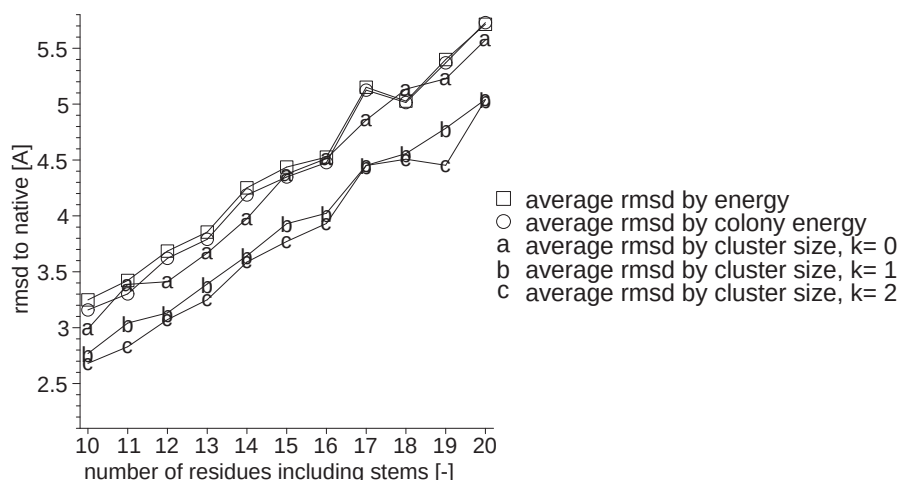


Figure 1. Average of rmsds as a function of total loop length for the loops from the PdbSelect25 proteins. Lines are added as a guide to the eye.

as, or better than, the largest cluster conformer found in step k . In this sense, repeatedly applying the clustering algorithm improves the largest cluster conformer on average.

For the large set of loops we investigated, the rmsd predicted by cluster size after iterative clustering, averaged over the ensembles of all loops of the same length, is a linear function of loop length. This indicates that the prediction quality is currently not limited by the dihedral angle sampling, but by the strategy for selecting a conformer that is likely to be close to native¹. This holds also for the results for the CASP6 targets¹.

4 Summary and Outlook

We briefly summarized results obtained with a new methodology that allows structure prediction of loops with flexible stem residues. The proposed methodology was applied to a large test set, allowing meaningful results of methods to select conformers from ensembles that are close to the native structure. We compared several selection criteria, namely ECEPP/3⁵ energy, colony energy⁷, and cluster size before and after application of a new clustering algorithm¹. The comparison shows that that energy is approximately as good as colony energy, cluster size before applying our clustering approach is on average better than both energy and colony energy, and cluster size after applying our clustering approach is the best criterion.

The loop prediction method developed here is ultimately going to be used in an existing ab initio protein prediction approach⁸. In this context, the loop prediction method must not assume information on the surrounding protein to be given, but loops and the remaining parts of the structure must be predicted simultaneously.

Acknowledgments

CAF gratefully acknowledges funding from the National Institute of Health (R01-GM-52032, R24-GM-069736). MM gratefully acknowledges funding by a Deutsche Forschungsgemeinschaft research fellowship (MO 1086/1-1).

References

1. M. Mönnigmann and C. A. Floudas. Loop structure prediction with flexible stem geometries. *Proteins: Structure, Function, and Bioinformatics*, 61:748–762, 2005.
2. M. A. DePristo, P. I. W. de Bakker, S. C. Lovell, and T. L. Blundell. Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. *Proteins: Structure, Function, and Bioinformatics*, 51:41–55, 2003.
3. W. Kabsch and Ch. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.
4. M. J. Bower, F. E. Cohen, and R. L. Dunbrack. Prediction of protein side chain rotamers from a backbone dependent rotamer library: a new homology modeling tool. *Journal of Molecular Biology*, 267:1268–1282, 1997.
5. G. Némethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *Journal of Physical Chemistry*, 96:6472–6484, 1992.
6. P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. *NPSOL 4.0 User's Guide*. Systems Optimization Laboratory, Dept. of Operations Research, Stanford University, CA., 1986.
7. Z. Xiang, C. Soto, and B. Honig. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7432–7437, 2002.
8. J. L. Klepeis and C. A. Floudas. ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical Journal*, 85:2119–2146, 2003c.

Modeling of Water Explicitly in the Replica-Exchange Simulation Method for Protein Folding

Mu Yuguang* and Yang Ye

School of Biological Sciences,
Nanyang Technological University, Singapore 637551

*Correspondent Author: E-mail: ygmu@ntu.edu.sg

Considering the replica-exchange simulation of protein in explicit water to be a two-Hamiltonian system with one Hamiltonian for conformational sampling in molecular dynamics simulation and another for controlling the exchanges between difference temperature replicas in Monte Carlo jumps, we introduce an approximation on the latter Hamiltonian using a continuum solvent model, surface-generalized Born model. Such replica exchange simulation with hybrid Hamiltonians method is applied to fold the C-terminus (residue 41-56) of protein G from extended structures. Promising results show that not only the total number of replica needed is largely reduced but also the folding efficiency is greatly enhanced. Combined with recently invented dihedral principle component analysis a general framework for *ab initio* folding a small protein merely from sequence knowledge is emerging.

PACS numbers: 87.15.Cc, 87.15.Aa

Predicting protein structure solely from its amino acid sequence has long been a great challenge in modern molecular biology^{1,2}. Although remarkable progress was made recently by Baker and his collaborators^{3,4}, the problem remains unsolved. Recent studies on peptides^{5,6} by molecular dynamics simulation tools using explicit water model, aided by advanced sampling strategies, such as replica exchange molecular dynamics (REMD) method⁷, show a prospective way towards solving the folding puzzle. However when REMD is applied to larger protein system a huge computation facility will be needed because the number of replicas needed increases simultaneously with the increase of the degrees of freedom of the system⁸. Endeavors to overcome the shortage of REMD have been undertook by several groups⁹⁻¹².

The Hamiltonian of a protein solvated in an aqueous environment can be written as $H_{sys} = H_p + H_{pw} + H_w$, where p , pw and w denote protein-protein, protein-water and water-water interactions respectively. In standard REMD scheme, replicas are propagating independently at certain temperatures with Monte Carlo (MC) exchange at certain intervals. The MC exchange corresponds to the Metropolis criterion following the detailed balance condition: $p_{i \leftrightarrow j} = \min(1, e^{\Delta\beta\Delta E})$, where $\Delta\beta = \beta_j - \beta_i$, $\beta = \frac{1}{k_B T}$, $\Delta E = E(\chi_j) - E(\chi_i)$ and χ_i, χ_j are the configurations of the neighboring replicas. The total energy, $E(\chi)$, should be the same as the system Hamiltonian and can also be decomposed into three terms: $E(\chi) = E_p + E_{pw} + E_w$ as well. In explicit water REMD simulation many water molecules are employed to solvate well the unfolded protein. This makes the total energy, E , and its differences between neighbours, ΔE , huge. In order to get a reasonable high exchange probability between the neighboring temperature ladders, $p_{i \leftrightarrow j}$, the temperature jumps, $\Delta\beta$, should be small. Therefore a large number of replicas are needed to cover a broad temperature range, which from 300K to 600K. Realizing that

the water-water interaction makes the largest contribution to the total energy a possible workaround could be that decouple the water-water interaction from the thermal baths and make the ΔE small though the total energy E nearly unchanged. Recent works done by Berne group⁹ and Simmerling group¹¹ were pursued in this direction using different strategies. However such decoupling not only causes unphysical Hamiltonian on most replicas in MD simulation, the resulting hot protein and cold water effect would also hinder the conformation sampling efficiency at high temperatures.

This letter presents an improved solution to this problem. The standard REMD method can be thought as a hybrid method of MD and MC simulation. The Hamiltonians for both MD and MC should be the same for the sake of consistency. However if one checks the roles of both Hamiltonians in details, such consistency is found to be not a necessity. The full system Hamiltonian, H_{sys} , is necessary for MD simulations to correctly sample conformational space. While for MC steps we find it is not necessary to use the same Hamiltonian, H_{sys} . In H_{sys} the water-water interaction is dominant but is not directly related to the protein folding. The key of the MC steps in REMD is to help protein jump out of its local minima. Evidently the REMD could be more efficient in folding protein if the Hamiltonian for MC steps represents the energies of the protein more directly. In light of that we introduce an approximation to the Hamiltonian for MC steps. Instead of using the H_{sys} as in the MD steps a second Hamiltonian with the same energies for protein-protein interaction but protein-water and water-water interactions approximated by a continuum solvent model, surface-generalized Born model (GBSA)¹³ is used in MC steps. In this way the dominate water-water interaction is averaged out and therefore a small ΔE is obtained. At current moment the implicit water model or continuum solvent model, although very computationally efficient, was found not to be able to fold peptide correctly¹⁴. Our method is a complement for such deficiency in the GBSA model.

We call this new version of REMD as REMD with hybrid Hamiltonians (REMDhH). Hybrid Hamiltonians costs the violation of detailed balance of the whole system to some degree. Considering the interested temperature range for REMD is from 300K to 600K the water behaves well as a liquid (in constant volume REMD) and should be quickly relaxed to its equilibrium state when the system jumps to a new temperature.

Here REMDhH is applied to fold the C terminus (residue 41-56) of protein G, a 16 amino acids peptide which was found to be able to fold into a beta-hairpin *in vitro*¹⁵. The GROMACS program suite¹⁶ and the full atomic OPLS-AA force field¹⁷ are used. The peptide is capped with the normal ACE and NME groups with 256 atoms in total. It is solvated by 5469 water molecules plus three K⁺ ions to neutralize the charged molecular system. The whole system consists of 16666 atoms in a cubic simulation box of 5.3 nm length. Sixteen replicas are used whose temperatures are 300.0, 317.2, 335.4, 354.6, 374.9, 396.4, 419.1, 443.1, 468.5, 495.4, 523.8, 553.8, 585.5, 619.1, 654.5 and 692.1 Kelvin. While 64 replicas had been used for the same peptide with less water molecules (1361) studied by standard REMD⁵. The GBSA energies are calculated using Tinker program¹⁸. The interface between GROMACS and Tinker is built by modifying GROMACS source code. A twin-range cutoff of 0.9/1.4 nm is used for the non-bonded interactions and a reaction-field correction with permittivity is employed. The integration

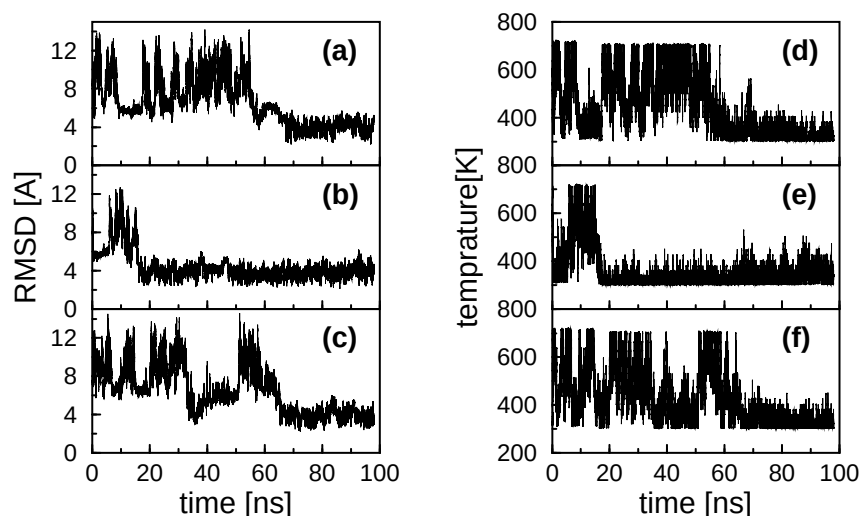


Figure 1. (a), (b) and (c) are the three folded replica trajectories of RMSD from NMR structure, (d), (e) and (f) are the trajectories of temperature jumps of the related replicas, respectively.

step in all simulations is 0.002 ps. Non-bonded pair lists are updated every 10 integration steps. The system is coupled to an external heat bath with a relaxation time of 0.1 ps. All bonds including hydrogen atoms are constrained in length. Each replica is run for 98 ns with replica exchange attempted every 2 ps. The final acceptance ration for replica exchanges is found to be 35%-43%.

To test the validity of the method the initial configurations of peptide are extended structures, which means only the information of amino acids sequence is supplied. Three folding events are found during the simulation. Figure 1a, 1b and 1c show the trajectories of root mean squared deviations (RMSD) from the NMR structure¹⁹ calculated with all atoms. The fastest folding happens in replica 9 (Figure 1b) around simulation time of 18 ns. The following folding events happen in replica 4 and 15 (Figure 1a and 1c) around simulation time of 60 ns. All the folded structures maintain stable until the end of simulations. Here we denote a structure as folded when its RMSD drops below 4Å. A more detailed analysis is given below. The folding processes revealed by REMD are a process of annealing and relaxation. The temperature trajectories (Figure 1d, 1e and 1f) of the three folded replicas indicate this view clearly. Accompanying with the decrease of RMSDs the temperatures are cooling down.

In order to find the quantitative reasons why the employed REMDhH method is superior to the standard REMD, the correlation between RMSDs and the total energy of the system with explicit water molecules, and the correlation between RMSDs and the GBSA energies are calculated and displayed in Figure 2a and 2b respectively. It is clearly shown that the correlation between the RMSDs and the total energies used by regular REMD is poor. Its correlation coefficient is 0.01. While the correlation between

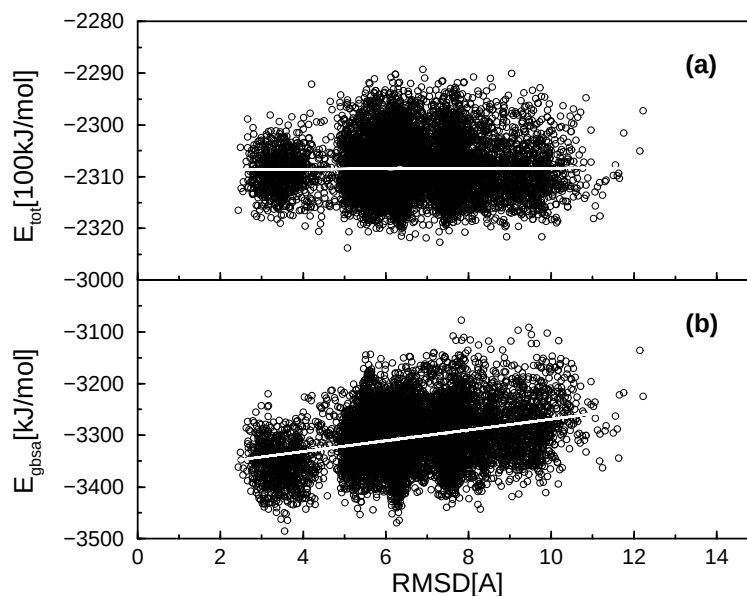


Figure 2. (a) Correlations of the total energies, including protein-protein, protein-water, and water-water interactions, with RMSD from NMR structure. The correlation coefficient is 0.01. (b) Correlations of the GBSA energies with RMSD. The correlation coefficient is 0.3. The white lines are the linear regression results.

the RMSDs and the energies from the GBSA model is much stronger whose correlation coefficient is 0.3. As mentioned earlier the folding process in REMD method could be regarded as an annealing process. To perform such annealing efficiently two factors for the MC steps must be considered. First the lower energy conformations could be identified in the high temperature conformational ensemble. Secondly these lower energy conformations have to be quickly annealed to the lower temperature simulations. The first factor requires that the energies for MC steps should interrogate the protein conformation directly. Otherwise the lower energy conformations would not easily be identified. The second factor claims that the fewer temperature ladders the quicker the annealing process which implies that the number of replicas in REMD should be minimized. The implementation of the current REMDhH could make both requirements satisfactory. If the REMD efficiency is approximated to be linearly correlated with the correlation coefficient of RMSD/energy for MC steps the speedup of folding by REMDhH compared with that of standard REMD is a factor of 30 for this system.

In protein structure prediction usually the folded structure is unknown. The RMSDs from the native structure and the fraction of native contacts which are commonly used as folding reaction coordinates are not available in such situations. In order to identify the global minimum new suitable reaction coordinates are desirable. We find that the recently invented dihedral principal component analysis (dPCA) method works²⁰ fine here. Figure 3 reveals the folding free energy curve for this beta-hairpin peptide obtained by projecting conformational ensemble of $T = 300K$ onto the first eigenvector of dPCA

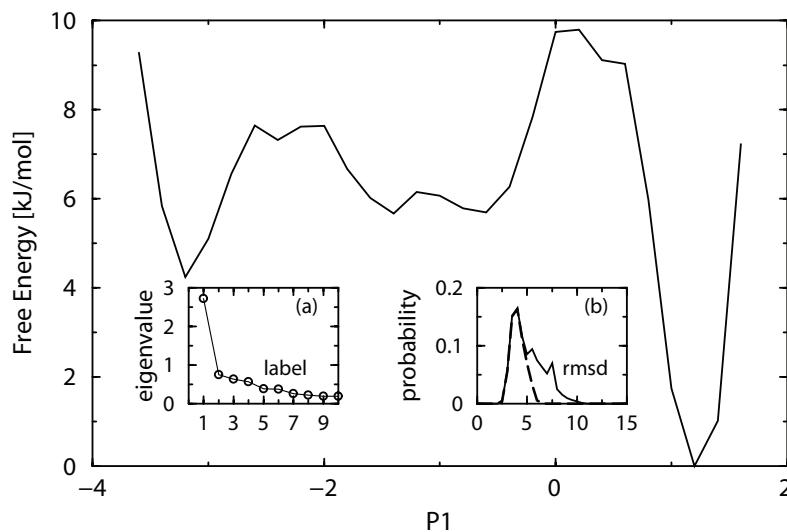


Figure 3. Folding free energy curve of the GB1 hairpin peptide obtained by projection onto the eigenvector of largest eigenvalue using dihedral PCA method. Inset (a) shows the first 10 largest eigenvalues of dPCA result. Inset (b) shows the distribution of RMSD of ensemble at $T = 300K$ (solid line) and the distribution of RMSD of the ensemble within the global minimum using such criteria $P1 > 0.7$ (dashed line).

whose eigenvalue is shown to be much larger than other eigenvalues in inset (a). The free energy curve is funnel-like and it is easy to identify the global minimum which is located around $P1 = 1.2$. Here $P1$ is the projection value on the first eigenvector. To check whether this global minimum is related to the native state the distribution of RMSDs from this ensemble is plotted in Figure 3 inset (b) by the dashed line. For comparison the distribution of RMSDs for all the structures obtained at $300K$ is shown by the solid line. Evidently the structures from the global minimum are native-like with average $RMSD = 3.5\text{\AA}$. And $P1$ is a better reaction coordinate than RMSD in the sense that the former can distinguish folded states from unfolded ones more clearly.

In summary a new version of REMD method is suggested. By decomposing REMD into two steps, one is MD step and the other is MC step, different Hamiltonians can be applied to them. In MD step the full Hamiltonian with all protein-protein, water-water, water-protein interactions included guarantees that the conformations of protein are sampled correctly. In MC step, however, only the protein-protein interaction is explicitly considered, the water-water and water-protein interactions are approximated by a continuum GBSA model. Such a hybrid Hamiltonian scheme reduces the number of replica greatly and meanwhile increases the folding efficiency by more than a factor of 10. On the other hand, dPCA analysis makes the global minimum easily identified if protein is folded in simulation. Combining REMDhH with dPCA a general framework for ab initio protein structure prediction is emerging. Together with the available of accurate force fields and powerful computational facility the era of solving protein folding by brute force MD simulations will be coming soon²¹.

Acknowledgments

The support of a Lee Kuan Yew Research Fellowship to Y.G.M. is acknowledged. The support of PhD study in Nanyang Technological University to Yang Ye is appreciated. The simulations were performed on the supercomputer of Bioinformation Research Center and the Frankfurt Center of Scientific Computing, which are acknowledged by the generous allocation of CPU time.

References

1. D. J. Wales and H. A. Scheraga, *Science* **285**, 1368 (1999).
2. J. N. Onuchic and P. G. Wolynes, *Curr. Opin. Struct. Biol.* **14**, 70 (2004).
3. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, *J. Mol. Biol.* **268**, 209 (1997).
4. C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, *Methods Enzymol* **383**, 66 (2004).
5. R. Zhou, B. J. Berne, and R. Germain, *Proc. Natl. Acad. Sci. USA* **98**, 14931 (2001).
6. A. E. Garcia and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **100**, 13898 (2003).
7. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.* **314**, 141 (1999).
8. H. Fukunishi, O. Watanabe, and S. Takada, *J. Chem. Phys.* **116**, 9058 (2002).
9. P. Liu, B. Kim, R. A. Friesner, and B. J. Berne, *Proc. Natl. Acad. Sci. USA* **102**, 13749 (2005).
10. T. Z. Lwin and R. Luo, *J. Chem. Phys.* **123**, 194904 (2005).
11. X. Cheng, G. Cui, V. Hornak, and C. Simmerling, *J. Phys. Chem. B* **109**, 8220 (2005).
12. E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, *Phys. Rev. Lett.* **96**, 028105 (2006).
13. Q. Liu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, *J. Phys. Chem. A* **101**, 3005 (1997).
14. R. Zhou and B. J. Berne, *Proc. Natl. Acad. Sci. USA* **99**, 12777 (2002).
15. F. J. Blanco, G. Rivas, and L. Serrano, *Nat. Struct. Biol.* **1**, 584 (1994).
16. H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, *Comp. Phys. Comm.* **91**, 43 (1995).
17. W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225 (1996).
18. R. V. Pappu, R. K. Hart, and J. W. Ponder, *J. Phys. Chem. B* **102**, 9725 (1998).
19. A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, *Science* **253**, 657 (1991).
20. Y. G. Mu, P. H. Nguyen, and G. Stock, *Proteins* **58**, 45 (2005).
21. M. Karplus and J. A. McCammon, *Nat. Struct. Biol.* **9**, 646 (2002).

Molecular Transport Through Channels and Pores

Walter Nadler¹ and Wolfgang R. Bauer²

¹ Department of Physics, Michigan Technological University,
Houghton, MI 49931, USA
E-mail: wnadler@mtu.edu

² Medizinische Universitätsklinik 1,
Josef Schneider Straße 2, D-97080 Würzburg, Germany
E-mail: w.bauer@medizin.uni-wuerzburg.de

Facilitated translocation of molecules through channels and pores is of fundamental importance for transmembrane transport in biological systems. Several such systems have specific binding sites inside the channel, but a clear understanding of how the interaction between channel and molecules affects the flow is still missing. We present a generic analytical treatment of the problem that relates molecular flow to the first passage time across and the number of particles inside the channel. Both quantities depend in different ways on the channel properties. For the idealized case of non-interacting molecules we find an increased flow whenever there is a binding site in the channel, despite an increased first passage time. In the more realistic case that molecules may block the channel, we find an increase of flow only up to a certain threshold value of the binding strength and a dependence on the sign of the concentration gradient, *i.e.* asymmetric transport. In all cases the reason for transport facilitation is an increased occupation probability of a particle inside the channel that overcomes any increase in the first passage time due to binding.

1 Introduction

Diffusion of molecules through channels and pores of an otherwise impermeable membrane is an important issue in biological transport at the cellular level^{1,2}. In recent years it has been noted that there are several cases where the molecules transported interact strongly with regions inside the channel³⁻⁷, apparently leading to an increase in transmembrane transport.

From an intuitive point of view, it is not clear at all why a strong interaction with the channel should facilitate transport. Indeed, one would expect that a strong binding is associated with a longer residence time inside the channel which reduces flow. Furthermore, molecules bound temporarily inside the channel may hamper transport of other molecules, especially when they are large⁵, and block the channel. So, why do traps and/or reaction sites within the channel facilitate molecular flow? These questions have to be addressed in the generic biological setting of a macroscopic concentration gradient across the membrane, see Fig. 1. An appropriate quantitative description should give the flow for a given concentration difference depending on the potential and other parameters describing the molecule-channel interaction. Physical insight can be gained if the flow can be related to other global properties of the system in question.

2 Theory

The dynamics of the density of the molecules inside the channel, $\rho(x, t)$, is determined by the Smoluchowski equation⁸,

$$\partial_t \rho(x, t) = D \partial_x [\partial_x - F(x)] \rho(x, t) , \quad (1)$$

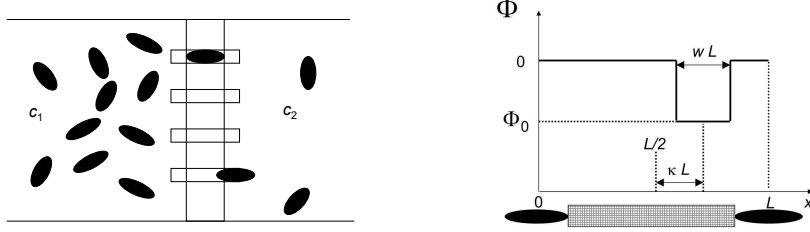


Figure 1. *(left)* Basic biological situation: A membrane separates two baths with molecular concentrations c_1 and c_2 . The baths are connected by channels (hatched rectangles) allowing only access of a single molecule. *(right)* Rectangular (box-like) shaped attractive potential of molecule-channel interaction. w is the relative width of the potential well, $|\Phi_0|$ is its depth. The relative shift of the well from its symmetric position is denoted by κ .

where x is the channel coordinate and D is the diffusion coefficient. $F(x)$ is the force describing the molecule-channel interaction that can always be derived from a potential function in one dimension, $F(x) = -\Phi'(x)$. At the ends of the channel we assume that baths hold the molecular densities constant at $\rho(0, t) \equiv c_1$ and $\rho(L, t) \equiv c_2$, respectively, as seen in Fig. 1.

Extending an old approach by Hardt⁹ we have shown recently¹⁰ that in this as well as in more general situations the flow J of non-interacting particles across some region is given by a macroscopic version of Fick's law

$$J = \frac{n}{\tau} (c_1 - c_2) \quad (2)$$

where τ is the mean first passage time (MFPT) to cross the region⁸ and n is a measure of the stationary state particle number in that region.

3 Results

The interplay between specific particle number and first passage time is illustrated by the rectangular potential well of depth Φ_0 sketched in Fig. 1 *(right)*.

3.1 Non-Interacting Particles

Using the channel average $\langle \cdot \rangle = L^{-1} \int_0^L dx$ we obtain the results¹¹

$$\tau = \frac{L^2}{2D} \langle e^\Phi \rangle \langle e^{-\Phi} \rangle \quad , \quad n = \frac{L}{2} \langle e^{-\Phi} \rangle \quad (3)$$

for MFPT and specific particle number, respectively. These results have important consequences for any form of the channel potential. Using the Cauchy-Schwarz inequality in the form $\langle f \rangle \langle g \rangle \geq \langle \sqrt{fg} \rangle^2$ on Eq. (3), we immediately see that *any* form of in-channel interaction that is non-constant leads to an increase in the MFPT, $\tau \geq L^2/2D$, and by this hampers particle flow. The physical reason is that potential barriers as well as potential wells have walls, and the particles have to get over these walls irrespective of whether they belong to wells or barriers. That is reflected also in the invariance of τ upon changing barriers to wells and vice versa by setting $\Phi(x) \rightarrow -\Phi(x)$. On the other hand, the specific

particle number increases only for predominantly attractive potentials, *i.e.* if the potential wells overcome the effects of barriers in Eq. (3). Both effects conspire to give the total flow as

$$J = \frac{D}{L} \langle e^{\Phi} \rangle^{-1} (c_1 - c_2) \quad (4)$$

So, for fully attractive potentials, *i.e.* $\Phi(x) \leq 0$, we always find an increase of the flow when compared to $\Phi(x) = 0$.

3.2 Blocking the Channel

The approach presented above can be readily extended to describe the effect of molecules that block the channel. As an extreme case we assume in the following that only a single molecule can occupy the channel, which is realistic in many cases. Before, the quantity $\rho(x, t)$ was the density of molecules in the channel at position x . We now interpret $\rho(x, t)$ as the *probability density* that a channel contains a particle at x . It is obvious that this density follows the same dynamics as described above in Eq. (1). However, the state variable x does not completely describe all states, but the empty channel has to be added as an additional state. This additional empty channel state leads to a *cyclic state model* described in more detail in Ref.¹¹.

Figure 2 (*left*) shows the relative increase of the flow, J/J_0 , for a particular set of parameters of a symmetric channel potential. Increase and subsequent decrease of the flow with increasing binding strength (negative Φ_0) can be seen clearly.

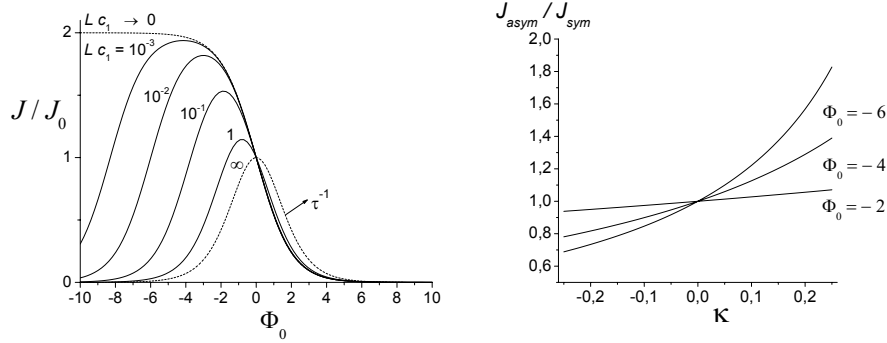


Figure 2. (*left*) Effect of the molecular channel interaction on flow. J_0 refers to vanishing interaction. The potential is box-like shaped, see Fig. 1 (*right*), and assumed to be centered ($\kappa = 0$) with relative width $w = 1/2$. $\Phi_0 > 0$ denotes a barrier and $\Phi_0 < 0$ is an attractive potential, *i.e.* a potential well. Unidirectional flow is considered, *i.e.* c_2 is set to zero, and the concentration c_1 is varied. The limiting cases of vanishing and very high concentration are also considered (dotted lines). Note that for the latter, flow is proportional the inverse first passage time τ . (*right*) Dependence of flow J_{asym} on the position of the binding site, κ , see Fig. 1 (*right*). J_{sym} is the flow for the symmetric potential ($\kappa = 0$). The parameters used are $Lc_1 = 0.1$, $c_2 = 0$, and $w = 1/2$, *i.e.* κ can vary from $-1/4$ to $1/4$. The flow increases when the binding site moves from the *cis* to the *trans* position with respect to the larger concentration.

The typical behavior of asymmetric transport through the channel, *i.e.* $\kappa \neq 0$, is illustrated in Fig. 2 (*right*) for a particular set of parameters. We see that flow is decreased if the binding site is close to the larger concentration, while the flow is increased otherwise.

Intuitively, flow increase due to the binding site being in the *trans* position, *i.e.* away from the higher concentration, can be viewed as the binding site "pulling" the molecules across the channel. Since in the blocking situation only one molecule is in the channel, exiting to the lower concentration side is faster than diffusing all the way back.

Note that this behavior is a major difference from the idealized case of non-interacting particles, where asymmetry did not matter for the transport. Interestingly, such a directional behavior has been observed most recently in the channel protein OmpF⁷.

4 Concluding Remarks

We presented an analytical approach to describe molecular transport through a membrane channel in the biological setting of a macroscopic concentration gradient across the membrane as depicted in Fig. 1. The goal was, in particular, to understand whether and how binding sites in a channel can facilitate transport, to understand the effect of channel blocking, and to explore asymmetric transport.

Transport facilitation could be explained by the fact that increased occupation probability inside a channel outweighs any slowing down of channel crossing. Previous approaches to describe asymmetric transport used the *flashing ratchet* paradigm¹². That approach, however, depends on non-equilibrium fluctuations¹³, while the only non-equilibrium aspect of the biological situation in Fig. 1 is the concentration gradient. What came as a surprise to us is that already at this level we were able to also explain asymmetric transport as a side effect of channel blocking. This is of particular importance in the light of recent experimental findings⁷.

Acknowledgments

This research was supported by NSF-grant No. CHE-0313618 and the DFG (SFB 688).

References

1. Cooper, K.E. *et al.* (1985) *Prog. Biophys. Molec. Biol.* **46**, 51.
2. Meller, A. (2003) *J. Phys. : Condens. Matter* **15**, R581.
3. Bezrukov, S.M. *et al.* (2000) *FEBS Letters* **476**, 224.
4. Hilty, C., & Winterhalter, M. (2001) *Phys. Rev. Lett.* **86**, 5624.
5. Nestorovich, E.M. *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **15**, 9789.
6. Schwarz, G. *et al.* (2003) *Biophys. J.* **84**, 2990.
7. Alcaraz, A. *et al.* (2004) *Biophys. J.* **87**, 943.
8. Gardiner, C.W. (1985), *Handbook of Stochastic Methods* (Springer, Berlin).
9. Hardt, S. (1981) *Bull. Math. Biol.* **43**, 89.
10. Bauer, W.R., & Nadler W. (2005), *J. Chem. Phys.* **122**, 244904-1.
11. Bauer, W.R., & Nadler W. (2006), *Proc. Natl. Acad. Sci. USA*, in press.
12. Frey, E., & Kroy, K. (2005) *Ann. Phys. (Leipzig)* **14**, 20.
13. Kostin, I. & Schulten, K. (2004) *Phys. Rev. Lett.* **93**, 238102.

Integrating Structural and Kinetic Enzymatic Information in Systems Biology

Matthias Stein, Razif R. Gabdoulline, and Rebecca C. Wade

EML Research gGmbH, Molecular and Cellular Modeling Group,
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany
E-mail: matthias.stein@eml-r.villa-bosch.de

The successful modelling of metabolic and signalling pathways in systems biology requires a detailed and consistent set of enzymatic kinetic parameters. Sometimes, these parameters are not available from the literature or were obtained under different experimental conditions. In many cases, the enzymes' activities are related to the molecular interaction field of the enzyme. PIPSA (Protein Interaction Property Similarity Analysis) is a method that enables the comparison of molecular interaction fields (for example the electrostatic potential) to classify proteins. Results obtained using the quantitative PIPSA (qPIPSA) method to derive estimates of missing enzymatic data are presented for a number of proteins involved in glycolysis.

1 Introduction

In systems biology, one aims at an understanding of biochemical processes in the context of a cell or organ. One approach to this goal is the mathematical modelling of protein networks by a set of coupled differential equations describing the variations of metabolite concentrations over time. Enzymatic kinetic parameters such as substrate binding affinities, catalytic turnover and initial metabolite concentrations are required when setting up a kinetic model for a signalling or metabolomic network. Despite recent achievements in large scale proteomics, many of the enzyme kinetic data have not been determined or are available only for a different organism or under different environmental conditions, see for example reference¹.

The derivation of quantitative structure-function relationships for the enzyme is one of the challenges of modern enzymology. Structure-based systems biology provides insight into complex enzymatic reactions at a molecular level. We here present quantitative Protein Interaction Property Analysis (qPIPSA) which relates electrostatic potentials of enzymes to their kinetic constants.

2 The Quantitative PIPSA Method

The PIPSA method has been described in detail in references^{2,3}. The molecular interaction fields of two proteins are compared in a skin region around the protein surface (see Figure 1).

PIPSA relies on protein structures and therefore a semi-automated procedure was set up to derive protein structural models by homology modelling when the structures have not been determined experimentally. For PIPSA, an experimentally well-characterized training set of enzymes is required. The retrieval of amino acid sequences, experimental enzymatic kinetic data from related organisms and the choice of an appropriate template structure is followed by multiple sequence alignment and comparative protein structure modelling.

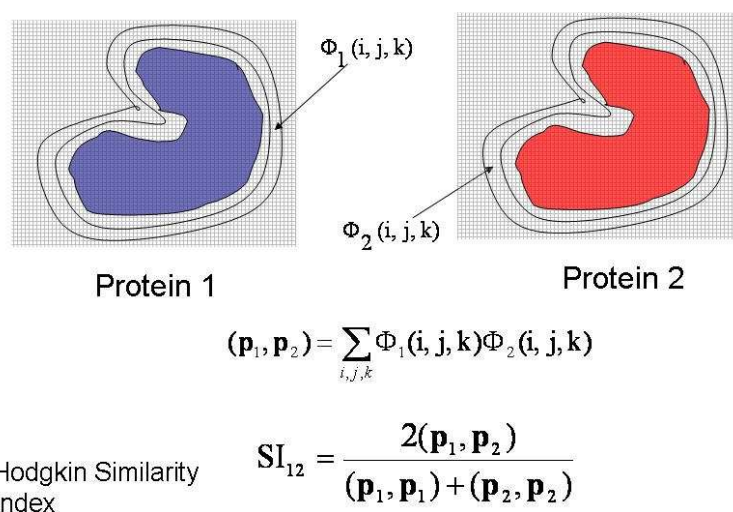


Figure 1. Schematic description of the PIPSA method. The molecular interaction fields of two proteins on a three-dimensional cubic grid are compared in a skin region. The similarity of the molecular interaction fields of the two proteins is calculated using the Hodgkin similarity index.

PIPSA can be applied to individual families of enzymes to generate kinetic parameters that are required for systems biology simulations or to a larger number of enzymes to derive a functional classification of enzymes.

3 Application to Glycolytic Enzymes

3.1 Glucokinases

Glucokinases catalyze the first step in glycolysis, namely the phosphorylation of glucose at the 6 position to yield glucose-6-phosphate. The substrate binding affinity (as given by K_m values) between enzymes from mammals, like *Homo sapiens* and *Rattus norvegicus*, and non-mammalian enzymes, from e.g. *Saccharomyces cerevisiae*, differ by several orders of magnitude although this is not mirrored in the variation of amino acid sequences. When the electrostatic potentials of a set of 8 experimentally characterized glucokinases from various organisms are compared, large variations can be detected. The analysis of the open (apo) form of the enzyme gives the best correlation with experimental K_m values, indicating that it is the open form of the enzyme not the closed form that is determining the substrate binding kinetics.

3.2 Triose Phosphate Isomerases

The electrostatic potentials of a series of twelve experimentally characterized organisms were analyzed⁴. This set was used to derive a correlation between the logarithm of k_{cat}/K_m values and the differences in electrostatic potentials. This correlation can be used to predict

the kinetic parameters of other enzymes. The correlation between predicted and experimental k_{cat}/K_m values indicated experimental outliers from rabbit and *Giardia lamblia*. In addition, patches of amino acid residues on the protein surface could be identified that are functionally important for substrate binding and catalysis, respectively.

3.3 The Full Glycolytic Pathway

The classical glycolysis consists of a series of ten chemical reactions to physiologically make use of the chemical energy stored in glucose molecules. The PIPSA method has been applied to the series of ten enzymes of the Emden-Meyerhof pathway and compared across a set of eleven organisms for which sequence information was available and for which sufficient sequence and structural homology was present in all ten enzymes of the pathway. This comparative metabolic pathway analysis shows that the interaction fields around the active sites are rather conserved across these organisms despite some amino acid sequence variation. The largest deviation from human enzymes is observed for glycolytic enzymes from the plants rice and *Arabidopsis*.

4 Concluding Remarks

Protein structure-based molecular systems biology provides insight into physiological processes at a very detailed level. This insight is complementary to the more abstract mathematical modelling and allows the functional classification of individual enzymes or the comparison of full pathways across various organisms. The generation of enzymatic parameters from protein structural modelling aids the construction and interpretation of kinetic models of networks of proteins for which not all of the critical parameters have been determined experimentally.

Acknowledgments

This work was supported by the Klaus Tschira Foundation and the BMBF within the German systems biology network ‘Hepatosys’ (Grant 0313076).

References

1. R. R. Gabdoulline, U. Kummer, L. F. Olsen, R. C. Wade, *Concerted Simulations Reveal How Peroxidase Compound II Formation Results in Cellular Oscillations*, Biophys. J. **85**, 1421–1428 (2003).
2. N. Blomberg, R. R. Gabdoulline, M. Nilges, R. C. Wade, *Classification of Protein Sequences by Homology Modeling and Quantitative Analysis of Electrostatic Similarity*, Protein **37**, 379–387 (1999).
3. R. C. Wade, R. R. Gabdoulline, F. De Rienzo, *Protein Interaction Property Similarity Analysis*, Int. J. Quant. Chem. **83**, 122–127 (2001).
4. R. R. Gabdoulline, M. Stein, R. C. Wade, *Relating Kinetic Parameters of Enzymes and Their Interaction Fields*, manuscript in preparation, (2006).

Secondary Structure of Compstatin Analogues: Insights from Molecular Dynamics Simulations in Explicit Water

Phanourios Tamamis¹ and Georgios Archontis¹

Department of Physics, University of Cyprus, PO20537, CY1678, Cyprus

E-mail: archonti@ucy.ac.cy

Compstatin is a cyclic, 13-residue peptide, which binds to the protein C3 and inhibits the activation of the complement system. To obtain a better understanding of the structural properties of the free peptide and the relation between sequence, structure and activity, we conduct multi-ns explicit-water simulations of three related systems. i) A native analogue of compstatin that is acetylated at the N-terminal end (NAT), ii) the inactive single mutant Q5G (G5), and iii) the more active double mutant V4W/H9A (W4A9). In the simulations the 5–8 region adopts a β -turn conformation with high probability, in agreement with earlier NMR studies. The propensity of this turn is not correlated with the activity of the three analogues. The rest of the molecules are mostly in a random-coil state, with β -turns occasionally formed outside the region 5–8.

1 Introduction

The complement system constitutes the initial defense against foreign pathogens¹. The development of drugs controlling complement activation is of significant medical interest, as it could treat a number of pathological conditions, including rheumatoid arthritis and rejection of xenotransplantation². A candidate against the unregulated activation of the complement system is the cyclic peptide compstatin, which has the sequence Ile1-Cys2-Val3-Val4-Gln5-Asp6-Trp7-Gly8-His9-His10-Arg11-Cys12-Thr13-NH₂ and contains the disulfide bridge Cys2–Cys12 (see fig. 1, left panel). Earlier NMR studies of the free peptide showed that the segment 5–8 adopts a predominantly type-I β -turn conformation in solution, whereas the rest of the molecule is highly flexible³. Additional systematic mutational studies determined that the two cysteines 2 and 12, Val3 and the residues 5–8 of the β -turn are critical for activity, but the turn itself is not sufficient for activity. In contrast, residues Val4, His9, His10 and Arg11 are not required for activity^{4,5}.

Apart from the formation of the 5–8 β -turn, the information from the NMR spectra on intramolecular interactions is limited, due to extensive conformational averaging. To gain insight on these interactions, we conduct multi-ns molecular dynamics simulations of a native analogue (NAT) (acetylated at the N-terminal end and more active by a factor of three), the single inactive mutant Q5G (G5)³, and the double mutant V4W/H9A (W4A9), which has a 45-fold higher activity than the native compound^{4,5}.

2 Methods

The simulation methodology has been described in detail in ref.⁶. Briefly, the total system consisted of one peptide and 7921 water molecules (23975 atoms in the case of NAT). The interatomic interactions were taken from the CHARMM22 all-atom force field⁷ and the water was represented by a modified TIP3P model⁸. All simulations employed cubic

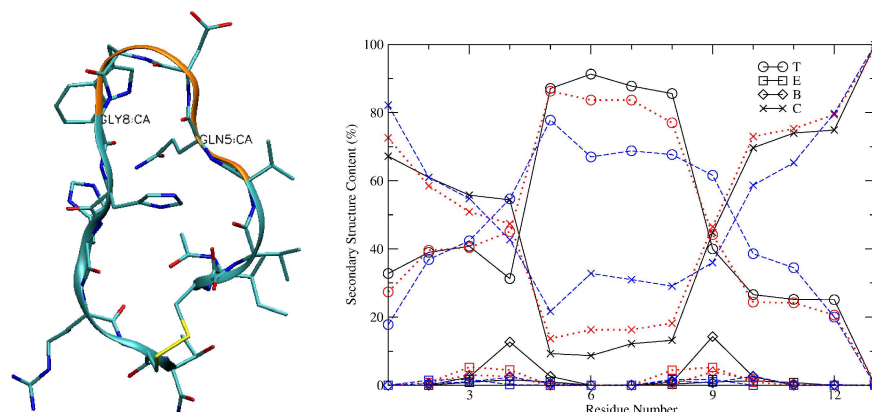


Figure 1. **Left:** Typical simulation structure of NAT. **Right:** Average residue secondary-structure population. The continuous, dotted and long-dashed lines correspond, respectively, to G5, NAT and W4A9.

boundary conditions and were performed with the program CHARMM⁹, version c31b2. For NAT, W4A9 and G5 we conducted, respectively, 22, 16 and 11 independent 2-ns simulations, starting from several different NMR structures (entry 1A1P³).

3 Results

To determine the average secondary structure of the three systems, we analyzed snapshots extracted at 5-ps intervals with the algorithm STRIDE¹⁰. The left panel of fig. 1 shows a representative simulation structure of the NAT analogue. The most important observed secondary structural elements correspond to random coil (C), β -turn (T), β -bridge (B) and β -sheet (E). The right panel of fig. 1 shows the corresponding average populations.

In all three analogues the region 5–8 folds into a β -turn with probability 85% (G5) to 70% (W4A9). The highest turn population corresponds to the inactive peptide (G5) and the lowest to the more active double mutant W4A9. Thus, the 5–8 turn propensity is not proportional to the peptide activity and its formation is not sufficient for activity, as shown by G5. A similar observation has been made with the inactive mutant V3A, which also has a high 5–8 turn propensity⁴. The probability of type I/III turns (family α_R - α_R ¹²) is 36.6%–33.5% for NAT, in good agreement with the NMR estimate. In G5 and W4A9 the probability is, respectively, higher and smaller than NAT⁶.

There is a small to moderate probability for the occurrence of turns elsewhere in the sequence. The most important of these is formed by the segment 6–9 in analogue W4A9, and is usually fused with the turn 5–8. Additional β -turns do occur in segments 1–4, 2–5, 8–11 and 9–12, but with significantly lower probability⁶. The N-terminal (1–4) and C-terminal (9–13) regions are largely unstructured. This is in agreement with the observed

conformational averaging in the NMR spectra, which has demonstrated that the region outside 5–8 is highly flexible.

Earlier MD simulations with a generalized Born approximation for the solvent¹¹ and a polar hydrogen energy function⁹ suggested the frequent formation of β -hairpin and α -helical motifs¹². In the current simulations, the 3–9 region of NAT forms a β -hairpin with a $\approx 5\%$ probability (fig. 1). In the other two analogues, the corresponding probability is 1.7% (W4A9) and 1.5% (G5). The inactive (G5) analogue forms the two main chain hydrogen bonds Val3(O)-His10(N) and Gln5(N)-Gly8(O), creating an isolated bridge between residues 4 and 9 (fig. 1). We conducted test simulations of the NAT system with the same GB energy function^{9,11,12}, in which the peptide formed β -hairpin and α -helical elements lasting for tens of ns (not shown). Such discrepancies between explicit and implicit solvent treatments have been observed with various GB implementations^{13,14}, possibly due to overestimation of the intramolecular interactions by GB. At the same time, the current explicit solvent simulations could underestimate somewhat the formation of β -hairpin and α -helical elements (despite their success in reproducing the NMR 5-8 β -turn population), due to the finite length of the simulations and force-field inaccuracies.

In conclusion, our simulations suggest that the free peptides are flexible, with a propensity for a 5–8 β -turn which is not proportional to activity and is highest for the inactive peptide (G5). The remainder of the molecule is predominantly in a random coil state, with additional β -turns occasionally observed outside the region 5–8.

Acknowledgments This work was partly funded by the ERYAN/0603 grant. We acknowledge useful discussions with Profs. Spyros Skourtis, Dimitrios Morikis, and John Lambris.

References

1. H. J. Muller-Eberhard. *Complement-Chemistry and pathways*. Raven Press, New York, 1992.
2. A. Sahu and J. D. Lambris. *Immunopharmacology*, 49:133–148, 2000.
3. D. Morikis, N. Assa-Munt, A. Sahu, and J. D. Lambris. *Prot. Sci.*, 7:619–627, 1998.
4. D. Morikis, M. Roy, A. Sahu, A. Troganis, P. A. Jennings, G. C. Tsokos, and J. D. Lambris. *J. Biol. Chem.*, 277:14942–14953, 2002.
5. B. Mallik, M. Katragadda, L. A. Spruce, C. Carafides, C. G. Tsokos, D. Morikis, and J. D. Lambris. *J. Med. Chem.*, 48:274–286, 2005.
6. P. Tamamis, S. Skourtis, D. Morikis, J. D. Lambris, and G. Archontis. Conformational analysis of compstatin analogues with molecular dynamics simulations in explicit water. In preparation.
7. A. D. Mackerell et al. *J. Phys. Chem. B*, 102:3586–3616, 1998.
8. E. Neria, S. Fischer, and M. Karplus. *J. Chem. Phys.*, 105:1902–1921, 1996.
9. B. R. Brooks et al. *J. Comput. Chem.*, 4:187–217, 1983.
10. D. Frishman and P. Argos. *PROT*, 23:566–579, 1995.
11. B. Dominy and C. L. Brooks III. *J. Phys. Chem. B*, 103:3765–3773, 1999.
12. B. Mallik, J. D. Lambris, and D. Morikis. *Proteins: Struct., Func. and Bioinf.*, 52:130–141, 2003.
13. R. Zhou and B. Berne. *Proc. Natl. Acad. Sci. USA*, 99:12777–12782, 2002.
14. H. Nymeyer and A. Garcia. *Proc. Natl. Acad. Sci. USA*, 100:13934–13939, 2003.

Using Simple Fluid Wetting as a Model for Cell Spreading

Matthew Reuter[‡] and Caroline Taylor[§]

Department of Chemistry, Michigan Technological University
Houghton, Michigan 49931, USA
E-mail: cmtaylor@mtu.edu

Cell morphology in contact with external media is dictated by the cell's interactions with the external media. One component that influences morphology is the thermodynamic aspect of cell spreading, which can be seen as analogous to wetting in a three-phase equilibrium. Results of classical density functional theory within the mean field approximation for a binary fluid are summarized to motivate simple changes to the model through the effective mass coefficient. We find that by altering the effective interactions we can alter the stable morphology, increasing contact by increasing component similarity. Thus we are able to propose a systematic approach to modeling cell spreading from simple mean field results.

1 Background

The spreading of a cell on a surface, though complex, is driven primarily by two components. One is a mechanical force from actin filament formation. The other is a more thermodynamic driving force, associated with balancing surface interactions, that allows analogy to simple wetting phenomena. This influence is visible in cell-cell contacts, such as found in tissues, and is of particular importance in biomedical applications¹, where patterned substrates are often employed in the design of biomimetic materials². The rate and extent of adsorption depends on the surface film's composition^{3,4}, which is exploited in micro-patterning applications⁵.

To develop the thermodynamic basis for cellular spreading from a line of three-phase contact, the cell-substrate-surround system can be envisioned as a constrained three-phase equilibrium, where there are two true phases and a third (the cell) which mimics one. Viewed in such a manner, there is readily conceivable an analog to the line tension which behaves characteristically in the spreading process, as is the case for a drop on a surface.

There are three general morphologies available to a drop in contact with an interface, shown in Figure 1. These are determined by the balance of interactions between the phases and characterized by a contact angle determined directly via Young's Law⁶. The drop can "bead" (Fig. 1(a)) or spread only minimally (non-wetting), as the system attempts to minimize the unfavorable interactions between the phases. Alternatively, it can spread unrestrictedly across the surface and wet it completely (Fig. 1(c)). Between these two regimes there is an intermediate state, in which the drop only partially wets the surface (Fig. 1(b)). In these instances, there is a line of three-phase contact along the drop boundary. Immediately before complete wetting, there is another possibility, that of pre-wetting (Fig. 1(d)), in which a layer of finite thickness first coats the interface before a discontinuous transition to infinite thickness occurs. Because the extent of cellular spreading is bounded by the cell membrane, it is to this final possibility that comparisons are most natural.

[‡]current address: Department of Chemistry, Northwestern University, Evanston, IL 60208

[§]The author also holds a visiting appointment at the James Franck Institute, University of Chicago, Chicago, Illinois 60637, USA.

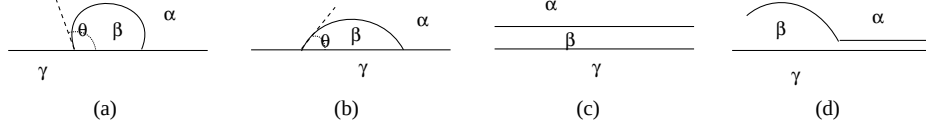


Figure 1. Wetting states possible for a drop on a surface: (a) non-wetting; (b) partial wetting; (c) complete wetting; (d) pre-wetting.

2 Simple Fluid wetting in a binary system

A binary three-phase system consists of two components ($i = 1, 2$) in three phases (α, β, γ), all of which are in equilibrium with each other, as shown in Figure 2, and is described by a free energy functional dependent upon the spatial component densities, $F[\rho_1, \rho_2]$. At the intersection of the three two-phase boundaries ($I = \alpha\beta, \beta\gamma, \alpha\gamma$), there is a line where all three phases are in contact. About this boundary there is a two-dimensional region of inhomogeneities which gives rise to an excess concentration along the line of three-phase contact. This results in an excess contribution to the free energy, yielding the line tension.

A commonly assumed model free energy density, Ψ , accounts for the variation of the densities in the region of the phase interfaces to second order:

$$\Psi = F[\rho_1, \rho_2] + \frac{1}{2} \sum_{i,j} m_{ij} \nabla \rho_i \cdot \nabla \rho_j, \quad (1)$$

where m_{ij} is a constant mass interaction term. The surface tension, σ_I , is determined by integrating this excess across the interface. The overall excess free energy is found similarly, and the surface excesses subtracted to yield the line tension:

$$\tau = \min_{\rho} \left[\int_A \Psi[\rho_1, \rho_2] da - \sum_I (\sigma_I R_I) \right] = \int_A (\Psi[\rho_1, \rho_2] - 2F[\rho_1, \rho_2]) da. \quad (2)$$

The latter form of the line tension, due to Kerins and Boiteux⁷, converges to the preceding variational form. Solving the Euler-Lagrange equations subject to the bulk conditions is

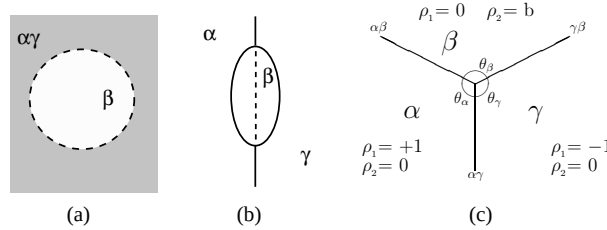


Figure 2. Model system for a drop of phase β at the interface between phases α and γ . (a): Top view of the drop showing the three-phase contact (dashed line); (b): Side view of the drop on the interface; (c): Region normal to the line of three-phase contact, showing the bulk component densities in each of the phases.

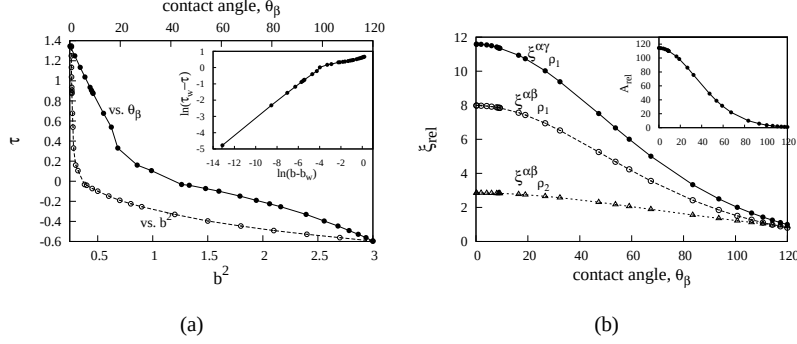


Figure 3. Results of solving the model binary system. (a) Line tension approaching wetting; (b) Response of the contact regions approaching the transition, relative to the $b = \sqrt{3}$ values. Used by permission from Ref. ⁹.

equivalent to minimizing the the free energy per unit area and yields the equilibrium density profiles without restricting them to a particular form.

The model system studied previously by Taylor and Widom^{8,9} is shown in Figure 2(c). It is described by a model free energy functional, $F[\rho_1, \rho_2; b]$ ¹⁰, which depends upon the relative component densities and a parameter (b , the bulk density of one of the components) which can be varied to induce a wetting transition, and $m_{ij} = \delta_{ij}$.

The line tension and contact region in this model exhibit an interesting behavior approaching the wetting transition, both changing qualitatively at the same point. In the line tension, shown in Figure 3(a), the slope changes near 20°. A shift between clearly different structures occurs in the that region ($b^2 \sim 0.3$, $\theta_\beta \sim 20^\circ$), resulting in an average behavior which satisfies the mean field prediction. The spatial extent of the inhomogeneous region of the phase contact (measured by the correlation length ξ), Figure 3(b), supports this, exhibiting a distinct plateau in the growth rate of the contact area at the same region.

Thus in even a so simple model there is non-uniformity, and it is this observation which motivates our current work, as it facilitates extension to systems where such behavior is involved. Such an approach allows the model to be applied to more complex systems by incorporating simple modifications.

3 Extending the model

The simple picture above is modified by altering the balance of the components within the square-gradient approximation, by scaling the off-diagonal values of the effective mass interaction parameter, m_{ij} . The rest of the model is as in the work of Taylor and Widom⁹.

A first test employs a value of $m_{12} = 0.75$ and maintains the self-interaction constant at unity. In this manner, the components are made to appear more similar, and the static contact angle ($b = \sqrt{3}$, $m_{12} = 0.75$) is 114°, showing a greater extent of spreading ($b = \sqrt{3}$, $m_{12} = 0$ gives 120°). This suggests that the variable morphology can be described by such a change in the simple model, although whether a wetting transition can be induced by such an external parameter is still undetermined.

From the phenomenological first steps, the mass interaction coefficient must be related

to physically-relevant values. The obvious way to do so is to establish a reference state chosen based on the problem, and relate m_{ij} to it, $m_{ij} = 1 - \frac{|\epsilon_i^o - \epsilon_j^o|}{\epsilon_i^o + \epsilon_j^o}$. In addition, an alternative form of the free energy functional itself is being constructed, with additional components that more closely mimic the real system¹¹.

4 Concluding Remarks

The use of a simple wetting model is not uncommon as an initial extension in studies of cell spreading, and it is promising that altering the effective component interactions does alter the morphology. The benefit of this model is that it is relatively tractable, while maintaining many of the characteristics of the unmanageably complex real system, and is readily extensible to higher levels of complexity (*i.e.* made more physical). It is thus fruitful to consider the system on as simple a level as possible, and incorporate higher levels of complexity systematically.

Within this classical wetting framework, a cell spreading on a surface can be represented as a set of three two-phase interfaces and simulations of slabs containing these interfaces used to determine the interfacial tensions (from the pressure tensor). From these, the analogues to the contact angle and line tension can be calculated via the spreading coefficient. Thus we may also carry out simple simulations that test the added complexity in the model and guide its development.

Acknowledgments

Funding for this work is provided by the Research Excellence Fund of the State of Michigan and Michigan Technological University. We are grateful to Benjamin Widom, Walter Nadler, and Prabhakar Bhimalapuram for many helpful and illuminating discussions.

References

1. D. Falconnet, G. Csucs, H. M. Grandin, and M. Textor. *Biomater.*, 27:3044–3063, 2006.
2. M. N. De Silva, J. Paulsen, M. J. Renn, and D. J. Odde. *Biotech. Bioeng.*, 93:919–927, 2006.
3. Y. Kikkawa, M. Fujita, T. Hiraishi, M. Yoshimoto, and Y. Doi. *Biomacromolecules*, 5:1642–1646, 2004.
4. B. Nguyen, A. Upadhyaya, A. van Oudenaarden, and M. P. Brenner. *Biophys. J.*, 86:2740–2474, 2004.
5. D. Lehnert, B. Wehrle-Haller, C. David, U. Weiland, and C. Ballestrem. *J. Cell Sci.*, 117:41–52, 2003.
6. J. S. Rowlinson and B. Widom. *Molecular theory of capillarity*. Clarendon Press, Oxford, 1982.
7. J. Kerins and M. Boiteux. *Physica A*, 117:575–592, 1983.
8. C. M. Taylor and B. Widom. *Mol. Phys.*, 103:647–655, 2005.
9. C. M. Taylor and B. Widom. *Phys. A*, 358:492–504, 2005.
10. I. Szleifer and B. Widom. *Mol. Phys.*, 75:925–943, 1992.
11. M. Reuter and C. M. Taylor. work in progress, 2006.

Parallel Cost-Sensitive Support Vector Machine Software for Classification

Tatjana Eitrich¹ and Bruno Lang²

¹ Central Institute for Applied Mathematics,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: t.eitrich@fz-juelich.de

² Applied Computer Science and Scientific Computing Group,
Department of Mathematics, University of Wuppertal, Germany
E-mail: lang@math.uni-wuppertal.de

1 Support Vector Machines

Support vector machines (SVMs) are well-known data mining methods for classification and regression problems¹. Their popularity is mainly due to their applicability in various fields of data mining, such as text mining², biomedical research³, and many more. Their accuracy is excellent and in many cases they outperform other machine learning methods such as neural networks. SVMs have their roots in the field of statistical learning which provides the reliable generalization theory⁴. Several properties that make this learning method successful are well-known, e.g. the kernel trick⁵ for nonlinear classification and the sparse structure of the final classification function. In addition, SVMs have an intuitive geometrical interpretation, and a global minimum can be located during the SVM training phase. In comparison to genetic algorithms or neural networks, less experience is required for using them, which helps researchers to get started with SVM software quite fast.

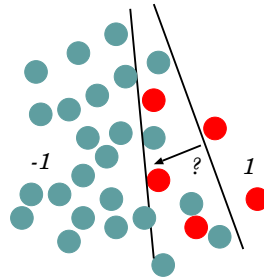


Figure 1. The problem of unbalanced and cost-sensitive classification.

2 Cost-Sensitive Support Vector Machine

Data sets with different class distributions lead to the effect that conventional machine learning methods are biased towards the larger class^{6,7}. To overcome this problem and

to obtain sensitive but also accurate classifiers we extended and improved the standard SVM formulation. In addition we use techniques addressing the problem of unbalanced classification, such as oversampling and threshold moving⁸. In Fig. 1 an unbalanced toy problem is shown. Recent results for a CYP P450 drug classification problem are given in ref 9.

Parallel Support Vector Machine

The main drawback of current SVM models is their high computational complexity for large data sets¹⁰. This can in fact restrict the applicability of SVMs since the amount of data for classification modeling increases dramatically. Therefore the development of highly scalable parallel SVM algorithms is a new important topic of current SVM research. Some algorithms for parallel SVM learning already do exist, but most of them are limited to heuristics for distributed training on reduced data sets^{11,12}. These are not useful as stand-alone systems for high quality learning on large data. We have implemented a parallel support vector machine software well suited for multi-processor shared memory (SMP) clusters that become more and more available. Our algorithm can be used in serial and parallel mode. The parallel implementation provides pure MPI and OpenMP modes as well as a hybrid mode which combines fine and coarse grained parallelization aspects to a well scalable SVM learning method¹³. The fine grained inner parallel scheme is shown in Fig. 2.

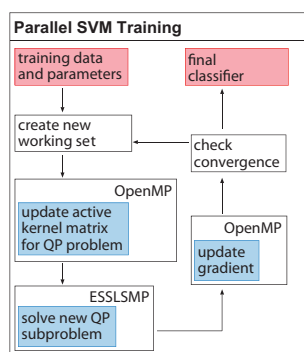


Figure 2. Parallel training algorithm based on a decomposition method for SVM training.

Summary and Future Work

We obtained a flexible cost-sensitive parallel SVM software that can be used on high-end machines with SMP architectures to process the large data sets that arise more and more in bioinformatics and other fields of research. Future work will be on enhanced parameter tuning.

Acknowledgments

The computations were performed on the IBM p690 cluster JUMP at Research Centre Jülich. The authors would like to thank the ZAM team at Jülich for support.

References

1. N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.
2. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML'98), Chemnitz*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer, 1998.
3. H. Yu, J. Yang, W. Wang, and J. Han. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. In *2nd IEEE Computer Society Bioinformatics Conference (CSB 2003), 11-14 August 2003, Stanford, CA, USA*, pages 220–228. IEEE Computer Society, 2003.
4. V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.
5. B. Schölkopf. The kernel trick for distances. In *NIPS*, pages 301–307, 2000.
6. C. Elkan. The foundations of cost-sensitive learning. In *Proc. of the 17. International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.
7. M. A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC*, 2003.
8. R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Proceedings of the 15th European Conference on Machine Learning (ECML'04), Pisa, Italy, 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 39–50. Springer, 2004.
9. T. Eitrich, A. Kless, C. Druska, J. Grotendorst, and W. Meyer. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. Technical Report FZJ-ZAM-IB-2006-09, FZJ, 2006.
10. N. Chen, W. Lu, J. Yang, and G. Li. *Support vector machine in chemistry*. World Scientific Pub Co Inc, 2004.
11. H. P. Graf, E. Cosatto, L. Bottou, I. Dourdanovic, and V. Vapnik. Parallel support vector machines: the cascade SVM. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 521–528. MIT Press, Cambridge, MA, 2005.
12. J.-X. Dong, A. Krzyzak, and C. Y. Suen. A fast parallel optimization for training support vector machines. In P. Perner and A. Rosenfeld, editors, *Proceedings of 3rd International Conference on Machine Learning and Data Mining*, pages 96–105, 2003.
13. T. Eitrich, W. Frings, and B. Lang. HyParSVM – a new hybrid parallel software for support vector machine learning on SMP clusters. In *Proceedings of the Euro-Par Conference, Dresden*, 2006.

A Study on the Structure and Dynamics of Spin Labeled Proteins by MD Simulations and EPR Spectroscopy.

Prasad Gajula¹, Igor V. Borovykh¹, Christian Beier¹,
Peter Gast², Martina Huber², and Heinz -Jürgen Steinhoff¹

¹ Dept.of Physics, University of Osnabrueck,Barbarastr-7, 49076 Osnabrück, Germany

² Dept.of Bio/Mol.physics, Huygens Laboratory, Leiden University, Leiden, The Netherlands

E-mail: prasad.gajula@uos.de

Investigations of conformational changes is a prerequisite to understand a protein's biological function. To explore these conformational changes a strategy is developed combining molecular dynamics (MD) simulations and electron paramagnetic resonance (EPR) spectroscopy of spin labeled proteins. The reaction center (RC) of *Rhodobacter sphaeroides*, a photosynthetic protein, is chosen as a model system to study the structure and dynamics that determine electron transport function. Only in combination EPR experiments and MD simulations positively identify the spin label linker position as Cys156 on subunit H. The distance between the primary acceptor (QA) and the bound spin label determined by EPR experiments and MD simulations deviate by about 7% indicating that the conformations of RC in the crystallized and in the solubilized states are slightly different.

1 Introduction

Proteins execute and control essentially all functions in living organisms by changing into different conformational sub-states¹. Understanding the underlying mechanisms involved in these conformational transitions is one of the main challenges of biophysical research. EPR spectroscopy is one of the few techniques available to study the structure and dynamics of a protein under physiological conditions. To overcome experimental limitations and to improve the analysis of the experimental results molecular dynamics simulations are applied. In the present report, the combination of MD simulations with EPR spectroscopy serves as a powerful tool to explore the conformation and conformational dynamics of the reaction center (RC) protein of *Rhodobacter sphaeroides*, a protein involved in photosynthesis. A cysteine specific spin probe, a so called methanethiosulfonate spin label which contains an unpaired electron on its nitroxide group, is attached to RC (for a detailed review on site-directed spin labeling see²). The EPR spectrum, which is sensitive to the spin label reorientational dynamics, reports on the interaction of the spin label side chain with the protein backbone or neighboring side chains, the conformational dynamics of the protein, and the rotational diffusion of the protein. Inter-residue distances can be determined by analysis of the dipolar interaction of two unpaired electrons, e.g., a spin label and a native radical center. MD simulations of the spin labeled RC, on the other hand, provide trajectories of the spin label side chain dynamics from which EPR spectra are calculated³.

2 Methods and Materials

The photosynthetic RC is an integral membrane protein complex, that is composed of three polypeptide subunits (L, M and H) and ten non covalently bounded cofactors⁴. This

complex (Fig.1) with particular spatial configuration plays a central role in the process of energy conversion. EPR experiments were performed on spin labeled RC. MD simulations were performed to identify which of the native cysteines the spin label binds preferentially. Simulation of EPR spectra based on MD trajectories allows a direct comparison of experimental spectra with the simulated one. For a detailed review on the method, see³.

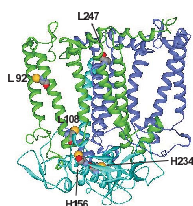


Figure 1. Reaction centre of Rb. Sphaeroides with 5 native cysteines at positions 92, 108, 247 on subunit L and at 156, 234 on subunit H. The spin label can only bind to one of these 5 cysteines.

3 Results and Discussion

3.1 Identification of the spin label linker position

Biochemical analysis and EPR experiments on the isolated LM complex shows that the spin label is bound to either of the two cysteines on subunit H. Further MD simulation analysis was performed to identify the exact position of the spin label on subunit H.

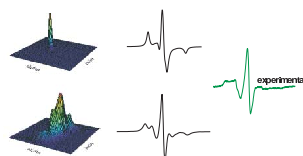


Figure 2. Orientation distribution of the nitroxide ring (left) and calculated EPR spectra (middle) for spin labels attached to position 234, C234R1, (top) and 156, C156R1, (bottom). The experimental spectrum is shown in green (right). The population distributions projected onto the alpha-beta Euler angle plane reveals the different restrictions of the reorientational dynamics of the nitroxides at two sites. These different nitroxide dynamics lead to distinguishable degrees of motional averaging of the g- and hyperfine tensor anisotropies and are reflected in the EPR spectral shapes.

The conformational space occupied by the nitroxide of spin label C234R1 during a 6 ns MD simulation run shows considerable restriction of the spin label motion. Consequently the EPR spectrum calculated for an isotropic distribution of protein orientations and with the potential determined from the shown orientation distribution (Fig.2) of the nitroxide coincides with a powder spectrum. In contrast, the space covered by the nitroxide of the spin label side chain at position C156 reveals considerable dynamics. The Euler angle projection shows a broad distribution of the angle beta with distinct population maxima.

The apparent hyperfine splitting and the spectral shape of C156R1 are similar to that of the experimental spectrum. Based on this result we conclude that cysteine 156 was labeled in the present experiment⁵.

3.2 Distance Measurements

Changes in RC conformation related to electron-transfer events could, in principle, be detected by measuring the distance between a native cofactor and a spin label. For that purpose a two-frequency pulsed electron paramagnetic resonance method (double electron electron spin resonance - DEER) was used⁶.

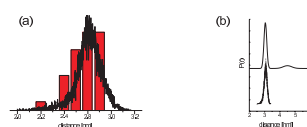


Figure 3. Distance between oxygen, O5 of QA and the centre of the N-O bond of the spin label for all frames of the MD run represented as distance distribution and the histogram of results of energy minimization on 30 structures (shown in red). (b) Distance distribution determined from analysis of experimental data (DEER). To facilitate comparison the distance distribution determined by MD simulation shifted by + 0.28 nm is also shown(bottom).

A distance of 3.05 nm is found in between the light generated semiquinone anion radical of the primary acceptor (QA) and the spin label at position H156 from DEER experiments. MD simulations are performed to interpret the distance. A 6 ns MD run comprising the entire RC protein yields a distance distribution that is close to the experimental one. However, the average distance value found by the MD simulation is smaller than that obtained by DEER by approximately 0.2 nm. Energy minimization on randomly selected points of the MD trajectory confirms this result yielding a slightly broader distribution width. The possible reasons for the difference in the distance values determined by two methods could be (i) incomplete sampling of the nitroxide accessible space during the 6ns MD or (ii) small differences of the conformations of RC in the frozen samples used in experiments and of the crystal structure used for MD simulations.

4 Conclusion and Outlook

Combining the molecular dynamics simulations and EPR spectroscopy we identify the spin label binding site at position 156 of subunit H of RC protein. The average distance between the native cofactor QA and the spin label bound to position 156 determined by EPR and MD deviate by about 7% indicating that the conformations of RC in the crystallized and in the solubilized states are slightly different.

References

1. Frauenfelder H. and McMahon B. , PNAS **95**, 9 (4795-4797)1998.
2. Steinhoff, H.-J. Frontiers in Bioscience 7, c97-110 (2002)
3. Steinhoff,H.-J.,Hubbell,W.L. Biophysical Journal **71** (2201)1996.
4. Stowell,M.H.B., *et.al.* Science 276, 812-816 (1997)
5. Gajula, P., Borovykh, I.V., Beier, C. , Gast, P., Steinhoff , H.-J. (in preparation)
6. Borovykh, I.V., Ceola, S., Gajula, P., Gast, P., Steinhoff , H.-J., Huber, M. JMR **180**, 178-185 (2006)

An Improved Evolutionary Strategy for Protein Structure Prediction

Srinivasa Murthy Gopal¹, Kyu Hwan Lee², and Wolfgang Wenzel¹

¹ Institut für Nanotechnologie,
Forschungszentrum Karlsruhe, 76021 Karlsruhe, Germany
E-mail: {gopal, wenzel}@int.fzk.de

² Korea Institute of Science and Technology,
Supercomputational Materials Simulation lab,
130650 Seoul, South Korea
E-mail: khlee@fireworks.kist.re.kr

We propose a new evolutionary strategy for protein structure optimization in a free energy landscape. This strategy is an improved version of existing basin hopping technique where multiple independent trajectories are used. In the new method, these trajectories depend on each other. This technique is implemented as a simple master-client model for distributed computing. We demonstrate advantage of purposed method in terms of computational effort and structure diversity.

1 Introduction

Protein structure prediction(PSP) is regarded as one of the grand computational challenge. Ab initio protein structure prediction is one promising approach to obtain the native protein conformation from first principles. We use a free energy forcefield¹ with combination of several optimization methods² to predict the native structure of protein. We identify the native structure as the lowest energy conformation in our forcefield. The basin hopping technique(BHT)³ has been our work horse for the structure optimization. We have predicted native structures of several proteins including 20 amino acid trp-cage protein, 40-amino acid headgroup of the HIV accessory protein using the BHT⁴. Though BHT proved to be a good optimization method, it has several drawbacks. Typically the independent BHT trajectories find identical structures corresponding to one local(global) minima of free energy landscape. There exists also a problem of BHT simulations going astray. We had previously identified this problem and proposed an evolutionary strategy⁵ which eliminates the problems associated with BHT. In this current work we propose an improved version for protein structure prediction.

2 Method

We have generalized the BHT approach to a population of size N which is iteratively improved by P concurrent dynamical processes. The population is evolved towards a optimum of the free energy surface with a simple evolutionary strategy(ES). The strategy balances the energy improvement as well as the population diversity. The conformations are drawn from the population and subjected to an annealing cycle. At the end of each cycle the

resulting conformation is either integrated into the *active* population or discarded. The algorithm was implemented as a simple master-client model in which the idle clients request a task from the master. The master maintains the active conformation of the population and distributes the work to the clients. Each step in the algorithm has three phases.

- *Selection* A conformation is drawn randomly from the *active* population. We have used a uniform probability distribution with active population of 20 conformers.
- *Annealing cycle* We used a geometric cooling schedule with T_{start} drawn from a uniform or exponential distribution and T_{end} fixed at 2K. The number of steps per cycle is increased as $10^5 \times \sqrt{N_{cycle}}$.
- *Population Update* We have adjusted the acceptance criterion for newly generated conformations to balance the population diversity and energy enrichment. We define the close structures as conformation which have bRMSD(back bone RMSD) of 3Å to each other. The master performs one of the following operations.
 - *Add* The new conformation is not *close* to any structure in the population, it is added to the pool.
 - *Replace* If the new conformation is *closest* to some structure in the population, it replaces that structure provided its re-weighted energy (see below) is less than the closest one.
 - *Merge* If the new conformation has several *close* structures, it replaces this group of structures provided its re-weighted energy is less than the best one of the group.

We have used an energy criterion for the *Replace* and *Merge* operations during population update. We re-weight the energy of the new conformation(E_{rew}) as

$$E_{rew} = A \times \tanh D \quad \text{where} \quad D = \frac{E_{best} - E_{new}}{A}$$

E_{new} is the actual energy of the new conformation, E_{best} is the current best low energy. We have also optimized the number of concurrent processes with respect to the size of *active* population. We have investigated the folding of a small beta peptide, tryptophan zipper (PDBID:1LE0) for this purpose. We fixed the size of active population to 20 and used 10,20,30,40,50,60 concurrent processes(Fig 1). We found an optimal number of processes to be approximately equal to two times the active population. We confirmed this for population size of 5 and 10.

3 Results

We have used the improved evolutionary strategy for predicting the native structure of a 12 amino acid tryptophan zipper(PDBID:1LE0). The table below shows the top 10 structures, their energies, secondary structure and bRMSD with respect to experimental structure.

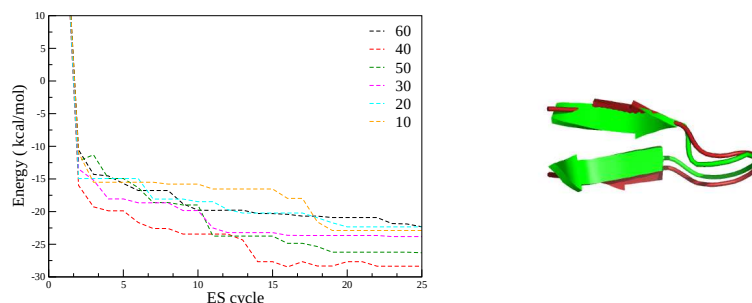


Figure 1. Left : The lowest Energy vs the ES cycle for different processes. Right : The predicted(red) and experimental(green) structures.

Name	Secondary structure	Energy (kcal/mol)	bRMSD Å
EXP	CEEECSSEEEEC		
1	CEEECSSEEEEC	-28.360	1.220
2	CCEECSSEEEEC	-21.350	3.800
3	CEEETTEEECCC	-19.470	3.790
4	CESSSSSEEEEC	-19.130	3.270
5	CCEECSSEEEEC	-19.040	3.630
6	CCCCCTTTTCCC	-18.820	6.170
7	CCCTTTTCCCCC	-18.450	4.510
8	CCCBTTBCCCC	-18.120	3.360
9	CCHHHHHHHHHC	-17.850	6.880
10	CCCCCTTTTCCC	-17.600	6.370

4 Conclusions

We have developed and applied an improved evolutionary strategy which evolves a set of conformers to population of low energy diverse structures. We have optimized the algorithm and used it for tertiary structure prediction of a small beta hairpin. The method presented here is well suited for the distributed computational architecture.

Acknowledgments

We thank the BMBF, the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2, WE 1863/14-1) and the Kurt Eberhard Bode Stiftung for financial support. We are grateful to S. Gregurick and J. Moult, for the contribution of part of code used in these simulations. We are thankful to KIST for computational resources.

References

1. T. Herges, W. Wenzel *An all-atom forcefield for tertiary structure prediction of helical proteins*, Biophysical Journal **87**, 3100 (2004) A. Verma and W. Wenzel (in preparation)
2. A. Schug, A. Verma, W. Wenzel and G. Schoen *Biomolecular structure prediction with stochastic optimization methods.*, Adv. Eng. Materials **7**, 1005 (2005)
3. A. Nayeem, J. Vila, and Harold A. Scheraga *A comparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [met]-enkephalin*, J. Comp. Chem **12**, 594 (1991)
4. A. Schug, A. Verma, K. H. Lee and W. Wenzel *Basin Hopping Simulations for All-Atom Protein Folding*, J. Chem. Phys. (**in press**), (2006)
5. A. Schug and W. Wenzel *An Evolutionary Strategy for All-Atom Folding of the 60-Amino-Acid Bacterial Ribosomal Protein L20*, Biophysical Journal **90**, 4273 (2006)

Ab initio Simulations of Cu Binding Sites in the N-Terminal Region of PrP

S. Furlan¹, F. Guerrieri^{2,3}, G. La Penna¹
S. Morante^{2,3}, and G.C. Rossi^{2,3}

¹ CNR - ISMac - Sezione di Genova
E-mail: {furlan, lapenna}@ge.ismac.cnr.it

² Dipartimento di Fisica - Università di Roma "Tor Vergata"

³ INFN - Sezione di Roma "Tor Vergata"
E-mail: {guerrieri, morante, rossig}@roma2.infn.it

The prion protein (PrP) binds Cu^{2+} ions in the octarepeat domain of the N-terminal tail up to full occupancy at pH=7.4. Recent experiments show that the HGGG octarepeat subdomain is responsible for holding the metal bound in a square planar coordination. By using first principle *ab initio* molecular dynamics simulations of the Car-Parrinello type, the Cu coordination mode to the binding sites of the PrP octarepeat region is investigated. Simulations are carried out for a number of structured binding sites. Results for the complexes $\text{Cu}(\text{HGGGW})+(\text{wat})$, $\text{Cu}(\text{HGGG})$ and the dimer $[\text{Cu}(\text{HGGG})]_2$ are presented. While the presence of a Trp residue and a H_2O molecule does not seem to affect the nature of the Cu coordination, high stability of the bond between Cu and the amide Nitrogens of deprotonated Gly's is confirmed in the case of the $\text{Cu}(\text{HGGG})$ system. For the more interesting $[\text{Cu}(\text{HGGG})]_2$ dimer a dynamically entangled arrangement of the two monomers, with intertwined N-Cu bonds, emerges. This observation is consistent with the highly packed structure seen in experiments at full Cu occupancy.

1 Introduction

The Prion Protein (PrP) is a cell surface glycolipid protein, highly expressed in the central nervous system of many mammals. Its physiological rôle is still unclear, but it has been shown that it can selectively bind Cu^{2+} . A number of Cu binding sites (no less than four) have been identified along the whole protein.

Mature PrP (comprising a.a. 23-231) has a flexible, disordered, N-terminal (23-120) and a globular C-terminal (121-231). The N-terminal domain of human PrP contains four repeats of the eight peptides, PHGGGWGQ, from residue 60 to 91, called *octarepeat*^a. It has been shown that each octarepeat is able to bind one Cu^{2+} ion^b.

Cu-octarepeat interaction is cooperative, and could possibly have a rôle in disease related PrP aggregation. Several techniques have been used in the study of the octarepeat structural arrangements, including X-ray crystallography¹ and EXAFS spectroscopy². Experiments have shown that only the oligopeptide HGGGW is directly involved in Cu^{2+} coordination. The crystallographic data shows that the Cu^{2+} ion is penta-coordinated, displaying a square planar equatorial coordination with three Nitrogens and one Oxygen from the HGGGW peptide, and an axial Oxygen from a water molecule.

The equatorial coordination of the N3O1 type is with the $\delta 1$ Nitrogen of the His, the deprotonated amide Nitrogens of the two following Gly's and one carbonyl Oxygen from

^aIn other species this number may be different. For instance, *Bos gaurus*, has 5 octarepeats.

^bOther possible binding sites have been found in the C-terminal region.

the second Gly of the sequence. The axially bound water molecule is kept in position by a hydrogen bond to the H ϵ 1 of the Trp indole ring.

We have investigated the coordination mode of Cu to the octarepeat via *ab initio* molecular dynamics simulations *à la* Car–Parrinello. In order to have a system of tractable size, we have not considered the whole octarepeat, but only smaller portions of it, complexed with Cu and water in various combinations. More precisely we have investigated the following systems: the Cu(HGGGW)(H₂O) complex, a box containing Cu(HGGG) and 41 water molecules and a system composed by two HGGG oligopeptides, both in the presence and in the absence of an associated pair of Cu²⁺ ions.

The downsizing of the system here consistent with the experimental findings that only the HGGGW portion is directly involved in Cu²⁺ coordination^{1,3}.

2 Methods

Car–Parrinello MD simulations have been carried out employing the Quantum-ESPRESSO package⁴. We have used Vanderbilt’s ultrasoft pseudopotentials and the PBE exchange–correlation functional⁵. Periodic boundary conditions have been imposed on the super-cell, with a minimum separation of 5 and 8 Å between replica, for neutral and charged systems, respectively. The energy cutoff was of 25 Ry, while the hard cutoff for the augmented charge density was of 250 Ry. All the simulations were spin-restricted. In some particularly interesting cases, namely the monomer Cu(HGGG) and the dimer [Cu(HGGG)]₂, we have performed spin unrestricted simulations, with S=1/2 and S=1, respectively.

Every simulation consisted of the following steps.

1. Electronic energy minimization with fixed atomic positions.
2. Energy minimization with respect to atomic and electronic degrees of freedom to attain the equilibrium geometry.
3. Two subsequent molecular dynamics simulations of 0.25 ps each at $T = 100$ and 200 K, respectively ^c.
4. Molecular dynamics simulations of about 2 ps at $T = 300$ K, using the same thermostat as in 3.

Thermalization is necessary to slowly approach room temperature and avoid that temperature oscillations obscure the electronic properties of the ground state. Equations of motion have been integrated using the usual velocity-Verlet algorithm with a time step of 0.12 fs. Simulations have been carried out on Linux-clusters using 8-16 processors, depending on the size of the systems ^d. The CPU time per step was, on average, of 9.6 s when running on 16 nodes at 2.7 GHz.

3 Results

Simulations of Cu(HGGGW)(H₂O) show that the water molecule in the axial coordination site is not bonded to Cu, but rather it is hydrogen-bonded to the Trp side chain. Indeed we

^cA Nosé–Hoover thermostat at the required temperature was coupled to ionic degrees of freedom.

^dFor a reference, the monomer Cu(HGGG) comprises 49 atoms and 157 valence electrons.

don't see a propensity for a chemical Cu-O(water) bond. Crystallographic packing effects can explain the presence of water seen in the crystal. We have performed simulations in super-cells of increasing size, finding that the volatility of the water molecule actually increases. We have thus decided to discard the Trp from further simulations.

Analysis of several trajectories of the complex Cu(HGGG) at $T = 300$ K show that the N3O1 coordination is stable, with the bond between Cu and amide Nitrogens from deprotonated Gly's stronger than that between Cu and the $\delta 1$ Nitrogen of the His. The Cu-O bond with the carbonyl group of the second Gly is weak, and affected by the greater mobility of the peptide C-terminal.

Simulations of Cu(HGGG) immersed in a box with 41 water molecules confirm the picture in which deprotonation of Gly's is favored by the greater stability of the bond with Cu, even if at physiological pH Gly's are expected to be protonated.

Study of the system [Cu(HGGG)]₂ in vacuum shows that, on the time scale of the ps, formation of a dimer is possible when the tetrapeptides bind copper. Exchange of ligands (visible after 0.86 ps at $T = 300$ K) between the two Cu ions is seen to keep the two tetra-peptides close to each other. Also for the dimer we observe a preference for an N amidic binding, while the bond with the imidazole ring of the His can be broken. In spin restricted simulations we observe a rather small Cu-Cu average distance (from 2.1 to 3 Å) which, in spin unrestricted ones, becomes slightly larger (2.4 to 3.4 Å), and thus nearer to experimental observations. In the absence of Cu, long-range electrostatic and dispersive interactions are unable to keep the two (HGGG) close together. We have not yet studied the effects of hydrophobic interactions.

Acknowledgments

The use of *Cineca* cluster in Bologna, Fermi cluster at *E. Fermi Research Center* in Roma, BEN cluster at *ECT** in Trento and *CERM* cluster in Firenze is gratefully acknowledged. We wish to thank the Organizers for the lively atmosphere of the Meeting.

References

1. C.S. Burns, E. Aronoff-Spencer, C.M. Dunham, P. Lario, N.I. Avdievich, W.E. Antholine, M.M. Olmstead, A. Vrielink, G.J. Gerfen, J. Peisach, W.G. Scott and G.L. Millhauser, *Molecular Features of the Copper Binding Sites in the Octarepeat Domain of the Prion Protein*, *Biochemistry* **41**, 3991–4001 (2002).
2. S. Morante, R. Gonzalez-Iglesias, C. Potrich, C. Meneghini, W. Meyer-Klaucke, G. Menestrina and M. Gasset, *Inter- and Intra-octarepeat Cu²⁺ Site Geometries in the Prion Protein* *J. Biol. Chem.* **279**, 11753–11759 (2004).
3. M.J. Pushie and A. Rauk, *Computational studies of Cu²⁺ [peptide] binding motifs: Cu[HGGG] and Cu[HG] as models for Cu²⁺ binding to the prion protein octarepeat region* *J. Biol. Inorg. Chem.* **8**, 53–65 (2003).
4. S. Baroni, A. Dal Corso, S. de Gironcoli, P. Giannozzi, C. Cavazzoni, G. Ballabio, S. Scandolo, G. Chiarotti, P. Focher, A. Pasquarello, K. Laasonen, A. Trave, R. Car, N. Marzari and A. Kokalj, <http://www.pwscf.org/>.
5. J.P. Perdew, K. Burke and M. Ernzerhof, *Generalized Gradient Approximation made simple* *Phys. Rev. Lett.* **77**, 3865 (1996).

Cross-Check Methods in Protein Simulations

Christoph Junghans^{1,2} and Ulrich H.E. Hansmann^{1,3}

¹ Computational Biology and Biophysics,
John von Neumann Institut für Computing,
Forschungszentrum Jülich, Germany
E-mail: {u.hansmann, c.junghans}@fz-juelich.de

² Institut für Theoretische Physik,
Universität Leipzig, Germany

³ Department of Physics,
Michigan Technological University,
Houghton, USA

The aim of this work^{1,2} is to study the behavior of three advanced Monte Carlo methods in protein simulations employing a realistic ECEPP/3-based all-atom model³. The implementation is based on the open source package SMMP⁴. The techniques applied were Wang-Landau sampling², parallel tempering^{6,7} and random tempering. All three techniques show very good agreement in the outcome, a cross-check of the simulation results is possible.

1 Introduction

We started from the question which present algorithm is the best. First we used parallel tempering and Wang-Landau sampling. After our first simulations² we found that both algorithms had nearly the same strength. Therefore, we moved to the question if the results correspond for all algorithms. This provides the opportunity to have a cross-check for your simulation data.

2 Model

The SMMP⁴ package we applied uses the well-known ECEPP/3 Potential³

$$E_{\text{tot}} = E_{\text{LJ}} + E_{\text{el}} + E_{\text{hb}} + E_{\text{tors}} .$$

3 Methods

3.1 Wang-Landau Sampling

This method is a combination of the Wang-Landau algorithm⁸ and the multicanonical algorithm⁵. First we use the Wang-Landau algorithm to evaluate the multicanonical weights $W_{\text{Muca}}(E)$ from the given estimator of the density of states $\Omega(E)$. The second step was a normal multicanonical simulation with fixed weights.

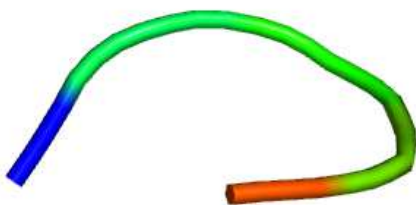


Figure 1. Backbone of the vacuum ground state of Met-enkephalin

3.2 Parallel Tempering

We also used the well known parallel tempering technique, which is a multi Markov chain process that provides the local updates at fixed temperature and global exchange updates between different temperatures.

3.3 Random Tempering

The latest method we studied was a uncommon type of simulated tempering⁹. We used the Wang-Landau algorithm⁸ to estimate the temperature depending weights for the global update which changes the temperature.

4 Object of study

We used the pentapeptide Met-enkephalin (see Fig. 1)



with a ground state energy at -12.43 kcal/mol. Because of its simplicity this peptide becomes a check point for every new algorithm.

5 Results

As Fig. 2 shows the mean value of the energy is sampled very good for all algorithms. The normalized specific heat shows small variation (see Fig. 3).

6 Conclusion

All techniques, parallel tempering, Wang-Landau sampling and random tempering allow to obtain information over a large temperature range from a single simulation. All are at least two orders of magnitude faster than a canonical simulation at a low temperature of $T = 100$ K.

Hence, choice between the algorithms will depend on the equipment available and the personal preferences of the researcher. As all three techniques are numerically different a cross-check of the simulation results in protein studies is possible.

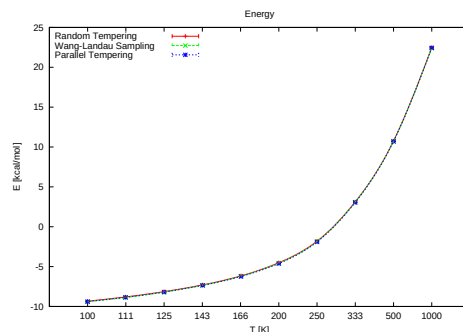


Figure 2. Averages of energy simulated with different methods show very good agreement.

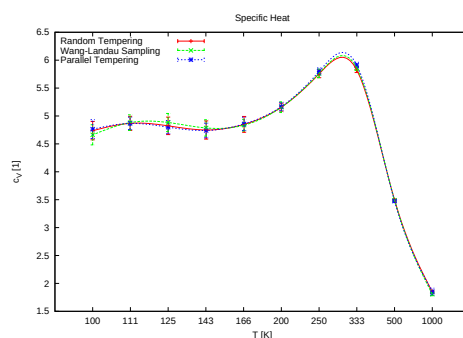


Figure 3. Specific heat simulated with different methods show small variations, but still good agreement.

References

1. C. Junghans, in Technical Report IB-2005-13, R. Esser (Ed.), John von Neumann Institute for Computing, Jülich, 2005.
2. C. Junghans and U.H.E. Hansmann, in press: Int. J. Mod. Phys. C (2006).
3. G. Némethy et al., J. Phys. Chem. **96** (1992) 6472.
4. F. Eisenmenger et al., Comp. Phys. Comm. **138** (2001) 192, (available from www.phy.mtu.edu/biophys/smmp.htm).
5. B.A. Berg and T. Neuhaus, Phys. Rev. Lett. **68** (1992) 9.
6. K. Hukushima and K. Nemoto, Jour. Phys. Soc. (Jap.) **65** (1996) 1604.
7. U.H.E. Hansmann, Chem. Phys. Lett. **281** (1997) 140.
8. F. Wang and D.P. Landau, Phys Rev. E **64** (2001) 056101.
9. E. Marinari and G. Parisi, Europhys. Lett. **19** (1992) 451.

The Influence of Molecular Surface Composition on the Outcome of Poisson Boltzmann Calculations Performed to Obtain Solvation Free Energies

Parimal Kar¹, Yanjie Wei¹, Ulrich H.E. Hansmann^{1,2}, and Siegfried Höfner¹

¹ Michigan Technological University,
Department of Physics, 1400 Townsend Drive, Houghton, MI 49931, USA
E-mail: {pkar, yawei, hansmann, shoefing}@mtu.edu

² John von Neumann Institute for Computing, FZ Jülich
52425 Jülich, Germany
E-mail: u.hansmann@fz-juelich.de

Poisson Boltzmann is a widely used method to describe solvation effects in structural biology. In the present study we want to focus on the Boundary Element Method within the various set of different Poisson Boltzmann Techniques. Here a fundamental dependence arises from the fact that initially the molecular surface must be computed and discretized in terms of small-sized simple geometric elements, i.e. triangles, tesseræ, curved triangles etc. The current study investigates the influence of this critical aspect of molecular surface discretization on the final results of PB/BEM-solvation free energy calculations. Different algorithms are compared to each other as well as to high-level reference states obtained from Quantum Chemistry calculations. Surface resolutions are systematically increased until a sufficient level of convergence is reached. The trends are analyzed with respect to different molecular surface algorithms and varying size and shape of proteins. General conclusions are drawn concerning key-requirements for the molecular surface composition in order to achieve a maximum level of technical robustness.

1 Introduction

Many important processes in biochemistry involve proteins in solution. Theoretical approaches aimed at understanding the structural and mechanical basis of these processes will therefore have to also consider the effect coming from the environment, i.e. from water, membranes, ionic solutions etc. One way of describing solvation effects in the context of biomolecular simulation is based on the Poisson Boltzmann treatment (PB) of macromolecules^{1,2}. Especially in the latter description — the *Boundary Element Method* (BEM) — a particular sensitivity arises with respect to the underlying molecular surface computation, which is used for defining the boundary between molecular volume and the environment. While in analogous models used in Quantum Chemistry³ the average area of these boundary elements has been advised to be on the order of $0.4 \text{ \AA}^2$, similar rules of thumb seem to be missing for classical PB/BEM approaches of protein structures. We therefore want to investigate in this present study how molecular surface decomposition will affect the net result of PB/BEM. Particular emphasis will be placed on the question if with increasing resolution into boundary elements stable converged PB/BEM results are achievable and how commonly employed molecular surface programs compare to each other in this regard.

2 Methods

Molecular surfaces are computed with two independent programs, the Connolly program⁴ and the SIMS program⁵. PB/BEM energies are calculated with program POLCH⁶. Public data bases⁷ are used for the download of protein structures in PDB format. A diverse set of different proteins is selected reaching from small-sized peptides of 44 residues to up to large-sized globular proteins of about 500 residues. PDB structures are visualized with program MOLDEN, after having removed HETATM lines, CONNECT lines, ANISOU lines and the footer section, and having deleted all but the 'A' chain in multiple chain pdb files so that column number five is exclusively left with 'A' labels. MOLDEN-visualization of the cleaned PDB files involves selecting force field **Tinker Amber** and writing out a PDB file from within MOLDEN with the option 'Write_With_Hydrogens'. Since MOLDEN always uses the default HIP type in AMBER jargon, we need to convert the HIS residues to HIP ones, as well as to change CYS residues engaged in disulfide bonds to CYX-type residues. After that we make sure that the initial residue is not PRO, and delete initial PROs in case. AMBER non-bonded parameters⁸, i.e. charges and van der Waals radii are assigned to all the atoms in the protein structures. The vdW radii are increased by a factor of 1.12 and the charges are scaled down by another factor of 0.9 which was found previously to lead to better agreement with high level Quantum Chemistry reference calculations. Before we execute Connolly's MSROLL program, inner/outer dielectric constants at the molecular boundary must be defined. In the case of water we use the standard values of 1.0 and 80.0. We execute the MSROLL program with different values for the *f* argument, where *f* defines the fineness, in other words, the resolution of the surface. When *f* is decreased, the resolution of the surface becomes better but computational cost will increase. For some very small values of *f* the Connolly program fails. In such cases we change parameter *l* which defines a minimal size of triangles to be considered distinct. The *p* parameter of MSROLL, that is the value of the probe sphere radius remains constant 1.5 (in Angstrom) in all the calculations. We record the analytically calculated surface area and volume and the employed *f* value and transfer the result file tmp.vet into a human readable format, called COMB_CONN.dt. Then we clean the COMB_CONN.dt file from critical entries, which are triangles that almost coincide and the like. In the next step, we compile the Poisson-Boltzmann program and record the number of boundary elements (triangles) and the number of atoms we have employed. Then we actually run the PB calculation, recording the PB energy and the number of iterations needed for the PB/BEM calculation. Now we refine the -*f* parameter and repeat the whole procedure. This is continued until convergence of the PB/BEM result is reached. The smaller the -*f* value becomes, the more BE will be produced, hence the more complicated and long-lasting the PB calculation will be.

Using the SIMS program is similar except that the molecular surface and the solvent excluded volume are computed numerically and that the probe radius, the smooth radius and the dot-density are defined within a small argument file to the program. Analogous to modifying the -*f* parameter in MSROLL is changing the dot-density in SIMS. Higher values of the dot-density will provoke higher surface resolutions. The remainder of the procedure described above is identical when using SIMS. We record dot-density, number of BE, PB energy, number of iterations, SASA and volume for comparison.

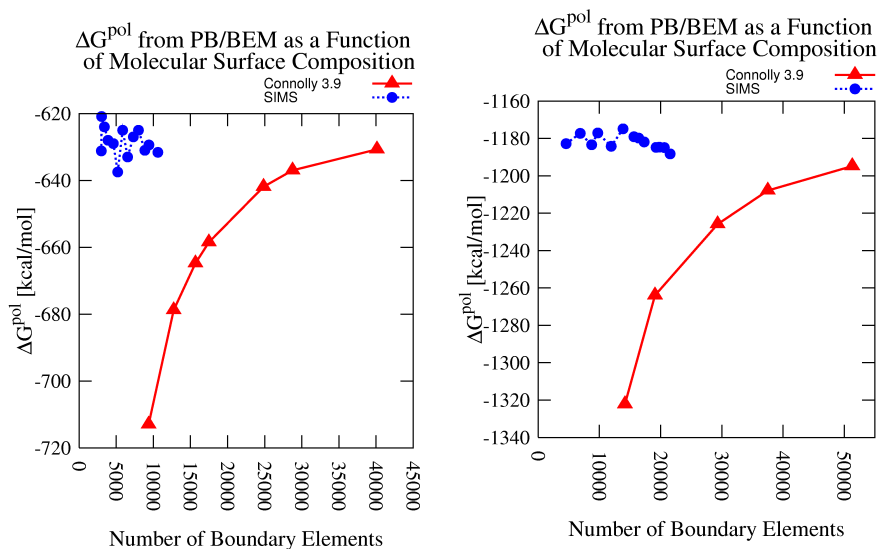


Figure 1. Variation of PB/BEM results as a function of molecular surface composition. Left panel: pdb code 1OTF, 62 residues; Right panel: pdb code 1WN2, 121 residues; Connolly's MSROLL program (red triangles) as well as the SIMS program (blue spheres) lead to the same PB/BEM result upon convergence. The SIMS program, however, does so with smaller numbers of BEs to consider.

3 Results and Conclusions

A diverse ensemble of different pdb structures is examined for sensitivity of PB/BEM calculations with respect to molecular surface composition. As an example two typical cases are shown in Figure 1 (pdb codes 1OTF and 1WN2 with 62 and 121 residues respectively).

The general trend is similar in either case. Connolly's MSROLL program (red triangles) reaches a plateau value in a continuous mode, while the SIMS program (blue spheres) approaches converged PB/BEM results in an alternating fashion when the surface resolution is steadily increased. Both programs lead to the same net PB/BEM result, hence the description should be of equal quality. Considerable dependence on this critical aspect is demonstrated clearly (± 100 kcal/mol uncertainty when working off the convergence domain). The SIMS program appears to reach convergence much faster than the Connolly program, which would have important implications on computational overall performance. The required average size of boundary elements for achieving sufficient levels of numerical accuracy lies in the range of $0.4 \text{ \AA}^2$ or below, hence is comparable to recommended values from Quantum Chemistry.

Acknowledgments

This work was supported in part by the National Institutes of Health Grant GM62838.

References

1. J. Warwicker, H. C. Watson, *Calculation of the electric potential in the active site cleft due to α -helix dipoles*, J. Mol. Biol. **157**, 671–679 (1982).
2. R. J. Zauhar, R. S. Morgan, *A new method for computing the macromolecular electric potential*, J. Mol. Biol. **186**, 815–820 (1985).
3. *The PCM model*, European Summerschool in Quantum Chemistry, Book II, Björn O. Roos and Per-Olof Widmark, Eds **561–597**, 2000 (.)
4. M. L. Connolly, *Computation of molecular volume*, J. Am. Chem. Soc. **107**, 1118–(1985).
5. Y. N. Vorobjev, J. Hermans, *SIMS: computation of a smooth invariant molecular surface*, Biophys. J. **73**, 722–732 (1997).
6. S. Höfinger, *Solving the Poisson Boltzmann Equation with the Specialized Computer Chip MD-GRAPE-2*, J. Comp. Chem. **26**, 1148–1154 (2005).
7. http://mbs.cbrc.jp/pdbreprdb/cgi/reprdb_menu.pl all default criterions except **Number of residues less than 40 excluded – NO, Include MUTANT – NO, Exclude COMPLEX, Exclude FRAGMENT, Include NMR – NO, Include Membrane Proteins – NO**
8. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*, J. Am. Chem. Soc. **117**, 5179–5197 (1995).

Generalized Born and Nonpolar Models in Solvation Free Energies.

Piotr Kmiec^{1,2}, Jaroslaw Kalinowski², and Bogdan Lesyng²

¹ Interdisciplinary Centre for Mathematical
and Computational Modelling,
Zwirki i Wigury 93, 02-089 Warsaw, Poland
E-mail: {pkmiec}@icm.edu.pl

² Department of Biophysics, Warsaw University,
Zwirki i Wigury 93, 02-089 Warsaw, Poland
E-mail: {pkmiec, jkalin, lesyng}@icm.edu.pl

Studies of biomolecular systems at a mesoscopic level require knowledge of effective, mesoscopic interaction potentials. Such potentials are being developed, accounting for continuum models of solvent. One of the approaches is a mean-field electrostatic Generalized Born (GB) model. For review of mesoscopic models and theories, see¹. The GB model and methods of computing GB radii are briefly presented, together with an outline of an SCC-DFTB method and CM3 SCC-DFTB charges². These charges were specifically designed to provide high quality description of electrostatic properties on a basis of the approximate density functional theory calculations. Differences between the experimental solvation energies and the GB electrostatic energies to the solvation energies give nonpolar (hydrophobic) contributions. The hydrophobic energies were fitted to solvent accessible surface area as well as to a polynomial series depending of the Born radii. Both approaches give satisfactory results, however, the latter is faster. Our model was tested using a set of experimental hydration energies for simple molecules³. Analytical approximations of the nonpolar contribution to the mesoscopic interaction potentials (hydrophobic contribution to the free energy of solvation) are proposed.

1 Introduction

The purpose of our study is to design fast, mesoscopic molecular interaction potentials, which are capable of reproducing solvation free energies. The electrostatic part of the free energy of solvation is approximated with the Generalized Born (GB) model using the CM3/SCC-DFTB atomic charges² as the source of the original electrostatic field.

2 Class IV charges

In brief, CM3/SCC-DFTB charges are given by the formula:

$$q_k = q_k^0 + \sum_{k' \neq k} T_{kk'}(B_{kk'}) = q_k^0 + \sum_{k' \neq k} D_{Z_k Z_{k'}} B_{kk'} + C_{Z_k Z_{k'}} (B_{kk'})^2 \quad (1)$$

where q_k^0 are typical Mulliken charges, and $T_{kk'}$ is a quadratic function of Mayer's bond orders $B_{kk'}$. The method for calculating CM3/SCC-DFTB charges was described in detail elsewhere².

3 Free energy of solvation

The solvation free energy G_{solv} can be decomposed into an effective electrostatic energy G_B , a nonpolar (hydrophobic) part G_{np} and a cross term G_{cross} , which typically is neglected. We use the Generalized Born Model (GB)⁴ to represent the electrostatic part.

$$G_B = -\frac{1}{2} \left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{ex}} \right) \sum_{k,l} \frac{q_k q_l}{\sqrt{r_{kl}^2 + R_k R_l e^{-\frac{r_{kl}^2}{4R_k R_l}}}} \quad (2)$$

where r_{kl} is the distance between atoms, ϵ_{in} , ϵ_{ex} are the dielectric constants of interior and exterior of the molecule, respectively. q_k , q_l are the atomic charges. R_k is an effective Born radius. The exact electrostatic energy, U_{single}^k , of a model molecule containing a single charge on the atom k is:

$$U_{single}^k = -\frac{1}{2} \left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{ex}} \right) \frac{q_k^2}{R_k} \quad (3)$$

A nonpolar contribution to the solvation free energy, G_{np} , is usually approximated using solvent accessible surface area (SASA). One of such methods, that bears similarity to our approach, is a Surface Generalized Born – Nonpolar model⁵, in which the nonpolar component (G_{np}) is described as a polynomial of A_i , where A_i is the SASA of the atom i .

$$G_{np} = \sum_{i=1}^N (\gamma(k_i) A_i + \alpha(k_i)) \quad (4)$$

$\gamma(k_i)$ parameters are interpreted as a surface tension coefficients, and $\alpha(k_i)$ are adjustable parameters dependent on atom types (k_i).

4 Numerical implementation and results

Born radius of an atom in a molecule contains a qualitative information of how deeply the atom is buried inside the molecule. Taking advantage of this fact, in our approach we use polynomials the inverse of R_k , instead of using A_i . Coefficients of these polynomials depend on the atom types Z_k .

$$G_{np} = \sum_k \left(a_{1,Z_k} \frac{1}{R_k} + a_{0,Z_k} \right) \quad (5)$$

Since the cross term is neglected in our model, $G_{np} = G_{solv} - G_B = G_{exp} - G_B$, an estimate of the nonpolar contributions to the free energy of solvation G_{solv} can be computed by subtracting the electrostatic contributions from the experimental values of the free energy of solvation. These approximate values of the nonpolar terms were used for parameterization of our model. It was carried out using a set of more than one hundred molecules selected from³. For each molecule CM3/SCC-DFTB charges were computed and used to estimate the GB energies. At the same time, electrostatic contribution was calculated by solving the Poisson-Boltzmann equation, and removing the Coulomb part of the energy. Accuracy of the GB approximation was validated by comparison with the PB results, see Figure 1. Coefficients in equation 5 were chosen to yield an optimal fit. Resulting nonpolar energies against the experimental estimates are presented in Figure 2.

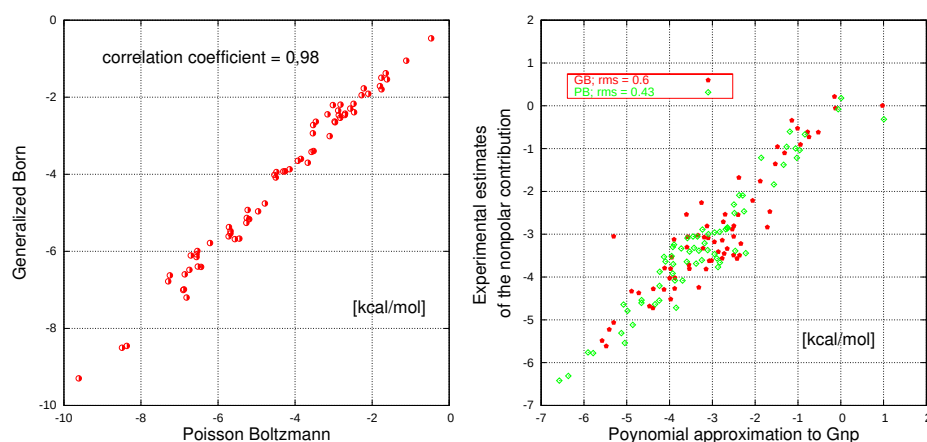


Figure 1. Electrostatic contribution to the free energy of solvation (left). Nonpolar contribution to the free energy of solvation (right).

5 Concluding remarks

Fitting the nonpolar term G_{np} to reproduce the free energies of solvation was the main goal of this work. We proposed a fast and simple model. Our results suggest it is sufficient to account only for the first order polynomials.

Acknowledgments

These studies were supported by the European CoE MAMBA and National CoE BioExploratorium.

References

1. Gruziel M., Kmiec P., Trylska J., Lesyng B., *Selected Microscopic and Mezoscopic Modelling Tools and Models - an Overview* in, Challenges and Advances in Chemistry and Physics, J. Leszczynski, series editor, Springer, in press.
2. J. Kalinowski, B. Lesyng, J.D. Thompson, C. Cramer, D.G. Truhlar, *Class IV Charge Model for the Self-Consistent Charge Density-Functional Tight-Binding Model*, J. Phys. Chem A **108**, 2545–2549 (2004).
3. Jason D. Thompson, Gregory D. Hawkins, Candee C. Chambers, David G. Giesen, Paul Winget, Christopher J. Cramer, and Donald G. Truhlar, Minnesota Solvation Database (version 2003.1).
4. Onufriew, A.; Bashford, D.; Case, D. A. *A Modification of the Generalized Born Model Suitable for Macromolecules*, J. Phys. Chem. B **104**, 3712–3720 (2000).
5. Gallicchio, E., Zhang, L.Y. and Levy, R.M. *The SGB/NP model for the solvent potential of mean force in water based on the surface generalized born model and a non-polar hydration free energy estimator* J. Comput. Chem. **23**, 517–529 (2002).

Determination of Short-Range Potentials for Physics-Based Protein-Structure Prediction

Urszula Kozłowska^{1,2}, Katarzyna Wachucik², Adam Liwo^{1,2}, and Harold A. Scheraga¹

¹ Baker Laboratory of Chemistry and Chemical Biology,
Cornell University, Ithaca, NY 14853-1301, U.S.A.,
E-mail: has5@cornell.edu

² Faculty of Chemistry, University of Gdańsk,
Sobieskiego 18, 80-952 Gdańsk, Poland

By integrating *ab initio* energy surfaces of the model terminally-blocked amino-acid residues: glycine, alanine, and proline, we calculated the potentials of mean force corresponding to the bending of the $C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond angles for the purpose of using them in our united-residue UNRES force field. The potentials for the glycine and alanine residues were found to be bimodal as are the corresponding statistical potentials determined from the Protein Data Bank.

1 Introduction

Local interactions play a very significant role in determining protein structure, because they determine the geometry of secondary-structure elements (e.g., the α -helices and the β -sheets), as well as that of turns and loop regions. Therefore, accurate representation of local-interaction terms in the empirical force fields for physics-based protein-structure prediction is of utmost importance. In this communication, we present the determination of the potentials of mean forces for the bending of the $C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond angles for the purpose of using them in our united-residue physics-based UNRES force field¹⁻⁴ for protein-structure prediction from *ab initio* energy surfaces of terminally-blocked amino-acid residues, which will replace the statistical potentials determined in our earlier work² from PDB statistics.

2 Methods

The model system used to calculate the potentials of mean force corresponding to the bending of virtual-bond angles θ is shown in Figure 1. The neighboring amino-acid residues are included because, from the PDB statistics, it follows² that the distribution of virtual-bond angles θ depends on the values of the neighboring virtual-bond dihedral angles γ and, consequently, the bending potentials in UNRES are functions of these angles.

The potentials of mean force of virtual-bond angle bending, $F_{XYZ}(\theta, \gamma_1, \gamma_2)$ (where X, Y, and Z are the types of residue involved; see Figure 1), were computed in this work from the energy maps of terminally-blocked X, Y, and Z residues based on the theory presented in Ref. 5 extended by applying the harmonic approximation to each point of the map of the central residue Y. We used the energy maps of terminally-blocked Gly, Ala,

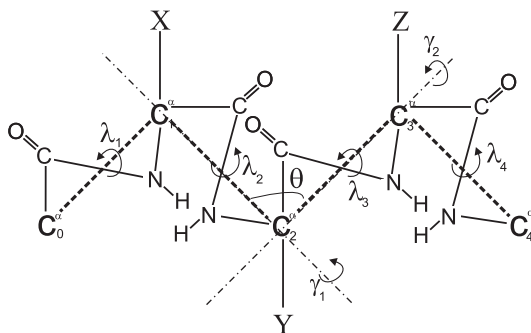


Figure 1. Illustration of a model system for the calculation of the potentials of mean force of the bending of the virtual-bond angles θ . The variables to integrate over are the torsional angles of rotation, $\lambda_1 - \lambda_4$ of the peptide groups about the $C^\alpha \cdots C^\alpha$ virtual-bond axes, while the virtual-bond angle θ and the virtual-bond dihedral angles γ_1 and γ_2 are the primary variables

and Pro calculated in Ref. 5 at the *ab initio* MP2/6-31G** level with the grid of 15° ; each point of a map corresponds to a structure minimized in all degrees of freedom except the angles λ . In this work, we computed the Hessian matrices at each point of a map and used them together with minimum-energy values to compute $F_{XYZ}(\theta, \gamma_1, \gamma_2)$. All potentials of mean force were computed at $T=298^\circ\text{K}$.

3 Results and Discussion

The plots of $\overline{F}_Y(\theta)$ (Boltzmann-averaged over the angles γ_1 and γ_2 for $X=\text{Gly}$, Ala, and Pro) are shown in Figure 2. In this Figure, the statistical potential $\overline{F}_{Ala}(\theta)$ determined in Ref. 2 is also presented for comparison.

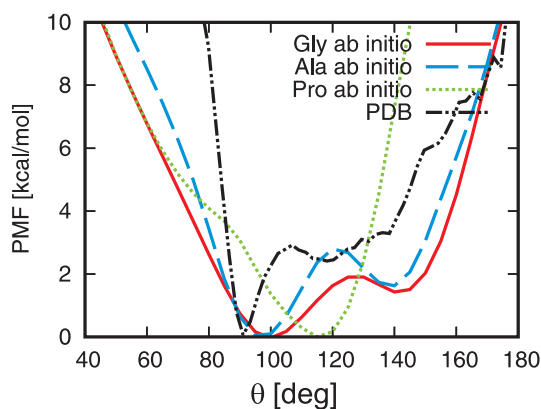


Figure 2. Potentials of mean force of virtual-bond-angle bending Boltzmann-averaged over the virtual-bond-dihedral-angles γ_1 and γ_2 compared with the statistical potential computed from the PDB statistics for Ala-type virtual-bond angles.

It can be seen from Figure 2 that the $\overline{F}_{Gly}(\theta)$ and $\overline{F}_{Ala}(\theta)$ potentials have two minima, one for $\theta = 100^\circ$ and the second one for $\theta = 140^\circ$. Analysis of the complete $F_{XYZ}(\theta, \gamma_1, \gamma_2)$ potentials shows that the first minimum corresponds to the α -helical and the second one to the extended region. This finding agrees with the analysis of the statistical potentials from the PDB². However, the second minimum is not pronounced in the statistical potentials which is caused by the fact that most of the data pertained to α -helical structures. The present potentials are not biased to any organized structure and are, therefore, expected to improve the performance of the UNRES force field.

Acknowledgments

Support from the National Institutes of Health (grant GM-14312), the National Science Foundation (grant MCB00-03722), the NIH Fogarty International Center (grant TW7193), the Polish Ministry of Education and Science (grant 3 T09A 134 27) as well as use of supercomputing resources at the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center, the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw is gratefully acknowledged.

References

1. A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J. Comput. Chem.*, 18:849–873, 1997.
2. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Ołdziej, and H. A. Scheraga. A united-residue force field for off-lattice protein-structure simulations. II: Parameterization of local interactions and determination of the weights of energy terms by Z-score optimization. *J. Comput. Chem.*, 18:874–887, 1997.
3. A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.*, 115:2323–2347, 2001.
4. S. Ołdziej, A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga. Optimization of the unres force field by hierarchical design of the potential-energy landscape: 2. Off-lattice tests of the method with single proteins. *J. Phys. Chem. B*, 108:16934–16949, 2004.
5. S. Ołdziej, U. Kozłowska, A. Liwo, and H. A. Scheraga. Determination of the potentials of mean force for rotation about $C^\alpha \cdots C^\alpha$ virtual bonds in polypeptides from the *ab initio* energy surfaces of terminally-blocked glycine, alanine, and proline. *J. Phys. Chem. A*, 107:8035–8046, 2003.

Large Scale Molecular Dynamics Simulations for the Derivation of Solvation Free Energies of Strongly Hydrophobic Molecules

Reema Mahajan¹, Dieter Kranzlmüller², Jens Volkert², and Siegfried Höfner³

¹ Indian Institute of Technology,
Department of Chemical Engineering,
Hauz Khas, New Delhi-16, India
E-mail: reema.mahajan@gmail.com

² Johannes Kepler University of Linz,
Institute of Graphics and Parallel Processing
Altenberger Straße 69, A-4040, Linz, Austria
E-mail: {kranzlmuller,volkert}@gup.uni-linz.ac.at

³ Michigan Technological University,
Department of Physics,
1400 Townsend Drive, Houghton, MI 49931, USA
E-mail: shoefing@mtu.edu

Liquid water is believed to be of crucial biological importance when it comes to solvating hydrophobic molecules. The so-called hydrophobic effect is a key-determinant in the protein folding problem and a deeper understanding of it would certainly be of great value to the biophysics community. The process of solvating hydrophobic molecules in water may be studied with the help of computer simulations that establish an atomistic picture of the liquid state of water. A study has already been carried out in the past using Molecular Dynamics simulations and a polarizable force field description for the water molecules. The focus of that study was on the temperature dependence of the free energy change observed for the onset and growth of a repulsive sphere that represented the hydrophobic species. The obtained data was in good agreement to predictions from the Lum Chandler Weeks theory on the unit volume entropy. However, only small-sized hydrophobic solutes could be studied and the critical domain where LCW theory claims a change in underlying physics could not be examined. Therefore an extension of this previous study is currently underway using again a high level polarizable force field and a largely extended water box of dimension 60x60x60 Å. The technical complexity of such an approach is discussed and an outlook is given on the various implications of such an investigation.

1 Introduction

About 50 years have passed by since Walter Kauzmann has published his extensive review on thermodynamic principles of protein unfolding¹. One of his conclusions has been that the so-called “Hydrophobic Bonds” would constitute the major stabilizing factor for protein assembly and protein stability. It is however till today that the actual kernel of hydrophobicity must still be considered an unsolved problem. Among other theoretical attempts Lum Chandler Weeks (LCW) theory of hydrophobicity has become fashionable. LCW theory divides small length scale hydrophobicity from large length scale hydrophobicity formally (volume dependence versus surface area dependence) as well as in terms of underlying physics (entropy drop versus entropy gain). The switch between these two

domains is said to occur at dimensions of $r=10 \text{ \AA}$ of spherical model solutes. Since an unbiased experimental verification of such a picture of hydrophobicity is extremely difficult it appears to be only natural to address the underlying physics behind LCW theory with computational approaches.

2 Methods

Suitable computational techniques for this kind of problem are free energy calculations, which have been introduced and methodically refined to directly study ΔG trends for various physico-chemical processes². The obvious physical process to study here was the ΔG corresponding to the introduction, onset and growth of an artificial repulsive sphere located in the center of a simulation cell otherwise filled with water molecules. The spherical repulsive cavity would then represent a model hydrophobic solute and the associated free energy change was the energy necessary to embed this hydrophobic solute in water. Exactly the same type of ΔG has also become known as the cavitation free energy³.

A previous study has already been carried out employing the high level polarizable AMOEBA force field⁴ for computer simulation of liquid water at seven individual temperatures in the range of 277K to 370K for calculating the free energy cost of creation of small-sized cavities (work currently under review). Ewald summation had been applied within the Molecular Dynamics simulations (MD) that form the basis of the Free Energy Perturbation (FEP) calculations in the framework of the Overlapping Spheres Technique (OST)⁵. Our previous study could reconfirm the Chandler picture of the unit volume entropy ΔS for small-sized hydrophobic solutes (from the temperature profiles of ΔG). The computational demand of this study was on the order of three CPU years (Itanium 1.4 GHz). Volumes of up to $r \approx 3 \text{ \AA}$ could be introduced, hence the small cavity regime could be examined.

In extension to this previous study our present efforts focus on largely increased volumes of up to $r=10 \text{ \AA}$ to probe the critical domain where the switch in physical exemplification is thought to occur. We therefore need to increase our water box substantially and start out with a trial system of dimension $60 \text{ \AA} \times 60 \text{ \AA} \times 60 \text{ \AA}$. The CPU requirements of such an approach must be tremendous. In this current work we estimate the required computational resource allocation of such a large scale attempt and try to find a suitable compromise in case the resource allocation turns out to be simply un-feasible.

3 Results and Conclusions

The estimated computational resource allocation for the case of a MD/FEP study based on AMOEBA water boxes of dimension $60 \text{ \AA} \times 60 \text{ \AA} \times 60 \text{ \AA}$ is summarized in Table 1 (middle column). Compared to the previously required computing resources (see Table 1, left column) this is however far beyond any acceptable limit. We therefore shrank the size of the water boxes layer by layer and estimated resulting overall resource allocations for various constellations. A cubic arrangement of about 1331 water molecules (see Table 1, right column) was finally chosen for production runs because it represents a good compromise between intended problem size and acceptable resource allocation. Such large scale computations can be conveniently carried out on large GRID systems in reasonable time lines.

	Previously (B ≤ 3 Å)	Estimate (B ≈ 10 Å)	Estimate (B ≈ 8 Å)
# Water Molecules	216	8000	1331
Box Size	(18.6Å) ³	(60Å) ³	(34.2Å) ³
Ewald, Cut-Off Radius	14 Å	14 Å	14 Å
Time Step	1 fs	1 fs	1 fs
ΔB in OST	± 25 x 0.0032 Å	± 25 x 0.0032 Å	± 25 x 0.0032 Å
# Intervals in OST per Å	12.5	12.5	12.5
Maximum Allowed Perturbation	2 RT	2 RT	2 RT
Annealing Period	2 ps	2 ps	2 ps
# Temperatures Considered	7	3	3
Intended MD/FEP Period	100 ps	100 ps	100 ps
# Trajectories per Temperature	29	120	90
Estimated CPU Time per Trajectory (Itanium 2, 1.4 Ghz)	5 days	5 years	2 months
# Trajectories Total	203	360	270
Estimated Total CPU Time	2.78 years	1800 years	45 years

Table 1. Experienced and estimated resource allocation for large scale MD/FEP simulations on cavitation free energy calculations in AMOEBA water. Previously consumed resources are shown in the column to the left, the middle column gives an estimate for a huge water box aimed at studies of cavities of radii up to B ≈ 10 Å, and the right column shows estimates for a largely increased water box for cavitation studies of radii up to B ≈ 8 Å. This final column to the right represents a good compromise between intended study size and feasible computational throughput.

Acknowledgments

This work was supported in part by the National Institutes of Health Grant GM62838.

References

1. W. Kauzmann, *Some factors in the interpretation of protein denaturation*, Adv. Protein Chem. **14**, 1–63 (1959).

2. T. Simonson, G. Archontis, M. Karplus, *Protein-Ligand Recognition: Free Energy Simulations Come of Age*, *Acc. Chem. Res.* **35**, 430–437 (2002).
3. J. Tomasi, B. Mennucci, R. Cammi, *Quantum Mechanical Continuum Solvation Models*, *Chem. Rev.* **105**, 2999–3094 (2005).
4. P. Ren, J. W. Ponder, *Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation*, *J. Phys. Chem. B* **107**, 5933–5947 (2003).
5. P. M. Postma, H. J. C. Berendsen, J. R. Haak, *Thermodynamics of cavity formation in water*, *Farad. Symp. Chem. Soc.* **17**, 55–67 (1982).

Simulation of Thick Polymers

T. Neuhaus¹, O. Zimmermann¹, and
Ulrich H. E. Hansmann^{1,2}

¹ John von Neumann Institute for Computing,
Research Center Jülich, 52425 Jülich, Germany
E-mail: {t.neuhaus, o.zimmermann}@fz-juelich.de

² Department of Physics, Michigan Technological University,
1400 Townsend Drive, Houghton, MI 49931-1295, USA
E-mail: hansmann@mtu.edu

We simulate $3d$ flexible off-lattice ring polymers of length L up to $L = 4000$ for various values of the global radius of curvature $R_{grc} = 0.25, 0.48, 1.0$ and $R_{grc} = 2.0$. We utilize two different ensembles: one with a δ -function constraint on the radius, and another one with a θ -function. For both cases global radius of curvature provides a valid regularisation of polymers with thickness $D = 2R_{grc}$. The Flory type critical exponent ν_{SAW} of self avoiding non-interacting rings at $D = 2$ is determined to be $\nu_{SAW} = 0.5869(5)$ from the radii of gyration chain length L -scaling, while other D -values produce consistent results. We also study low temperature configurations of marginally compact Lennard Jones homo-polymers on a ring and identify some conformational simplices.

1 Introduction

The mechanism by that proteins fold into their biologically active form is despite decades of research still not resolved. Most of the ongoing work focuses on the sequence specificity of the final structure, i.e. emphasizes the differences of proteins to polymers. For instance, in the funnel picture¹ it is assumed, that the energy landscape of a functional protein resembles a funnel leading to the biologically active structure, while a random chain leads to the rough energy landscape of spin glasses. Only recently, a different approach has been considered. Banavar et. al.²⁻⁴ have claimed that ground state conformations of *short* and *thick* homo-polymers exhibit only a finite set of conformational simplices i.e., helices, sheets and some others, which are independent from the actual form of the employed chain interactions. In light of the claims by Banavar et. al. it seems to be useful to study in more detail the physics of thick polymers.

2 Models

A natural starting point are homo-polymers i.e., $3d$ off-lattice chains with a constant finite thickness D : tubes. We will consider non-interacting flexible tubes - without stiffness -, and in a more ambitious approach: thick homo-polymers with an attractive interaction on the chain. Tubes are regularised with the help of a new quantity: the *global radius of curvature*⁵. The partition function of thick polymers is denoted $Z(D)$ or $Z(R_0)$ with $D = 2R_0$ and is obtained upon inserting a θ -function constraint

$$\sum_{conf} 1 \rightarrow \sum_{conf} \theta(R_{grc} - R_0) \quad (1)$$

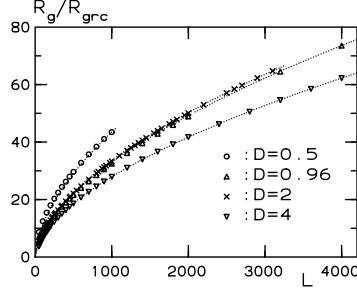


Figure 1. All $R_g := \sqrt{R_g^2}$ data of eq.(4) in units of R_{grc} at $R_{grc} = 0.25, 0.48, 1.0$ and $R_{grc} = 2.0$ with $D = 2R_{grc}$ as a function of L . The scaling curves use $\nu_{SAW} = 0.5874$ and are adjusted to match the data points at largest L . Around $D \approx 1 - 2$ the polymer swelling just is $\text{const} D$.

into the partition function. For the interacting Lennard Jones homo-polymers we have

$$Z_{LJ}(R_0) = \sum_{rings} \delta(R_{grc} - R_0) e^{-\frac{1}{T} \sum_{i < j+1} V_{LJ}(r_{ij})} \quad (2)$$

where T is the temperature. The Lennard Jones potential has the form

$$V_{LJ}(r_{ij}) = 4 \left(\left(\frac{1.6}{2^{1/6} r_{ij}} \right)^{12} - \left(\frac{1.6}{2^{1/6} r_{ij}} \right)^6 \right). \quad (3)$$

It has a minimum value of -1 . at $r_{min} = 1.6$.

3 Scaling of Thick Ring Polymers

In the Monte Carlo simulation we measure the mean radius of gyration R_g squared, which is defined as

$$R_g^2 = \frac{1}{L} \left\langle \sum_i (\vec{x}_i - \vec{x}_{cms})^2 \right\rangle. \quad (4)$$

We display in Fig.1 all of our R_g simulation data in units of R_{grc} . The curves in the plot correspond to the finite length scaling forms $R_g \propto L^{\nu_{SAW}}$ with a fixed value $\nu_{SAW} = 0.5874$. None of the data sets at either value of R_{grc} is consistent with a Gaussian behavior at $\nu = 0.5$. A detailed analysis of effective exponents yields

$$\nu_{SAW} = 0.5869(5), \quad (5)$$

which puts non-interacting tubes into the universality class of self-avoiding walks.

4 Interacting LJ Homopolymers

We also explore the properties of near ground state conformations of interacting flexible thick tubes at a value $R_{grc} = 1.4$. Here we display in Fig.2 near ground state configurations for interacting Lennard Jones ring homo-polymers at length $L = 95$.



Figure 2. Near ground state conformations of a single Lennard Jones ring homo-polymer from $T = 0.1$ simulations at $R_{grc} = 1.4$ for $L = 95$. The left top corner (at position 1) displays the complete ring, while as counted from left to right , and from top to bottom, positions 2, 3, 4 and 5 display sub-structures of the same ring.

5 Concluding Remarks

There exists the conjecture, that ground state - or near ground state - chain configurations of short thick tubes with attractive homo-polymer interactions would fold only into a finite set of conformational simplices. The degree of universality in these structures is currently under investigation.

Acknowledgments

The calculation were done on computers of the Neumann Institute for Computing, Forschungszentrum Jülich, Jülich, Germany. This work was supported under NSF grant (CHE-0313618).

References

1. P. E. Leopold, M. Montal and J. N. Onuchic, (1992) Proc. Natl. Acad. Sci. USA 89, 8721-8725.
2. J. R. Banavar and A. Maritan, Rev. of Mod. Phys. **75** (2003) 23.
3. J. R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan and A. Trovato, J. Phys.: Cond. Matt. **15**:1787 (2003).
4. J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno and A. Trovato, Phys. Rev. E **70**:041905 (2004).
5. O. Gonzalez and H. Maddocks, PNAS **96**:4769 (1999).

Symplectic Algorithms of Molecular Dynamics. An Application to the United-Residue Model.

Franciszek Rakowski^{1,3}, Paweł Grochowski^{1,2}, and Bogdan Lesyng² Adam Liwo⁴
Harold A. Scheraga⁵

¹ Interdisciplinary Center for Mathematical and Computational Modelling (ICM),
Pawińskiego 5a, 02-106 Warsaw, Poland
E-mail: rakowski@icm.edu.pl

² Department of Biophysics, Faculty of Physics, Warsaw University,
Żwirki i Wigury 93, 02-089 Warsaw, Poland

³ Faculty of Physics, Warsaw University of Technology,
Pl. Politechniki 1, 00-661 Warsaw, Poland

⁴ Faculty of Chemistry, University of Gdańsk,
Sobieskiego Str. 18, 80-952 Gdańsk, Poland

⁵ Baker Lab. of Chemistry and Chem. Biology,
Cornell University, Ithaca, New York 14853-1301.

In order to describe the dynamics of a molecular system, governed by an effective Hamiltonian based on a coarse-grain United-Residue (UNRES) potential energy model¹, multiple time-step (MTS) symplectic algorithms have been developed and implemented. The theoretical background of symplectic flows generated by the Hamiltonian equations of motions is outlined and the principles of symplectic integrators are presented. Decomposition of the UNRES forces was proposed, and an optimal MTS scheme for real molecular systems has been implemented. This algorithmic approach appears to be optimal for the UNRES model, which results in a noticeable speed up of computations. Taking advantage of the reduced model and the MTS algorithmic approach, one can carry out long-time MD simulations. The results are being prepared for publication⁵.

1 Hamiltonian Systems

Among many ordinary differential equation (ODE) systems, Hamiltonian systems are of particular importance in physics. The hamiltonian equations of motion are:

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i} \quad (1)$$

The phase-flow given by such a system conserves the Hamiltonian function (energy). In general, the Hamiltonian systems conserve the so-called symplectic structure (the symplectic structure is defined as a quadratic differential form, providing $d\omega^2 = 0^2$), therefore a series of values expressed by Poincaré integrals, e.g.:

$$\oint_{C(0)} \sum_i p_i dq_i = \oint_{C(t)} \sum_i p_i dq_i, \quad (2)$$

where $C(t)$ is an arbitrary closed curve in the phase space evolving from the initial state $C(0)$ according to the flow.

2 Integration of the Hamiltonian Systems.

An algorithm for integration of the Hamiltonian systems⁴ should conserve the symplectic structure with the phase-flow. It also conserves a modified, approximate Hamiltonian: $H^S = H + \tilde{H}$ which is located near the original one. In order to derive the symplectic algorithms one can use the Liouville operator (Poisson bracket) acting on the state vector to denote the Hamilton equations:

$$\dot{\Gamma} = \hat{L} \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}, \quad (3)$$

with the formal solution for a time interval Δt :

$$\Gamma(t + \Delta t) = e^{\Delta t \hat{L}} \Gamma(t). \quad (4)$$

Application of the Trotter factorization scheme and time discretisation give the recipe for the Velocity Verlet algorithm:

$$\Gamma(t + \Delta t) \approx e^{\frac{\Delta t}{2} \hat{L}_U} e^{\Delta t \hat{L}_T} e^{\frac{\Delta t}{2} \hat{L}_U} \Gamma(t), \quad (5)$$

where the $\hat{L} = \hat{L}_T + \hat{L}_U$ is decomposed into parts related to the kinetic energy and potential energy, respectively. The symplecticity of such an algorithm depends on the time-step Δt

3 Multiple Time-Step Algorithms.

In order to achieve the Multiple Time-Step (MTS) scheme for integrating the Hamilton equations of motion it is necessary to decompose the potential part of Liouville operator, into two parts: one generates the fast varying forces and the other one the slow varying forces:

$$\hat{L}_V = \hat{L}_{Vf} + \hat{L}_{Vs}. \quad (6)$$

Now we can construct the scheme which gives the family of MTS algorithms, each for different splitting number N :

$$\Gamma(t + \Delta t) \approx e^{\frac{\Delta t}{2} \hat{L}_{Vs}} \left(e^{\frac{\Delta t}{2N} \hat{L}_{Vf}} e^{\frac{\Delta t}{N} \hat{L}_T} e^{\frac{\Delta t}{2N} \hat{L}_{Vf}} \right)^N e^{\frac{\Delta t}{2} \hat{L}_{Vs}} \Gamma(t). \quad (7)$$

Now the symplecticity of a given algorithm depends on the splitting number N and time-step Δt .

4 Application of the MTS Algorithm to UNRES Model.

The United-Residue force field model describes a polypeptide chain in a reduced and physics-based approach. Residues are represented by α -carbon atoms, which are linked together by virtual bonds - this constitutes the backbone. The representation of the side chains is united, and it is referred to the C^α atoms. The interaction sites are located in the midpoints between consecutive C^α (the backbone) and at given distances from C^α atoms (the side chains).

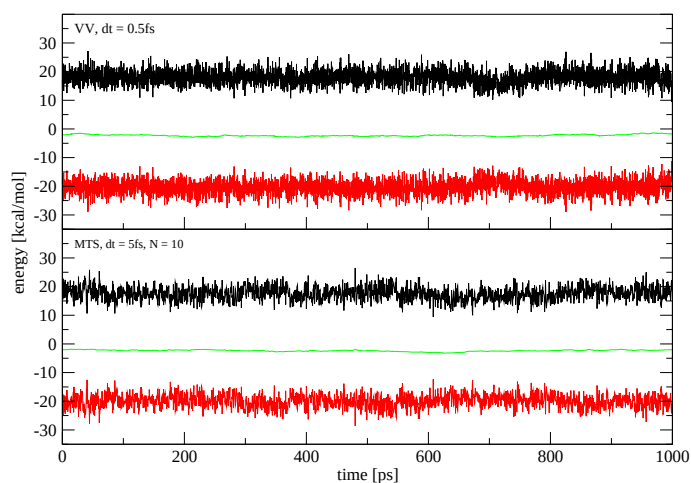


Figure 1. Energies obtained using the Velocity Verlet and MTS algorithms. The potential energy - lower line, total energy - middle line and kinetic energy - upper line, on each graph. The simulation was carried out for the Ala_{10} .

The UNRES potential function was decomposed into parts generating slow and fast varying forces. Based on this, the MTS algorithm was implemented^{3,5} and tested with several integration schemes - different pairs $(N, \Delta t)$, see Fig. 1, for example.

The Velocity Verlet algorithm and MTS algorithm gives the same stability of the MD run. The computational cost of performing equivalent MD runs is much lower when using the MTS algorithm, which is shown in Table 1:

Table 1. Values for the maximum amplitude of the energy fluctuation from the Ala_{10} 1 nano second MD run. Time step units are in mtu.(1mtu = 48.9fs), and energy in kcal/mol.

Velocity Verlet			MTS			
Δt	$ \delta E $	Exec. time	Δt	N	$ \delta E $	Exec. time
0.01	1.50	79.8 min	0.1	10	1.42	9.23 min

Acknowledgments

This work was supported by the Center of Excellence for Multi-scale Biomolecular Modelling, Bioinformatics and Applications (MAMBA) and National CoE BioExploratorium.

References

1. A. Liwo, St. Oldziej, M.R. Pincus, R.J. Wawak, S. Rackovsky and H.A. Scheraga *A United-residue Force-field for Off-lattice Protein structure Simulations. I: Functional Forms and Parameters of Long-range Side-chain Interaction Potentials from Protein Crystallographic Data*. J. Comput. Chem. **18**, 849-873 (1997)
2. V. I. Arnold *Mathematical Methods of Classical Mechanics*. Springer, New York, 1995.
3. Khalili M., Liwo A., Rakowski F., Grochowski P. and Scheraga H.A. *Molecular Dynamics with the United-Residue Model of Polypeptide Chains. I. Lagrange Equations of Motion and Tests of Numerical Stability in the Microcanonical Mode* J.Phys.Chem.B **109**, 13785–13797 (2005).
4. J. M. Sanz-Serna *Symplectic Integrators for Hamiltonian Problems: an Overview* Acta Numerica **1**, 243-286 (1992)
5. F. Rakowski, P. Grochowski, B. Lesyng, A. Liwo and H.A. Scheraga *Implementation of a Symplectic Multiple Time-step Molecular Dynamics Algorithm Based on the UNRES Mesoscopic Potential Energy Function*. to be published.

A Fast Wavelet Based Evaluation of Coulomb Potentials in Molecular Systems

Godehard Sutmann and Silke Wärow[†]

Central Institute for Applied Mathematics and
John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: g.sutmann@fz-juelich.de

1 Introduction

Long range interactions often play an important role in complex molecular systems. Due to charged patterns or dipolar structures, molecules interact via Coulomb interactions, thereby influencing dynamical processes or static structures. Modeling charged or polar systems on a microscopic level is often limited by the number of system constituents, since calculation of long range interactions are computationally very demanding, as they scale like $\mathcal{O}(N^2)$ since all interactions between particle pairs have to be considered explicitly. However, approximating infinitely large systems by relatively small ones and imposing periodic boundary conditions at the same time, results in evaluating lattice sums, which usually converge conditionally. Splitting of lattice sums and applying certain controllable approximations leads to handy expressions which, however, still scale like $\mathcal{O}(N^{3/2})$.

Due to this computational burden, research was focused on developing methods, which overcome the quadratic scaling and reduce it to an optimal $\mathcal{O}(N)$, or at least an $\mathcal{O}(N \log N)$ behavior. In general the proposed methods can be classified into mesh-free and mesh-based algorithms, both existing for periodic and open system geometries. Mesh-free algorithms, like the Fast Multipole Method (FMM)¹, should in principle be more promising, since the discretization error is avoided. The problem encountered here is however a very evolved and demanding work to implement the method properly. Alternatives may be found in mesh-based methods, where fast methods like multigrid techniques^{2,3} may be applied. All these methods, scaling like $\mathcal{O}(N)$ have in common that they work in a hierarchical way, i.e. collecting different contributions on different length scales or levels of resolution.

In the present article a new mesh-based method is proposed, which makes use of a fast Wavelet transform technique, thereby reducing the computational complexity to $\mathcal{O}(N)$.

2 Calculation of Potential Energies in Particle Systems

The potential at a charges' position with index i is given by

$$\phi(\mathbf{r}_i) = \sum_{j \neq i}^N \frac{q_j}{|\mathbf{r}_i - \mathbf{r}_j|} \quad (1)$$

[†]Present address: INFORM GmbH, Pascalstr.23, D-52076 Aachen, Germany

where N is the number of charges in the system, q_j the charge of particle j and $\mathbf{r}_j$ its position. The interaction energy of particle i with all other particles is then simply given by

$$U_i(\mathbf{r}_i) = q_i \phi(\mathbf{r}_i) \quad (2)$$

Eq. 1 may also be expressed in short hand notation as matrix-vector product

$$\Phi(\{\mathbf{R}\}) = \mathbf{A}(\{\mathbf{R}\})\mathbf{Q} \quad (3)$$

where $\Phi = \{\phi_1, \dots, \phi_N\}$, $\mathbf{Q} = \{q_1, \dots, q_N\}$ and matrix elements are given by $A_{ij} = 1/|\mathbf{r}_i - \mathbf{r}_j|$. The interaction energy could then simply be written as $\mathbf{U} = \text{diag}(\mathbf{Q}\Phi^T)$. The calculation of the potential via Eq. 3 uses a constant charge vector, but the matrix elements change from step to step if the particles move along their trajectories. Therefore this approach would imply to recalculate matrix elements, which is an $\mathcal{O}(N^2)$ operation. Therefore it would be desirable to shift the time dependence from the matrix to the vector, which would imply only an $\mathcal{O}(N)$ operation. This may be achieved by introducing a mesh, onto which the particle charges are distributed and which has a constant grid spacing throughout the time evolution of the system (even this constraint may be relaxed by rescaling of lengths). Therefore, the grid based summation would be

$$\hat{\Phi}(\{\hat{\mathbf{R}}\}) = \hat{\mathbf{A}}(\{\hat{\mathbf{R}}\})\hat{\mathbf{Q}} \quad (4)$$

where ' $\hat{\cdot}$ ' means mesh based quantities. Nevertheless, also with this modification, the complexity is still $\mathcal{O}(N^2)$ due to a dense matrix-vector product. To obtain an efficiency gain, it will therefore be necessary to transform the matrix into a sparse representation.

3 Principles of Wavelet Based Summations

In principle, wavelets are a tool to analyse data on different time and length scales⁴. In contrast to Fourier analysis, signals are developed in a basis, which have local support. Wavelet analysis is one class of multiscale analysis methods, which splits signals into fine and coarse contributions in a hierarchical scheme. If the original signal f is within a subspace V_{-1} of $L^2(\mathbb{R})$, then high- and low-frequency parts may be projected out via

$$f = \mathcal{P}_0 f + \mathcal{Q}_0 f \quad (5)$$

where V_{-1} is now the sum of orthogonal subspaces, containing low- (V_0) and high-frequency (W_0) parts

$$f \in V_{-1} = V_0 \oplus W_0 \quad (6)$$

This procedure may be continued up to a certain level M , where the subspaces W_i are subject to further splittings into high- and low-frequency contributions, thereby leading to the splitting scheme

$$f = \mathcal{P}_M f + \sum_{k=0}^M \mathcal{Q}_k f \quad (7)$$

Wavelet analysis works as a combined low- and high-pass filter, thereby generating coefficients, describing the strength of signal change on a given coarsening level. Small coefficients thus contain only little information to details on a given level. In image processing it

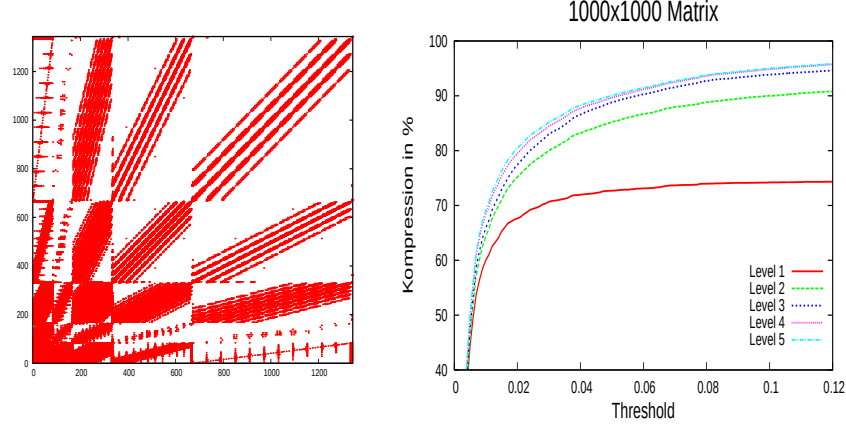


Figure 1. Left: Structure of the inverse distance matrix in wavelet space after thresholding. Right: Compression of original matrix as a function of threshold value x for different levels M of multiresolution analysis (cmp. Eq. 7). In both cases a Daubechies-12 wavelet basis was used⁵.

has been observed that after wavelet transformation of pixel distributions of images, most coefficients are very small. Introducing a threshold value, below of which the absolute value of coefficients is set to zero, the wavelet representation of pictures become sparse. A back-transform using the sparse matrix representation contains still the essential features of the picture.

This principle may also be applied to other matrices, where the number of small coefficients may depend on the symmetry of the matrix. In the present case it is applied to the inverse-distance matrix $\mathbf{A}$. Since the diagonal of that matrix is on one hand formally infinity, but on the other hand contains contributions to the self-energy of the particle, it is not of interest and therefore is simply set to zero. If charges are distributed to the grid by the nearest grid point method⁶, no further corrections to self-energies have to be applied. It is, nevertheless, stressed that this approximation induces relatively large discretization errors.

A wavelet transform may be represented by an orthogonal matrix $\mathcal{W}$, which leads in the present case to the representation

$$\mathcal{W}\Phi = \mathcal{W}\mathbf{A}\mathcal{W}^T \mathcal{W}\mathbf{Q} \quad (8)$$

which can be written as

$$\tilde{\Phi} = \tilde{\mathbf{A}}\tilde{\mathbf{Q}} \quad (9)$$

where ' $\tilde{\cdot}$ ' means a wavelet transformed quantity. Applying a threshold operation $\mathcal{T}_x$ to $\tilde{\mathbf{A}}$, where x is the threshold size, transforms $\tilde{\mathbf{A}}$ into a sparse representation (cmp. Fig. 1). Since the distance between grid points in the geometrical domain is kept constant during time, the wavelet transform of $\hat{\mathbf{A}}$ has only to be performed in the beginning of the simulation. The time dependence is completely shifted to the vector $\hat{\mathbf{Q}}$, thereby reducing the transformation to a N -dimensional quantity. Applying a fast wavelet transform the computational complexity is reduced to $\mathcal{O}(N)$. Finally, to obtain the potential on every grid

point, a fast backtransform of $\mathcal{O}(N)$ complexity is applied in order to obtain $\hat{\Phi}$. The potential at the charges' position is then obtained via interpolation from the grid points, which may be performed via polynomial approximation of the potential surface. In numerical tests, it was found that the sparse matrix-vector product may also be performed with $\mathcal{O}(N)$ complexity while keeping the approximation error constant. The latter, of course, depends on the size of the threshold x , but it may be bound to relative errors of 10^{-2} and below.

4 Conclusions

A fast method of order $\mathcal{O}(N)$ for the calculation of electrostatic potentials has been described. It is based on a wavelet transformation technique which, in combination with a threshold operation, transforms the inverse distance matrix, $\mathbf{A}$, into a sparse representation, thereby dramatically reducing the computational time of matrix-vector multiplication. The proposed method may be extended to calculate interaction energies and forces between particles. To increase the accuracy, other schemes may be applied to smear charges to the grid and to interpolate the potential back. Here it is necessary to treat self energies and near-field contributions in a proper way. Work in this direction is in progress.

References

1. L. Greengard. *The rapid evaluation of potential fields in particle systems*. MIT press, Cambridge, 1988.
2. R. D. Skeel, I. Tezcan, and D. J. Hardy. Multiple grid methods for classical molecular dynamics. *J. Comp. Chem.*, 23:673–684, 2002.
3. G. Sutmann and B. Steffen. A particle-particle particle-multigrid method for long-range interactions in molecular simulations. *Comp. Phys. Comm.*, 169:343–346, 2005.
4. G. Beylkin. *Wavelets, Multiresolution Analysis and Fast Numerical Algorithms*, 1991. A draft of INRIA Lecture Notes.
5. I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.
6. R. W. Hockney and J. W. Eastwood. *Computer simulation using particles*. McGraw-Hill, New York, 1981.

A Highly Accurate and Optimal Method to Calculate Long Range Interactions

Matthias Bolten and Godehard Sutmann

John von Neumann Institute for Computing
Central Institute for Applied Mathematics
Research Centre Jülich
52425 Jülich, Germany
E-mail: {m.bolten,g.sutmann}@fz-juelich.de

1 Systems with Long-Ranged Interactions

Generally, interactions between particles may be classified into short- and long-range interactions. Potentials, which are long-ranged exhibit a functional dependence with a decay slower than $1/r^d$, where d is the dimension of the physical space. Examples for this type of potential are the electrostatic or the gravitational potentials.

Considering long range interactions in atomistic simulations without simplifying assumptions, is a computationally very demanding task. Since all particle-particle interactions are of importance, no cut-off radius can be introduced, beyond which interactions are artificially set to zero. Therefore, a straight forward implementation of long range interaction potentials results in a computational complexity of $\mathcal{O}(N^2)$, if N is the number of particles. More advanced algorithms have been developed for periodic or open system geometries, which scale like $\mathcal{O}(N \log N)$ (e.g. P3M³ for periodic systems), or $\mathcal{O}(N)$ (e.g. FMM¹ for open systems).

In the present article a fast and accurate procedure of complexity $\mathcal{O}(N)$ is described, which is based on a multigrid technique.

2 Particle-Mesh Methods

In order to calculate the electrostatic energy of particle i , the sum

$$U_i = \frac{1}{4\pi} \sum_{j=1, j \neq i}^N q_i q_j \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}$$

has to be evaluated, where $|\cdot|$ is the Euclidean distance and $\mathbf{x}_k$ and q_k denote the position and the charge of particle k .

Instead of doing so, the potential due to all charges except for charge i may also be calculated using Poisson's equation

$$\Delta\Phi(\mathbf{x}) = f(\mathbf{x}) = \sum_{j=1, j \neq i}^N q_j \delta(\mathbf{x} - \mathbf{x}_j). \quad (1)$$

In this case, the electrostatic energy of particle i is readily calculated as

$$U_i = q_i \Phi(\mathbf{x}_i).$$

To calculate the energies for all particles, N Poisson equations would have to be solved. Therefore, in order to reduce the computational time, the right hand side of Eq. 1 is modified so that all particles in the system are included at the same time. Solving this equation yields a potential energy surface to which all particles contribute. In order to calculate energies for single particles, the self-energy contribution has to be corrected. Usually, equations of type Eq. 1 are solved using finite difference schemes, i.e. the PDE is discretized on a uniform grid and the charges are assigned to the mesh using an appropriate scheme like nearest-grid-point (NGP) or cloud-in-cell (CIC)³. The linear system can then be solved using standard methods, e.g. Particle-Mesh (PM) method, as described by Hockney and Eastwood³.

PM works well for fine grids which separate the particles from each other. In order to work with coarser grids which contain particles in neighboring or even the same boxes, a splitting similar to the splitting used in the Ewald summation can be used. The potential is splitted using the identity

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1-f(r)}{r}$$

where a possible choice is $1 - f(r) = \text{erf}(r)$, which corresponds to the potential of a Gaussian charge distribution. The Particle-Particle-Particle-Mesh method developed by Hockney and Eastwood³ uses charge assignments to the mesh (with e.g. CIC), splitting of the potential into near- and far-field and the appropriate near-field correction to calculate the forces and energies due to the slow-decaying Coulomb-potential. If working in periodic systems, an efficient method to solve the linear system is based on a Fourier-transform technique where the fast Fourier transform can be applied, which results in a computational complexity of $\mathcal{O}(N \log N)$.

3 Multigrid Methods

Multigrid methods are known to be optimal solvers for linear systems arising from the discretization of elliptic PDEs. An introduction can be found for example in the books of Hackbusch² or Trottenberg et al.⁶. The idea of multigrid methods is quite simple: If a linear system is solved using an iterative solver like Jacobi or Gauß-Seidel, the iteration converges rather slowly, but the high frequency components of the error are smoothed very fast. To speed up the solution process, the system is being solved on a hierarchy of grids, so that each frequency component is treated on the appropriate grid.

In order to achieve an optimal scaling (i.e. $\mathcal{O}(N)$), in the present work a multigrid method is applied to solve Poisson's equation. A Gauß-Seidel relaxation scheme is used in combination with a forth order discretization of the Laplace operator⁵. In case of systems with open boundaries, Dirichlet conditions are imposed on the surface of the (finite) computational domain. For a fast evaluation of boundary potentials a multipole expansion with computational complexity $\mathcal{O}(N)$ is used as described in the paper of Sutmann and Steffen⁴.

4 B-Spline Densities

Since a multigrid method is used to solve the PDE, there is no need to work in Fourier space. Therefore the use of an FFT can be avoided completely. As a splitting scheme to

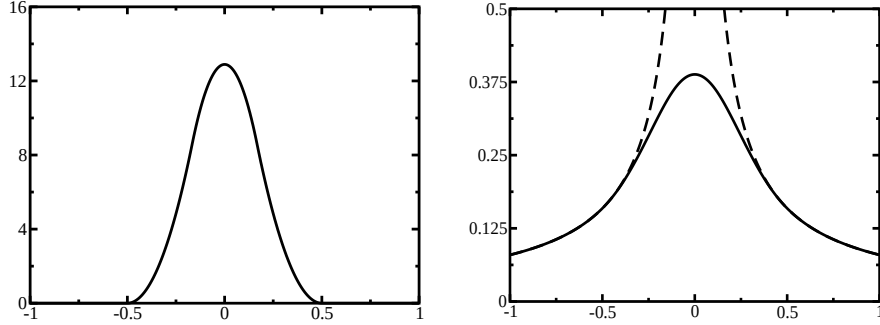


Figure 1. Left: radial shape of a charged centered, quadratic and normalized B-spline density. Right: potential of a B-spline density. The potential of a point charge, exhibiting a divergence at $r = 0$, is shown for comparison.

separate near and far field contributions a density with finite support is used. Instead of using the conventional choice $1 - f(r) = \text{erf}(r)$ the present work makes use of a form for $1 - f(r)$ which corresponds to the potential of a uniformly charged B-spline density.

Considering the centered, quadratic and normalized B-spline with width $2R$ given by

$$\rho_B(\mathbf{x}) = \begin{cases} \frac{-486|\mathbf{x}|^2 + 162R^2}{32\pi R^5} & : |\mathbf{x}| < \frac{R}{3} \\ \frac{486|\mathbf{x}|^2 - 972|\mathbf{x}|R + 486R^2}{64\pi R^5} & : \frac{R}{3} \leq |\mathbf{x}| \leq R \\ 0 & : R < |\mathbf{x}| \end{cases},$$

the corresponding potential is calculated via

$$\Phi_B(\mathbf{x}) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{\rho_B(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} d\mathbf{x}'$$

and can be given in analytical form as

$$\Phi_B(\mathbf{x}) = \begin{cases} \frac{3(81|\mathbf{x}|^4 - 90|\mathbf{x}|^2 R^2 + 65R^4)}{4\pi 80R^5} & : |\mathbf{x}| < \frac{R}{3} \\ \frac{-243|\mathbf{x}|^5 + 810|\mathbf{x}|^4 R - 810|\mathbf{x}|^3 R^2 + 405|\mathbf{x}| R^4 - 2R^5}{4\pi |\mathbf{x}| 160R^5} & : \frac{R}{3} \leq |\mathbf{x}| \leq R \\ \frac{1}{4\pi |\mathbf{x}|} & : R < |\mathbf{x}| \end{cases}$$

The B-spline densities are centered on the charges in the system and sampled onto the mesh in order to compute the right hand side of the PDE. The influence of this charge discretization to particles closer than distance R is taken into account by direct summing up the near-field contributions by explicit particle-particle calculations. Therefore, in this step the charge assignment to the grid and the associated error is avoided.

5 Conclusion

In the present work a mesh-based particle simulation method was presented which avoids the Fourier-space and the FFT completely using a multigrid technique to solve the Poisson equation. Overall the method results in an optimal computational complexity of $\mathcal{O}(N)$. First numerical experiments show very good accuracy. At the moment work is in progress to reliably estimate errors and to parallelize the method.

References

1. L. Greengard, and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comp. Phys. **73**, 325–348 (1987).
2. W. Hackbusch, *Multi-Grid Methods and Applications*, (Springer-Verlag, Berlin, 1985).
3. R. W. Hockney, and J. W. Eastwood, *Computer simulation using particles*, (Institute of Physics, Bristol, 1988).
4. G. Sutmann, and B. Steffen, *A particle-particle particle-multigrid algorithm for long range interactions in molecular systems*, Comp. Phys. Comm. **169**, 343–346 (2005).
5. G. Sutmann, and B. Steffen, *High-Order Compact Solvers for the Three Dimensional Poisson Equation*, J. Comp. Appl. Math. **187**, 142–170 (2006).
6. U. Trottenberg, C. Oosterlee, and A. Schüller, *Multigrid*, (Academic Press, San Diego, 2001).

Evolutionary Analysis in Pathogenesis-Related Proteins

Nicole M. Scherer¹, Claudia E. Thompson², Loreta B. Freitas²,
Sandro L. Bonatto³, and Francisco M. Salzano²

¹ Bioinformatics Department, Heinrich-Heine-Universität Düsseldorf
Universitätsstrasse 1, Geb. 25.02.02, 40225 Düsseldorf, Germany
E-mail: scherer@cs.uni-duesseldorf.de

² Departamento de Genética, Instituto de Biociências, UFRGS
Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil
E-mail: francisco.salzano@ufrgs.br

³ Centro de Biologia Genômica e Molecular, Faculdade de Biociências, PUCRS
Av. Ipiranga 6681, 90619-900 Porto Alegre, RS, Brazil

Pathogenesis-related proteins (PRs) are expressed by host plants following infections by fungi, bacteria or viruses, or after induction by abiotic stress factors. PRs enhance the host capability to limit subsequent infections. They comprise a wide range of forms, such as hydrolases, transcription factors, protease inhibitors, enzymes associated with various metabolic pathways, and allergenic products. Their functional motifs are related to a number of eukaryotic proteins, involved in very distinct functions. It is possible that their defensive functions evolved after their emergence as gene families. We believe that natural selection is a fundamental factor in the establishment of this resistance form in plants. Our analysis on the primary structure of representatives of 14 PR families identified several target sites for adaptive evolution. Testing how these changes structurally affect the protein molecules should help to relate adaptive mutations with their biological functions.

1 Introduction

Plants are likely to be infected by a large number of pathogens, like bacteria, fungi and viruses. Hence, plants have evolved a variety of mechanisms to prevent pathogen colonization and disease. Our study deals with a special class of defense proteins called pathogenesis-related proteins (PRs). Until now, 17 PR families have been identified¹. They are induced after pathogenic infection or environmental stress, and exhibit antifungal, antibacterial, insecticidal, nematocidal and antiviral effects. These provide the host plant enhanced capability to limit subsequent infections inhibiting pathogen growth, multiplication or spread. Their toxicity is due to hydrolyticity, proteinase-inhibitory and membrane-permeabilizing ability².

In 2000, Bishop *et al.*³ combined sequence evolution analysis and knowledge of the structure of the chitinase type I gene family to understand the variable effectiveness of specific chitinases against different pathogens. They found an excess of amino acid replacements in the active site and substrate binding cleft, which cannot be explained by a relaxation of selection pressure alone and therefore indicates positive selection.

In this study, we first conduct a phylogenetic analysis of pathogenesis-related proteins with protein-coding DNA sequences of 14 PR families⁴. We then apply a maximum likelihood framework based on codon substitution models to identify adaptive evolution. Finally, the sites identified as positively selected are further investigated on the structure of the protein.

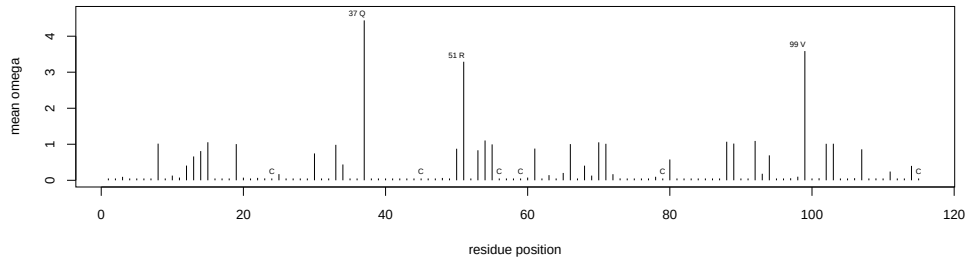


Figure 1. Values for ω ratios for sites along the PR-4 sequence under model M2 (Positive Selection) in CODEML⁸. Cysteine residues involved in disulphide bridges are indicated by the letter C

2 Methods

Not all mutations occurring in the DNA sequences lead to amino acid substitutions in the protein sequence. Those mutations which do change the coded amino acid are called non-synonymous. Synonymous mutations, in contrast, are silent and do not alter the amino acid sequence. Thus, they are considered free from selection pressure. Positive selection at the molecular level is normally tested by the comparison of non-synonymous (d_N) and synonymous (d_S) substitution rates in protein-coding genes. The ratio $\omega = d_N/d_S$ is an indicator for selective pressure. Positive selection is indicated by $\omega > 1$. On the other hand, $\omega < 1$ indicates purifying selection and $\omega = 1$ neutral evolution.

For each PR family, typical patterns and the reference sequences were used to search for homologous sequences in the Swiss-Prot/TrEMBL (UniProt Knowledgebase⁵) protein sequence database. We used ClustalX⁶ to align the sequences of each dataset. The protein and DNA sequence alignments were compared and manually edited with respect to the codon frame. Then, we reconstructed the phylogenetic trees of each PR family using IQPNNI⁷. Based on these trees, we applied CODEML⁸ to obtain the values of ω for each site under different models. Nested models like M1 (nearly neutral) and M2 (positive selection) are compared in a likelihood ratio test (LRT). If M1 is rejected in favor of M2 we assume that the sites with $\omega > 1$ are under positive selection. We applied SWISS-MODEL⁹ to infer the protein structure of all our PR dataset, except for PR-7, for which no appropriate template was available. Finally, we calculated the Euclidean distance from the alpha-carbon of each amino acid to the position of its homologous site in the template.

3 Results

The maximum likelihood analysis of the 14 PR datasets indicates positive selection in eight families of pathogenesis-related proteins. Here, we present the results for PR-4 (a wound-induced chitin-binding protein) as an example. Figure 1 shows the ω values for each residue position obtained under model M2 in CODEML⁸. The positively selected sites (37 Q, 51 R, 99 V) are labeled. Residues 37 Q and 99 V are placed in alpha-helices, and 51 R belongs to a gamma-turn. The members of this family have six conserved cysteine residues that form three disulphide bridges¹⁰. The six cysteine residues are among the sites with ω ratios close to zero due to purifying selection.

4 Discussion

The Euclidean distances calculated on the superimposed structures should provide a means to examine the relationship between the positively selected sites and their function on the protein. In the inferred structures of the PRs, the differences observed in the positively selected sites are due to differences in residue size. The deviations resulting from *indels* are greater than those resulting from substitutions. Using only this approach we could not see a correlation between these mutations and their effects in the protein structure. In order to make inferences about the positively selected sites, it is necessary to include the biochemical and biophysical variables in the analysis, and also to treat the *indels* separately. In the future, other modeling approaches should be used to overcome these difficulties.

Acknowledgments

We would like to thank Roland Fleißner, Simone Linz and Bui Quang Minh for their helpful comments and suggestions on the poster and this manuscript. This work was supported by grants from PRONEX, CNPq, FINEP, CAPES, FAPERGS and PROPESQ-UFRGS.

References

1. <http://www.bio.uu.nl/~fytopath/PR-families.htm>
2. Aglika Edreva. *Pathogenesis-related proteins: research progress in the last 15 years*. Gen. Appl. Plant Physiology **31**, 105–124 (2005)
3. J. G. Bishop, A. M. Dean and T. Mitchell-Olds. *Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution*, PNAS **97**, 5322–5327 (2000)
4. N. M. Scherer, C. E. Thompson, L. B. Freitas, S. L. Bonatto, and F. M. Salzano. *Patterns of molecular evolution in pathogenesis-related proteins*, Genet. Mol. Biol. **28**, 645–653 (2005).
5. A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, L. S. Yeh. *The Universal Protein Resource (UniProt)*, Nucleic Acids Res. **33**, D154–D159 (2005) <http://www.uniprot.org>
6. R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, J. D. Thompson. *Multiple sequence alignment with the Clustal series of programs*. Nucleic Acids Res. **31**, 3497–3500 (2003)
7. L. S. Vinh and A. von Haeseler. *IQPNNI: Moving fast through tree space and stopping in time*, Mol. Biol. Evol. **21**, 1565–1571 (2004)
8. Z. Yang. *PAML: a program package for phylogenetic analysis by maximum likelihood*, Comput. Appl. Biosci. **13**, 555–556 (1997)
9. T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch. *SWISS-MODEL: an automated protein homology-modeling server*, Nucleic Acids Research **31**, 3381–3385 (2003).
10. B. Svensson, I. Svendsen, P. Hojrup, P. Roepstorff, S. Ludvigsen and F. M. Poulsen. *Primary structure of barwin: a barley seed protein closely related to the C-terminal domain of proteins encoded by wound-induced plant genes*. Biochemistry **31**, 8767–8770 (1992)

A Parameterized Procedure for Consensus Sequences

- Validation of the Method

Sławomir Walkowiak and Bogdan Lesyng

Department of Biophysics, Warsaw University,
Zwirki i Wigury 93, 02-089 Warsaw, Poland
E-mail: {swalk, lesyng}@icm.edu.pl

A consensus sequence for a given protein family is an averaged, representative sequence which describes common features of the family. Once generated it can be used as a query to protein databases. The quality measure of a constructed consensus sequence can be defined as a number of newly found sequences, or as an average similarity to all family members. In this work we made an attempt to compare these two ways of estimation of the quality measure of consensus sequences, and/or find any correlation between them. Four kinase families were chosen, and for each of them seven consensus sequences were constructed. Each of newly generated sequences was designed based on different sets of threshold parameters. Each of the constructed sequences was used as a database query, and also for each of them an average similarity to its family was determined. The average level of similarity was computed by scoring all possible pairwise alignments with the aid of a semihomology algorithm¹, estimating statistical significance of the number of identities and the identity distribution. Although we haven't found any unique set of the threshold parameters which could provide best results in both methods, we observed some correlation between the chosen sets of the threshold parameters and the 'scores' obtained for the consensus sequences.

1 Introduction

A consensus sequence can be defined as a best, unique sequence representation for the given protein family. The construction process of the sequence depends on an arbitrarily chosen set of control parameters describing the residue type at a specific position. Various sets of the threshold parameters defining gaps, conservative or non specific residue types, usually provide different consensus sequences².

2 Estimation of the Quality of Consensus Sequence

In this study we estimate the quality measure of consensus sequences in two ways. Firstly a consensus sequence was used as a query for protein databases and an average similarity to its family was determined. Secondly we computed an average level of similarity to all given protein family members. This procedure was applied for each of seven different parameter sets describing different consensus sequences for four kinase families.

3 Consensus Sequence as a Database Query

Each consensus sequence was used in a BLAST³ query, and from the BLAST output sets we selected only those hits for which e-values were higher than a preset threshold value. In order to cutoff statistically insignificant hits, the obtained sets were then compared with the results for the template sequences, for each of the protein families. Table 1 presents the results.

Family	parameter set	no. of new sequences	$-\langle \log(e - value) \rangle$
RKT	6	23	14,17
RKT	3	23	14
RKT	2	23	14
RKT	1	22	13,82
RKT	7	43	7,21
JAK	6	37	17,7
JAK	2	35	16,86
JAK	1	36	16,81
JAK	3	34	16,44
JAK	7	49	5,8
cAMP	6	15	29,7
cAMP	2	13	29,31
cAMP	1	13	29,31
cAMP	3	24	26,58
cAMP	7	34	12,56

Table 1. Results of using the consensus sequences for the database query. For the fourth chosen family, heksokinases, we didn't find any new hits.

4 Calculation of an Average Similarity Level of the Consensus Sequences

An average similarity level of a consensus sequence for a given parameter set was computed by carrying out all possible pairwise alignments within its protein family. Each particular alignment was then scored using a semihomology algorithm¹ along with two other methods, computing a statistical significance value of obtaining a given number of identities, and to analyze the identity distribution⁴. The semihomology algorithm is based on a three dimensional diagram of all possible transitions and transversions between residues for all existing codons, and claims a single mutation as the main cause of evolutionary changes. Statistical significance of the number of the identities allows to compare results of the aligned sequences of different lengths. Estimation of the identity distribution, provides information about importance of the quality measure (briefly speaking: more regularity - more importance).

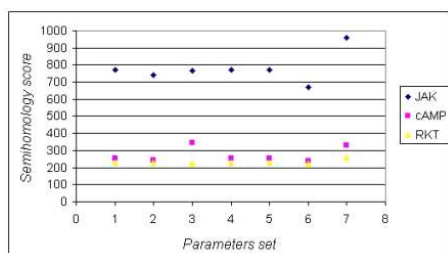


Figure 1. An average semihomology score for given parameters set.

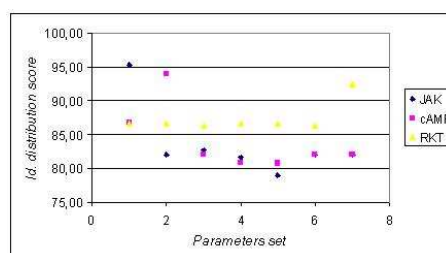


Figure 2. An average of the normalized identity distribution score for given parameter sets.

5 Concluding Remarks

Some correlations between the parameter sets and the 'scores' were obtained for the consensus sequences. A most optimal set of the parameters for the studied estimation procedures couldn't be determined. Further studies using a larger group of protein families should provide more convincing results.

Acknowledgments

These studies were supported by the European CoE MAMBA and National CoE BioExploratorium.

References

1. Leluk, J.
A new algorithm for analysis of the homology in protein primary structure, Computers & Chemistry **22**, 123–131 (1998).
2. Leluk, J., Fogtman, A., Lesyng, B.
Construction of Consensus Sequences of B-Spectrin Protein Family with variable Threshold Parameters and Validation of the Applied Approach, (2005).
3. Altschul, S.F. Gish, W., Miller, W., Myers, E.W., and Lipman, D.J.
Basic local alignment search tool, J. Mol. Biol. **215**, 403–410 (1990).
4. Leluk, J., Mikoajczyk, A.
A new approach to sequence comparison and similarity estimation, XXIX Kongres FEBS, 26.06 -1.07.2004, Warszawa, Eur. J. Biochem. **271**, Supplement, 29 (2004).

Prediction of Parallel and Antiparallel Beta Sheets Based on Sequence Profiles Using Support Vector Machines

Longhui Wang¹, Olav Zimmermann¹, and Ulrich H.E. Hansmann^{1,2}

¹ John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {l.wang, olav.zimmermann, u.hansmann}@fz-juelich.de

² Dept. of Physics, Michigan Technological University, 1400 Townsend Drive, Houghton, MI
49931, USA
E-mail: hansmann@mtu.edu

In this paper, we use SVM to construct classifiers to distinguish parallel and antiparallel beta sheets. Sequences are encoded as psiblast profiles. With seven-cross-validation carried on a 2686 non-homologous protein dataset, we obtain MCC (Matthew Correlation Coefficient) of 0.539836. The result shows that this two categories are separable by sequence profile.

1 Introduction

Beta-sheets are important secondary structures in the three-dimensional structures of peptides and proteins¹. Analysis of their properties may therefore help us to understand the general mechanism of folding in proteins. In this paper, we assume that we know which part of the amino acid sequence is beta-sheet, and then use a SVM based classifier to distinguish parallel and antiparallel beta sheets based on sequence profiles generated by PSI-BLAST².

2 Dataset

The original dataset includes 2686 non-homologous proteins with pairwise sequence identity less than 25%. The secondary structure is assigned from the experimentally determined tertiary structure by DSSP³. There are eight secondary structure classes: H (α -helix), G (3_{10} -helix), I (π -helix), E(β -strand), B(isolated β -bridge), T(turn), S(bend) and -(rest). There are 412,748 residues in the whole dataset. In our experiments, we only use E(β -strand), B(isolated β -bridge) are excluded. There are 89,500 β -strand residues in the dataset, 22.28% of them are parallel, 78.29% of them are antiparallel, and include mixed samples. Residues from mixed samples are assigned as parallel.

3 Method

3.1 Coding Scheme

We use PSI-BLAST profiles (PSSM) to describe each amino acid. The PSSM has $20 \times N$ elements with N the length of protein sequence. Each element of PSSM is an integer

between -7 and 7, and is scaled to [0,1] range by the following function:

$$f(x) = \begin{cases} 0.0 & \text{if } x \geq 5; \\ 0.5 + 0.1x & \text{if } -5 < x < 5; \\ 1.0 & \text{if } x \leq -5. \end{cases}$$

Here x is the value from the raw profile matrix⁴.

For each protein, we perform 3-iteration blast on non-redundant(nr) database to get PSSM. The window length is 15 amino acid residues, and is fixed during the whole process. This value is commonly used from protein secondary structure prediction based on Support Vector Machine and Neural Network.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method proposed by Vapnik and his coworkers⁴. Some samples are inseparable in low dimensional space, SVM use a kernel function to map them into high dimensional space, and seek a hyper-plane to divide them. It performs very well on a lot of pattern recognition problems. In our paper, we use C-SVM. In this kind of SVM, we need to select a kernel function and the regularization parameter C for the classifier. We use radial basis function (RBF):

$$K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2)$$

Different C and γ correspond to different classifiers, so we have to select best value for them our problem. And we use libsvm toolbox to construct classifiers⁵, which can be download from <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

4 Experiment and Results

To select the best C and γ for SVM, we randomly group the dataset (2686 proteins) into 56 subsets, and take seven of them to do seven-cross-validation. Each of the seven subsets has similar parallel and antiparallel composition as the original one. Detailed information about these seven subsets are omitted. All the results are evaluated by percentage and Matthew Correlation Coefficient (MCC). Initial test shows C=512 and $\gamma = 0.003175$ as the best parameters. With this values, we do seven-cross-validation on 2686 proteins. Results are shown in Table 1.

From the results in Table 1, we can see that the total prediction accuracy on the whole dataset for parallel and antiparallel beta sheets is 85.35%, with MCC of 0.539836, which shows that sequence profiles are efficient to distinguish parallel and antiparallel beta sheets using SVM classifiers.

5 Discussion

The presented results illustrate that parallel and antiparallel beta sheets residues can be separated only using sequence information. Changing window length and adding penalty

Test set	Number of res	Antiparallel β -sheet%	Total predict Accuracy	MCC
Test1	11831	0.781844	0.860367	0.557849
Test2	13046	0.751265	0.84622	0.558996
Test3	11089	0.785824	0.847777	0.500422
Test4	13920	0.788578	0.863721	0.549801
Test5	12704	0.784635	0.854534	0.531594
Test6	14215	0.785227	0.858248	0.548517
Test7	12695	0.763056	0.840567	0.522223
Total	89500	0.7772	0.853552	0.539836

Table 1. Results on whole dataset

parameters in SVM may further improve the results. We should also do the analysis about which kind will the mixed samples be predicted to. An upcoming paper will discuss the results in more details⁶.

Acknowledgment

Ulrich H.E. Hansmann is supported by a research grant (GM62838) of the National Institute of Health (USA).

References

1. S. S Zimmermann and H. A. Scheraga, *Local interactions in bends of proteins. Fast parallel algorithms for short-range molecular dynamics*, Proc. Natl. Acad. Sci. USA **9**, 4126–4129 (1977).
2. F. A. Stephan, L. T. Madden, A. A. Schaffer, J. Zhang, W. Miller and D. J. Lipman, *Capped BLAST and PSI-BLAST: a new generation of protein database search programs*, Nucleic Acids Res. **25**, 3389–3402 (1997).
3. W. Kabsh and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*, Biopolymers. **22**, 2577-2637 (1983).
4. H. Kim and H. Park, *Protein secondary structure prediction based on an improved support vector machines approach*, Protein Engineering. **16(8)**, 553–560 (2003).
5. C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*. (2001).
6. L. Wang, O. Zimmermann and U.H.E. Hansmann, *Prediction of parallel and antiparallel beta sheet on sequence profiles using support vector machines*. In Preparation (2006).

Side-Chain Ordering in PolyGlutamic Acid

Yanjie Wei¹, Walter Nadler¹, and Ulrich H.E. Hansmann^{1,2}

¹ Department of Physics, Michigan Technological University,
Houghton, MI 49931, USA
E-mail: {yawei,wnadler,hansmann}@mtu.edu

² John-von-Neumann Institute for Computing,
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: u.hansmann@fz-juelich.de

We report results from multicanonical simulations of Poly-Glutamic acid chains of length 10 residues. A decoupling of backbone and side chain ordering is observed in its folding process. While the details of the two transitions vary between the peptide in gas phase and in an implicit solvent, our results indicate that side chain ordering occurs only after the backbone topology is completely formed, independent of the particular environment.

1 Introduction

The folding of proteins involves several structural transitions, such as α -helix or β -sheet formation, and the ordering of side chains. The role of these transitions and their order are still only poorly understood. Computer experiments offer one way to study these questions but are often hampered by poor convergence of the simulations¹. We used multicanonical sampling² to study the role of side chain ordering. A homopolymer, Poly-Glutamic acid, was used because its long side chains can interact through hydrogen bonds. In both gas phase and solvent we observed a two-step process starting with a coordinated helix-formation and collapse of the molecule. At much lower temperatures side chain ordering takes place. The two-step process itself is independent on the specific environment, indicating the de-coupling of backbone and side-chain ordering could be a common characteristic of protein folding. A more thorough discussion of the results presented here can be found in Ref.³

2 Methods

Our investigations rely on simulations of *Gl*₁₀ with the ECEPP/3 force field⁵ as implemented in the 2005 version of the program package SMMP^{4,7}. The interactions between the atoms within the homopolymer chain are approximated by a sum consisting of electrostatic energy, Lennard-Jones, hydrogen-bonding, and torsion energy terms.

In order to enhance sampling we rely on the multicanonical approach² as described in Ref⁴. Configurations are weighted with an iteratively determined term $w_{MU}(E)$ such that the probability distribution is approximately constant. Multicanonical simulations of 5×10^6 sweeps were performed both in vacuum and in an implicit solvent⁸. The thermodynamic averages of any quantity $\langle O \rangle$ are obtained by re-weighting⁹:

$$\langle O \rangle (T) = \frac{\int dx O(x) e^{-E(x)/k_B T} / w_{MU}[E(x)]}{\int dx e^{-E(x)/k_B T} / w_{MU}[E(x)]} \quad (1)$$

where x counts the configurations of the system.

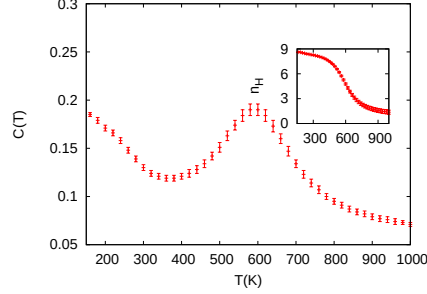


Figure 1. Specific heat $C(T)$ as a function of temperature T for Glu₁₀ in gas phase as obtained from a multicanonical simulation with 5×10^6 sweeps. The inlay shows the average number of helical residues $\langle n_H \rangle (T)$.

3 Results and Discussions

In the vacuum simulation, two peaks are observed in the specific heat plot (Fig. 1) indicating two transitions. The first peak at $T_1 = 590$ K corresponds to a helicity change (see the inlay to Fig. 1) that separates a high temperature region where the backbone has no ordering from a region where temperatures are low enough to allow the formation of backbone hydrogen bonds and subsequent growth of an α -helix.

In Fig. 2 we display the average number n_{hb}^S of side-chain and n_{hb}^B of backbone hydrogen bonds as function of temperature. The change in the number of side-chain hydrogen bonds and the fluctuations in this figure at $T_2 = 164$ clearly show that the corresponding peak in the specific heat indicates indeed a second transition. In the low temperature phase the side chains nestle along the cylinder formed by the helix and are stabilized by the side chain hydrogen bonds (not shown here). Such a structure was already described in Ref.¹⁰.

We have also extended our investigation to that of solvated Glu₁₀. Again, we observe two peaks in the specific heat (see Fig. 3). The peak at the higher temperature, $T_1 = 477$ K, marks again the collapse and subsequent formation of an α -helix (see the inlay of Fig. 3). But the transition temperature is 113 K lower than in the gas phase. Also, the peak in the

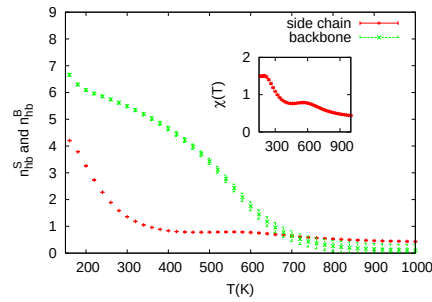


Figure 2. Average number of side chain hydrogen bonds $\langle n_{hb}^S \rangle (T)$ and backbone hydrogen bonds $\langle n_{hb}^B \rangle (T)$ as functions of temperature T for Glu₁₀ in gas phase as obtained from a multicanonical simulation with 5×10^6 sweeps. The inlay shows the fluctuation $\chi(T)$ in the number of side chain hydrogen bonds.

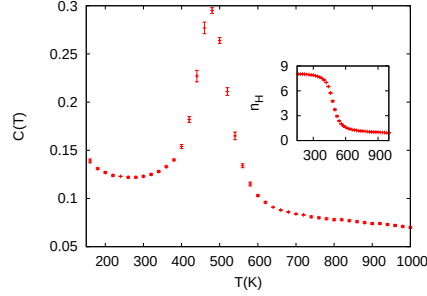


Figure 3. Specific heat $C(T)$ as function of temperature T for solvated Glu₁₀ as obtained from a multicanonical simulation with 5×10^6 sweeps. The inlay shows the average number of helical residues $\langle n_H \rangle (T)$

specific heat is higher and narrower, indicating a sharper, more well-defined transition.

The second peak in the specific heat that is observed in Fig. 3 at the lower temperature $T_2 = 111$ K is more narrow and smaller than the corresponding one for the molecule in gas phase. As with the coil-helix transition temperature T_1 , this transition is also shifted to lower temperatures in the solvent, albeit by the smaller amount of 53 K.

As in the case of Glu₁₀ in gas phase, the side chains are ordered at temperatures below T_2 . However, in contrast to the gas phase, the mechanism of side chain ordering is different in water. In Fig. 4, although no pronounced decrease in the solvation energy can be seen around T_2 , the *fluctuations* of the solvation energy show a pronounced increase. The increase of the solvent energy fluctuations continues to lower temperatures, while the fluctuations of the internal energy show a plateau and decrease below T_2 . The peak in the specific heat around T_2 is due to increasing anti-correlations below T_2 . In the low temperature phase the side chains order themselves by extending into the solvent (not shown here), in contrast to the gas phase behavior.

Hence, while the structure of the solvated molecule also evolves in a two-step process, the mechanism that leads to the side chain ordering is different from the gas phase. There

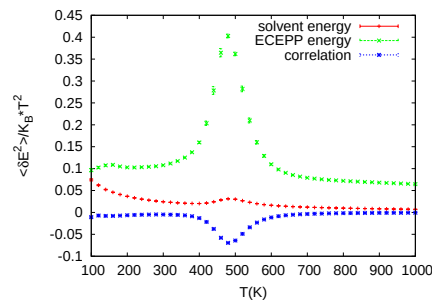


Figure 4. Fluctuations of the solvation energy $\langle \delta E_{solv}^2 \rangle (T)$ and intra-molecular energy $\langle \delta E_{ECEPP/3}^2 \rangle (T)$ for solvated Glu₁₀ as obtained from a multicanonical simulation with 5×10^6 sweeps. In addition, the cross-correlation $\langle \delta E_{solv} \delta E_{ECEPP/3} \rangle (T)$ is shown. Note that $\delta E = E - \langle E \rangle$.

the ordering of side chains was due to formation of hydrogen bonds between the side chain. In solvent, however, the polar side chains interact directly with the surrounding water that screens them from forming hydrogen bonds among each other. Our results indicate that the de-coupling of backbone and side-chain ordering does not depend on the details of the environment. Hence, it is reasonable to assume that this could be a common characteristic in protein folding.

Acknowledgments

We thank H. Scheraga for discussions and for drawing our attention to Ref.¹⁰. We also acknowledge support by a research grant (CHE-9981874) of the National Science Foundation (USA).

References

1. V. Aleksenko, W. Kwak and U.H.E. Hansmann, *Physica A*, **350**, 28 (2005).
2. B.A. Berg and T. Neuhaus, *Phys. Lett. B* **267**, 249 (1991).
3. Y. Wei, W. Nadler and U.H.E. Hansmann, Side-chain and backbone ordering in a Polypeptide, submitted for publication
4. U.H.E. Hansmann and Y. Okamoto, *J. Comp. Chem.* **14**, 1333 (1993).
5. M.J. Sippl, G. Némethy, and H.A. Scheraga, *J. Phys. Chem.* **88**, 6231 (1984), and references therein.
6. F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *Comp. Phys. Comm.* **138**, 192 (2001).
7. F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *Comp. Phys. Comm.* **174**, 422 (2006).
8. T. Ooi, M. Obatake, G. Nemethy, and H.A. Scheraga, *Proc. Natl. Acad. Sci. USA* **8**, 3086 (1987).
9. A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988); *Phys. Rev. Lett.* **63**, 1658(E) (1989), and references given in the erratum.
10. J.F. Yan, G. Vanderkooi and H.A. Scheraga, *J. Chem. Phys.* **49**, 2713 (1968).

Already published:

**Modern Methods and Algorithms of Quantum Chemistry -
Proceedings**

Johannes Grotendorst (Editor)

Winter School, 21 - 25 February 2000, Forschungszentrum Jülich

NIC Series Volume 1

ISBN 3-00-005618-1, February 2000, 562 pages

out of print

**Modern Methods and Algorithms of Quantum Chemistry -
Poster Presentations**

Johannes Grotendorst (Editor)

Winter School, 21 - 25 February 2000, Forschungszentrum Jülich

NIC Series Volume 2

ISBN 3-00-005746-3, February 2000, 77 pages

out of print

**Modern Methods and Algorithms of Quantum Chemistry -
Proceedings, Second Edition**

Johannes Grotendorst (Editor)

Winter School, 21 - 25 February 2000, Forschungszentrum Jülich

NIC Series Volume 3

ISBN 3-00-005834-6, December 2000, 638 pages

out of print

**Nichtlineare Analyse raum-zeitlicher Aspekte der
hirnelektrischen Aktivität von Epilepsiepatienten**

Jochen Arnold

NIC Series Volume 4

ISBN 3-00-006221-1, September 2000, 120 pages

**Elektron-Elektron-Wechselwirkung in Halbleitern:
Von hochkorrelierten kohärenten Anfangszuständen
zu inkohärentem Transport**

Reinhold Löwenich

NIC Series Volume 5

ISBN 3-00-006329-3, August 2000, 146 pages

**Erkennung von Nichtlinearitäten und
wechselseitigen Abhängigkeiten in Zeitreihen**

Andreas Schmitz

NIC Series Volume 6

ISBN 3-00-007871-1, May 2001, 142 pages

**Multiparadigm Programming with Object-Oriented Languages -
Proceedings**

Kei Davis, Yannis Smaragdakis, Jörg Striegnitz (Editors)
Workshop MPOOL, 18 May 2001, Budapest
NIC Series Volume 7
ISBN 3-00-007968-8, June 2001, 160 pages

**Europhysics Conference on Computational Physics -
Book of Abstracts**

Friedel Hossfeld, Kurt Binder (Editors)
Conference, 5 - 8 September 2001, Aachen
NIC Series Volume 8
ISBN 3-00-008236-0, September 2001, 500 pages

NIC Symposium 2001 - Proceedings

Horst Rollnik, Dietrich Wolf (Editors)
Symposium, 5 - 6 December 2001, Forschungszentrum Jülich
NIC Series Volume 9
ISBN 3-00-009055-X, May 2002, 514 pages

Quantum Simulations of Complex Many-Body Systems:

From Theory to Algorithms - Lecture Notes

Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 10
ISBN 3-00-009057-6, February 2002, 548 pages

Quantum Simulations of Complex Many-Body Systems:

From Theory to Algorithms- Poster Presentations

Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands
NIC Series Volume 11
ISBN 3-00-009058-4, February 2002, 194 pages

**Strongly Disordered Quantum Spin Systems in Low Dimensions:
Numerical Study of Spin Chains, Spin Ladders and
Two-Dimensional Systems**

Yu-cheng Lin
NIC Series Volume 12
ISBN 3-00-009056-8, May 2002, 146 pages

**Multiparadigm Programming with Object-Oriented Languages -
Proceedings**

Jörg Striegnitz, Kei Davis, Yannis Smaragdakis (Editors)
Workshop MPOOL 2002, 11 June 2002, Malaga
NIC Series Volume 13

ISBN 3-00-009099-1, June 2002, 132 pages

**Quantum Simulations of Complex Many-Body Systems:
From Theory to Algorithms - Audio-Visual Lecture Notes**

Johannes Grotendorst, Dominik Marx, Alejandro Muramatsu (Editors)
Winter School, 25 February - 1 March 2002, Rolduc Conference Centre,
Kerkrade, The Netherlands

NIC Series Volume 14

ISBN 3-00-010000-8, November 2002, DVD

Numerical Methods for Limit and Shakedown Analysis

Manfred Staat, Michael Heitzer (Eds.)

NIC Series Volume 15

ISBN 3-00-010001-6, February 2003, 306 pages

**Design and Evaluation of a Bandwidth Broker that Provides
Network Quality of Service for Grid Applications**

Volker Sander

NIC Series Volume 16

ISBN 3-00-010002-4, February 2003, 208 pages

**Automatic Performance Analysis on Parallel Computers with
SMP Nodes**

Felix Wolf

NIC Series Volume 17

ISBN 3-00-010003-2, February 2003, 168 pages

**Haptisches Rendern zum Einpassen von hochaufgelösten
Molekülstrukturdaten in niedrigaufgelöste
Elektronenmikroskopie-Dichteverteilungen**

Stefan Birmanns

NIC Series Volume 18

ISBN 3-00-010004-0, September 2003, 178 pages

Auswirkungen der Virtualisierung auf den IT-Betrieb

Wolfgang Gürich (Editor)

GI Conference, 4 - 5 November 2003, Forschungszentrum Jülich

NIC Series Volume 19

ISBN 3-00-009100-9, October 2003, 126 pages

NIC Symposium 2004

Dietrich Wolf, Gernot Münster, Manfred Kremer (Editors)

Symposium, 17 - 18 February 2004, Forschungszentrum Jülich

NIC Series Volume 20

ISBN 3-00-012372-5, February 2004, 482 pages

**Measuring Synchronization in Model Systems and
Electroencephalographic Time Series from Epilepsy Patients**

Thomas Kreutz

NIC Series Volume 21

ISBN 3-00-012373-3, February 2004, 138 pages

**Computational Soft Matter: From Synthetic Polymers to Proteins -
Poster Abstracts**

Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)

Winter School, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn

NIC Series Volume 22

ISBN 3-00-012374-1, February 2004, 120 pages

**Computational Soft Matter: From Synthetic Polymers to Proteins -
Lecture Notes**

Norbert Attig, Kurt Binder, Helmut Grubmüller, Kurt Kremer (Editors)

Winter School, 29 February - 6 March 2004, Gustav-Stresemann-Institut Bonn

NIC Series Volume 23

ISBN 3-00-012641-4, February 2004, 440 pages

**Synchronization and Interdependence Measures and their Applications
to the Electroencephalogram of Epilepsy Patients and Clustering of Data**

Alexander Kraskov

NIC Series Volume 24

ISBN 3-00-013619-3, May 2004, 106 pages

High Performance Computing in Chemistry

Johannes Grotendorst (Editor)

Report of the Joint Research Project:

High Performance Computing in Chemistry - HPC-Chem

NIC Series Volume 25

ISBN 3-00-013618-5, December 2004, 160 pages

**Zerlegung von Signalen in unabhängige Komponenten:
Ein informationstheoretischer Zugang**

Harald Stögbauer

NIC Series Volume 26

ISBN 3-00-013620-7, April 2005, 110 pages

Multiparadigm Programming 2003

Joint Proceedings of the

**3rd International Workshop on Multiparadigm Programming with
Object-Oriented Languages (MPOOL'03)**

and the

**1st International Workshop on Declarative Programming in the
Context of Object-Oriented Languages (PD-COOL'03)**

Jörg Striegnitz, Kei Davis (Editors)

NIC Series Volume 27

ISBN 3-00-016005-1, July 2005, 300 pages

**Integration von Programmiersprachen durch strukturelle Typanalyse
und partielle Auswertung**

Jörg Striegnitz

NIC Series Volume 28

ISBN 3-00-016006-X, May 2005, 306 pages

**OpenMolGRID - Open Computing Grid for Molecular Science
and Engineering**

Final Report

Mathilde Romberg (Editor)

NIC Series Volume 29

ISBN 3-00-016007-8, July 2005, 86 pages

Computational Nanoscience: Do It Yourself!

Lecture Notes

Johannes Grotendorst, Stefan Blügel, Dominik Marx (Editors)
Winter School, 14. - 22 February 2006, Forschungszentrum Jülich
NIC Series Volume 31
ISBN 3-00-017350-1, February 2006, 528 pages

NIC Symposium 2006 - Proceedings

G. Münster, D. Wolf, M. Kremer (Editors)
Symposium, 1 - 2 March 2006, Forschungszentrum Jülich
NIC Series Volume 32
ISBN 3-00-017351-X, February 2006, 384 pages

All volumes are available online at

www.fz-juelich.de/nic-series/.