Publication Series of the John von Neumann Institute for Computing (NIC)
NIC Series Volume 36

John von Neumann Institute for Computing (NIC)

From Computational Biophysics to Systems Biology (CBSB07)

Celebrating 20 years of NIC

edited by

Ulrich H. E. Hansmann
Jan Meinke
Sandipan Mohanty
Olav Zimmermann

Symposium, 02. - 04. May 2007
Forschungszentrum Jülich
Proceedings

organized by

John von Neumann Institute for Computing

NIC Series Volume 36

ISBN: 978-3-9810843-2-0

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:	Martina Kamps
Distributor:	NIC-Secretariat Research Centre Jülich 52425 Jülich Germany
Internet:	www.fz-juelich.de/nic/
Printer:	Graphische Betriebe, Forschungszentrum Jülich

NIC Series Volume 36
ISBN: 978-3-9810843-2-0
Published in full on the internet
Persistent Identifier: urn:nbn:de:0001-00422
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0001-00422

Preface

In 2007 the John von Neumann Institute for Computing (NIC) celebrates its 20th anniversary. During its two decades of existence, computer simulations have become an invaluable tool for researching biological systems. This was also the topic of the second annual workshop “From Computational Biophysics to Systems Biology” (CBSB07). About 110 researchers and students from all over the world met from May 2nd to May 4th, 2007 at the Research Centre Jülich and discussed physics-based approaches to systems biology. They emphasized the increasing importance of computing in Biology, Medicine and other life sciences. Complex diseases such as cancer or Alzheimer’s, to name only two, can be understood only if we obtain an insight in the working of cells on the molecular level. The analysis of the huge amount of experimental data, together with the problem that some processes or molecules cannot be detected in experiments, require the use of simulations. However, the large number of components in a cell, the difficulties in understanding their dynamics, and the complex interactions between them, make such simulations extremely challenging. Problems such as the all-atom simulation of chromatin dynamics or of ribosomes and other supermolecular systems will require access to computers with hundreds of Teraflops. In a ”town hall” meeting the participants emphasized the role of supercomputers in obtaining a detailed understanding of the working of cells. They agreed that further progress will require the supercomputer centers to provide both increased computing capacity and improved support and accessibility.

As in the previous year, the participants explored in scientific presentations and numerous informal discussions a wide range of topics ranging from single macromolecules to the working of entire cells. Topics included protein folding, miss-folding and aggregation; the interaction between proteins and other molecules; the assembly of nano-structures, multi-protein, protein-DNA/RNA complexes; and the modeling of cellular systems at a molecular level. This proceeding volume collects selected presentations from the 3-day long workshop that may serve as starting point for further discussions. It is divided into articles by invited speakers and such originally presented as posters or in contributed talks, as the interdisciplinary nature of the articles often defies a simple classification according to subject areas.

Besides the editors, Thomas Neuhaus, Tatjana Eitrich, Everaldo Arashiro, and Xiaolin Xiao helped organizing the workshop. For their most valuable help with the local arrangements we are greatly indebted to Helga Frank, Erika Wittig and Martina Kamps. We also wish to thank IBM for generous support.

Jülich, July 2007

U. H. E. Hansmann, J. Meinke, S. Mohanty, O. Zimmermann

Contents

Invited Talks

Contributed Talks and Posters

The Slip-Length Effects in Molecular Dynamics of Bead-Like Models of Proteins

Piotr Szymczak[1] **and Marek Cieplak**[2]

[1] Institute of Theoretical Physics
Warsaw University
00-681 Warsaw, ul. Hoża 69, Poland
E-mail: Piotr.Szymczak@fuw.edu.pl

[2] Institute of Physics, Polish Academy of Sciences
02-668 Warsaw, Al. Lotników 32/46, Poland
E-mail: mc@ifpan.edu.pl

We explore the role of the slip- or no-slip boundary condition for beads representing hydrophobic and hydrophylic amino acids in coarse grained models of proteins with an implicit solvent. The nature of the boundary conditions affects both the hydrodynamic friction forces and the Langevin noise term. We find that various large conformational changes in model proteins, such as folding and streching, are overall fairly stable against making the noise inhomogeneous in this way. However, there are noticeable differences in the scenarios of the processes.

1 Introduction

Coarse-grained models of proteins appear to play an increasingly important role in molecular dynamics studies of large conformational changes in proteins, especially when it comes to big proteins and their complexes. These models allow one to access time scales which are orders of magnitude longer than those available in all-atom simulations. Among the coarse-grained models, the Go-like implementations[1] link the properties of a protein directly to its native geometry and are probably the easiest to use. In the simplest version of a Go-like model[2–4] the protein is represented by its C^α atoms and the effective C^α–C^α contact potentials are used with minima at the experimentally established distances in the native state whereas non-native contacts are usually represented by hard core repulsion in order to prevent entanglements. Another term in the Hamiltonian takes care of the local stiffness of the backbone either by introducing biases in the bond and dihedral angles[5] or by favoring native senses of local chiralities[6]. Here, we choose the latter.

This description has also to be augmented by interactions with an implicit solvent which acts as a thermostat and imposes fluctuational forces on the protein. A common choice[2,7–9] is to take the Lennard-Jones potential for the interactions in the native contacts and to account for thermal and hydrodynamic friction forces using Langevin dynamics. The equation of motion for the i'th amino acid reads then

$$m_i\ddot{\mathbf{r}}_i = -\gamma_i\dot{\mathbf{r}}_i + \mathbf{F}_i^c + \mathbf{\Gamma}_i \ . \tag{1}$$

where $\mathbf{F}_i^c$ is the net force due to the molecular potentials, γ_i is the friction coefficient and $\mathbf{\Gamma}_i$ is a white noise term with the variance obeying $\langle \mathbf{\Gamma}_i(t)\mathbf{\Gamma}_j(t')\rangle = 2k_BT\gamma_i\delta(t-t')\mathbf{I}\delta_{ij}$ where $\mathbf{I}$ is the identity matrix and k_B is the Boltzmann constant.

The folding and unfolding properties of such a system clearly depend on the native geometry. But how important are other parameters in this model? Folding is not affected

by the precise choice of the contact potential between amino acids i and j and in this respect the Lennard-Jones potential, $4\epsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$, is almost undistinguishable from, say, the 10-12 potential[12]. As to the amplitude, ϵ, of the potential, assuming it to be uniform appears to be optimal since other models correlate with the experimental results on stretching worse[13].

We now consider other parameters in the model. Obviously, the values of the masses, m_i, of the amino acids affect the equations of motion for the individual beads. We have found, however, that using a uniform mass $m_i = m$ (where m is a typical amino acid mass $m \approx 2 \times 10^{-22} g$) does not affect the folding times in any noticeable manner[10]. This is consistent with both the overall coarse-grained character of the Go model and the fact that the motion is considered in the overdamped regime.

In this paper, we explore the effects of making the damping parametr γ non-uniform. In general, its value depends on the amino acid, especially on its size and position on the hydrophobicity scale. According to the classical Stokes equation, the friction coefficient is given by $\gamma = 6\pi\eta a$, where a is the hydrodynamic radius of the bead and η - the viscosity of the solvent. This formula actually applies when no-slip boundary conditions are in effect. However, the details of molecular interaction at the microscale determine the nature of boundary conditions used in a continuum description of the drag force. In particular, hydrophobicity of the surface at the molecular scale results in an increased slip length[16,17]. In order to account for the slip effects, Stokes's law may be rewritten as[18] $\gamma = c\pi\eta a$ where c is a constant determined by the choice of boundary conditions at its surface. The limiting cases are 6 for no-slip and 4 for perfect slip boundary conditions. To study the possible impact of friction nonuniformity due to the hydrophobicity we analyze a model, in which all hydrophobic aminoacids are supposed to move under the perfect slip conditions (c=4) whereas for hydrophilic aminoacids, no-slip boundary conditions are assumed to hold (c=6). Thus in the simplest division of amino acids into two classes one would have two possible values of γ. To some extent, hydrophobicity effects are already implicit in the native geometry and, therefore, in the contact potentials. Here, however, we focus on the impact of the non-uniform frictional forces and Langevin noise.

Unlike the masses, the values of γ are relevant in the overdamped regime so one might expect sensitivity to their choices. We demonstrate, however, that this sensitivity is weak. We consider several conformational processes in a protein: folding, stretching at a constant velocity, stretching at a constant force, and stretching by a uniform fluid flow. All details of the molecular dynamics simulations are as described in ref.[9], including the criteria for selecting native contacts, and the time unit, τ, is of order of a nanosecond[19,20]. We find that a non-uniform γ affects the folding times and folding scenarios. It affects the stretching by fluid flow, but not the set of metastable conformations. Finally, it essentially does not influence stretching at constant speed, but delays rupture by a constant force. As an illustration, we focus on ubiquitin (the Protein Data Bank code 1ubq) and crambin (1crn). Following ref.[15] the residues Ala, Val, Cys, Ile, Leu, Met, Phe, Tyr, and Trp are considered hydrophobic and the remaining as hydrophylic. The average c factor in γ is 0.84 and 0.89 of the uniform no-slip value for crambin and ubiquitn respectively.

2 Folding

When studying folding, we monitor establishment of contacts starting from an unfolded state with no contacts. A contact between aminoacids i and j is said to be established when the corresponding value of the distance between them, r_{ij}, crosses the threshold value of $1.5\sigma_{ij}$ (which is close to the inflection point of the Lennard-Jones potential) for the first time. Folding is considered to be achieved when all contacts are established simultaneously. The folding time, t_{fold}, of the protein is then determined as a median value in a set of many trajectories.

The kinetics of folding can be quantified with the use of the so-called scenario diagrams[8] in which one plots an average time to establish a contact against the contact order, i.e., against the sequential distance, $|j-i|$, between the amino acids that form a native contact. Figures 1 and 2 compare the folding scenarios for the Go model with uniform γ to the nonuniform case for ubiquitin and crambin respectively. Making comparisons requires introduction of time scaling factors to account for different average values of γ in the models. It has been established that, in the overdamped regime, folding times vary linearly with γ[19,2,9] in the uniform model. By introducing a set of hydrophobic aminoacids with lower γ values we are effectively decreasing the mean friction coefficient, by 16% in crambin, which results in a reduced folding time. This reduction is observed in Figure 1 (although, instead of the 16%, the observed decrease in folding times is only by $\sim 11\%$). Moreover, in addition to this overall shift, there are some finer differences between the scenarios, the most notable of which is the delay in formation of the longest ranged contacts in the non-uniform model (*cf.* the inset of Figure 1).

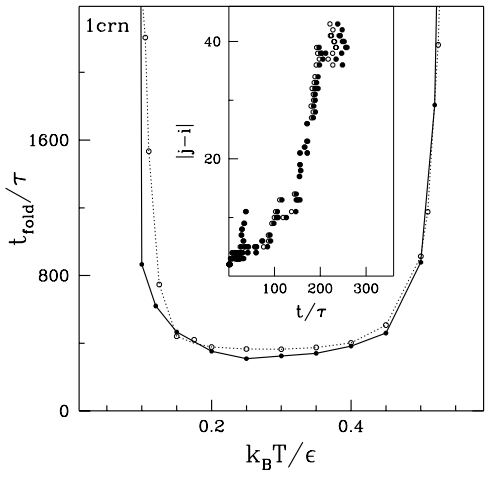


Figure 1. The dependence of t_{fold} on T for uniform (open circles) and nonuniform (solid circles) friction along the chain. The data points are based on at least 200 trajectories and are gathered at the optimal temperature for folding – 0.3 ϵ/k_B. The inset shows the similarly denoted average time intervals needed to establish contacts at a sequential distance $|j-i|$ for the first time. The uniform-case data points are reduced by 16% to account for the shift in $<\gamma>$.

Figure 2 shows the folding scenario diagram for ubiquitin. It is seen that the non-uniformity in the noise affects essentially the whole scenario, not only at large $|j-i|$ as in crambin, even if one accounts for the shift in $<\gamma>$. The optimal folding time decreases from 781 to 693 τ in this case.

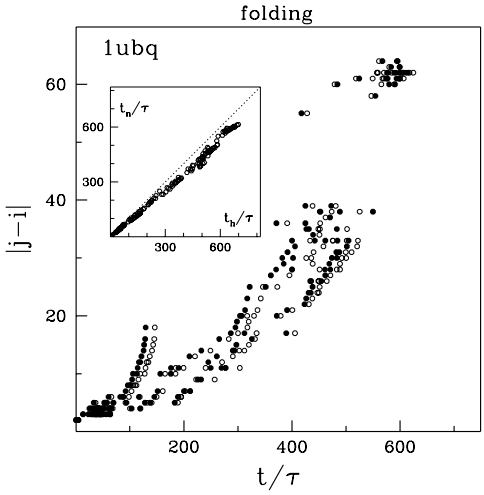


Figure 2. The folding scenario diagram for ubiquitin. The symbols used are as in Figure 1. The uniform-case data points are reduced by 12% to account for the shift in $<\gamma>$. The inset correlates the nonuniform-case times, t_n, to establish a contact with the uniform, or homogeneous, case, t_h. The times t_h are not scaled and are seen to be generally larger than t_n.

3 Stretching

Next, we consider mechanical stretching of proteins. Figure 3, for ubiquitin, shows that stretching at constant speed results in a force of resistance, F, that varies only in small details between the uniform and non-uniform models. The force is shown against the displacement, d, of a spring that is attached to a terminus of the protein on one end and moves with a speed of 0.005 Å/τ at the other end[10]. It should be noted that, even in the uniform γ case, the $F-d$ curves depend on γ very little[10] and, therefore, we do not expect to see any systematic shifts with a changing $<\gamma>$.

When stretching protein at constant force[21,22], one applies the stretching force to the protein terminus and monitors the end-to-end distance, L. The kinetics of the process may be quantified by determining statistics of the unfolding times[23,21] and by using the scenario diagrams[23]. Here, we consider the latter. In the unfolding diagrams, one shows when, on average, a contact at the sequential distance $|j-i|$ is broken for the last time (i.e., it crossed the distance od 1.5 σ_{ij} for good). Figure 4 shows the unfolding scenarios for ubiquitin. It is seen that the order of rupturing events does not change significantly when friction becomes nonuniform. However the specific times of the events get shifted noticeably: lack of uniformity in γ delays rupture because a local strengthening in the

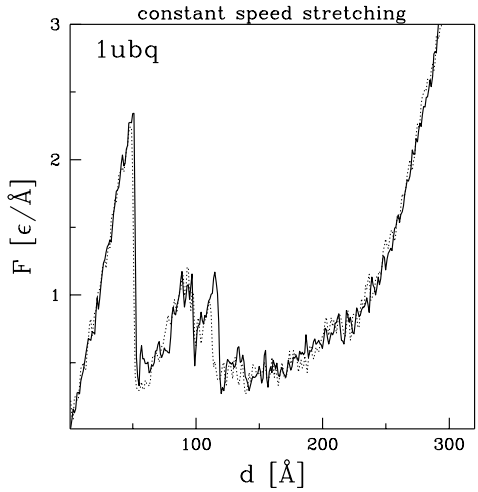


Figure 3. Single-trajectory force-extension curves for the ubiquitin unfolding at a constant speed for uniform (dotted line) and non-uniform (solid line) friction along the chain. The biggest differences are seen around the third force peak.

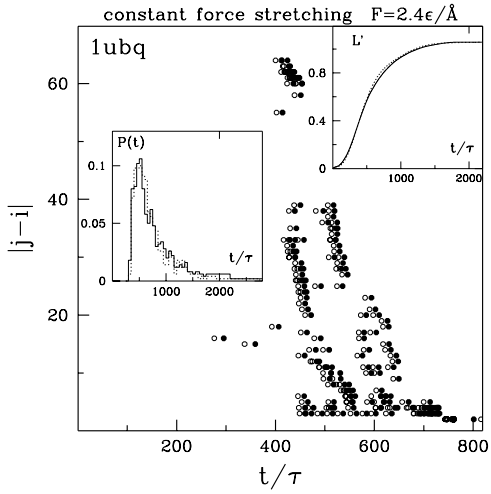


Figure 4. Unfolding scenarios at a constant large force for ubiquitin. The open and solid symbols correspond to the uniform and non-uniform friction respectively. The uniform system shown in this figure has γ equal to $< \gamma >$ of the non-uniform system. The data points are based on 500 trajectories. The inset on the left shows the histograms of the unfolding times – $P(t)$ is normalized to 50. The inset on the right shows the average normalized end-to-end length, $L' = \frac{L-L_f}{L_u-L_f}$ where L_f and L_u stand for the folded and unfolded end-to-end distances, respectively. In both insets, the dotted (solid) line corresponds to the uniform (non-uniform) system.

damping force generates pinning centers. In order to demonstrate it in a clean fashion, we do not perform any rescaling of unfolding times, but instead compare the non-uniform

system to a uniform one in which γ is reduced by 21% . In this way, $< \gamma >$ in both systems match. The insets of Figure 4 demonstrate that in the case of nonuniform friction the distribution of unfolding times has fatter tail and that the sigmoidal trajectory-averaged end-to-end distance acquires larger values at longer times. Both facts are consistent with the emergence of the delays seen in the scenario diagram.

Finally, we consider stretching of a single protein in a uniform flow through drag on individual beads (amino acids). Although the process has not yet been realized experimentally, the simulations[24–26] suggest that uniform flow unfolding leads to a richer set of metastable conformations than the constant force pulling and thus may offer potentially wider diagnostic tools to investigate structure of proteins compared to experiments based on the atomic force microscopy.

Note that in this case making γ_i nonuniform along the chain results in a change in a net hydrodynamic force acting on a protein which is equal to (when one neglects hydrodynamic interactions) $F = \sum_{i=1}^{N} \gamma_i U$ where U is the flow velocity. Thus a reduction in values of γ reduces the overall pulling force due to drag and should result in delay of unravelling processes. This is indeed seen in Figure 5 which shows the behavior of L as a function of time for ubiquitin. Despite the delay, one observes the same metastable conformations in both model systems. It is worth to note that in this case the nonuniformity in individual γ_i values plays a less important role. In fact the uniform system with the friction coeffcients γ suitably rescaled (as described above) shows the unfolding trajectories almost indistinguishable from those of the nonuniform model.

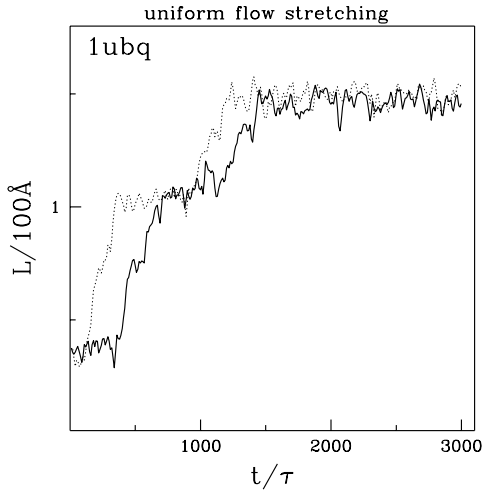


Figure 5. The end-to-end distance for single trajectories in flow-stretching of ubiquitin. The dotted and solid lines correspond to homogeneous and inhomogeneous friction along the sequence.

Acknowledgements

This work was funded by the European program IP NaPa through Warsaw University of Technology.

References

1. S. Takada, *Proc. Natl. Acad. Sci. (USA)* **96**, 11698-11700, 1999.
2. T. X. Hoang and M. Cieplak, *J. Chem. Phys.* **112**, 6851-6862, 2000.
3. J. Karanicolas and C. L. Brooks III, *Prot. Sci.* **11**, 2351-2361, 2002.
4. J. I. Sulkowska and M. Cieplak, *J. Phys.: Cond. Mat.* **19**, 283201, 2007.
5. C. Clementi, H. Nymeyer, and J. N. Onuchic, *J. Mol. Biol.* **298**, 937, 2000.
6. J. I. Kwiecinska and M. Cieplak. *J. Phys.: Cond. Mat.* **17**, S1565-S1580, 2005.
7. M. Cieplak and T. X. Hoang, *J. Biol. Phys.* **26**, 273-294, 2000.
8. T. X. Hoang and M. Cieplak, *J. Chem. Phys.* **113**, 8319-8328, 2001.
9. M. Cieplak and T. X. Hoang, *Biophys. J.* **84**, 475-488, 2003.
10. M. Cieplak, T. X. Hoang, and M. O. Robbins, *Proteins: Struct. Funct. Bio.* **56**, 285-297, 2004.
11. J. Tsai, R. Taylor, C. Chothia, and M. Gerstein, *J. Mol. Biol.* **290**, 253-266, 1999.
12. M. Cieplak and T. X. Hoang, *Physica A* **330**, 195-205, 2003.
13. J. I. Sułkowska and M. Cieplak (submitted).
14. D. K. Klimov and D. Thirumalai, *Phys. Rev. Lett.* **79**, 317-320, 1997.
15. O. D. Monera, T. J. Sereda, N. E. Zhou , C. M. Kay and R. S. Hodges J. Pept. Sci. **1**, 319-329, 2004.
16. M. Cieplak, J. Koplik, J. R. Banavar, *Phys. Rev. Lett.* **86**, 803, 2001.
17. L. Bocquet and J-L. Barrat, *Soft Matter*, DOI: 10.1039/b616490k, 2007.
18. J. R. Schmidt and J. L. Skinnera J. Chem. Phys. **119**, 8062-8068, 2004.
19. T. Veitshans, D. Klimov, and D. Thirumalai, *Folding Des.* **2**, 1-22, 1997.
20. P. Szymczak and M. Cieplak, *J. Chem. Phys.* **125**, 164903, 2006.
21. J. M. Fernandez and H. Li, *Science*, **303**, 1674, 2004.
22. M. Schlierf, H. Li, J. M. Fernandez *Proc. Natl. Acad. Sci (USA)* **101**, 7299, 2004.
23. P. Szymczak and M. Cieplak, *J. Phys.: Cond. Mat.* **18**, L21-L28, 2006.
24. A. Lemak, J. R. Lepock, and J. Z. Y. Chen, *Proteins: Structure, Function and Genetics* **51**, 224, 2003.
25. A. Lemak, J. R. Lepock, J. Z. Y. Chen, *Phys. Rev. E* **67**, 031910, 2003.
26. P. Szymczak and M. Cieplak, *J. Chem. Phys* **125**, 164903, 2006.

Advances in De Novo Protein Design

Christodoulos A. Floudas[1], **Ho Ki Fung**[1], **and Martin S. Taylor**[2]

[1] Department of Chemical Engineering,
Princeton University, Princeton, NJ 08540
E-mail: floudas@titan.princeton.edu

[2] School of Medicine,
Johns Hopkins University, Baltimore, MD 21205
E-mail: mstaylor@jhmi.edu

A new de novo protein design framework and its applications to the redesign of compstatin, human beta defensin-2, and the C-terminal analogs of Complement 3a is presented.

1 Introduction

De novo protein design searches for amino acid sequences that are compatible with a three-dimensional protein backbone template. Traditionally the backbone coordinates were treated as fixed in order to reduce the search space and make the design problem tractable. However, this is a highly questionable assumption as proteins are known to exhibit backbone flexibility. In de novo design, backbone flexibility was incorporated through either the consideration of multiple backbones with sequence search performed on each of them under the fixed template assumption, or the parameterization of backbone[16]. Recently we have developed a novel framework which performs de novo design on a truly flexible backbone template, which is defined by continuous C^{α}-C^{α} distances and dihedral angles between upper and lower bounds, through NMR structure refinement.

2 Our De Novo Protein Design Framework

Our two-stage de novo protein design framework not only selects and ranks amino acid sequences for a particular fold using a novel integer linear programming (ILP) model, but also validates the specificity to the fold for these sequences based on the full-atomistic forcefield AMBER[1]. The two stages are outlined as below:

2.1 Stage One: In Silico Sequence Selection

The ILP model we use for sequence selection into a single template structure, which is the most computationally efficient one among 13 equivalent formulations we studied[9], takes the form:

$$\begin{aligned} \min_{y_i^j, y_k^l} & \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}(x_i, x_k) w_{ik}^{jl} \\ \text{subject to} \quad & \sum_{j=1}^{m_i} y_i^j \;=\; 1 \;\forall\, i \\ & \sum_{j=1}^{m_i} w_{ik}^{jl} \;=\; y_k^l \;\forall\; i, k > i, l \\ & \sum_{l=1}^{m_k} w_{ik}^{jl} \;=\; y_i^j \;\forall\; i, k > i, j \\ & y_i^j \;,\; y_k^l \;,\; w_{ik}^{jl} \;=\; 0 - 1 \;\forall\; i, j, k > i, l \end{aligned} \tag{1}$$

Set $i = 1, \ldots, n$ defines the number of residue positions along the template. At each position i there can be a set of mutations represented by $j\{i\} = 1, \ldots, m_i$, where for the general case $m_i = 20 \forall i$. The equivalent sets $k \equiv i$ and $l \equiv j$ are defined, and $k > i$ is required to represent all unique pairwise interactions. Binary variables y_i^j and y_k^l are introduced to indicate the possible mutations at a given position. Specifically, variable $y_i^j(y_k^l)$ will be one if position $i(k)$ is occupied by amino acid $j(l)$, and zero otherwise. The composition constraints require that there is exactly one type of amino acid at each position. The pairwise energy interaction parameters E_{ik}^{jl} were empirically derived by solving a linear programming parameter estimation problem, which restricts the low energy high resolution decoys for a large training set of proteins to be ranked energetically less favorable than their native conformations[2].

Besides the basic model (1), we also developed a weighted average model and a binary distance bin model[10] for de novo design based on a flexible template with multiple crystal or solution structures.

2.2 Stage Two: Approximate Method for Fold Validation

Driven by the full atomistic forcefield AMBER[1], simulated annealing calculations are performed for an ensemble of several hundred random structures generated for each sequence from stage one using CYANA 2.1[3,4] within the upper and lower bounds on C^{α}-C^{α} distances and dihedral angles input by the user. This feature allows our framework to observe true backbone flexibility[5]. The TINKER package[6] is subsequently used for local energy minimization of these conformers. A fold specificity factor is finally computed for each sequence using the following equation:

$$f_{specificity} \;=\; \frac{\sum_{i \in \text{new sequence conformers}} \exp[-\beta E_i]}{\sum_{i \in \text{native sequence conformers}} \exp[-\beta E_i]} \tag{2}$$

3 Case Studies

3.1 Compstatin

Compstatin (PDB code: 1A1P) is a synthetic 13-residue cyclic peptide that inhibits the cleavage of C3 to C3a and C3b in the human complement system and thus hinders complement activation. It is a novel drug candidate for treating inappropriate complement activation that has shown highly promising results in numerous pre-clinical trials conducted recently. The de novo design on compstatin is aimed at acquiring the sequences for the best

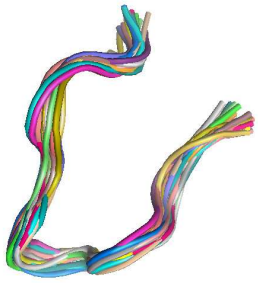


Figure 1. Flexible template of compstatin for de novo protein design as illustrated by the overlapping of its 21 NMR structures available from the Protein Data Bank.

inhibitors to C3. It was performed based on the flexible template of all 21 NMR structures available from the Protein Data Bank (Fig. 1).

As for the mutation set of the design, Cys^2 and Cys^{12} were kept invariant to maintain the disulfide bridge in order to aid the formation of the hydrophobic cluster and prohibit the termini from drifting apart. The type I β-turn, Gln^5-Asp^6-Trp^7-Gly^8, was also fixed as it was not found to be a sufficient condition for activity. Val^3 and Trp^7 were not mutated either as they were found to interact directly with C3. For the varied positions, positions 1, 4 and 13 were allowed to select only from the hydrophobic amino acid set (A,F,I,L,M,W,V,Y). In addition, this set included threonine for position 13 to allow for the selection of the wild type residue at this position. For positions 9, 10, and 11, all residues were allowed, excluding cysteine and tryptophan. This mutation set leads to a problem with complexity 3.0×10^6. Results for the design can be found in[10].

3.2 Human Beta Defensin-2

Human Beta Defensin-2 (hβD-2) is a cysteine-rich 41-residue cationic peptide found in the human immune system. It belongs to the class of small, cationic peptides known as defensins. hβD-2 is crucial to innate immunity[11]. It possesses antimicrobial property derived from the electrostatic force between the positive charge on the defensin molecule and the negative charge of the anionic head group of the microbe's membrane lipids. This electrostatic force disrupts the microbe's cell membrane and thus kills the cell[11].

Three different sets of flexible design templates were employed for the de novo design of hβD-2. The first one corresponds to the X-ray crystal structure elucidated by[11](PDB code: 1FD3) at a resolution of 1.35Å. The other two were generated using molecular dynamics simulation with generalized Born implicit solvation (Fig. 2) and molecular dynamics simulation with explicit water molecules (not shown).

In the design of hβD-2, SASA patterning was applied to restrict the sequence search space for the de novo design of hβD-2. The 41 positions of hβD-2 are classified into the core, surface, and intermediate categories which determine the mutation set for each position. This corresponds to the full-squence design of the antimicrobial peptide with

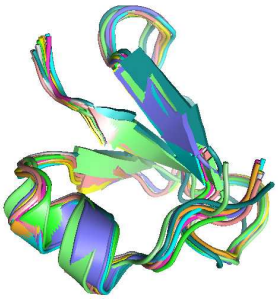


Figure 2. Overlay of the 10 structures of human beta-defensin-2 used for the flexible design template from MD simulations with the GB implicit solvation model. The structures are snapshots with 1 ns increment.

problem complexity of 6.40×10^{37}.

It should be noted that in the sequence selection stage, biological constraints, which were obtained from a homology search using PSI-BLAST, were imposed to ensure certain conserved properties among the sequence solutions. The constraints added cover charge characteristics, hydrophobic content, and amino acid occurrence frequencies of the human beta defensin-2 homologs. Results for the de novo design can be found in[12].

3.3 Complement 3a

Complement 3a (C3a) is a 77-residue small cationic peptide derived from the cleavage of the amino-terminus of the α-chain of complement component C3. It is a potent mediator which controls the pro-inflammatory activities of the complement system[13]. Having small molecular size and high potency, C3a proves to be a strong candidate as a superior therapeutic agent. Our de novo design aims at obtaining a potential peptide-drug candidate based on the C-terminal sequence of the C3a fragment of C3.

Like the design of human beta defensin-2, three different sets of flexible templates were employed. One corresponds to the single crystal structure elucidated by[14], and the other two were generated using molecular dynamics simulations, one with the generalized Born implicit solvation model and the other with explicit water molecules (Fig. 3). The basic sequence selection model (1) was used for the single crystal structure template, whereas both the weighted average formulation and the binary distance bin formulation[10] were employed for the flexible templates from molecular dynamics simulation.

Table 1 shows the mutation set of the de novo design.

Results of the de novo design are tabulated in[15]. Several 15-residue peptides were selected to be synthesized based on our predictions from the de novo design framework. The best sequence was experimentally validated to be close in performance to the superpotent peptide synthesized by[13] in Ca^{2+} mobilization assay.

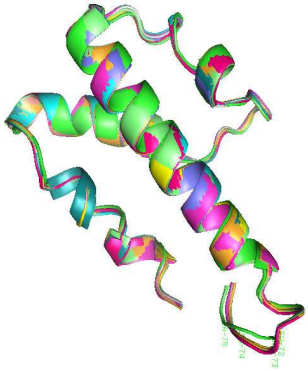


Figure 3. Overlay of the 10 structures of Complement 3a used for the flexible design template from MD simulations with the GB implicit solvation model.

Table 1. Mutation set of *in silico* sequence selection of C3a.

Positions	Native residue	Allowed mutations
63	L	A,I,L,M,F,Y,W,V
64	R	all except C and P
65	R	R,N,D,Q,E,G,H,K,S,T
66	Q	R,N,D,Q,E,G,H,K,S,T
67	H	R,N,D,Q,E,G,H,K,S,T
68	A	all except C and P
70	A	R,N,D,Q,E,G,H,K,S,T
71	S	R,N,D,Q,E,G,H,K,S,T
72	H	R,N,D,Q,E,G,H,K,S,T

4 Conclusions

In this paper, we presented the advances in our de novo protein design models, as well as our predictions on compstatin, human beta defensin-2, and C3a.

Acknowledgments

CAF gratefully acknowledges support from the National Science Foundation, the National Institutes of Health (R01 GM52032, R24 GM069736), and the US Environmental Protection Agency (GAD R 832721-010). This work has not been reviewed by and does not represent the opinions of USEPA.

References

1. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *A 2nd Generation*

Force-Field For The Simulation Of Proteins, Nucleic-Acids, And Organic-Molecules, J. Am. Chem. Soc. **117**, 5179–5197, 1995.

2. R. Rajgaria, S. R. McAllister, and C. A. Floudas, *A Novel High Resolution C^{α}-C^{α} Distance Dependent Force Field Based on a High Quality Decoy Set*, Proteins **65**, 726–741, 2006.
3. P. Guntert, C. Mumenthaler, and K. Wuthrich, *Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA*, J. Mol. Bio. **273**, 283–298, 1997.
4. P. Guntert, *Automated NMR structure calculation with CYANA. Methods Mol Biol*, J. Mol. Bio. **278**, 353–378, 2004.
5. C. A. Floudas, *Research Challenges, Opportunities and Synergism in Systems Engineering and Computational Biology*, AIChE Journal **51**, 1872–1884, 2005.
6. J. Ponder, *TINKER, software tools for molecular design. 1998.* (Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 1998).
7. J. L. Klepeis, C. A. Floudas, D. Morikis, C. G. Tsokos, E. Argyropoulos, L. Spruce, and J. D. Lambris, *Integrated Structural, Computational and Experimental Approach for Lead Optimization: Deisgn of Compstatin Variants with Improved Activity*, J. Am. Chem. Soc. **125**, 8422–8423, 2003.
8. J. L. Klepeis, C. A. Floudas, D. Morikis, C. G. Tsokos, and J. D. Lambris, *Design of Peptide Analogs with Improved Activity using a Novel de novo Protein Design Approach*, Ind. Eng. Chem. Res. **43**, 3817–3826, 2004.
9. H. K. Fung, S. Rao, C. A. Floudas, O. Prokopyev, P. M. Pardalos, and F. Rendl, *Computational Comparison Studies of Quadratic Assignment Like Formulations for the In Silico Sequence Selection Problem in De Novo Protein Design*, J. Comb. Optim. **10**, 41–60, 2005.
10. H. K. Fung, M. S. Taylor, and C. A. Floudas, *Novel Formulations for the Sequence Selection Problem in De Novo Protein Design with Flexible Templates*, Optim. Methods & Software **22**, 51–71, 2007.
11. D. Hoover, K. Rajashankar, R. Blumenthal, A. Puri, J. Oppenheim, O. Chertov, and J. Lubkowski, *The Structure of Human β-Defensin-2 Shows Evidence of Higher Order Oligomeration*, J. Biol. Chem. **275**, 32911–32918, 2000.
12. H. K. Fung, C. A. Floudas, M. S. Taylor, L. Zhang, and D. Morikis, *Towards Full-Sequence De Novo Protein Design with Flexible Templates for Human Beta-Defensin-2*, Biophys. J., submitted for publication (2007).
13. J. A. Ember, N. L. Johansen, and T. E. Hugli, *Designing Synthetic Superagonists of C3a Anaphylatoxin*, Biochemistry **30**, 3603–3612, 1991.
14. R. Huber, H. Scholze, E. P. Paques, and J. Deisenhofer, *Crystal Structure Analysis and Molecular Model of Human C3a Anaphylatoxin*, Hoppe-Seylers Z Physiol Chemie **361**, 1389–1399, 1980.
15. H. K. Fung, C. A. Floudas, M. S. Taylor, L. Zhang, and D. Morikis, *Redesigning Complement 3a based on Flexible Templates from both X-ray Crystallography and Molecular Dynamics Simulation*, in preparation (2007).
16. H. K. Fung, and C. A. Floudas, *Computational De Novo Peptide and Protein Design: Rigid Template versus Flexible Templates*, Curr. Protein & Peptide Sci., submitted for publication (2007).

Chromatin Dynamics *in silicio*

Jörg Langowski

Division Biophysics of Macromolecules, German Cancer Research Center
Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany
E-mail: jl@dkfz.de

The packing of the genomic DNA in the living cell is essential for its biological function. While individual aspects of the genome architecture, such as DNA and nucleosome structure or the arrangement of chromosome territories are well studied, much information is missing for a unified description of cellular DNA at all its structural levels. Computer modeling can contribute to such a description. We present here Monte-Carlo and Brownian dynamics simulations of DNA and the chromatin fiber, using an elastic chain approximation for the DNA and a simple hard-core description for the histone proteins. The model allows for the prediction of possible higher-order chromatin structures and their mechanical properties. As examples, we show the unrolling of DNA from the histone core, the response of the 30 nm chromatin fiber to mechanical stretching and possible regimes of stable and unstable packing of chromatin.

Further, we show dynamics simulations of the nucleosome on the microsecond time scale, using a new coarse-grained model. Finally, describing the chromatin fiber as an elastic chain, we implement models for the transport of proteins in the cell nucleus, reproducing the anomalous subdiffusion found experimentally.

1 Introduction

One of the major challenges in modern molecular and structural biology is the structural and dynamical organization of the cell nucleus[1]. Genomic information is encoded into DNA, a long filamentous macromolecule which is compacted into chromatin by its association with histone proteins. Chromatin also forms a long flexible chain with a diameter of about 30 nm and constitutes about 5 - 10% of the total volume of the nucleus. In every human cell, 6×10^9 base pairs of DNA – that is, a total length of about 2 meters – must be packed into a more or less spheroid nuclear volume about 10 - 20 μm in diameter. This compaction must occur in such a way that the DNA molecule is still easily accessible to enzymes acting on it, such as replication, transcription and repair machineries, or regulatory factors. In addition, more and more supramolecular entities are being identified, including nucleoli, PML bodies, Cajal bodies, spliceosomes etc. that fulfill important biological roles in transcription, splicing, replication or repair mechanisms, but whose structural association with other parts of the nucleus is hardly understood. Important biological questions, such as the gene distribution inside the interphase nucleus, 'memory' of chromosome positions during cell division or the flexibility, accessibility of the folded chromatin chain, interaction of distant parts of the genome or the transport of nuclear factors to their binding site can only be understood through a detailed description of the higher order folding of the DNA molecule in the cell nucleus.

DNA organization in the cell is a phenomenon that needs to be described on many length and time scales (Fig. 1). Such a multiscale modelling problem must be approached by some adequate approximation, in which we will have to define subunits of the molecule that behave like rigid objects on the size and time scale considered. These objects inter-

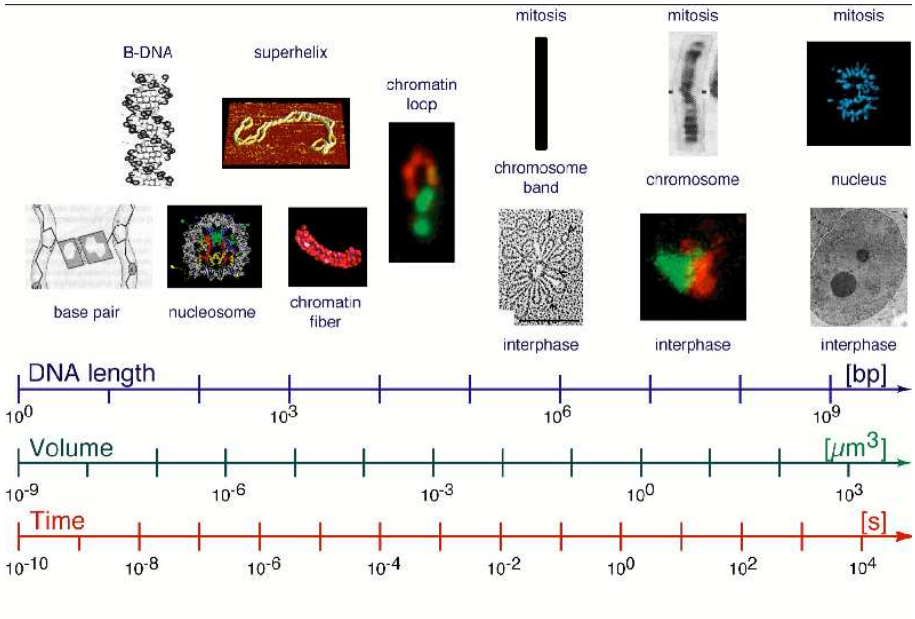


Figure 1. Overview of time and length scales relevant for genome organization.

act through potentials that may in principle be derived from the interatomic force fields; however, in practice one mostly uses potentials that have been determined experimentally.

2 Polymer Chain Models for DNA and Chromatin

The motif of the 'linear elastic filament' in genome organization is repeated on many length scales (Fig. 1): DNA, as well as the chromatin fiber and to some extent its higher order structures may be approximated by a flexible wormlike chain (WLC). Thus, we may develop models of DNA and the chromatin fiber based on a coarse-grained description using a linear segmented chain. Segments are assumed to behave like rigid cylinders on the time and length scale considered; they are connected by elastic joints, with bending, torsional and stretching potentials approximated by Hookean springs with spring constants that are known independently. [a]

2.1 Segmented Wormlike Chain

Fig. 2 schematizes a segmented chain geometry. A vector $\mathbf{s}_i$ defines the direction and length of each segment i, $\mathbf{f}_i$ is a unit vector normal to the segment and $\mathbf{g}_i$ is an auxiliary

[a] For DNA, the approximation of Hookean bending elasticity has recently been challenged by the finding that short DNA fragments have a cyclization probability much higher than expected for a homogeneously elastic wormlike chain[2], and atomic force microscopy observations that sharp bends occur more frequently than expected for a Gaussian distribution of bending angles[3].

vector that is used to take into account permanent bending of the DNA. The details of this chain geometry are given in[4].

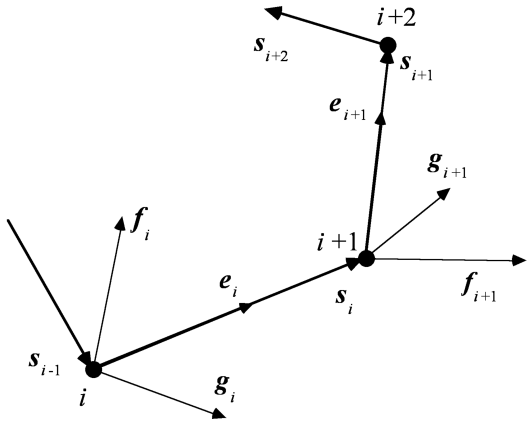


Figure 2. Section of a segmented polymer chain as used in the DNA and chromatin models described here.

2.2 Intersegment Interactions

Adjacent segments (i and $i+1$ in Fig. 2) interact with each other through bending, twisting and stretching potentials. Independent of the form of the local intersegment potential, the WLC approximation always holds for sufficiently long chains, as has recently been shown by Wiggins and Nelson[5]. Thus, if one does not consider tight local bends, the potentials between adjacent segments can be approximated by Hookean springs. Furthermore, potentials must be defined for long-range interactions between non-neighboring segments: in the case of 'naked' DNA, this interaction is the electrostatic repulsion between the negatively charged sugar-phosphate backbones and can be described by a screened Coulomb potential[4]. If the DNA is associated with proteins as in the case of chromatin, the geometry of these complexes and their specific interaction must be taken into account. In the chromatin chain model developed by Wedemann et al.[6], nucleosomes are approximated by rigid ellipsoids and their interaction by a Gay-Berne potential (an anisotropic Lennard-Jones potential[7]).

2.2.1 Bending Rigidity

In the WLC model the length of the segments should be chosen well below the *persistence length* L_p, which is a measure of the bending flexibility of the chain molecule. It is defined as the correlation length of the direction of the chain measured along its contour:

$$\langle \vec{u}\,(s)\,\vec{u}\,(s+s')\rangle = e^{-s/L_p} \tag{1}$$

Here $\vec{u}(s)$ is a unit vector in the direction of the chain ($\boldsymbol{e}_i$ in Fig. 2) and *s* resp. *s'* is the position along the chain contour, the angular brackets indicating the average over all positions and chain conformations. Molecules shorter than L_p behave approximately like a rigid rod, while longer chains show significant internal flexibility.

The bending elasticity A - the energy required to bend a polymer segment of unit length through an angle of 1 radian - is related to the persistence length by $L_p = A/k_BT$, k_B being Boltzmann's constant and T the absolute temperature. Thus, the energy required to bend two segments of the chain of length *l* by an angle θ with respect to one another is:

$$E_b = \frac{k_BT}{2}\frac{L_p}{l}\theta^2 \tag{2}$$

For DNA, L_p has been determined in a number of experiments (for a compilation, see [2]). While some uncertainties remain as regards the value at very high or low salt concentrations, the existing data agree on a consensus value of L_p = 45-50 nm (132-147 bp) at intermediate ionic strengths (10-100 mM NaCl and/or 0.1-10 μM Mg^{2+}). For high values of θ, the potential may deviate from the simple harmonic form (see footnote a and ref.[3]).

2.2.2 Torsional Rigidity

The torsional rigidity C, defined as the energy required to twist a polymer segment of unit length through an angle of 1 radian, may be related in an analogous way to a *torsional persistence length* L_T, defined as the correlation of a vector normal to the chain axis and with fixed orientation relative to the molecular structure of the polymer chain. The torsional rigidity C has been measured by various techniques, including fluorescence polarization anisotropy decay[8–10] and DNA cyclization[11–13], and the published values converge on a value of L_T = 65 nm (191 bp).

2.2.3 Stretching Rigidity

The stretching elasticity of DNA has been measured by single molecule experiments[14,15] and also calculated by molecular dynamics simulations[16,17]. The stretching modulus σ of DNA is about 1500 pN, where $\sigma = F \cdot L_0/\Delta L$ (ΔL being the extension of a chain of length L_0 by the force F). The stretching energy of a segment of length l that is stretched by Δl is:

$$E_{str} = \frac{1}{2}\frac{\sigma}{l}\Delta l^2 \tag{3}$$

DNA stretching does not play a significant role in chromatin structural transitions, since much smaller forces are already causing large distortions of the 30 nm fiber (see below).

2.2.4 Intrachain Interactions

The average DNA helix diameter used in modeling applications such as the ones described here includes the diameter of the atomic-scale B-DNA structure and – approximately – the thickness of the hydration shell and ion layer closest to the double helix. Both for the calculation of the electrostatic potential and the hydrodynamic properties of DNA (i.e. the friction coefficient of the helix for viscous drag) a helix diameter of 2.4 nm describes the

chain best[18–20,4]. The choice of this parameter is supported by the results of chain knotting[21] or catenation[22], as well as light scattering[23] and neutron scattering[19] experiments.

As pointed out in[4,24] DNA intrachain electrostatic repulsion can be adequately described by a Debye-Hückel electrostatic potential between two uniformly charged non-adjacent segments (*i*, *j*) in a 1-1 salt solution:

$$E_{ij}^{(e)} = \frac{\nu^2}{D} \iint d\lambda_i d\lambda_j \frac{e^{-\kappa r_{ij}}}{r_{ij}} \tag{4}$$

Here, the integration is done along the two segments, λ_i and λ_j are the distances from the segment beginnings, r_{ij} is the distance between the current positions at the segments to which the integration parameters λ_i and λ_j correspond; κ is the inverse of the Debye length, so that $\kappa^2 = 8\pi e^2 I/k_B T D$, I is the ionic strength, e the proton charge, D the dielectric constant of water, ν the linear charge density which for DNA is equal to $\nu_{DNA} = -2e/\Delta$ where Δ = 0.34 nm is the distance between base pairs. More details as to the normalization of the linear charge density etc. have been given in our earlier paper[4].

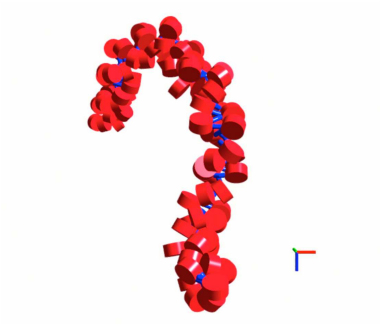


Figure 3. Example of a Monte-Carlo equilibrated structure of a chromatin fiber consisting of 100 nucleosomes (red), linker segments (blue) of repeat length l = 205 nm.

3 Monte-Carlo Model of the Chromatin Chain

As an example of the application of a polymer chain model to genome structure, we describe here the simulation of nanomechanical properties of the chromatin fiber by a Monte-Carlo model. The flexibility of the chromatin fiber has been measured in a set of experiments, either by relating the spatial distance of markers on the DNA to their genomic distance[25–27] or by direct measurements of cyclization probabilities[28,29]. The persistence obtained cover a large range from unrealistically low values of about 30 nm[28,29] to values of up to 200 nm[26]. In our recent work[30], we show that depending on the local structure of the DNA on the nucleosome, the nucleosome repeat and the presence or absence of linker histone H1, this wide range of persistence lengths may be reproduced.

In the model the chromatin fiber is approximated as a flexible polymer chain consisting of rigid ellipsoidal disks, 11 nm in diameter and 5.5 nm in height. These disks are connected by linker DNA, represented by two cylindrical segments. Incoming and outgoing

linker DNA are set 3.1 nm apart of each other. This geometry used is essentially the "two angle" model developed earlier by Woodcock et al.[31].

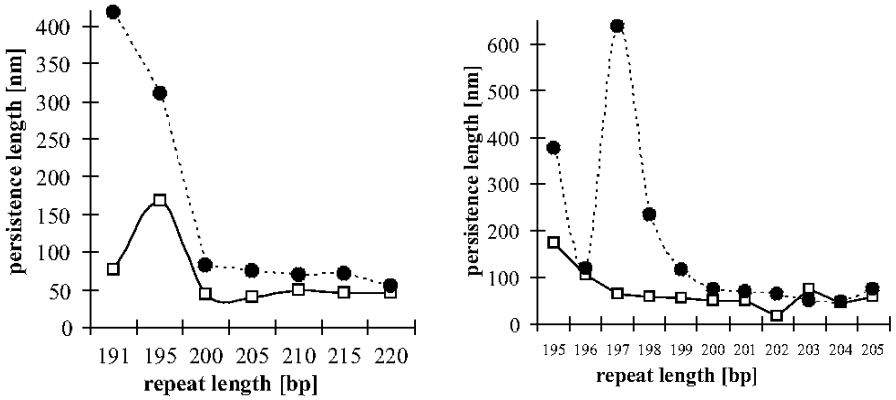


Figure 4. Persistence length of modelled 30 nm chromatin fibers with different nucleosomal repeats in the presence and absence of linker histone H1. The twisting angle between adjacent nucleosomes is adjusted to the canonical value of 360° per 10.5 bp. The persistence lengths of fibers with linker histone (closed symbols, dashed lines) are higher than for fibers without linker histone (open symbols, solid lines). This effect is stronger for short repeats and weakens with increasing repeat length. The peaks show that the twisting angle strongly influences the stiffness of the fiber, leading to a non-monotonous variation of L_p with nucleosome repeat.

A typical conformation of a 100 nucleosome chain after 3.10^6 MC steps is shown in Fig. 3. Simulations were done with either a condensed fiber as a starting conformation or an initial conformation where all segments are ordered in a straight line.

The bending and stretching rigidities of the modelled chromatin fiber are then computed over the trajectory from the fluctuations in the bending angle or the fluctuation in the overall fiber length. The results show that the bending and the stretching stiffness of the chromatin fiber strongly depend on the local geometry of the nucleosome. Both the persistence length L_p, characterizing the bending stiffness of the fiber, and the stretching modulus ϵ, which describes the stretching stiffness of the fiber, decrease if either the linker lengths or the opening angle are increased, or the twisting angle is reduced. This behavior is independent of the presence of the linker histone H1. The latter decreases the opening angle α between the entry and exit of the linker DNA and as a result leads to a more condensed fiber structure for high salt concentrations[32]. This is in agreement with our simulations, since the presence of the linker histone-induced stem motif yields higher persistence lengths thus stiffer fibers (Fig. 4).

The other major result of the simulation comes from comparing the persistence length of the modelled fibers to that of a hypothetical rod from a isotropic elastic material having the same stretching rigidity as the chromatin fiber. Such a rod would have a bending rigidity 4-10 times higher than that actually measured, or simulated here. Thus, the chromatin fiber is less resistant to bending than to stretching. This property of the chromatin fiber is important for its ability to condense and decondense, for example to prevent or allow transcriptional access. Chromatin fibers thus seem to be packed more easily via dense

loops than by a linear compression. The formation of such dense loops of hairpin structures of interdigitated chromatin arrays has been recently suggested[33], and some hairpin conformations could also be seen in our simulations (data not shown).

References

1. T. P. O'Brien, C. J. Bult, C. Cremer, M. Grunze, B. B. Knowles, J. Langowski, J. McNally, T. Pederson, J. C. Politz, A. Pombo, G. Schmahl, J. P. Spatz, and R. Van Driel, *Genome function and nuclear architecture: from gene expression to nanoscience*, Genome Res, **13**, no. 6, 1029–41, 2003.
2. T. E. Cloutier and J. Widom, *Spontaneous sharp bending of double-stranded DNA*, Molecular Cell, **14**, no. 3, 355–362, 2004.
3. P. A. Wiggins, T. Van der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker, and P. C. Nelson, *High flexibility of DNA on short length scales probed by atomic force microscopy*, Nature Nanotechnology, **1**, no. 2, 137–141, 2006.
4. K. Klenin, H. Merlitz, and J. Langowski, *A Brownian dynamics program for the simulation of linear and circular DNA and other wormlike chain polyelectrolytes*, Biophys J, **74**, no. 2 Pt 1, 780–788, 1998.
5. P. A. Wiggins and P. C. Nelson, *Generalized theory of semiflexible polymers*, Phys Rev E Stat Nonlin Soft Matter Phys, **73**, no. 3 Pt 1, 031906, 2006.
6. G. Wedemann and J. Langowski, *Computer simulation of the 30-nanometer chromatin fiber*, Biophys J, **82**, no. 6, 2847–59., 2002.
7. J. G. Gay and B. J. Berne, *Modification of the overlap potential to mimic a linear site–site potential*, Journal of Chemical Physics, **74**, 3316–3319, 1981.
8. M.D. Barkley and B.H. Zimm, *Theory of twisting and bending of chain macromolecules: analysis of the fluorescence depolarization of DNA*, Journal of Chemical Physics, **70**, 2991–3007, 1979.
9. B. S. Fujimoto and J. M. Schurr, *Dependence of the torsional rigidity of DNA on base composition*, Nature, **344**, no. 6262, 175–7., 1990.
10. J.M. Schurr, B.S. Fujimoto, P. Wu, and L. Song, "Fluorescence studies of nucleic acids: dynamics, rigidities and structures.", in: Topics in Fluorescence Spectroscopy, J.R. Lakowicz, (Ed.), vol. 3, pp. 137–229. Plenum Press, New York, 1992.
11. D Shore and R. L. Baldwin, *Energetics of DNA twisting. I. Relation between twist and cyclization probability*, Journal of Molecular Biology, **179**, 957–981, 1983.
12. D. S. Horowitz and J. C. Wang, *Torsonal Rigidity of DNA and Length Dependence of the Free Energy of DNA Supercoiling.*, J. Mol. Biol., **173**, 75–91, 1984.
13. W. H. Taylor and P. J. Hagerman, *Application of the method of phage T4 DNA ligase-catalyzed ring-closure to the study of DNA structure I.NaCl-dependence of DNA flexibility and helical repeat.*, J. Mol. Biol., **212**, 363–376, 1990.
14. P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J.-L. Viovy, D. Chatenay, and F. Caron, *DNA: An extensible molecule*, Science, **271**, 792–794, 1996.
15. S. B. Smith, Y. Cui, and C. Bustamante, *Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules*, Science, **271**, no. 5250, 795–799, 1996.
16. F. Lankas, J. Sponer, P. Hobza, and J. Langowski, *Sequence-dependent Elastic Properties of DNA*, J Mol Biol, **299**, no. 3, 695–709, 2000.

17. F. Lankas, Iii Te Cheatham, N. Spackova, P. Hobza, J. Langowski, and J. Sponer, *Critical effect of the n2 amino group on structure, dynamics, and elasticity of DNA polypurine tracts*, Biophys J, **82**, no. 5, 2592–609., 2002.
18. J. J. Delrow, J. A. Gebe, and J. M. Schurr, *Comparison of hard-cylinder and screened Coulomb interactions in the modeling of supercoiled DNAs*, Biopolymers, **42**, no. 4, 455–70, 1997.
19. M. Hammermann, N. Brun, K.V. Klenin, R. May, K. Toth, and J. Langowski, *Salt-dependent DNA superhelix diameter studied by small angle neutron scattering measurements and Monte Carlo simulations*, Biophys J, **75**, no. 6, 3057–3063, 1998.
20. K. V. Klenin, M. D. Frank-Kamenetskii, and J. Langowski, *Modulation of intramolecular interactions in superhelical DNA by curved sequences: a Monte Carlo simulation study*, Biophysical Journal, **68**, no. 1, 81–88, 1995.
21. V. V. Rybenkov, N. R. Cozzarelli, and A. V. Vologodskii, *Probability of DNA knotting and the effective diameter of the DNA double helix*, Proceedings of the National Academy of Sciences of the USA, **90**, 5307–5311, 1993.
22. V.V. Rybenkov, A.V. Vologodskii, and N.R. Cozzarelli, *The effect of ionic conditions on the conformations of supercoiled DNA .1. Sedimentation analysis*, J Mol Biol, **267**, no. 2, 299–311, 1997.
23. M. Hammermann, C. Steinmaier, H. Merlitz, U. Kapp, W. Waldeck, G. Chirico, and J. Langowski, *Salt effects on the structure and internal dynamics of superhelical DNAs studied by light scattering and Brownian dynamics*, Biophys J, **73**, no. 5, 2674–87, 1997.
24. H. Merlitz, K. Rippe, K. V. Klenin, and J. Langowski, *Looping dynamics of linear DNA molecules and the effect of DNA curvature: a study by Brownian dynamics simulation*, Biophysical Journal, **74**, no. 2 Pt 1, 773–779, 1998.
25. R. K. Sachs, G. van den Engh, B. Trask, H. Yokota, and J.E. Hearst, *A random-walk/giant-loop model for interphase chromosomes*, Proceedings of the National Academy of Sciences of the USA, **92**, 2710–2714, 1995.
26. C. Mnkel, R. Eils, S. Dietzel, D. Zink, C. Mehring, G. Wedemann, T. Cremer, and J. Langowski, *Compartmentalization of interphase chromosomes observed in simulation and experiment*, Journal of Molecular Biology, **285**, no. 3, 1053–1065, 1999.
27. K. Bystricky, P. Heun, L. Gehlen, J. Langowski, and S. M. Gasser, *Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques*, Proc Natl Acad Sci U S A, **101**, no. 47, 16495–500, 2004.
28. L. Ringrose, S. Chabanis, P. O. Angrand, C. Woodroofe, and A. F. Stewart, *Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances*, The EMBO Journal, **18**, no. 23, 6630–6641, 1999.
29. J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, *Capturing Chromosome Conformation*, Science, **295**, no. 5558, 1306–1311, 2002.
30. F. Aumann, F. Lankas, M. Caudron, and J. Langowski, *Monte Carlo simulation of chromatin stretching*, Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), **73**, no. 4, 041927, 2006.
31. C. L. Woodcock, S. A. Grigoryev, R. A. Horowitz, and N. Whitaker, *A chromatin folding model that incorporates linker variability generates fibers resembling the native structures*, Proceedings of the National Academy of Sciences of the USA, **90**, no.

19, 9021–9025, 1993.
32. K. van Holde and J. Zlatanova, *What determines the folding of the chromatin fiber*, Proceedings of the National Academy of Sciences of the USA, **93**, 10548–10555, 1996.
33. S. A. Grigoryev, *Keeping fingers crossed: heterochromatin spreading through interdigitation of nucleosome arrays*, Febs Letters, **564**, no. 1-2, 4–8, 2004.

Causality and Correlation Analyses of Molecular Dynamics Simulation Data

Adam Gorecki[1,2], **Joanna Trylska**[2], **and Bogdan Lesyng**[1]

[1] Department of Biophysics and CoE BioExploratorium, Faculty of Physics, University of Warsaw
Zwirki i Wigury 93, 02-089 Warsaw, Poland

[2] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
Pawinskiego 5A, 02-106 Warsaw, Poland
E-mail: {agorecki,joanna,lesyng}@icm.edu.pl

The subject of this study is the application and tuning of existing statistical analysis methods for molecular dynamics (MD) data analysis. Special attention is focused on detecting causality relationships (time precedence) between events in MD, based on time series derived from trajectories. The problems of time-series preprocessing, such as normalization and filtering, and the choise of the most appropriate causality detection method is discussed. Features and characteristics of two existing and widely applied methods: Directed Transfer Function[3] and conventional Granger causality approaches[2] are described. We suggest an adaptation of the conventional Granger method for MD analysis. The adapted Granger method is tested using the MD/SCC-DFTB simulation data of the proton transfer reaction in malonaldehyde[4] and a coarse-grained MD simulation of HIV-1 protease[5].

1 Introduction

Detecting causality relationships between conformational changes in biomolecular systems simulated with molecular dynamics (MD) methods is of crucial importance for describing their mechanisms and understanding the logic of their functioning. An attempt to approach this problem was presented in our recent study[1]. We followed the Granger causality methodology[2] and applied a Multi-Variate Autoregressive Model (MVAR) with Directed Transfer Function(DTF), which was used successfully in EEG time-series analyses[3]. However, the method still requires some tuning, and in this presentation we deal mostly with a conventional Granger approach[2]. We analyse also two following problems - normalization of the data and the noise filtering.

2 The Causality Analysis Model

Classical correlation analysis detects linear coupling between variables at the same time but it cannot detect linear couplings with a time shift or nonlinear couplings. One of the more advanced solutions is the MVAR model, which can detect time-shifted linear couplings:

$$\mathbf{X}(t) = \sum_{i=1}^{p} \mathbf{A}(i)\mathbf{X}(t-i) + \mathbf{E}(t) \tag{1}$$

where: $\mathbf{X}(t) = \{X_1(t), \ldots, X_k(t)\}$ - vector of analysed k variables at time t, called also channels; $t - i \equiv t - i \cdot dt$ - a notation for the time shift of i steps backward;

$\mathbf{A}(i),\ i = 1, \ldots, p$ - fitted MVAR coefficients, matrices of $k \times k$ dimension; p - model order; $\mathbf{E}(t)$ - white noise vector of k dimension.

The $\mathbf{A}(i)$ are fitted to satisfy condition (1) and keep $\mathbf{E}(t)$ components, $E_i(t)$, linearly uncorrelated, with proper mean and standard deviation. The standard deviations of $E_i(t)$ correspond to the white noise levels in each variable and determine $\mathbf{A}(i)$ estimation.

The raw MVAR coefficients are usually not representative for our purposes. Results of the MVAR fit usually depend on the model order (real signals do not satisfy a strict MVAR model), they can be also different for subsets of time series. In our purposes the MVAR model plays a role of a searching engine, not the system parametrization method.

Some time ago we tested[1] a more comfortable representation of the MVAR analysis namely, the Directed Transfer Method[3]. This method was designed for EEG analysis and it is based on the frequency representation of signal transmission, well optimized for linear systems with a clear linear filter interpretation, and noise level independent. However, the method requires variables of the same units and a similar signal type (such as electric potentials recorded as EEG signals). It is sensitive to scaling variables and gives ambiguous results for variables expressed in different units. This problem is very important in MD data analysis because MD simulation observables, such as distances, angles, combinations of different degrees of freedom, energies, etc., have different units. Normalization of variables is connected with choosing the appropriate noise level for the rescaled variables, which determines the MVAR fit.

In this study we test an older, but equivalent to DTF, model: the conventional Granger causality[2] approach. This method is based on comparing of the MVAR fit error *with* and *without* selected variable information. To estimate causal influence of X_j variable on variable X_i, we should select MVAR-model order p, and perform the following MVAR fits.

From the fit for full variables set $\mathbf{X}(t) = \{X_1(t), \ldots, X_k(t)\}$

$$\mathbf{X}(t) = \sum_{i=1}^{p} \mathbf{A}(i)\mathbf{X}(t-i) + \mathbf{E}(t), \tag{2}$$

we compute the residual variance matrix $\mathbf{V} = \langle \mathbf{E}^T(t)\mathbf{E}(t) \rangle$.
From the fit for variables set with X_j excluded:

$$\mathbf{X}^{(j)}(t) = \{X_1(t), ..., X_{j-1}(t), X_{j+1}(t), ..., X_k(t)\}$$

$$\mathbf{X}^{(j)}(t) = \sum_{i=1}^{p} \mathbf{A}^{(j)}(i)\mathbf{X}^{(j)}(t-i) + \mathbf{E}^{(j)}(t), \tag{3}$$

we compute the residual variance matrix: $\mathbf{V}^{(j)} = \langle (\mathbf{E}^{(j)}(t))^T \mathbf{E}^{(j)}(t) \rangle$.
Note that $\dim \mathbf{V} = k \times k$ and $\dim \mathbf{V}^{(j)} = k - 1 \times k - 1$.
The causality measure for $X_j \to X_i$ direction is defined as:

$$J_{ij} = 1 - \frac{V_{ii}}{V_{ii}^{(j)}} \in [0; 1]. \tag{4}$$

The J_{ij} matrix is usually asymmetric and represents the strength of a delayed linear coupling between pairs of variables $X_j \to X_i$, where 0 corresponds to no coupling, 1 - to X_i fully determined by X_j with the linear relationship.

The conventional Granger method results are not sensitive to rescaling and unit choice, which results from the construction of this method. This feature is very important for our applications. Optional renormalization can be applied for better numerical MVAR fitting convergence and accuracy.

3 The Conventional Granger Method - Examples of Applications

3.1 Malonaldehyde Molecular Dynamics Trajectory

Malonaldehyde molecule is shown in Fig.1a. The analysed trajectory was derived from a combined quantum/classical dynamics of the proton transfer between O1 and O2 oxygen atoms[4], and contains 1000000 observations probed every 10 fs (every 10 steps of dynamics). Previous analyses show that the proton transfer is a cooperative reaction and the

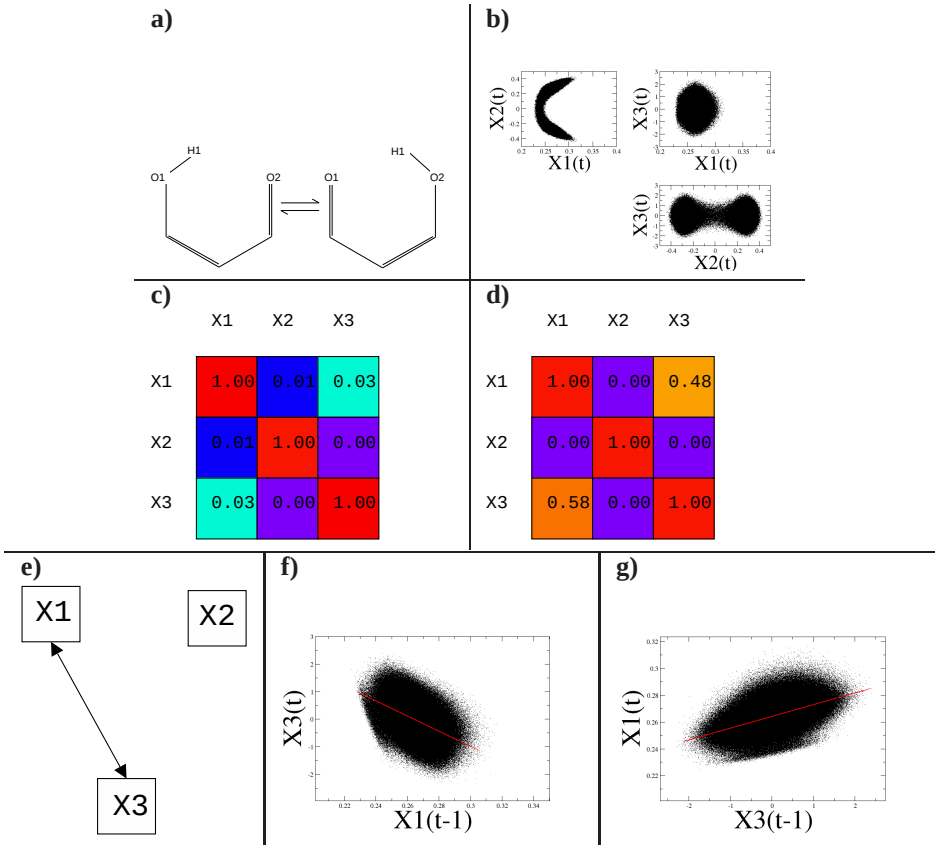


Figure 1. Malonaldehyde example: a) the malonaldehyde molecule; b) relationships between values of X_1, X_2, X_3 variables in the same time t; c) classical correlation matrix for the X_1, X_2, X_3 variables defined in Eqn.5 ; d) the Granger causality matrix J_{ij} for the X_1, X_2, X_3; e) causality diagram corresponding to J_{ij} with the shown feedback $X_1 \leftrightarrow X_3$; f,g) relationships between values of the X_1, X_3 variables with the time shift ± 1.

required condition for the proton hopping is the small O1 and O2 distance. We selected following variables for the causality search:

$$X_1 = |O_1O_2|,\ X_2 = \frac{|HO_1| - |HO_2|}{|HO_1| + |HO_2|},\ X_3 = (v_{\vec{O2}} - v_{\vec{O1}}) \cdot \vec{O_1O_2}/|O_1O_2| \quad (5)$$

The X_2 variable is called the „reaction coordinate" and in our example describes the relative proton position. The X_3 variable is the projection of relative velocities of O1 and O2 atoms on the $\vec{O_1O_2}$ direction. Simple correlation matrix (Fig.1c) does not show any interesting correlation because the variables measured in the same moment are either independent or couplings are nonlinear (Fig. 1b). The Granger method applied to malonaldehyde data shows strong bidirectional causality relationship (feedback) between variables X_1 and X_3. It is an example of existing of non-instant correlations between the spatial and velocity degrees of freedom.For automatic detection of the X_1 and X_2 couplings, we need to apply nonlinear and/or instant causality extensions of the conventional Granger approach.

3.2 HIV-1 Protease Molecular Dynamics Trajectory

We have analysed the coarse-grained dynamics trajectory of HIV-1 protease (Fig.2a) performed in the NVE ensemble[5]. The trajectory contained 10000 frames, delayed by 10 ps. The conformation of the protein was represented by reduced variables: 10 PCA projections derived from the Essential Dynamics method[7]. No instant linear correlations between the variables were seen in this parametrization because they are already included in the PCA eigenvectors. The Granger method detected only weak couplings (Fig.2b) and for the analysis we have selected those with $J_{ij} > 0.1$ (diagram in Fig.2c). J_{16} is the largest element in the first row, which indicates the $X_6 \rightarrow X_1$ coupling. This coupling is strongly nonlinear (Fig.2d) and difficult to detect by linear methods, but the Granger method detected it as a result of some linear correlation. The coupled PCA motions are shown in Fig.2ef - the eigenvector corresponding to X_1 is the most significant movement which characterizes the flap opening. The second one, corresponding to X_6, is also a component of flap movement. Both components change distance between the 17 and 39 residues of two symmetric chains. The movement in X_6 direction is preceding the X_1 flap opening movements, but the Granger method detects it as not very important statistically. A similar correlation was described in the cited articles[5,6].

4 The Influence of Smoothing

Smoothing algorithms (e.g. Savitzky-Golay filter[9]), are usually based on linear filters. The output signal at time t is a linear combination of input values over a window of times: $X_i'(t) = \sum_{k=-k_0}^{k=+k_0} c(k) X_i(t+k)$. Linear filtering of the signal obviously interferes with the MVAR model; the MVAR model will detect filter parameters when run on a too densely probed data. In some cases, filtering can remove some noise from the signal, and can help in detecting of the expected couplings. Sampling of the data for the MVAR analysis should have lower density than the window size used for the preceding smoothing operation.

We applied smoothing of the data by Savitzky-Golay polynomial filter[9], with $k_0 = 4$ (then window was 9 frames long) and order of polynomial 2, then probed data every 10 steps. This operation doesn't improve causal relations visibility.

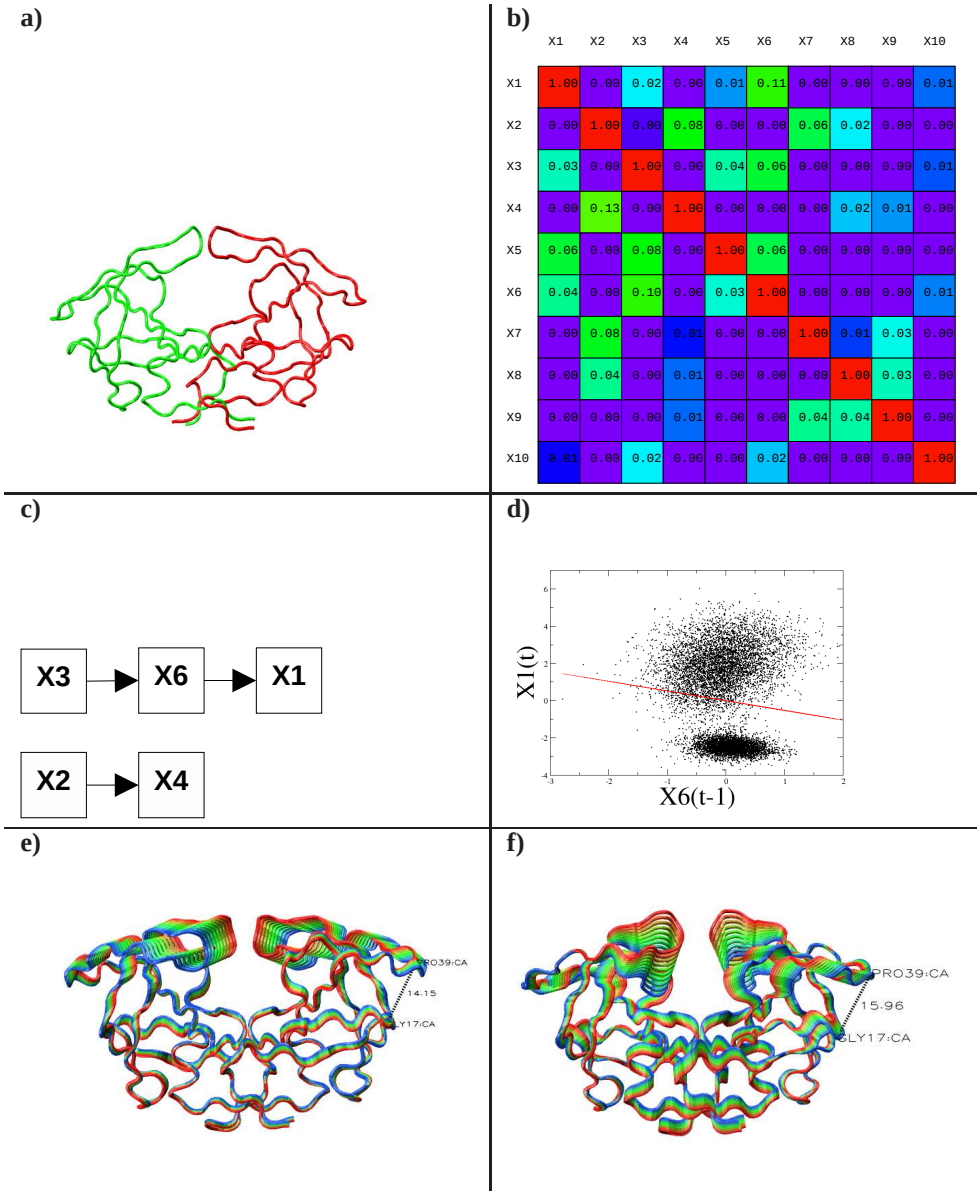


Figure 2. HIV-1 protease example: a) the ribbon model of HIV-1 PR homodimer; b) the Granger causality matrix J_{ij} for the $X_1, \ldots, X_{10}$ variables (reduced coordinates which result from the projection of the MD coordinates on ten most significant PCA components); c) causality diagram corresponding to J_{ij} which shows causality relation for $J_{ij} \geq 0.1$; d) the relationship $X_6 \rightarrow X_1$ with the time shift of 1 e,f) the PCA eigenvectors (movement directions) corresponding to the X_1 and X_6 projections, respectively. Motions occur from blue to red, and back.

5 Conclusions

We have been developing and applying two versions of the causality analyses: the MVAR/DTF and the conventional Granger approaches. MVAR/DTF was presented in[1]. This study deals mostly with the conventional Granger approach. We analysed MD simulation data of malonaldehyde and HIV-1 protease.

The conventional Granger method based on the Multi-Variate Autoregression Model is a quite efficient tool for the molecular dynamic data analysis, because it is independent on normalization and can be applied for signal channels characterized with different units. We have been developing the generalization of this method for detecting of non-linear couplings. For the studied examples prefiltering of the data before the MVAR analysis did not improve the sensitivity of the method in detecting the expected couplings. It interferes with the MVAR and should be use with some care.

Acknowledgments

We would like to acknowledge R. KUS from the Department of Biomedical Physics for providing his very fast MVAR calculation software, L. WALEWSKI of the Department of Biophysics of Warsaw University for providing the molecular dynamics trajectories of proton transfer in the malonaldehyde molecule, V.TOZZINI for providing the molecular dynamics trajectories of HIV-1 protease. We are also grateful to M.KAMINSKI, R.KUS and K.BLINOWSKA from the Department of Biomedical Physics of Warsaw University for helpful discussions and consultations on the applications of the MVAR and DTF methods. The authors were supported by Warsaw University (115/E-343/ICM/BST-1076/2005). JT acknowledges support from University of Warsaw (115/30/E-343/S/2007/ICM BST 1255), Polish Ministry of Science and Higher Education (3 T11F 005 30, 2006-2008), Fogarty International Center (NIH Research Grant # R03 TW07318) and Foundation for Polish Science. AG thanks Polish Ministry of Science and Higher Education (N202 079 32/1841,2007-2008).

References

1. A. Gorecki, J. Trylska, B. Lesyng, Europhys. Let., 2006, 75, 503-509.
2. C. W. J. Granger, Econometrica, 1969, 37, 424-438.
3. K.J. Blinowska, R. Kus, M.J. Kaminski, Phys. Rev. E, 2004, 70, 050902.
4. L. Walewski, P. Bala, M. Elstner, Th. Frauenheim, B. Lesyng, Chem. Phys. Let., 2004, 397, 45 1-458.
5. V. Tozzini, J. Trylska, Chia-en Chang, J. A. McCammon, J. Struct. Biol., 2007, 157, 606-615.
6. V. Tozzini, J. A. McCammon, Chem. Phys. Lett., 2005, 413:123-128.
7. A.Amadei, A.B.M.Linssen, H.J.C.Berendsen, Proteins:Struct.Funct.Genet., 1993, 17, 412.
8. R. Hegger, H. Kantz, and T. Schreiber, CHAOS 9,1999 , 413 .
9. Numerical Recipes, chapter 14.8, `http://www.nr.com`.

Mesoscopic Dynamics with the UNRES Force Field – a Tool for Studying the Kinetics and Thermodynamics of Protein Folding

Adam Liwo[1,2], Cezary Czaplewski[1,2], Stanisław Ołdziej[1,2], Maurizio Chinchio[2], Ana V. Rojas[2,3,4], Mey Khalili[5], Mariusz Makowski[1,2], Sebastian Kalinowski[1], Urszula Kozłowska[2], Rajesh K. Murarka[2], and Harold A. Scheraga[2]

[1] Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland
E-mail: {adam, czarek, stan, momo, bes}@chem.univ.gda.pl

[2] Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14843-1301, USA
E-mail: {m.chinchio, uad2, rkm25, has5}@cornell.edu

[3] Department of Physics and Astronomy and

[4] Center for Computation and Technology, Louisiana State University,
Baton Rouge, LA 70803-4001, USA
E-mail: arojas1@lsu.edu

[5] University Chemical Laboratories, Lensfield Road, Cambridge, UK CB2 1EW
E-mail: mk520@cam.ac.uk

All-atom simulations of protein folding starting from arbitrary structures are currently possible only for small proteins and, consequently, united-residue models of polypeptide chains are used in the field. The UNRES model and the respective force field developed in our laboratory belong to this class. As opposed to knowledge-based or heuristic coarse-grained force fields, UNRES was carefully derived as a restricted free energy function of a polypeptide chain, in which fine degrees of freedom not included in the model have been integrated out. This article summarizes recent developments and applications of the force field.

1 Introduction

United-residue or mesoscopic models of proteins receive considerable attention because, in contrast to all-atom simulations, they enable us to study the folding of large proteins at micro- or millisecond time scale. Most of the existing approaches are knowledge-based, which makes explicit use of the information from protein structural databases during the process of conformational search. These approaches are very successful in predicting the structure of proteins[1] but their application to study folding dynamics is limited.

During the last few years we have been developing a mesoscopic physics-based force field termed UNRES (UNited RESidue)[2]. Its initial applications were energy-based predictions of protein structure defined as the lowest-energy conformation[2,3]. Despite considerable success in this field, we realized that such energy-based prediction does not correspond to the process of formation of the native structure, which, following Anfinsen's thermodynamic hypothesis[4] is a collection of very similar conformations occupying the basin with the lowest free energy and, consequently, we changed the approach to ensemble-based global optimization. We also extended the application of UNRES to study the thermodynamics, dynamics, and kinetics of protein folding.

2 The UNRES Force Field

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, as shown in Figure 1.

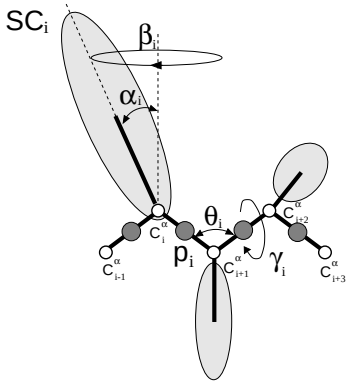


Figure 1. The UNRES model of the polypeptide chain. Dark circles represent united peptide groups (*p*), open circles represent the C^{α} atoms, which serve as geometric points. Ellipsoids represent side chains, with their centers of mass at the SC's. The p's are located half-way between two consecutive C^{α} atoms. The virtual-bond angles θ, the virtual-bond dihedral angles γ, and the angles α_{SC} and β_{SC} that define the location of a side chain with respect to the backbone are also indicated.

The effective energy function is a sum of different terms corresponding to interactions between the SC ($U_{SC_iSC_j}$), SC and p ($U_{SC_ip_j}$), and p ($U_{p_ip_j}$) sites, as well as local terms corresponding to bending of virtual-bond angles θ (U_b), side-chain rotamers (U_{rot}), virtual-bond torsional (U_{tor}) and double-torsional (U_{tord}) terms, virtual-bond-stretching (U_{bond}) terms, correlation terms ($U_{corr}^{(m)}$) pertaining to coupling between backbone-local and backbone-electrostatic interactions[5] (where m denotes the order of correlation), and a term accounting for the energetics of disulfide bonds (U_{SS}). Each of these terms is multiplied by an appropriate weight, w. The energy function is given by equation 1.

$$\begin{aligned} U = & w_{SC}\sum_{i<j} U_{SC_iSC_j} + w_{SCp}\sum_{i\neq j} U_{SC_ip_j} + w_{pp}\sum_{i<j-1} U_{p_ip_j} \\ & + w_{tor}\sum_i U_{tor}(\gamma_i) + w_{tord}\sum_i U_{tord}(\gamma_i,\gamma_{i+1}) \\ & + w_b\sum_i U_b(\theta_i) + w_{rot}\sum_i U_{rot}(\alpha_{SC_i},\beta_{SC_i}) + \sum_{m=3}^{6} w_{corr}^{(m)} U_{corr}^{(m)} \\ & + w_{bond}\sum_{i=1}^{nbond} U_{bond}(d_i) + w_{SS}\sum_i U_{SS;i} \end{aligned} \tag{1}$$

The prototype of equation 1 is the restricted free energy (RFE) corresponding to a given coarse-grained conformation of a polypeptide chain in water in which all secondary

degrees of freedom, such as the solvent coordinates, internal rotation angles of the side chains, and the angles for the rotation of the peptide groups about the $C^{\alpha} \cdots C^{\alpha}$ virtual bonds have been integrated out[5]. We expressed the RFE[5] in terms of Kubo's cluster cumulants[6] and the cluster cumulants into a generalized cumulant series. The fact that the UNRES energy function has the sense of a free energy implies its dependence on temperature, which we introduced recently[7] by multiplying the energy-term weights by appropriate temperature factors.

3 Search of Conformational Space with UNRES

In our earlier implementations of UNRES[2], we were looking for the global minimum of the energy. We developed a global-optimization tool termed a Conformational Space Annealing (CSA) method[8], based on the framework of genetic algorithms. This approach produces the lowest energy minima and can be identified with finding a "folded" structure at the temperature of 0 K. This approach ignores the conformational entropy and, consequently, some proteins whose native-like structures were found as lowest in energy in CSA searches failed to fold to the native structures in canonical MD simulations with the UNRES force field[9].

To obtain a protocol closer to the real folding process, we developed[9] mesoscopic molecular dynamics, in which the UNRES energy function (equation 1) plays the role of the potential energy; we refer to this approach as UNRES/MD. We use either the equations of motion with explicit stochastic and friction terms to account for non-conservative forces coming from the solvent (Langevin dynamics; appropriate to study folding dynamics) or couple the Newton-like equations of motion with the Berendsen thermostat, which results in faster search of the conformational space. Recently[10], we extended the UNRES/MD approach to study the folding of multichain proteins and also developed a procedure to handle dynamic formation and breaking of disulfide bonds during UNRES/MD runs[11]. For better search of the conformational space we implemented[12] the multiplexing replica-exchange method[13] in UNRES/MD (this algorithm is referred to as UNRES/MREMD). We parallelized UNRES/MREMD to run on sytems with 1,000 processors or more; an example of scalability is shown in Figure 2.

4 Applications of UNRES

4.1 Folding Dynamics

UNRES/MD provides an about 4000- and 200-fold speed-up compared with all-atom MD with explicit and implicit solvent, respectively[9], thus making *ab initio* folding simulations possible. We studied[9] the folding of a number of α-helical proteins and an $\alpha + \beta$-protein (1E0G). With the version of the UNRES available at that time the simulated folding of the α-helical proteins followed the diffusion-collision model, while the folding of 1E0G started from the formation of all-α-helical structures with subsequent approach, straightening, and final packing of the ends into a two-stranded β-sheet.

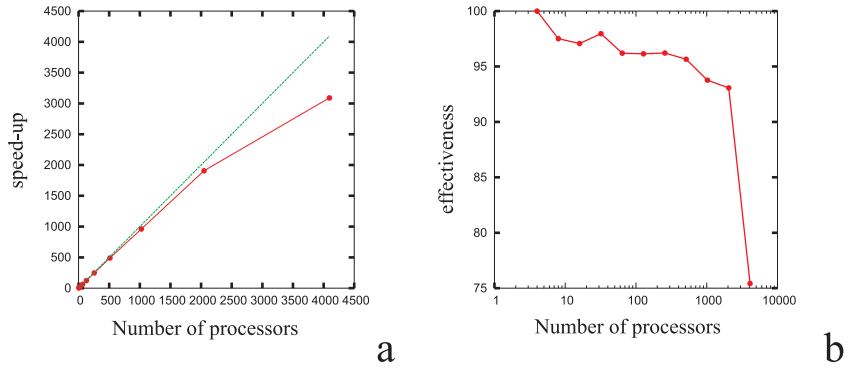


Figure 2. (a) Speedup of the UNRES/REMD algorithm as a function of the number of processors (red solid line and circles) compared with the theoretical line corresponding to 100 % speedup (green dashed line). (b) Efficiency (speedup/number of processors) vs. the number of processors for the same system. The protein studied was 1SAP proteins and the simulations were run on the Jülich Blue Gene supercomputer.

4.2 Folding Kinetics

The advantage of parallel computing combined with the speed up offered by UNRES compared to an all-atom approach enables us to run multiple Langevin trajectories simultaneously from which information about the mechanism and kinetic equations can be extracted. We carried out such a study[14] on the N-terminal fragment of the B-domain of staphylococcal protein A (a 46-residue protein with a three-helix-bundle structure). We found two folding routes: a fast one proceeding directly to the native structure and a slow one passing through an intermediate with mispacked α-helices. We also found that the variation of the fraction of native structure obeys biexponential kinetics, which is in full agreement with the presence of two folding routes.

4.3 Folding Thermodynamics

UNRES/MREMD is a robust tool to compute the thermodynamical and structural characteristics of proteins at various temperatures and, thereby, to determine the thermodynamics of protein folding. We implemented[7] the weighted histogram analysis method (WHAM)[15] method to process the results of MREMD simulations. The computed curves of heat capacity and ensemble-averaged native-likeness (e.g., RMSD from the experimental structure) are good measures of the quality of the force field[7, 12].

4.4 Prediction of Protein Structure

We define the native structure as the most probable conformational ensemble at a temperature below that of the folding transition. We developed a protocol[7] in which we run UNRES/MREMD simulations, then determine the heat-capacity curve and, finally, run a cluster analysis and select the clusters with the greatest probability at a temperature below the folding-transition temperature. Figure 3 shows the results of the implementation of this protocol to predict the structure of target T0300 in the CASP7 blind-prediction test.

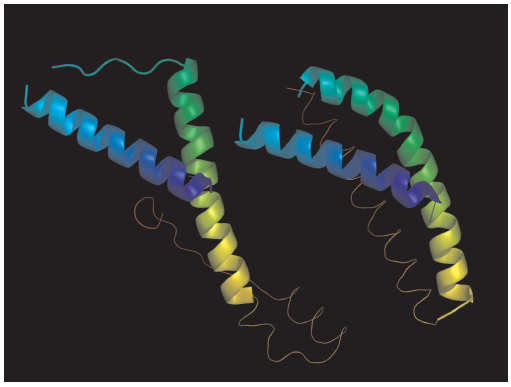


Figure 3. Comparison of the experimental (left) and predicted structure of the 2–65 fragment of the CASP7 target T0300 (RMSD=6 Å).

5 Concluding Remarks

The UNRES model has two advantages: it is a coarse-grained model with which to study protein folding in real time and is based on clear physical principles. The present UNRES force field is still in the process of tuning but can already be used to carry out reliable prediction of protein structure, folding thermodynamics, and folding kinetics. Further applications include studying large-scale motions of proteins and protein complexes and membrane proteins.

Acknowledgments

This work was supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB05-41633), the NIH Fogarty International Center (TW7193), and grant DS 8372-4-0138-7 from the Polish Ministry of Science and Higher Education. This research was conducted by using the resources of (a) our 800-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) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Jülich, Germany, (d) the Beowulf cluster at the Department of Computer Science, Cornell University, (e) the resources of the Center for Computation and Technology at Louisiana State University, which is supported by funding from the Louisiana legislature (f) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, (g) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (h) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

References

1. J. Moult. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Phil. Trans. R. Soc. B*, 361:453–458, 2006.

2. H. A. Scheraga, A. Liwo, S. Ołdziej, C. Czaplewski, J. Pillardy, D. R. Ripoll, J. A. Vila, R. Kaźmierkiewicz, J. A. Saunders, Y. A. Arnautova, A. Jagielska, M. Chinchio, and M. Nanias. The protein folding problem: global optimization of force fields. *Frontiers in Bioscience*, 9:3296–3323, 2004.
3. S. Ołdziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kaźmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y.-K. Kang, K. D. Gibson, and H. A. Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field – test with CASP5 and CASP6 targets. *Proc. Natl. Acad. Sci. U.S.A.*, 102:7547–7552, 2005.
4. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.
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. R. Kubo. Generalized cumulant expansion method. *J. Phys. Soc. Japan*, 17:1100–1120, 1962.
7. A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H.A. Scheraga. Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. *J. Phys. Chem. B*, 111:260–285, 2007.
8. 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.
9. A. Liwo, M. Khalili, and H. A. Scheraga. Molecular dynamics with the united-residue (UNRES) model of polypeptide chains; test of the approach on model proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2362–2367, 2005.
10. A. V. Rojas, A. Liwo, and H.A. Scheraga. Molecular dynamics with the united-residue force field: Ab initio folding simulations of multichain proteins. *J. Phys. Chem. B.*, 111:293–309, 2007.
11. M. Chinchio, C. Czaplewski, A. Liwo, S. Ołdziej, and H. A. Scheraga. Dynamic formation and breaking of disulfide bonds in molecular dynamics simulations with the UNRES force field. *J. Chem. Theor. Comput.*, 2007, submitted.
12. M. Nanias, 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.
13. Y. M. Rhee and V. S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys J.*, 84:775–786, 2003.
14. 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.
15. 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, 1992.

Mixed QM/MM Calculations in Biological Systems

Gustavo de M. Seabra[1], **Juan Torras**[2], **Ross C. Walker**[3], **and Adrian Roitberg**[1]

[1] Quantum Theory Project, University of Florida
P.O. Box 118435, Gainesville, FL 32611-8435, USA
E-mail: roitberg@ufl.edu

[2] Departament d'Enginyeria Química, EUETII, Universitat Politècnica de Catalunya
Pça. Rei 15, 08700-Igualada, SPAIN

[3] San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, La Jolla, CA, 92093-0505, USA

The computational study of a variety of important processes, such as processes involving bond breaking and forming and electron reorganization, requires the application of quantum mechanical methods. In biological systems, this situation is further complicated because the influence of the environment must be taken into account. In these situations, hybrid quantum mechanics / molecular mechanics methods can be used. We describe here two new QM/MM implementations based on the Amber Molecular Dynamics packages that make it possible to use of a variety of quantum mechanical methods with any of the techniques available in Amber. Results are provided for the application of these implementations to two different systems.

1 Introduction

Due to the large system size, computational studies of biologically relevant systems such as proteins, nucleic acids and carbohydrates need to resort to approximated Molecular Mechanics (MM) methods which apply parameterized force fields to describe molecular properties such as bond lengths, angles, dihedrals, electrostatic and van der Waals forces. The use of such parameterized methods greatly reduces the computational complexity, allowing the study of processes including ligand binding, enzyme reaction mechanisms, protein folding, refolding, and denaturation, providing invaluable help in the analysis of complex experimental data and structures.

Despite continuous efforts to develop more reliable force fields for use in MM calculations, including the use of QM calculations and genetic algorithms in the parameterization of the force field, classical mechanics methods lack the ability to treat fundamentally quantum processes, such as bond breaking and forming and charge fluctuations as a function of geometry, or to describe parts of the potential energy surface far from equilibrium. In some cases, although computationally expensive, it is possible to treat a model system purely by QM methods, but the effect of the environment must be either neglected or simulated by a continuum dielectric approximation. An alternative that allows the explicit inclusion of environment effects while treating the most relevant part of the system with full quantum mechanics was first explored by Warshel and Levitt as early as 1976,[1] and is the use of hybrid quantum mechanics / molecular mechanics (QM/MM) calculations whereby a subsection of the system is treated by QM methods, the remainder (environment) is treated by standard molecular mechanics (MM) methods, and a coupling potential is used to connect the two regions.

Recent work in our group has focused in the implementation and application of hybrid QM/MM techniques for the study of biologically relevant systems. In particular, discussed here are two recently developed interfaces: the new native, (semi-empirical) QM/MM support available in Amber 9,[2] and the integration of the Amber MD program with a QM program through the PUPIL interface. These interfaces offer the advantage of blending seamlessly with the Amber program, allowing the application of any of the advanced sampling methods available in Amber to QM/MM problems.

2 QM/MM

In a hybrid QM/MM calculation, the system is partitioned into two regions: A QM region, typically consisting of a relatively small number of atoms relevant for the specific process being studied, and a MM region with all the remaining atoms. The total Hamiltonian ($\hat{H}$) for such a system is written as:

$$\hat{H} = \hat{H}^{QM} + \hat{H}^{MM} + \hat{H}^{QM/MM}, \tag{1}$$

where $\hat{H}^{QM}$ and $\hat{H}^{MM}$ are the Hamiltonians for the QM and MM parts of the system, and are calculated using either the QM method chosen or the usual force field equations, respectively. The remaining term, $\hat{H}^{QM/MM}$, describes the interaction between the QM and MM parts and typically contains terms for electrostatic, van der Waals and bonded interactions across the region boundaries:

$$\hat{H}^{QM/MM} = \hat{H}^{QM/MM}_{vdW} + \hat{H}^{QM/MM}_{elect} + \hat{H}^{QM/MM}_{bonds}. \tag{2}$$

In the approach used for both Amber QM/MM interfaces presented here, the van der Waals ($\hat{H}^{QM/MM}_{vdW}$) is treated as usual by the MM program, using the standard 12-6 Lennard-Jones parameters from the force field in use. It has been shown that this choice does not introduce significant errors in the calculation.[3] Also, both interfaces described here use a link atom scheme if bonds are broken across the QM–MM boundary. Finally, the remaining term ($\hat{H}^{QM/MM}_{elect}$), describing the electrostatic interaction between the MM and QM zones, depends on the specific interface.

3 Semi-Empirical QM/MM in Amber

One of the most used programs for MD simulations is the *sander* program, part of the Amber suite. The *sander* interface for QM/MM has been recently rewritten, placing strong emphasis on speed and accuracy, allowing the simulation of systems with a reasonably sized quantum mechanical region (around 300 atoms) for long (nanoseconds) timescales. It now natively includes a number of semi-empirical Hamiltonians for QM/MM simulations including MNDO,[4,5] AM1,[6] PM3,[7,8] PDDG/PM3[9] and SCC-DFTB.[10] Our group has been involved in the implementation of the SCC-DFTB method.

The Self-Consistent-Charge Density-Functional Tight Binding (SCC-DFTB) is an approximate method based on Density Functional Theory (DFT) from a second order expansion of the DFT total energy in the charge density fluctuations, and has been described in detail elsewhere.[11,10] It has been successfully applied to the study of a variety of systems, and has been shown to yield results comparable in accuracy to *ab-initio* MP2 calculations

with large basis sets.[12] (For a recent review on applications of SCC-DFTB to biological systems, the reader is referred to Ref.[13].)

The integration of SCC-DFTB with Amber 9 has been described recently.[14] In this implementation, the electrostatic interaction energy between the QM and MM regions ($\hat{H}_{elect}^{QM/MM}$ in Equation 2) is calculated as a Coulomb interaction between the atomic Mulliken charges calculated from the SCC-DFTB (q_α) and the classical RESP charges on the MM atoms (Q_A), as in Equation 3.

$$E_{elect}^{QM/MM} = \sum_{\alpha}^{QM} \sum_{A}^{MM} \frac{Q_A \, q_\alpha}{r_{\alpha A}}. \tag{3}$$

Showing how the QM/MM implementation can directly tap into the methods in Amber, the free energy surface (FES) for the capped alanine dipeptide (ACE-ALA-NME, AD) in vacuum was generated using Replica Exchange Molecular Dynamics,[15] with the AD treated by SCC-DFTB. 6 replicas were used, at temperatures of 161.2K, 219.9K, 300.0K, 419.3K, 558.4K and 761.8K. Exchanges were attempted 10,000 times, with 0.5ps between attempts, and a time step of 1fs with SHAKE was used. The Langevin thermostat with a

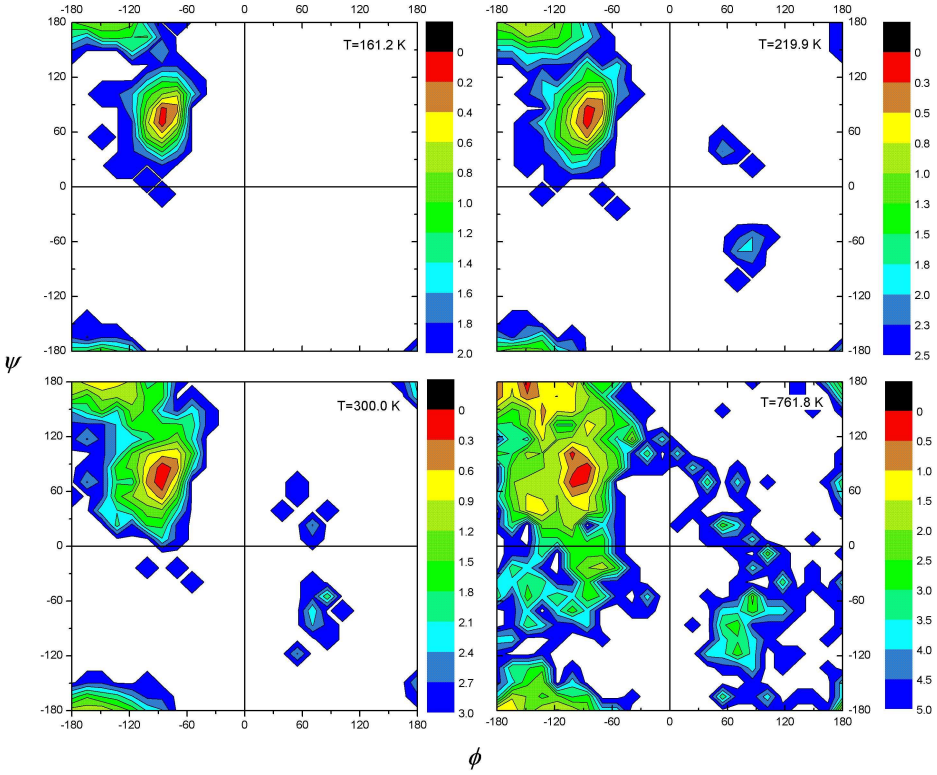


Figure 1. Free Energy Surfaces for alanine dipeptide obtained from the SCC-DFTB/REMD calculation.

collision frequency of 2.0ps^{-1} was used to regulate the temperatures.

Figure 1 shows the FES for four different temperatures, as calculated from $G = -RT\ln(P)$, where G is the Gibbs free energy, R is the gas constant and T is the temperature, and P is the (normalized) probability P of finding the AD in a conformation at a particular region in (ϕ,ψ)-space from the MD trajectories. The surfaces show a clear minimum around ϕ/ψ values of ($-83°$,$76°$), which corresponds to the known C7$_{eq}$ minimum for AD in vacuum, and other structures are populated as temperature increases. The relative energies compare well with calculated *ab-initio* values calculated previously.[16]

4 Amber-Gaussian QM/MM Through the PUPIL Interface

The interface described in the previous session has the advantage of being convenient and easy to use, especially for users experienced with the Amber MD program. At the current stage of development, however, the QM region is fixed throughout the calculation, and limited to semi-empirical Hamiltonians.

The PUPIL (Program for User Program Interfacing and Linking) is a free, open source package created to facilitate the interfacing of *arbitrary* MD and QM programs.[17,18] Currently, interfaces with DLPOLY, MNDO97 and Siesta have been developed. We have

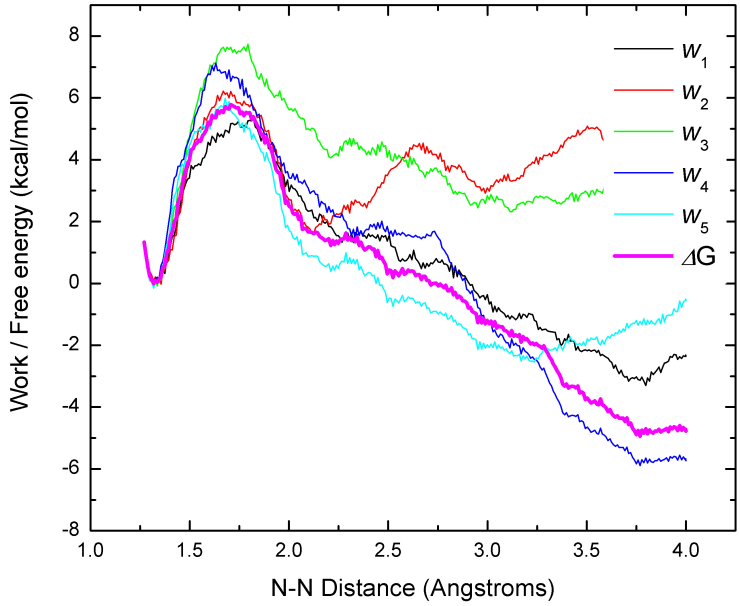


Figure 2. Work of breaking the N-N bond in Angeli's salt for 5 different simulations, and the free enrgy obtained by the Jarzynsky equation.

recently included interfaces with *sander* and Gaussian03 to the PUPIL system. In PUPIL, the equations of motion are propagated as usually by the MD program and the forces over the QM region are provided by the QM program, while the PUPIL interface manages execution of the two programs, as well as the required information flow. The influence of the MM atoms on the QM zone is taken into account by electronic embedding, where the MM atoms are considered as point charges fixed at the positions of the MM atoms, and with the charge values from the force field parameters in use.

In the Gaussian03 interface, the force contribution from the quantum atoms to the total force acting upon the classical atoms (F_i^{QM}) is obtained by dividing the electronic density into a grid, and calculating the interaction of the MM charge with each point in the grid:

$$F_i^{QM} = \sum_j^{cube} \mathbf{r_{ij}} \frac{q_{PC_i}\, dq_j}{|r_{ij}|^3}, \tag{4}$$

where

$$dq_j = \rho_j\, dx\, dy\, dz. \tag{5}$$

As an example, the free energy change associated to a specific process can be calculated using the Jarzynski relationship:[19–21]

$$e^{-\beta \Delta G(d)} = \left\langle e^{-\beta W(d)} \right\rangle_{d=d_\circ}, \tag{6}$$

where the brackets indicate an average taken over a large number of independent realizations of the process (in this case, molecular dynamics trajectories) all starting at different points belonging to the same equilibrated ensemble.

Figure 2 shows work results five trajectories breaking the N-N bond in the monoprotonated Angeli's salt (O2N-NHO$^-$) in explicit water, together with the free energy calculated using Equation 6. The Angeli's salt was treated quantum mechanically by Gaussian03 at the UB3LYP/6-311+G(d) level, and put in a large box of TIP3P water molecules. The free energy barrier for breaking the N-N bond is calculated to be 5.84 kcal/mol, which is in good agreement with previous estimations using B3LYP with the same basis set and implicit solvation.[22]

5 Conclusions

This communication describes and exemplifies two new QM/MM implementations. Both have the advantage of being based on the Amber molecular dynamics package, allowing their application in tandem with any of the advanced sampling methods available in Amber. Results of calculations using both implementations are provided.

Acknowledgments

The authors want to acknowledge Erik Deumens and S.B. Trickey for their part in the initial development of the PUPIL package. Computer resources were provided by the Large Allocations Resource Committee through TeraGrid grant TG-MCA05S010 to AER, and

the TeraGrid grant TG-CHE060072T to GS and JTC, the High-Performance Computing Center at the University of Florida and a local cluster built with a generous IBM-SUR grant. RCW acknowledges NIH Grant GM57513 for funding this work. RCW and AER would like to thank the San Diego Supercomputer Center for their continued support of this project and future developments in Amber 10 through their Strategic Applications Collaboration program.

References

1. A. Warshel and M. Levitt, *Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme*, Journal of Molecular Biology, **103**, no. 2, 227–249, 1976.
2. "Amber 9", `http://amber.scripps.edu`.
3. Demian Riccardi, Guohui Li, and Qiang Cui, *Importance of van der Waals Interactions in QM/MM Simulations*, Journal of Physical Chemistry B, **108**, 6467–6478, 2004.
4. Michael J. S. Dewar and Walter Thiel, *Ground states of molecules. 38. The MNDO method. Approximations and parameters*, J. Am. Chem. Soc., **99**, no. 15, 4899–4907, 1977.
5. Michael J. S. Dewar and Walter Thiel, *Ground states of molecules. 39. MNDO results for molecules containing hydrogen, carbon, nitrogen, and oxygen*, J. Am. Chem. Soc., **99**, no. 15, 4907–4917, 1977.
6. Michael J. S. Dewar, Eve G. Zoebisch, Eamonn F. Healy, and James J. P. Stewart, *Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model*, J. Am. Chem. Soc., **107**, no. 13, 3902–3909, 1985.
7. James J. P. Stewart, *Optimization of parameters for semiempirical methods I. Method*, Journal of Computational Chemistry, **10**, no. 2, 209–220, 1989.
8. James J. P. Stewart, *Optimization of parameters for semiempirical methods II. Applications*, Journal of Computational Chemistry, **10**, no. 2, 221–264, 1989.
9. Matthew P. Repasky, Jayaraman Chandrasekhar, and William L. Jorgensen, *PDDG/PM3 and PDDG/MNDO: Improved semiempirical methods*, Journal of Computational Chemistry, **23**, no. 16, 1601–1622, 2002.
10. M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and Gotthard Seifert, *Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties*, Physical Review B: Condensed Matter, **58**, no. 11, 7260–7268, 1998.
11. Thomas Frauenheim, D. Porezag, Marcus Elstner, G. Jungnickel, J. Elsner, M. Haugk, A. Sieck, and Gotthard Seifert, *An ab initio two-center tight binding approach to simulations of complex materials properties*, Mat. Res. Soc. Symp. Proc., **491**, 91–104, 1998.
12. H. G. Bohr, K. J. Jalkanen, M. Elstner, K. Frimand, and S. Suhai, *A comparative study of MP2, B3LYP, RHF and SCC-DFTB force fields in predicting the vibrational spectra of N-acetyl-L-alanine-N '-methyl amide: VA and VCD spectra*, Chemical Physics, **246**, no. 1-3, 13–36, 1999.

13. M. Elstner, *The SCC-DFTB method and its application to biological systems*, Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta), **116**, no. 1-3, 316–325, 2006.
14. Gustavo M. Seabra, Ross C. Walker, Marcus Elstner, David A. Case, and Adrian E. Roitberg, *Implementation of the SCC-DFTB Method for Hybrid QM/MM Simulations within the Amber Molecular Dynamics Package*, Journal of Physical Chemistry A, **111**, no. 26, 5655–5664, 2007.
15. Ulrich H. E. Hansmann, *Parallel tempering algorithm for conformational studies of biological molecules*, Chemical Physics Letters, **281**, no. 1-3, 140–150, 1997.
16. Rubicelia Vargas, Jorge Garza, Benjamin P. Hay, and David A. Dixon, *Conformational Study of the Alanine Dipeptide at the MP2 and DFT Levels*, Journal of Physical Chemistry A, **106**, no. 13, 3213–3218, 2002.
17. Juan Torras, E. Deumens, and S. B. Trickey, *Software Integration in Multi-scale Simulations: the PUPIL System*, Journal of Computer-Aided Materials Design, **13**, no. 1, 201–212, 2006.
18. Juan Torras, Yao He, Chao Cao, Krishna Muralidharan, E. Deumens, H. P. Cheng, and S. B. Trickey, *PUPIL: A systematic approach to software integration in multi-scale simulations*, Computer Physics Communications, **177**, no. 3, 265–279, 2007.
19. C. Jarzynski, *Nonequilibrium equality for free energy differences*, Physical Review Letters, **78**, no. 14, 2690–2693, 1997.
20. A. Crespo, M. A. Marti, D. A. Estrin, and A. E. Roitberg, *Multiple-Steering QM-MM Calculation of the Free Energy Profile in Chorismate Mutase*, J. Am. Chem. Soc., **127**, no. 19, 6940–6941, 2005.
21. Hui Xiong, Alejandro Crespo, Marcelo Marti, Dario Estrin, and Adrian Roitberg, *Free Energy Calculations with Non-Equilibrium Methods: Applications of the Jarzynski Relationship*, Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta), **116**, no. 1-3, 338–346, 2006.
22. A.S. Dutton, J.M. Fukuto, and K.N. Houk, *Mechanisms of HNO and NO Production from Angeli's Salt: Density Functional and CBS-QB3 Theory Predictions*, J. Am. Chem. Soc., **126**, no. 12, 3795–3800, 2004.

Evolution of Experimental and Theoretical Determinations of Protein Structure and Protein Folding Pathways

Harold A. Scheraga, Adam Liwo, Cezary Czaplewski, and Stanisław Ołdziej

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

Physical chemical studies of hydrogen bonding and hydrophobic interactions, and experimental studies of the structure and folding pathways of bovine pancreatic ribonuclease A motivated the development of a theoretical approach to compute protein structure and protein-folding pathways.

1 Introduction

This article traces the development of our experimental and theoretical efforts to gain an understanding of the underlying physics that controls the progression from a newly-synthesized polypeptide chain to the three-dimensional structure of a native biologically-active fibrous or globular protein. Our earliest involvement with this problem was concerned with the influence of hydrogen bonds and hydrophobic interactions on protein structure and reactivity. This work led to our efforts to determine protein structure and folding pathways, first by experimental methods, and subsequently by theoretical methods.

2 Internal Bonding in Proteins

Internal hydrogen bonds influence the observed pKs of ionizable groups[1] and even the reactivity of covalent bonds[2], e.g., peptide bonds. Figure 1 provides an example of a hydrogen bond between a tyrosyl donor and a glutamate acceptor. The observed pKs of

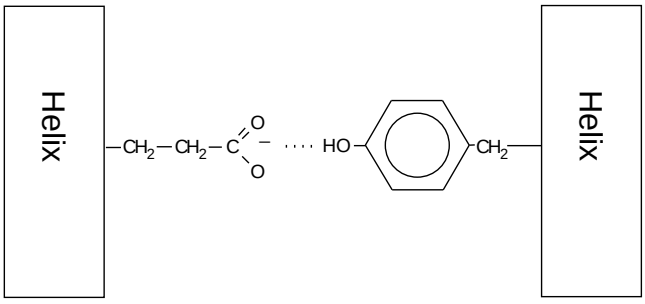


Figure 1. Example of a tyrosyl · · · glutamate hydrogen bond.

these groups are modified by the free energy to form (or break) such a hydrogen bond. Therefore, in comparison with a non-hydrogen-bonded model compound, the observed pKs of such hydrogen-bonded tyrosyl and glutamate groups will be raised or lowered, respectively. Consequently, such departures of the pKs from those of model compounds are diagnostic for the presence of such hydrogen bonds.

Hydrophobic interactions can provide a nonpolar environment which will also influence the pKs of nearby ionizable groups. A theory[3] for the thermodynamics of hydrophobic interactions, based on the structures of liquid water and of aqueous hydrocarbon solutions was presented in 1962, and upgraded[4] in 2004. By themselves, hydrogen bonds in proteins in water are not very strong because of the necessity to shed water in order to form a hydrogen bond between the polar groups of a protein. However, as illustrated in Figure 2, the presence of nearby nonpolar groups can provide hydrophobic interactions[5] with the nonpolar parts of residues such as lysine and glutamic acid and also restrict the internal rotational freedom of the ionizable side chains. In addition, nonpolar groups can restrict the access of water to the polar parts of ionizable side chains. Thus, the cooperativity of nonpolar groups and hydrogen bonding of ionizable side chains can strengthen the hydrogen bonds.

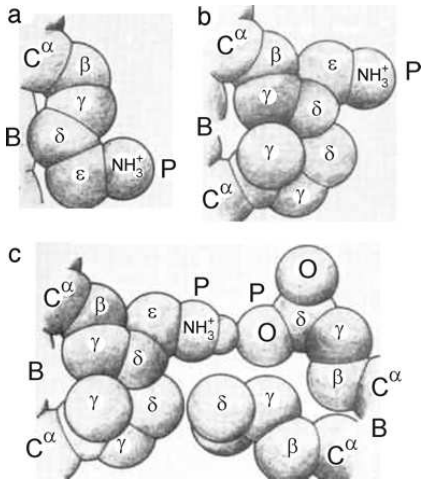


Figure 2. Illustration of various hydrophobic interactions of a polar side chain with its surroundings. B refers to the backbone, and P to the polar head.

3 Location of Hydrogen Bonds in Proteins

Before the advent of X-ray crystallography, NMR, and recombinant DNA methods to determine protein structure, experimental studies to locate hydrogen bonds between ionizable groups, as indicated by the dotted lines in Figure 3, provided distance constraints on the folding of a protein backbone. Such studies, carried out on the 124-residue protein bovine pancreatic ribonuclease A (RNase A), showed that 3 of its 6 tyrosyl residues had

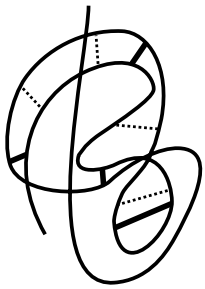


Figure 3. Schematic representation of a protein. Solid and dotted lines represent disulfide bonds and non-covalent interactions, respectively.

abnormally-high[6] pKs, and that 3 of its 11 carboxyl groups had abnormally-low[7] pKs. Further, the UV absorption spectrum of tyrosine was perturbed[8] at low pH where carboxyl groups ionize. This evidence suggested the existence of three tyrosyl $\cdots$ carboxylate hydrogen bonds, and subsequent physical and biochemical experiments[7] identified one pairing, namely

Tyr25 $\cdots$ Asp14
Tyr92 $\cdots$ Asp38
Tyr97 $\cdots$ Asp83

out of the 19,800 possible ways to pair 3 of 6 Tyr and 3 of 11 carboxyl groups of RNase A. The identification of these 3 interactions was subsequently verified by the X-ray structure. These three non-covalent interactions (dotted lines in Figure 3) and the four disulfide bonds (solid-line crosslinks in Figure 3) provide 7 distance constraints on the folding of the backbone. However, 7 distance constraints are not sufficient to provide an accurate description of the backbone of a 124-residue protein such as RNase A. To determine the backbone structure, as is now done by NMR, many more distance constraints would be required. In fact, it is possible to specify the number of distance constraints required[9] in order to determine the structure within any desired RMSD from the native structure.

4 Initial Considerations of a Theoretical Approach to Structure Simulation

On the other hand, even 7 known distances could serve as restraints on a potential energy function to compute the native structure of a protein. This provided the motivation to develop[10] a theoretical approach to compute protein structure, first by making use of distance restraints and, subsequently, to rely on a physics-based potential function without the need to incorporate distance restraints. At about the same time, Anfinsen[11] identified spontaneous protein folding, and introduced the thermodynamic hypothesis for a theoretical approach, and we expanded our interest from determining structure to also determining folding pathways (first by experiment, and later by theory).

5 Experimental Studies of Oxidative Folding of RNase A

Our experimental study of the oxidative folding of RNase A led to the mechanism shown in Figure 4. Figure 4(a) shows that a pre-equilibrium exists between the unfolded forms

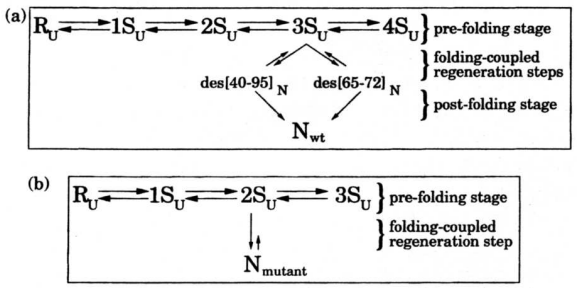


Figure 4. (a) Oxidative folding of wild-type RNase A. (b) Oxidative folding of two three-disulfide mutants of RNase A (C40A/C95A and C65S/C72S).

of reduced RNase A, R, and various ensembles of disulfide-bonded intermediates[12]. The rate-determining step[13] is the reshuffling of the three-disulfide ensemble, 3S, by SH/SS interchange, to form two main intermediates, des $[40\text{-}95]_N$ and des $[65\text{-}72]_N$ which each contain three native disulfide bonds but lack the 40-95 and 64-72 disulfide bonds, respectively. These two native intermediates[14, 15] rapidly form the native structure of the wild-type protein. As shown in Figure 4(b), two very minor pathways exist[16, 17] in which the 2S ensemble undergoes oxidation to des $[40\text{-}95]_N$ and des $[65\text{-}72]_N$, which could be detected only with the aid of mutants which lacked the 40-95 and 65-72 disulfide bonds, respectively.

The overall scheme for the oxidative folding of RNase A is shown in Figure 5. Of the 28 possible 1S species, 40% have the native 65-72 disulfide bond, and 10% have the non-native 58-65 disulfide bond, and the remaining 26 species accumulate only to the extent of $<$10% each in folding of the whole protein[18]. The 65-72 disulfide bond persists increasingly in the remainder of the pathway[19] to form des $[40\text{-}95]_N$. Interestingly, the same 40:10 ratio that is found in the protein is also found when a fragment of reduced RNase A from Cys 58 to Cys 72 is oxidized[20, 21]. This result is attributed to preferential native-forming interactions[22] and not to entropic effects in the 65-72 loop, since both possible loops (58-65 and 65-72) have the same size; it is this kind of physics that is revealed by such experimental studies, and by our concomitantly developed molecular mechanics approach.

6 All-atom Determination of Protein Structure and Folding Pathways

Progressing from our initial work[10] in 1965, with a hard-sphere potential, we developed an all-atom ECEPP (Emperical Conformational Energy Program for Peptides) force field[23],

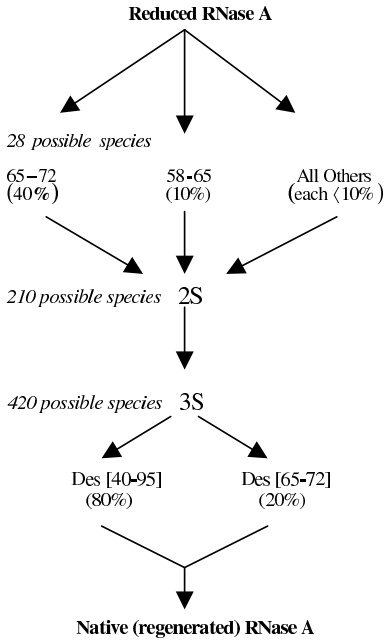


Figure 5. Overall scheme of oxidative folding of RNase A

and improved it in subsequent years, the latest version of which[24] appeared in 2006. The ECEPP force field and various procedures that we developed for global optimization of the potential energy were described in a recent review[25]. The largest protein whose structure we have computed with the all-atom force field is the 46-residue three-helix bundle, protein A[26], and the 36-residue villin headpiece[27].

Our initial attempt[28] to compute an all-atom folding pathway made use of the stochastic difference equation method of Elber[29]. The pathways were computed from each of a large ensemble of unfolded states of protein A to the final folded state, and then averaged. It was found that the C-terminal helix folded first, followed by the N-terminal helix, and then the middle helix. Various folding pathways have been proposed for protein A, and it has been found[30] that environmental factors and different components in the various force fields used may account for the reported differences.

7 A Hierarchical Approach to Protein Structure and Folding Pathway Prediction

In order to compute protein structures larger than those of protein A, we have developed a hierarchical procedure which initially makes use of a united-residue (UNRES) model of a polypeptide chain[31–37] together with a conformational space annealing (CSA) procedure[38] to search the UNRES conformational space to find the *region* in which the global minimum might lie. Then the lowest-energy structures are converted from the UNRES representation

to an all-atom one[39,40] whose ECEPP energy (including hydration) can then be globally optimized.

The UNRES model consists of virtual-bond chains for the backbone and side chains, with the less-important degrees of freedom (rotation of the peptide groups around their virtual C^{α}–C^{α} bonds, internal rotations about side-chain bonds, etc.) averaged out. The force centers are the positions of the averaged-out peptide groups and the ends of the virtual bonds at the center of gravity of the side chains. The UNRES energy consists of interactions between these force centers, the energies to vary the positions and rotational states of side chains, the energies to vary the angle between successive backbone virtual bonds, the torsional angles around the backbone virtual bonds, and double torsions around two neighboring virtual bonds, and multi-body interactions. The CSA procedure is essentially a genetic algorithm in which a finite set of widely- dispersed UNRES minima are forced to coalesce to the region of the global minimum.

Performance in successive blind tests from CASP3 to CASP7 has provided sufficient confidence to encourage us to develop[41,42] and apply[43–45] a molecular dynamics treatment based on UNRES. Our recent work with this molecular dynamics approach is being discussed at this workshop by A. Liwo[46].

8 Conclusions

The evolution of the development of our experimental and theoretical approaches to gain an understanding of the fundamental physics that controls protein structure and folding pathways has been traced. It is elaborated upon in the accompanying article by Liwo et al[46]. Current work is focused on the use of the molecular dynamics approach with UNRES, and the improvement of this methodology including introduction of entropic effects[37].

Acknowledgments

This research was supported by grants from NIH (GM-14312, TW-7193) and NSF (MCB05-41633), and was carried out with resources of (a) our 820-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) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Juelich, Germany, (d) the National Center for Supercomputing Applications System at the University of Illinois at Urbana-Champaign, (e) the cluster at the Department of Computer Science, Cornell University, (f) the resources of the Center for Computation and Technology at Louisiana State University, which is supported by funding from the Louisiana legislature's Information Technology Initiative, (g) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdansk, (h) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdansk, and (i) the Cornell Theory Center which receives funding from Cornell University, New York State, Federal agencies, and corporate partners.

References

1. M. Laskowski, Jr. and H.A. Scheraga *Thermodynamic considerations of protein reactions. I. Modified reactivity of polar groups*, J. Am. Chem. Soc. **76**, 6305–6319, 1954.
2. M. Laskowski, Jr. and H.A. Scheraga *Thermodynamic considerations of protein reactions. II. Modified reactivity of primary valence bonds*, J. Am. Chem. Soc. **78**, 5793–5798, 1956.
3. G. Némethy and H.A. Scheraga *The structure of water and hydrophobic bonding in proteins. III. The thermodynamic properties of hydrophobic bonds in proteins*, J. Phys. Chem. **66**, 1773–1789, 1962.
4. J.H. Griffith and H.A. Scheraga *Statistical thermodynamics of aqueous solutions. I. Water structure, solutions with non-polar solutes, and hydrophobic interactions*, J. Molec. Str. **682**, 97–111, 2004.
5. G. Némethy, I.Z. Steinberg, and H.A. Scheraga *The influence of water structure and of hydrophobic interactions on the strength of side-chain hydrogen bonds in proteins*, Biopolymers **1**, 43–69, 1963.
6. C. Tansford, J.D. Hauenstein, and D.G. Rands *Phenolic hydroxyl ionization in proteins. II. Ribonuclease*, J. Am. Chem. Soc. **77**, 6409–6413 , 1958.
7. H.A. Scheraga *Structural studies of pancreatic ribonuclease*, Fed. Proc. **26**, 1380–1387, 1967.
8. H.A. Scheraga *Tyrosyl-carboxylate ion hydrogen bonding in ribonuclease*, Biochim. Biophys. Acta **23**, 196–197, 1957.
9. H. Wako and H.A. Scheraga *On the use of distance constraints to fold a protein*, Macromolecules **14**, 961–969 , 1981.
10. G. Némethy and H.A. Scheraga *Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method*, Biopolymers **3**, 155–184, 1965.
11. C.B. Anfinsen *Principles that govern folding of protein chain*, Science **181**, 223–230, 1973.
12. D.M. Rothwarf, Y.-J. Li, and H.A. Scheraga *Regeneration of bovine pancreatic ribonuclease A. Identification of two nativelike three- disulfide intermediates involved in separate pathways*, Biochemistry **37**, 3760–3766 , 1998.
13. 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.
14. S. Shimotakahara, C.B. Rios, J.H. Laity, D.E. Zimmerman, H.A. Scheraga, and G.T. Montelione *NMR structural analysis of an analog of an intermediate formed in the rate-determining step of one pathway in the oxidative folding of bovine pancreatic ribonuclease A: Automated analysis of 1H, 13C, and 15N resonance assignments for wild-type and [C65S, C72S] mutant forms*, Biochemistry **36**, 6915–6929 , 1997.
15. J.H. Laity, C.C. Lester, S. Shimotakahara, D.E. Zimmerman, G.T. Montelione, and H. A. Scheraga *Structural characterization of an analog of the major rate–determining disulfide folding intermediate of bovine pancreatic ribonuclease A*, Biochemistry **36**, 12683–12699 , 1997.

16. M. Iwaoka, D. Juminaga, and H.A. Scheraga *Regeneration of three- disulfide mutants of bovine pancreatic ribonuclease A missing the 65-72 disulfide bond: Characterization of a minor folding pathway of ribonuclease A and kinetic roles of Cys65 and Cys72*, Biochemistry **37**, 4490–4501 , 1998.
17. X. Xu and H. A. Scheraga *Kinetic folding pathway of a three-disulfide mutant of bovine pancreatic ribonuclease A missing the [40-95] disulfide bond*, Biochemistry **37**, 7561–7571 , 1998.
18. X. Xu, D.M. Rothwarf, and H.A. Scheraga *Nonrandom distribution of the one-disulfide intermediates in the regeneration of ribonuclease A*, Biochemistry **35**, 6406–6417 , 1996.
19. M.J. Volles, X. Xu, and H.A. Scheraga *Distribution of disulfide bonds in the two-disulfide intermediates in the regeneration of bovine pancreatic ribonuclease A*, Biochemistry **38**, 7284–7293 , 1999.
20. P.J. Milburn and H.A. Scheraga *Local interactions favor the native 8-residue disulfide loop in the oxidation of a fragment corresponding to the sequence Ser-50-Met-79 derived from bovine pancreatic ribonuclease A*, J. Protein Chem. **7**, 377–398 , 1988.
21. K.H. Altmann and H.A. Scheraga *Local structure in ribonuclease A. Effect of amino acid substitutions on the preferential formation of the native disulfide loop in synthetic peptides corresponding to residues Cys58-Cys72 of bovine pancreatic ribonuclease A*, J. Am. Chem. Soc. **112**, 4926–4931 , 1990.
22. R.P. Carty, M.R. Pincus, and H.A. Scheraga *Interactions that favor the native over the non-native disulfide bond among residues 58-72 in the oxidative folding of bovine pancreatic ribonuclease A*, Biochemistry **41**, 14815–14819 , 2002.
23. F.A. Momany, R.F. McGuire, A.W. Burgess, and H.A. Scheraga *Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids*, J. Phys. Chem. **79**, 2361–2381 , 1975.
24. Y.A. Arnautova, A. Jagielska, and H.A. Scheraga *A new force field (ECEPP-05) for peptides, proteins and organic molecules*, J. Phys. Chem. B. **110**, 5025–5044 , 2006.
25. 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. Nanias *The protein folding problem: Global optimization of force fields*, Frontiers in Bioscience **9**, 3296–3323, 2004.
26. 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.
27. D.R. Ripoll, J.A. Vila, and H.A. Scheraga *Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH*, J. Mol. Biol. **339**, 915–925 , 2004.
28. 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.
29. R. Elber, A. Ghosh, and A. Cardenas *Long Time Dynamics of Complex Systems*, Acc. Chem. Res. **35**, 396–403 , 2002.

30. A. Jagielska and H.A. Scheraga *Influence of temperature, friction, and random forces on folding of the B-domain of Staphylococcal Protein A: All-atom molecular dynamics in implicit solvent*, J. Comput. Chem. **28**, 1068–1082 , 2007.
31. 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.
32. 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 short-range interactions and determination of weights of energy terms by Z-score optimization*, J. Comput. Chem. **18**, 874–887 , 1997.
33. A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Ołdziej, R.J. Wawak, S. Rackovsky, M.R. Pincus, and H.A. Scheraga *A 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.
34. 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.
35. A. Liwo, P. Arłukowicz, C. Czaplewski, S. Ołdziej, J. Pillardy, and H.A. Scheraga *A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field*, Proc. Natl. Acad. Sci. U.S.A. **99**, 1937–1942, 2002.
36. S. Ołdziej, J. Lagiewka, A. Liwo, C. Czaplewski, M. Chinchio, M. Nanias, 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.
37. A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H.A. Scheraga *Modification and optimization of the united-residue (UNRES) potential-energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins*, J. Phys. Chem. B. **111**, 260–285 , 2007.
38. 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.
39. R. Kaźmierkiewicz, 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.
40. R. Kaźmierkiewicz, 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).
41. 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.

42. 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.
43. 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.
44. 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.
45. A. V. Rojas, A. Liwo, and H.A. Scheraga *Molecular dynamics with the united-residue force field. Ab initio folding simulations of multichain proteins*, J. Phys. Chem. B. **111**, 293–309, 2007.
46. A. Liwo, C. Czaplewski, S. Ołdziej, M. Chinchio, A.V. Rojas, M. Khalili, M. Makowski, S. Kalinowski, U. Kozłowska, R.K. Murarka, and H.A. Scheraga *Mesoscopic dynamics with the UNRES force field. A tool for studying the kinetics and thermodynamics of protein folding*, in "From Computational Biophysics to Systems Biology 2007", in press.

The E-Cell Project and Challenges in Computational Systems Biology

Koichi Takahashi[1,2]

[1] The Molecular Sciences Institute,
2168 Shattuck Avenue, Berkeley, CA 94704, USA.
E-mail: ktakahashi@molsci.org

[2] Institute for Advanced Biosciences,
Keio University, Fujisawa, 252-8520, Japan.

Summarizing our experience in launching and running the E-Cell Project over the past ten years, I will discuss some of major challenges we believe we will face in the next ten years of cell and systems biological simulation, including the following two; (1) Undeniably the last ten years of computational systems biology has been (re)discovery of the biggest bottleneck in biochemical modeling; the lack of high-throughput and reliable means of obtaining reaction rate coefficients. Computational aids in determination of reaction rate coefficients will be one area in which fruitful interactions between molecular biology, biophysical chemistry, and supercomputing are highly expected. (2) Macromolecular crowding is ubiquitous and found in all types of cellular organisms on the earth, which can, when coupled with localization and diffusion, alter biochemical dynamics, change equilibrium points, slow down and change the manner how big molecules diffuse, and amplify intrinsic noise. It is also a suspected physico-chemical factor behind the emergence of eukaryotic organisms. Development of formal treatment and computational methods for crowded intracellular media will be some of the most important tasks left for computational biologists.

1 Introduction

The E-Cell Project [a] was started in 1996 with the aim of establishing technological bases that will make possible predictive modeling and simulation of cellular systems at the molecular level. Although we are halfway towards the ultimate goal we defined 11 years ago, I believe we are now at a good position where we can speculate, and, to some extent, predict the next 10 years of cell simulation technology by extrapolating our experience in the last decade, with the hope that it will be of any help in thinking about challenges of what nature we will encounter and what will have been done if we would be overcoming these. But before that, in the next section, let us think about why we want to model biochemical and cellular systems, and what we could do in the first place. I will then briefly look back the history of the E-Cell Project and our simulation platform E-Cell System, and discuss some challenges I believe we will face in the process of establishing cell simulation technology.

2 Simulation and Science

Why do we want to model and simulate biological systems? Why do we want to do it at the molecular level, and what insight do we expect from it?

[a] `http://e-cell.org`

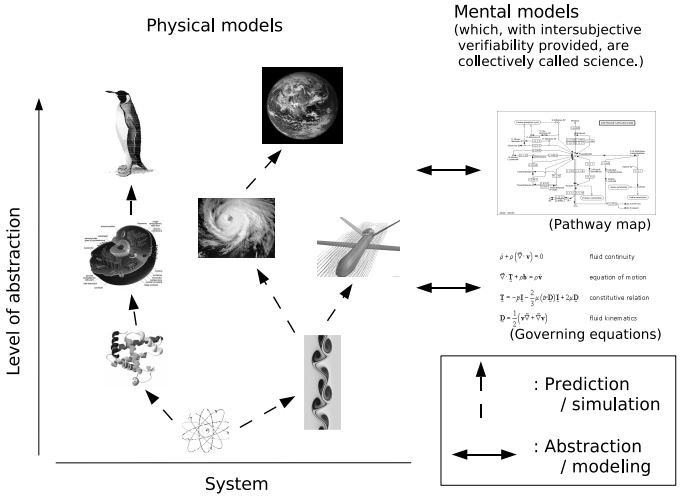


Figure 1. Physical models, mental models, and how these models are obtained or related by means of predictive and abstraction powers of modeling and simulation.

Science is a collection of mutually interconnected mental models that are intersubjectively verifiable and shared by the community of scientists. As a first approximation, we could define scientific research as forming of such mental models. In Figure 1, on the right hand side are two examples of such models, a metabolic pathway map and a set of equations for fluid dynamics. Time-space reproduction of molecules constituting the physical systems, such as those shown on the left hand side of the figure, do not by themselves represent understanding of how the systems work. Such physical models must somehow undergo a process of abstraction to construct models that have the power to predict and/or to explain behaviors of systems in different states or of different variations of the same type. Simulation allows us to leap forward exactly one step in the level of abstraction, from lower to higher. For example, equations about inter-atomic forces could be extrapolated to depict how proteins fold through the medium of numerical simulation. Or, atmospheric data gathered at a geographically sparse set of observation points for a certain period of time may lead to a model of a larger area, longer time-scales, and/or finer resolutions. Let us call this ability of simulation that enables us to interpolate or extrapolate physical models at certain levels of abstraction beyond our accessibility to the real-world target systems a *predictive power* of numerical simulation.

E-Cell Project is interested in modeling and simulating cellular systems at the level of the system-level output, or, physiology. If we want to build physiological models of the cell equipped with the predictive power, we have to model the system at one step lower in the hierarchy of abstraction. How big or small this 'one step' should be is subject to many technological factors. In principle, modeling becomes easier as we go lower in the abstraction level, since composite phenomena and complex components in the system tend to be decomposed to more basic principles and simpler components that are easier to model, and in many cases, less complicated to experimentally measure necessary parameters. This is

a crucially important aspect of the picture because the hardest part of cellular modeling and simulation lies in its ontological complexity, rather than emerging complexities seen in many other physical simulations[2]. On the other hand, how far we can go down given a desired observation frame highly depends on computational capacities and accuracy of numerical methods.

3 The E-Cell Project

What we aimed at when we started the E-Cell Project was to bring this predictive ability of numerical simulation into molecular biology. Our initial aim was to show it is not impossible to computationally reconstruct inner workings of cellular organisms at the whole cell-scale, and by doing so, to open a pathway that leads to establishment of cell simulation as a scientific method and as an inter-disciplinary technology that spans from measurement methods to numerical analysis.

We started the E-Cell project stimulated by the determination of the smallest known 580kb genome sequence of *Mycoplasma genitalium*. By the summer of 1997, we developed an early version of our modeling and simulation platform, the E-Cell System, and a model of a virtual cell of which 127 genes constitute its basic functions to sustain itself as a living organism; enzymes in the energy metabolism and phospholipid synthesis, and gene transcription and translation systems[1]. Since then, the E-Cell Project has been expanding areas of work to construct more detailed and precise models of cellular functions, such as the bacterial chemotaxis signaling pathway of *Escherichia coli* and metabolic pathways of human erythrocytes and mitochondria. Concomitant to the biological modeling projects was the development of the computational platform, the E-Cell System. We identified seven desirable features of cell simulation platforms to be truly useful, which constitute a quite different set of requirements than that of conventional physical simulators, such as the need for an integrative multi-algorithm, multi-timescale framework, object-orientation, dynamic model structure, and real-time user interactions. Due to the length limitation, interested readers are referred to other E-Cell Project publications[2] for more discussions. Since the development of the first version[1], our current version is E-Cell System version 3[3], which meets five of the seven requirements we defined. We have started the development of a new generation simulation kernel that will be the core of the E-Cell System version 4 to address the two remaining features, multi-spatial representations and dynamic model structure, in addition to a vastly improved support for parallel computation [b].

4 Some Important Challenges in Computational Systems Biology

Numerical simulation, as well as experimental measurement technologies, is placed at the center of disciplines playing important roles in computational systems biology, which is inter-disciplinary by nature. Here I pick up two (among many) major challenges in computational systems biology that we believe we will face in the course of pursuing cell simulation technology development.

[b] `http://e-cell.org/developers/e-cell-4`

4.1 Methods for Determination of Reaction Rate Coefficients

The last ten years have been rediscovery of the lack of high-throughput and reliable means of obtaining reaction rate coefficients, and this lack formed the biggest bottleneck in the biochemical modeling and simulation workflow. Many modeling projects got stuck as soon as they faced this lack of input parameters, and a common feature of the limited number of successful modeling projects has been the accumulation of a large body of kinetic studies for decades, each determined parameter for a particular enzyme corresponding to someone's doctoral degree.

When we think about these systems, however, we notice that there is no such measurable physical quantities like 'net reaction rate constants', but there is only coupling of two distinct physical occurrences; diffusive encounters of reactants and subsequent interactions between them. Let us think about bimolecular reactions, that are most commonly seen in cellular metabolic, signaling and gene expression pathways. Such reactions can formally be written as $A + B -> C$. Recall Arrhenius relation

$$k_{net} = A \exp\left(-\frac{E_a}{k_B T}\right), \tag{1}$$

where k_{net} is the net reaction rate, A is the frequency factor, E_a is the activation energy, k_B is the Boltzmann constant and T is temperature. It can be read that there are two factors that determine the net rates of reactions, the frequency of molecular collisions and the activation energy which is related to interactions between molecules. Let us then look at this from another point of view;

$$\frac{1}{k_{net}} = \frac{1}{4\pi D\sigma} + \frac{1}{k_{intrinsic}}, \tag{2}$$

where D is the diffusion coefficient, σ is the effective cross section, and $k_{intrinsic}$ is the intrinsic reaction rate. When both of A and B molecules diffuse, D becomes a relative diffusion coefficient. The first term in the right hand side indicates essentially the same as A, the frequency of collision, and the intrinsic reaction rate in the second term corresponds to the activation energy. Where did this $4\pi D\sigma$ come from? It came from the Smoluchowski theory of diffusion-limited reactions[7]. In the simplest case where we describe the system in one dimensional space,

$$k(t) = 4\pi\sigma^2 D \frac{\partial C}{\partial r} \tag{3}$$

where t is the time since one of the reactive species was injected to the media while the other species was in a well-mixed state, r is the distance between reactants, and C is the concentration of the reactant. At the equilibrium ($t \to \infty$), and in the low-density limit, we get; $\lim_{t\to\infty} k(t) = \lim_{t\to\infty} 4\pi\sigma D \left(1 + \frac{\sigma}{\sqrt{\pi D t}}\right) = 4\pi\sigma D$. While this result partially applies to more complex systems[8], when concentration gradients and noise caused by the slow diffusion of macromolecules become relevant, or when density of molecules is high (e.g. see next section), simulation is typically the only way to accurately investigate the system. However, this decomposition of net reaction rates into two simpler, physically better formulated components may be one possible approach for cell simulation to overcome the bottleneck of reaction rates. An array of experimental methods that can be used to see how big molecules in the cell diffuse, such as fluorescence correlation spectroscopy and

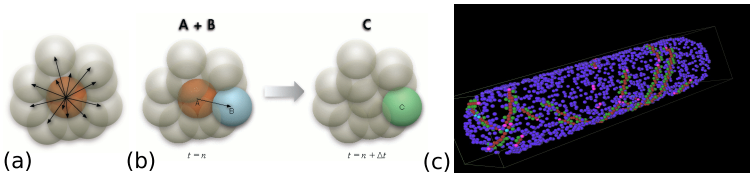


Figure 2. An extended lattice gas automata-like simulation method. (a) diffusion of a molecule to one of its 12 neighboring spheres in cubic-close packing. (b) a Monte-Carlo bi-molecular reaction $A + B \rightarrow C$ within the method with the time step Δt. (c) A sample simulation snapshot of MinD polymerization in *Escherichia coli* membrane. Reproduced with permission by Satya Arjunan[9].

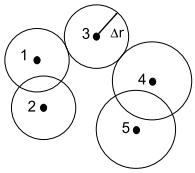


Figure 3. Basic idea used in the Green's Function Reaction Dynamics. Particles are 'protected' by protecting domains (spheres in this figure) to limit interactions up to two bodies. In this figure, the particles 1, 2, and 4, 5 form pairs, and the particle 3 remains single. Particles are propagated according to corresponding Green's functions for diffusion.

single particle tracking, are rapidly developing. At the same time, dynamics of protein-protein, protein-ligand, and protein-nucleic acid interactions are some of the areas where expected developments in high-performance scientific computing would be found most useful, ultimately leading to a high-throughput method of determination of reaction rates, possibly coupled with some form of high-sensitivity biosensors.

4.2 Macromolecular Crowding

Another area where supercomputing may stimulate scientific breakthroughs may consist in the striking nature of intracellular media. Extremely high densities of macromolecules (50-400 mg/ml, compare to 1-10 mg/ml typical *in vitro* conditions), called intracellular macromolecular crowding, is ubiquitous and found in all types of cellular organisms on the earth. Such non-ideal properties of the space in the cell result in different equilibrium points, altered reaction rates, slow and anomalous diffusion of macromolecules, and thus modified overall behaviors and dynamical characteristics of biochemical systems. Recently it was proposed that macromolecular crowding might be the physico-chemical culprit behind the emergence of eukaryotic cells[4].

Computational methods that could be used to study diffusion, localization and crowding of proteins in signaling pathways are reviewed somewhere else[5]. Here I mention two approaches the E-Cell Project is currently interested in. First of such approach belongs to a class of methods called cellular automata. Shown in Figure 2 is an extended lattice-gas automata. This class of method discretize the intracellular space with regular lattice, and propagate molecules to neighboring sites using Monte-Carlo calculations. When lattice

resolution is set adequately, the methods can give an approximate reproduction of molecular crowding. Lattice-based methods have a great affinity to modern digital computer architectures, and have promising scalability to supercomputers. Another class of method we are working to implement on the E-Cell System makes use of particles in continuum space and time[6]. This method called Green's Function Reaction Dynamics (GFRD) decomposes the multi-body problem that constitutes the biochemical system into a set of one and two-body problems (Figure 3), and corresponding Green's functions for diffusion-reaction are used to propagate the particles. GFRD is an accelerated form of Brownian Dynamics (BD) simulation method that is capable of effectively represent crowded space, and is typically up to five orders of magnitude faster than traditional approaches.

5 Conclusion

Understanding cellular systems as systems, not simply as a collection of molecular components, yet at the molecular level, is undoubtedly one of the most important scientific challenges in the 21st century, where computation may be the key to breakthroughs.

Acknowledgments

I thank Satya N.V. Arjunan for the permission to use figures from his work, and Nathan Addy for his help in editing. I am a Human Frontier Science Program Fellow. The E-Cell Project is supported by CREST/JST and the MEXT of Japan (Leading Project for Biosimulation and the 21st Century COE Program).

References

1. E-CELL: software environment for whole-cell simulation, M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K Yugi, J. C. Venter, and C. Hutchison, Bioinformatics **15**, 1, 1999.
2. Multi-algorithm and multi-timescale cell biological simulation, K. Takahashi, PhD Thesis, Keio University (2004).
3. Multi-algorithm, multi-timescale method for cell simulation, K. Takahashi, K. Kaizu, B. Hu, and M. Tomita, Bioinformatics **20**, 4, 1999.
4. Genomics and the Irreducible Nature of Eukaryote Cells, C. G. Kurland, L. J. Collins, and D. Penny, Science **312**, 5776, 2006.
5. Space in systems biology of signaling pathways – intracellular molecular crowding in silico, K. Takahashi, S. Arjunan, and M. Tomita, FEBS Letters **579**, 8, 2005.
6. Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space, J. S. van Zon and P. R. ten Wolde, J. Chem. Phys. **123**, 234910, 2005.
7. Diffusion-Limited Reactions, S. A. Rice, in Compr. Chem. Kinet., Vol. **25**, editec by C. H. Bamford, C.F.H. Tipper, R.G. Compton, Elsevier, New York, (1985).
8. Theory of reversible diffusion-influenced reactions, N. Agmon and A. Szabo, J. Chem. Phys. **92**, 5270, 1990.
9. A 3D pole-to-pole oscillation model of MinD on Escherichia coli membrane, S. N. V. Arjunan and M. Tomita, unpublished (2007).

Stretching RNA Hairpins

Changbong Hyeon[1] **and D. Thirumalai**[2,3]

[1] Center for Theoretical Biological Physics, University of California at San Diego
La Jolla, California 92093, USA

[2] Biophysics Program
Institute for Physical Science and Technology

[3] Department of Chemistry and Biochemistry, University of Maryland
College Park, MD 20742, USA
E-mail: thirum@glue.umd.edu

1 Introduction

Optical tweezer experiments have used mechanical force to trigger folding and unfolding of RNA molecules at a single molecule level[1,2]. Application of tension to a specific position of the molecule, induces sequence and structure-dependent response. The measured force-extension curve (FEC) is fit using the worm-like (WLC) model[3,4]. The stability of RNAs is inferred by integrating the FECs. For simple motifs, such as hairpins, it has been shown that the stability of the native structures can be accurately measured using mechanical unfolding trajectories which exhibit multiple transitions between the folded and the unfolded state when the force is held constant[1].

Mechanical force has also been used to probe unfolding and refolding kinetics of RNA. The cooperative reversible folding of hairpins has been shown by monitoring the end-to-end distance (R), a variable conjugate to the mechanical f, as a function of time. This procedure works best when RNA folding is described using two-state approximation. For multidomain ribozymes the folding/unfolding kinetics is complex and new tools are required to interpret the kinetic data. In a pioneering study Onoa *et. al.*[2] showed that the rips in FECs for the L-21 derivative of *Tetrahymena thermophila* ribozyme (*T.* ribozyme), composed of multiple domains, are a result of unfolding of individual intact domains that are stabilized in the native state by counterion-dependent tertiary interactions.

In this note, we probe the forced-unfolding dynamics of RNA hairpins using a simple model. Because these simulations can be used to directly monitor structures in the transition from folded to the fully stretched states, unfolding pathways can be unambiguously resolved. We use a *Self-Organized Polymer (SOP) model* for RNA that is based only on the self-avoiding nature of the RNA and the native structure. We apply the SOP model to probe forced-unfolding of a number of RNA structures of varying complexity. Many of the subtle features of the variations in the mechanical unfolding as a function of f and r_f can be illustrated using P5GA, a simple RNA. Our results show that the response of RNA to force is largely dependent on the architecture of the native state. More importantly, we have established that the characterization of the the energy landscape requires using force values (or loading rates) over a wide range.

2 Methods

Model: The SOP model for RNA that retains chain connectivity and favorable attractive interactions between sites that stabilize the native fold. Each interaction center represents the center of mass of a nucleotide. In terms of the coordinates $\{\mathbf{r}_i, i = 1, 2, \ldots N\}$ of RNA with N nucleotide the total potential energy in the SOP representation is

$$\begin{aligned} V_T &= V_{FENE} + V_{nb}^{(att)} + V_{nb}^{(rep)} \\ &= -\sum_{i=1}^{N-1} \frac{k}{2} R_0^2 \log(1 - \frac{(r_{i,i+1} - r_{i,i+1}^o)^2}{R_0^2}) \\ &+ \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \epsilon_h [(\frac{r_{ij}^o}{r_{ij}})^{12} - 2(\frac{r_{ij}^o}{r_{ij}})^6] \Delta_{ij} \\ &+ \sum_{i=1}^{N-2} \epsilon_l (\frac{\sigma^*}{r_{i,i+2}})^6 + \sum_{i=1}^{N-3} \sum_{j=i+3}^{N} \epsilon_l (\frac{\sigma}{r_{ij}})^6 (1 - \Delta_{ij}). \end{aligned} \quad (1)$$

The first term is for the chain connectivity. The finite extensible nonlinear elastic (FENE) potential[5] is used with $k = 20 kcal/(mol{\cdot}\text{Å}^2)$, $R_0 = 0.2\ nm$, and $r_{i,i+1}$ is the distance between neighboring beads interaction centers i and $i + 1$, $r_{i,i+1}^o$ is the distance in the native structure. The use of FENE potential is more advantageous than the standard harmonic potential especially when considering forced-stretching because the fluctuations of $r_{i,i+1}$ are strictly restricted around $r_{i,i+1}^o$ with variation of $\pm R_0$. The Lennard-Jones potential is used to account for interactions that stabilize the native topology. Native contact is defined for the pair of interaction centers whose distance is less than $R_C = 1.4$ nm in the native state for $|i - j| > 2$. If i and j sites are in contact in the native state, $\Delta_{ij} = 1$, otherwise $\Delta_{ij} = 0$. We used $\epsilon_h = 0.7\ kcal/mol$ for the native pairs, $\epsilon_l = 1$ $kcal/mol$ for non-native pairs. In the current version, we have neglected non-native interactions which will not qualitatively affect the results because, under tension, such interactions are greatly destabilized. To ensure the non-crossing of the chain, we set $\sigma = 7$Å. Only for $i, i + 2$ pairs we set $\sigma^* = 3.5$ Å to prevent the flattening of the helical structures when the overall repulsion is large. There are five parameters in the SOP force field (k, R_0, ϵ_h, ϵ_l, and R_c)[6]. Of these the results are sensitive to the precise values of ϵ_h/ϵ_l and R_c. We have discovered that the quantitative results are insensitive to R_c as long as it is in the physical range that is determined by the RNA contact maps. In principle, the ratio ϵ_h/ϵ_l can be adjusted to obtain realistic values of forces. For simplicity we choose a uniform value of ϵ_h for all RNA constructs. Surprisingly, the SOP force field, with the same set of parameters, can be used to obtain near quantitative results for RNA molecules of varying native topology.

Simulations: Using the SOP model, we simulated the mechanical unfolding and refolding of P5GA hairpin. In force-clamp simulations a constant force is applied to one end of the molecule while the other end is fixed. Finally, in force-quench computations the force on the molecule is reduced to the final value to initiate mechanical refolding. In both force-clamp and force-quench setup the dynamics of the linker (usually hybrid RNA/DNA handles) is not relevant whereas depending on the characteristics of the linkers the dynamics of linker may play an important role in the force-ramp experiments[7].

Time scales: Since a typical value for the mass of a nucleotide, $m \sim 300g/mol - 400g/mol$, the average distance between the adjacent nucleotides in the SOP representation of RNA is $a \approx 5$ Å, $\epsilon_h = 0.7\ kcal/mol$, the natural time is $\tau_L = (\frac{ma^2}{\epsilon_h})^{1/2} = 3 \sim 5ps$. We use $\tau_L = 4.0ps$ to convert simulation times into real times. To estimate the time scale for mechanical unfolding dynamics we use a Brownian dynamics algorithm[8,9] for which the natural time for the overdamped motion is $\tau_H = \frac{\zeta\epsilon_h}{k_B T} h\tau_L$ ($\tau_L = 4ps$). We used $\zeta = 100\tau_L^{-1}$ in the overdamped limit, that approximately corresponds to friction constant of a nucleotide in water. The equations of motion in the overdamped limit are integrated using the Brownian dynamics algorithm.

Force-induced transitions in a simple hairpin (P5GA): Liphart *et. al.* showed that the P5ab hairpin, the construct in which P5c stem-loop and the A-rich bulge in P5a are removed from the P5abc subdomain in T. ribozyme, reversibly folds in an all-or-none fashion upon application of constant force. The equilibrium between the native basin of attraction (**NBA**) and the unfolded basin of attraction (**UBA**) can be shifted by altering the value of the constant force, f_c. To probe the two-state behavior of hairpins under force we used a smaller 22-nt hairpin, P5GA (Protein Data Bank (PDB) id: 1eor)[10,11]. For the P5GA hairpin simulations over a wide range of forces can be performed in reasonable times. The topologically simple hairpin has a single tetra-loop and nine consecutive base pairs. In an earlier study[11] we showed, using a minimal three interaction site (TIS) model in which each nucleotide is represented by three sites, that the dynamical behavior of P5GA under tension is qualitatively similar to P5ab. The much simpler SOP representation of P5GA allows us to probe exhaustively the folding and unfolding kinetics of the hairpin that is manipulated by force-ramp, force-quench, and force-clamp.

The hallmark of P5ab[1] and P5GA[11], when a constant force is applied to either the 3' or the 5' ends, is the observation of bistable kinetics. When a constant f_c is applied to the 3' end P5GA makes transitions (Fig. 1-B) between the **UBA** ($R \approx 8nm$) to the **NBA** ($R \approx 2nm$). At $f_c =$ 14.0, 15.4, 17.5 pN a large number of transitions occur over $45ms$ duration which suggests that the hairpin dynamics is effectively ergodic. As in our previous study[11] the equilibrium constant between the folded and unfolded hairpin calculated using a long mechanical unfolding trajectory coincides with an independent ensemble average calculation i.e., time averages are roughly equivalent to ensemble averages. When $f_c = 14\,pN$ the residence time in the **NBA** is much greater than in the **UBA** while at $f_c = 16.8\ pN$ the **UBA** is preferentially populated (Fig. 1-B). The population of P5GA in the **NBA** changes when f_c is varied can be seen in the histogram ($P(R)$) of the end-to-end distance R (Fig. 1). At $f_c = 15.4\,pN$, which is slightly above the midpoint of the **NBA** $\Leftrightarrow$ **UBA** transition, several jumps between the **NBA** and **UBA** are observed. The $P(R)$ distribution reflects the bistable nature of the landscape. The free energy profile with respect to R is computed using $\Delta F(R) = -k_B T \log P(R)$. From $P(R)$ at $f_c = 15.4$ we can obtain the free energy of stability of the folded hairpin with respect to the unfolded state using $\Delta G \approx f_c \Delta R_{UF}$ where ΔR_{UF} is the distance between the folded and unfolded states of P5GA. Using $\Delta R_{UF} \approx 6nm$ we find that $\Delta G \approx 13kcal/mol$. The Vienna RNA package[12], which uses an entirely different free energy parameters for RNA, gives $\Delta G \approx 12.8kcal/mol$. This comparison shows that the SOP model can, for simple structures, give accurate results for stability. At $f_c = 15.4\,pN$, the transition barrier is about $\sim 1.5k_B T$. The **UBA** is more populated at this value of f_c.

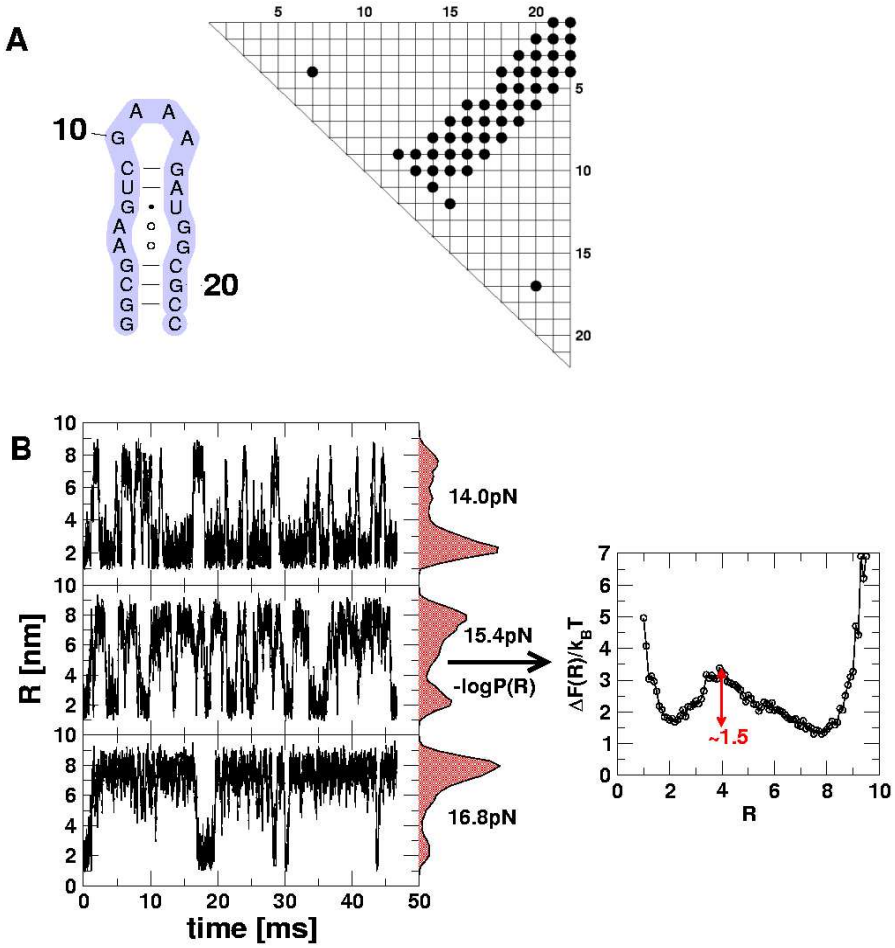


Figure 1. **A.** The secondary structure of P5GA hairpin and its contact map. **B.** The time-dependent fluctuations of P5GA hairpin between the folded ($R \approx 1.5nm$) and unfolded ($R \approx 8nm$) states. The end-to-end distance changes spontaneously between two values. The force-clamped dynamics of the P5GA hairpin is probed for $\sim 45\ ms$. The histograms $P(R)$ at $f_c = 14.0, 15.4, 16.8\ pN$ are shown in red. The free energy profile $\Delta F(R)$ as a function of R for $f_c = 15.4\ pN$ on the right shows two-state behavior.

The observed transition times are much shorter than the residence times in each basin of attraction which is also a reflection of the underlying cooperativity of the all-or-none of nature of hopping between **UBA** and **NBA**.

Force-ramp: We also performed force-ramp simulations by subjecting the P5GA hairpin to a continuously changing force, i.e., varying the loading rate (Methods). The simplicity of the SOP model allows us to use values of r_f that are comparable to those used in LOT experiments. At $r_f = 45\ pN/s$ ($\sim 10 r_f^{LOT}$) the force-extension curves show a transition

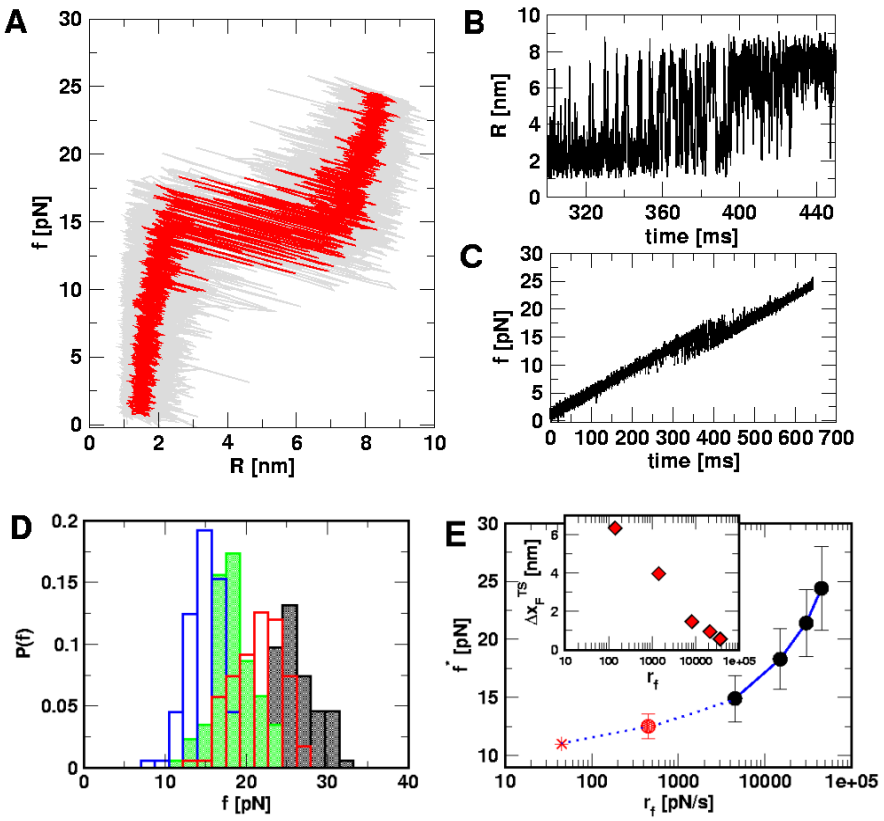


Figure 2. Force-ramp unfolding of P5GA hairpin. **A.** An example of FEC at the loading rate $r_f = 45\ pN/s$ ($k_s = 0.07\ pN/nm$, $v = 0.64\ \mu m/s$). The data points are recorded every 50 μs (grey color), but for better illustration the running average is displayed every 500 μs (red color). **B.** The end-to-end distance (R) as a function of time. **C.** Time dependence of f. The loading rate $r_f = df/dt$ is nearly a constant. **D.** Distribution of unbinding forces from 100 trajectories at four loading rates $r_f = r_f^o$, $\frac{10}{3}r_f^o$, $\frac{20}{3}r_f^o$, and $10r_f^o$ ($r_f^o = 4.5 \times 10^3$ pN/s ($k_s = 0.7\ pN/nm$, $v = 6.4\ \mu m/s$)). The red star is the rupture force value ($f^* = 12.5\ pN$) at $r_f = 45$ pN/s. **E.** Plot f^*, the most probable unfolding force, as a function of r_f. The position of transition state Δx_F^{TS} is computed using $\Delta x_F^{TS} = k_B T \frac{\Delta \log r_f}{\Delta f^*}$. The inset shows the variation of Δx_F^{TS} as a function of r_f.

to the **UBA** at around $f \sim 13\ pN$ (Fig. 2-A). As the force dynamically increases we observe bistable fluctuations in the FEC between the **NBA** and the **UBA** just as when force is held constant (Fig. 2-A). The conformational fluctuations between the two states are unambiguously seen in the time-dependence of the end-to-end distance ($R(t)$) (Fig. 2-B). As time progresses the force is ramped up which results in global unfolding ($R \approx 8\ nm$) for $t > 400\ ms$ (Fig. 2-B). During the time scale of simulation we find frequent and sharp transition between the **UBA** and the **NBA** (Fig. 2-B).

The location of the unfolding transition state Δx_F^{TS} for proteins and RNA is often estimated from force-ramp experiments using the variation of the most probable rupture

force with r_f ($[f^*, \log r_f]$ plot). The loading rate, $r_f = df(t)/dt$, and can be accurately estimated from the slope of the time dependence of $f(t)$ as a function of time. The slope of $f(t)$ as a function of t (Fig. 2-C) is nearly the same as $r_f \approx k_s \times v$, where v is the pulling speed. Strictly speaking, $r_f = k_{eff} \times v$ with $k_{eff}^{-1} = k_s^{-1} + k_{mol}^{-1} + k_{linker}^{-1}$ and k_s, k_{mol}, and k_{linker} are the spring constants of the optical trap, the RNA molecule, and linker, respectively. Typically $k_s \ll k_{mol}, k_{linker}$, thus $k_{eff} \approx k_s$[13]. Throughout the paper we obtain the loading rate using $r_f = k_s v$.

From the force distributions, computed at four different loading rates (Fig. 2-D), we observe that the most probable rupture force (f^*) does not increase logarithmically over a wide range of loading rates (Fig. 2-E). Only if the range of r_f is restricted f^* changes linearly with r_f[14]. The location of the transition state (Δx_F^{TS}) is usually calculated using $f^* \approx (\frac{k_B T}{\Delta x_F^{TS}}) \log r_f$[14] which may be reasonable as long as r_f range is small. However, the $[f^*, \log r_f]$ plot is highly nonlinear (Fig. 2-E). If we use linear regression to analyze the $[f^*, \log r_f]$ plot then $\Delta x_F^{TS} \sim 0.8\ nm$ for the distance between the **NBA** and the transition state. The small value of Δx_F^{TS} is a consequence of the large variation of Δx_F^{TS} as r_f is changed[11,15]. If the loading rate is varied over a broad range, the r_f-dependence of Δx_F^{TS} is manifested as a pronounced convex curvature[7] in the $[f^*, \log r_f]$ plot (Fig. 2-E). Based on the equilibrium free energy profile $F(R)$ as a function of R[11] we expect that for P5GA $\Delta x_F^{TS} \approx 3\ nm$ if r_f is small. Indeed, from the constant force simulation results in Fig. 1-B we find $\Delta x_F^{TS} \approx \frac{R_U - R_F}{2}$. Thus, the slope of $[f^*, \log r_f]$ should decrease (Δx_F^{TS} increases) as r_f decreases. To illustrate the dramatic movement in the transition state we have calculate Δx_F^{TS} using f^* values for two consecutive values of r_f. For example, using $f^* \approx 12.5\ pN$ at $r_f = 450\ pN/s$ and $f^* \approx 14.9\ pN$ at $r_f = 4.5 \times 10^3\ pN/s$ (a value that can be realized in AFM experiments) we obtain $\Delta x_F^{TS} \approx 4.0\ nm$. From the values of f^* at five values of r_f we find that Δx_F^{TS} can move dramatically (Fig. 2-E). In particular, we find that Δx_F^{TS} changes by nearly a factor of ten as the loading rate is decreased to values that are accessible in LOT experiments (see inset in Fig. 2-E). Because of the nearly logarithmic variation of Δx_F^{TS} on r_f over a narrow range of r_f we do not expect Δx_F^{TS} to change appreciably if r_f is lowered from $45pN/s$ to $\approx 5pN/s$. At high r_f or f_c the unfolded state is greatly stabilized compared to the folded state. From Hammond's postulate[16], generalized to mechanical unfolding[7], it follows that as f_c increases Δx_F^{TS} should move closer to the native state. The simulations are, therefore, in accord with the Hammond's postulate.

3 Conclusions

We have used the self-organized polymer representation of RNA hairpins to predict their mechanical unfolding trajectories. Constant force and force-ramp simulations show that dramatic changes in the force profiles take place as the loading rates and the values of the force are varied. If the force is varied over a wide range then regions of the energy landscape that cannot be accessed in conventional experiments can be probed. However, in order to realize the full utility of the single molecule force spectroscopy it becomes necessary to use force in distinct modes (constant force, force-ramp, and other combinations) along with reliable computations that can mimic the experimental conditions as closely as possible. We conclude the paper with the following additional remarks.

Transition state movements show changes from plastic to brittle behavior: The small size and simple architecture of RNA hairpins has allowed us to explore their response to force over a range of r_f that spans four orders of magnitude. The lowest r_f value is close to those used in LOT experiments. A key prediction of our simulations is that the location of the transition state for P5GA moves dramatically from about 6 nm at low r_f to about 0.5 nm at high r_f (see inset to Fig. 2-E). The large value of Δx_F^{TS} at low r_f suggests that P5GA is plastic while the small Δx_F^{TS} at high r_f is suggestive of brittle behavior. The mechanical properties of RNA structures can be drastically altered by varying the loading rate which is reminiscent of the changes in the visco-elastic behavior of polymeric materials that changes with frequency. The transformation from plastic to brittle behavior can be captured using the fragility index[15] used to describe mechanical unfolding of hairpins. Although we have discussed the r_f-dependent movement of the transition states using P5GA as an example we predict that this result is general and should be observed in other RNA structures as well.

Acknowledgements

This work was supported in part by a grant from the National Science Foundation through grant number NSF CHE-05-14056.

References

1. J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Reversible unfolding of single RNA molecules by mechanical force*, Science, **292**, 733–737, 2001.
2. B. Onoa, S. Dumont, J. Liphardt, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, *Identifying Kinetic Barriers to Mechanical Unfolding of the* T. thermophila *Ribozyme*, Science, **299**, 1892–1895, 2003.
3. C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, *Entropic Elasticity of λ-Phase DNA*, Science, **265**, no. 5178, 1599–1600, 1994.
4. J. F. Marko and E. D. Siggia, *Bending and Twisting Elasticity of DNA*, Macromolecules, **27**, 981–988, 1996.
5. K. Kremer and G. S. Grest, *Dynamics of entangled linear polymer melts: A molecular-dynamics simulation*, J. Chem. Phys., **92**, 5057–5086, 1990.
6. C. Hyeon, R. I. Dima, and D. Thirumalai, *Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins*, Structure, **14**, 1633–1645, 2006.
7. C. Hyeon and D. Thirumalai, *Forced-unfolding and force-quench refolding of RNA hairpins*, Biophys. J., **90**, 3410–3427, 2006.
8. D. L. Ermak and J. A. McCammon, *Brownian dynamics with hydrodynamic interactions*, J. Chem. Phys., **69**, 1352–1369, 1978.
9. T. Veitshans, D.K. Klimov, and D. Thirumalai, *Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties*, Folding Des., **2**, 1–22, 1996.
10. S. Rudisser and I. Tinoco, Jr., *Solution Structure of Cobalt(III)Hexammine Complexed to the GAAA Tetraloop, and Metal-ion Binding to GA Mismatches*, J. Mol. Biol., **295**, 1211–1223, 2000.

11. C. Hyeon and D. Thirumalai, *Mechanical unfolding of RNA hairpins*, Proc. Natl. Acad. Sci., **102**, 6789–6794, 2005.
12. I. V. Hofacker, *Vienna RNA secondary structure server*, Nucl. Acids. Res., **31**, no. 13, 3429–3431, 2003.
13. M. Manosas and F. Ritort, *Thermodynamic and Kinetic Aspects of RNA pulling experiments*, Biophys. J., **88**, 3224–3242, 2005.
14. E. Evans and K. Ritchie, *Dynamic Strength of Molecular Adhesion Bonds*, Biophys. J., **72**, 1541–1555, 1997.
15. M. Manosas, D. Collin, and F. Ritort, *Force-Dependent Fragility in RNA Hairpins*, Phys. Rev. Lett., **96**, 218301, 2006.
16. G. S. Hammond, *A correlation of reaction rates*, J. Am. Chem. Soc., **77**, 334–338, 1953.

Simulation of Linker Histone-Chromatin Interactions

Georgi V. Pachov[1], **Razif R. Gabdoulline**[1,2] **and Rebecca C. Wade**[1]

[1] EML Research gGmbH, Molecular and Cellular Modeling Group,
Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany

[2] Center for Modelling and Simulation in the Biosciences (BIOMS),
Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
E-mail: {*georgi.pachov, rebecca.wade*}@eml-r.villa-bosch.de

The dynamics and interactions between protein-protein and protein-DNA molecules appearing on different time and length scales in the cell are of fundamental interest. In chromatin, linker histone binds to the highly charged nucleosome facilitated by electrostatic attraction. However, its position and orientation with respect to the nucleosome core are unknown. Using an implicit representation of the solvent, rigid-body docking of a linker histone to the nucleosome has been performed by BD simulations. This reveals how the protein binds to DNA on the nucleosome. Two distinct binding sites on the linker histone have been identified by the docking simulation.

1 Introduction

In the cell nucleus, DNA wraps around histone proteins (forming nucleosome particles) and packs to a highly negatively charged structure, called chromatin fiber. The positioning of the nucleosomes within the fiber depends on the presence or lack of a linker histone protein[1]. However, the exact position and orientation of the linker histone with respect to the nucleosome particle is an unresolved problem, although some recent studies have proposed different binding modes[2–4].

Experimentally, it is known that the linker histone binds to the nucleosomal DNA[5] and to at least one of the linker DNAs. This, indicates the important role of the linker DNA's length in chromatin compaction[6]. Since the linker histone appears to be crucial for the opening and closing of the chromatin fiber during transcription and many other biological processes related to the linker histone and nucleosome in the higher order structure of the genome, we performed molecular docking simulations using a Brownian Dynamics (BD) algorithm[7]. This study aims to reveal the binding position and orientation of linker histone to the nucleosome.

2 Method

Using the crystal structure of the nucleosome core particle[8] (NCP, Protein Data Bank - PDB code 1kx5) and of the globular domain of the linker histone H5[9] (GH5, PDB code 1hst), we have performed molecular docking simulations with the Simulation of Diffusional Association (SDA) software package[10].

Chain B of the linker histone was used for the simulation. Since the histone tails and residues 119-128 in chain C are much floppier than the other parts of the nucleosome they were removed from it. In order to include linker DNAs, the tetranucleosome structure[12] was used. The DNAs from both nucleosome structures, the one without tails and one from

the tetranucleosome, were aligned with Pymol[13] and an additional 20 bp DNA was added to the nucleosome without tails. We will refer to this structure as tNCP. In addition, one of the linker DNAs in the tNCP was *shifted* away from the other linker DNA allowing more space for the linker histone to penetrate close to the nucleosomal DNA (see Fig. 1). This structure will be designated as sNCP. Partial charges and atomic radii were assigned with the PDB2PQR program[14] using the Amber force field. The electrostatic potential was computed for the NCPs and GH5 by solving the nonlinear Poisson-Boltzmann equation on grids with 257^3 and 200^3 points, respectively. The programs used were APBS[15] and UHBD[16]. The temperature was set to 300 K, the solvent dielectric constant to 78, the solute to 2 and the ionic strength to 100 mM.

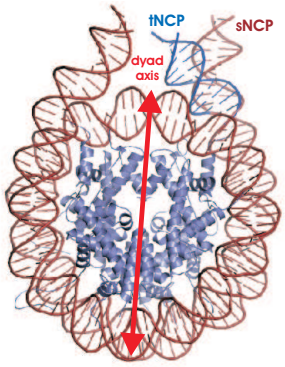


Figure 1. Difference between sNCP and tNCP structures. The linker DNA (ruby) in sNCP is shifted away in comparison with the linker DNA (blue) in tNCP. The red line shows the dyad axis of the nucleosome.

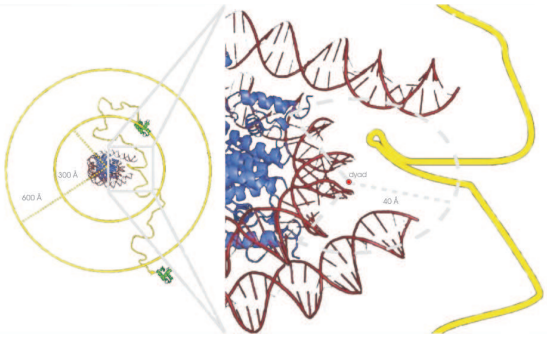


Figure 2. The picture shows the BD method, in which each trajectory of GH5 is started randomly on an inner sphere surface centered on the NCP and truncated on an outer sphere surface. The constraint region, where the docking solutions are recorded, is given in the zoomed picture. The red dot is the dyad point at the nucleosomal DNA.

To reduce the number of partial charge sites on the biomolecules, the ECM[17] program was applied. It derives effective charges fitted to reproduce the electrostatic potential in a uniform dielectric medium. On DNA, the charges were assigned to the *P* atoms only. The net formal charges are -237 e for the NCPs and 11 e for the GH5. An exclusion volume grid with 0.5 Å spacing was assigned to avoid van der Waals overlaps. The BD docking simulations were carried out with the SDA package modified in such a way that the docked complexes are recorded only if they satisfy predefined constraints (Fig. 2). For our case, we used two constraints: (i) a center-to-center between both particles ($\leq$ 73 Å) and (ii) dyad point-center distance between the dyad at the nucleosomal DNA and the center of the GH5 ($\leq$ 40 Å). In the simulations, the molecules are modeled as rigid bodies with the short-range attractive interactions neglected. The simulation method has been described in detail in the references[7,10,11]. The trajectories start at a center-to-center distance b=300 Å and finish at c= 600 Å (Fig. 2). The time step was set to 0.25 ps for center-to-center distances up to 130 Å and it increased linearly for larger distances. The interaction energies as well as the coordinates of the complexes satisfying the constraints were recorded.

For both sNCP and tNCP, five different runs with different random generators, i.e. different starting positions and orientations, were performed. For each nucleosome structure

25000 complexes were recorded and the 2500 lowest energy docked complexes were clustered with the program `clust`[18]. After clustering the representatives of 5 and 6 clusters were analyzed for the sNCP and tNCP, respectively.

3 Results and Discussion

The docking solutions for the linker histone GH5 were extensively investigated. Previous theoretical and experimental studies[2–4] identified the most important residues in binding, but the controversial results obtained left the problem open. Therefore, we aimed to find out which residues contribute to binding, how they influence the position and orientation of the linker histone with respect to the nucleosome and whether the distance between the linker DNAs, their mutual orientation and position affects the linker histone binding mode. The results are shown in Table 1 and Table 2 for sNCP and tNCP, respectively. The proximity of every residue considered as important to the nucleosome is given in two related ways as: (i) general - close to the nucleosomal DNA (nDNA), linker DNA (lDNA) or neither to the nDNA nor to the lDNA (Nn/lDNA) and (ii) average RMSD appearance $\bar{d}$ to the dyad point at the nucleosomal DNA (from all 5/6 representatives). The average interaction energy for each cluster varies between [-168;-104] kT and [-229;-128] kT for sNCP and tNCP, respectively.

Residue /	**nDNA** (N^{com})	**lDNA** (N^{com})	**Nn/lDNA** (N^{com})	$\bar{\mathbf{d}}$, Å
R42	3 (2357)	1 (72)	1 (68)	18.1
R94	3 (2357)	2 (140)	0 (0)	20.6
K97	3 (2357)	2 (140)	0 (0)	24.3
K85	2 (1981)	3 (516)	0 (0)	20.2
K40	1 (1676)	0 (0)	4 (821)	26.3
R47	1 (72)	3 (749)	1 (1676)	26.1
K69	1 (72)	3 (749)	1 (1676)	28.3
R73	1 (72)	3 (749)	1 (1676)	29.8
R74	1 (72)	2 (373)	2 (2052)	34.6
K52	2 (377)	2 (2052)	1 (68)	29.5
K55	1 (72)	2 (681)	2 (1744)	29.7
K59	0 (0)	1 (1676)	4 (821)	40.2

Table 1. Number of representatives of the clusters according to each residue located within 15 Å of the nucleosomal DNA (nDNA), linker DNAs (lDNA) and away (Nn/lDNA, $\leq$ 15 Å) from either nDNA or nDNA. All 2500 docked solutions (in brackets) are presented by 5 representatives. The distance $\bar{d}$ is the average RMSD from a specified residue (atom N_ζ for K and C_ζ for R) to the dyad point on the nucleosomal DNA for all representatives.

It is clearly seen for sNCP and tNCP that the site with residues Arg42, Arg94 and Lys97 appears to be closer to the nucleosomal DNA than to any of the linker DNAs (Fig. 3). Moreover, residue Arg42 takes the closest position to either nDNA or lDNA, i.e. less than 5 Å in comparison with the other residues[19]. On the other hand, the site with residues Arg47, Lys69, Arg73 and Arg74 shows a preference either for one of the linker DNA or for being away. This is in contrast to the theoretical model of Brown *et al.*[2]. The third distinct site in GH5 consisting of residues Lys52, Lys55 and Lys59 is mostly directed

Residue /	nDNA (N^{com})	lDNA (N^{com})	Nn/lDNA (N^{com})	$\bar{d}$, Å
R42	4 (1842)	2 (649)	0 (0)	24.6
R94	4 (1842)	2 (649)	0 (0)	24.5
K97	4 (1842)	1 (74)	1 (575)	24.3
K85	2 (480)	2 (1291)	2 (720)	27.5
K40	3 (554)	2 (1362)	1 (575)	22.7
R47	0 (0)	3 (1743)	3 (748)	31.46
K69	0 (0)	3 (1101)	3 (1390)	32.4
R73	0 (0)	3 (1101)	3 (1390)	33.7
R74	0 (0)	2 (526)	4 (1965)	33.7
K52	0 (0)	4 (1965)	2 (526)	27.3
K55	0 (0)	3 (1319)	3 (1172)	28.8
K59	1 (74)	4 (1965)	1 (452)	30.3

Table 2. The definition of the parameters is the same as in Table 1, but for the tNCP structure.

away from the NCP, although in the sNCP, Lys52 appears close to the nucleosome. Similar behaviour is observed for Lys40 as well. According to the docking data, it seems that the strongest contribution to binding close to the dyad point comes from residue Arg42 rather than residue Lys85 as stated by Fan and Roberts[3]. However, the general trend for the residues participating in the binding sites in both papers[2,3] is conserved in our simulations. It should be noted that there is a slight difference between the linker histone molecule used in the Brown *et al.*[2] model on one hand and in ours and in the Roberts model[3] on the other. In the former model, molecule A of the X-ray structure of GH5 resolved by Ramakrishnan *et al.*[9], has been used for homology modeling of H1[0] rather than molecule B used here and by Fan and Roberts[3]. The main difference is in the orientation of Lys85, which is directed to the Arg42 site in molecule B, while in molecule A it appears close to the helix 3 site. In addition, the relative orientation of Arg42 and Lys40 differs in both molecules by approximately 90°. Therefore, the binding sites identified in both studies do not differ

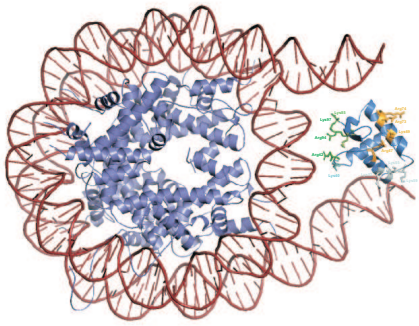


Figure 3. Representative of the largest (1676 solutions) and second lowest energy (-156 kT) cluster of the docked structures to sNCP.

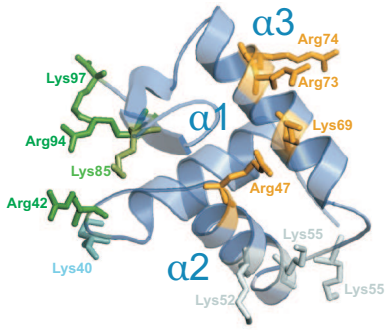


Figure 4. Linker histone structure with the analysed residues and binding sites according to Table 1 and Table 2. The orientation is the same as in Fig. 3

much from each other, when taking into account the above considerations. In fact, this is the reason why Fan and Roberts[3] consider Lys85 as a separate binding site while Brown *et al*.[2] assign Lys85 to the helix-3 binding site. Another difference in the final results comes from the assignment of Lys40 to a binding and nonbinding site in the Fan and Roberts[3] and the Brown *et al*.[2] models, respectively. The latter is deduced from a FRAP experiment performed with the $H1^0$ linker histone and is in agreement with our study only for the sNCP structure (see Table 1). However, the interpretation of the results for the tNCP has to be taken with care since this crystal structure does not include the linker histone. Apparently, due to the closeness of both linker DNAs in the tNCP structure, GH5 cannot penetrate to the nucleosomal DNA with the helix-3 site (see Table 2). On the other hand, both β sheets and the loop between helix-1 and helix-2, which include Arg42, Arg94, Lys97, Lys85 and Lys40 residues, need less space than helix-3 to fit in between the linker DNAs as indicated in Table 2. Therefore, from this point of view our results for sNCP seem more consistent with the experimental data[2].

4 Conclusion

The computational docking of a linker histone to the nucleosome has revealed two distinct binding sites on GH5 consisting of: (1) Arg42, Arg94, Lys97 with partial contribution of Lys85, which most probably bind to the nucleosomal DNA and (2) Arg47, Lys69, Arg73 and Arg74, which seem to contact one of the linker DNAs (see Fig. 4). The site including Lys52, Lys55 and Lys59 appears to be a nonbinding site with respect to the NCP, although some of the residues are located close to the sNCP in some of the solutions. In addition, the docking results clearly indicate the importance of the distance and orientation between both DNA arms, which might change the binding modes significantly. Therefore, including the flexibility in future investigations is necessary.

To obtain more insights into the exact contribution and position of each residue to binding, we aim to investigate the effects of single and double mutations in GH5 on binding. Moreover, simulation of diffusional motion of both molecules for computation of the association rates and encounter complex formation is under study. All this would be important for further development of models regarding the chromatin structure and compaction.

Acknowledgments

The work was supported by the German Research Foundation (DFG), the Center for Modelling and Simulation in the Biosciences (BIOMS) and the Klaus Tschira Foundation (KTF).

References

1. P. JJ. Robinson and D. Rhodes, *Structure of the '30 nm' chromatin fibre: A key role for the linker histone*, Curr. Opin. Str. Biol. **16**, 336–343, 2006.
2. D. T. Brown, T. Izard and T. Misteli, *Mapping the interaction surface of linker histone $H1^0$ with the nucleosome of native chromatin in vivo*, Nat. Str. Mol. Biol. **13**, 250–255, 2006.

3. L. Fan and V. A. Roberts, *Complex of linker histone H5 with the nucleosome and its implications for chromatin packing*, PNAS **103**, 8384–8389, 2006.
4. M. M. S. Bharath, N. R. Chandra and M. R. S. Rao, *Molecular modeling of the chromatosome particle*, Nucl. Acid Res. **31**, 4264–4274, 2003.
5. A. A. Travers, *The location of linker histone in the nucleosome*, Trends Biochem. Sci. **24**, 4–7, 1999.
6. P. JJ. Robinson and D. Rhodes, *EM measurements define the dimension of the "30-nm" chromatin fiber: Evidence for a compact, interdigitated structure*, PNAS **103**, 6506–6511, 2006.
7. D. L. Ermak and J. A. McCammon, *Brownian dynamics with hydrodynamic interactions*, J. Chem. Phys. **69**, 1352–1360, 1978.
8. C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder and T. J. Richmond, *Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å*, J. Mol. Biol. **319**, 1097–1113, 2002.
9. V. Ramakrishnan, J. T. Finch, V. Graziano, P. L. Lee and R. M. Sweet, *Crystal structure of globular domain of histone H5 and its implications for nucleosome binding*, Nature **362**, 219–223, 1993.
10. R. R. Gabdoulline, R. C. Wade, *Brownian dynamics simulation of protein-protein diffusional encounter*, Methods. **14**, 329–341, 1998.
`http://projects.villa-bosch.de/mcmsoft/sda/4.23/index.html`
11. R. R. Gabdoulline, R. C. Wade, *Protein-Protein association: Investigation of factors influencing association rates by brownian dynamics simulations*, J. Mol. Biol. **306**, 1139–1155, 2001.
12. T. Schalch, S. Duda, D. F. Sargent, T. J. Richmond, *X-ray structure of a tetranucleosome and its implications for the chromatin fibre*, Nature Lett. **436**, 138–141, 2005.
13. W. L. DeLano, *The Pymol Molecular Graphics System (2002)*.
`http://www.pymol.org`
14. T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, N. A. Baker, *PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations*, Nucl. Acids Res. **32**, W665–W667, 2004.
15. N. A. Baker, D. Sept, S. Joseph, M. J. Holst, J. A. McCammon *Electrostatics of nanosystems: application to microtubules and the ribosome*, Proc. Natl. Acad. Sci. **98**, 10037–10041, 2001.
16. M. E. Davis, J. D. Madura, B. A. Luty, J. A. McCammon, *Electrostatic and diffusion of molecules in solution: simulations with the University-of-Houston- Brownian Dynamics program*, Comput. Phys. Commun. **62**, 187–197, 1991.
17. R. R. Gabdoulline, R. C. Wade, *Effective charges for macromolecules in solvent*, J. Phys. Chem. **100**, 3868–3878, 1996.
18. Unpublished software developed by Domantas Motiejunas (MCM Group, EML Research gGmbH).
19. Unpublished data.

Conformational Study of Amyloid Beta (ABeta) Peptide

Priya Anand and F. S. Nandel

Department of Biophysics,
Panjab University, Chandigarh-160014, India
E-mail: priyaanand_27@hotmail.com

ALZHEIMER'S DISEASE is the most common cause of senile dementia. Pathological hallmarks of the disease include senile plaques and neurofibrillary tangles. The plaques result from the transformation of soluble Aβ protein monomer into insoluble aggregates. Determination of the molecular structure of Aβ fibrils by using molecular dynamic simulation softwares provide an insight into the precise arrangement of monomers that would allow targeting the critical steps in fibrilogenesis process. The optimized geometry for the (Aβ_{15-42}) monomer is a U-like structure.

1 Introduction

Alzheimer's disease (AD) is a neurological disorder, affecting approximately 12.5 and 47.2% of the population in the United States over the ages 65 and 85, respectively. Defining features are formation and progressive deposition of insoluble amyloid plaques and presence of neurofibrillary tangles. ABeta-peptide is 39-43 amino acids polypeptide with heterogeneous termini that is generated from the cleavage of a larger amyloid precursor protein (APP), which under appropriate cellular conditions, aggregate as senile plaques, which is a critical step in the neurodegenerative processes associated with AD.

A detailed understanding of the structural properties of amyloid monomer could play a crucial role in the understanding the molecular mechanism of the disease. Despite the limitations of conventional NMR or X-ray in the amyloid fibril structural study, some structural information has emerged from various techniques[1,2] mass spectroscopy, and solid state NMR spectroscopy. However, the structural characterization of the Abeta monomer still remains difficult due to its tendency to aggregate.

Complimentary to the experimental studies, computer simulations can yield valuable information on structure, and stability, of the monomer resulting in the possible fibril formation mechanism of the ABeta protein. We have undertaken molecular dynamics (MD) simulations study on monomer of different short length peptide i.e 1-16 (referred as N-terminal region) and 15-42 residue fragments in aqueous solution to complement, and to add further extend the experimental solution studies of the beta-peptides and to explore the structural stability, hydration and dynamics of peptide fibrils. The N-terminal region of ABeta has been shown to be flexible and accessible within amyloid fibrils and the remaining 29-42 hydrophobic is shown to adopts a beta-sheet conformation. N-terminal constitutes an attractive therapeutic target for active or passive immunization approaches, as illustrated by the ability of monoclonal antibiotics directed towards this region to dissociate amyloid fibrils.

2 Material and Methods

The peptide Aβ originally consists of 42 amino-acid residues: [Asp-Ala-Glu-Phe-Arg-His-Asp-Ser-Gly-Tyr-Glu-Val-His-His-Gln-Lys-Leu-Val-Phe-Phe-Ala-Glu-Asp-Val-Gly-Ser-Asn-Lys-Gly-Ala-Ile-Ile-Gly-Leu-Met-Val-Gly-Gly-Val-Val-Ala-Ala], which is usually expressed as Aβ_{1-42} . The conformational analysis of Aβ_{1-42} has been first carried out using quantum mechanical methods (QM) PCILO program (Pertubative Configuration Interaction Using Localized Orbitals) on Sun W, Ultra 5-10; sparc was used, for which the Aβ_{1-42} peptides were fragmented into smaller overlapping fragments with both the N- & C- terminal protected by acetyl and dimethylamine group respectively.

The (ϕ,ψ) values obtained from the quantum mechanical calculations were used as the starting geometry for molecular mechanics and molecular dynamics. Another structure used as the starting geometry for molecular dynamics was a linear extended starting structure, minimized structure obtained after running molecular dynamics in vacuum for 50ns at 300K using AMBER8 molecular dynamics.

For Molecular dynamics simulations Aβ_{1-42} was fragmented into N-terminal region with 1-16 residues peptide and C-terminal fragment with 15-42; (where no 1,16,15,42 indicates the residue no) peptide utilizing the program the GROMACS 3.3.1.

Energy minimization was carried out using steepest descent followed by conjugate gradient method. All atoms of the system were considered explicitly, and their interactions were computed by using the 43a1 force field with periodic boundary conditions. Peptide was solvated in a box by using SPIC216water molecules. All solute and solvent atoms were treated explicitly. Water molecules were added around the peptide to fill a octahedron box with walls at least 9nm, from any peptide atom. All simulations were performed using periodic boundary conditions to eliminate surface effects. Long-range electrostatic interactions were handled by using the particle mesh Ewald methodology. The solvated peptides were subsequently minimized. Position restraining done using weak coupling to a bath of constant temperature (To = 300 K, coupling time T = 0.1 ps), and the pressure controlled using weak coupling to a bath of constant pressure (Po = 1 bar, coupling time T = 0.5 ps). The production runs were done with the same pressure and temperature coupling constants as the restrained runs. From the trajectories obtained we extracted the potential energy, Root mean square deviation, radius of gyration and Surface area properties and number of hydrogen bonds formed during the simulation. For visualization of the structures and the interactions RASMOL, and VMD was used.

3 Results and Discussion

Molecular view suggests that optimized geometry for the (Aβ_{15-42}) monomer has a U-like structure (Fig A). Phi, psi values indicate that Aβ_{15-42} peptide does not adopt any regular secondary structure, instead a turn is present between residues 26-32. As displayed in figure, its amphipathic character is not lost i.e the longer arm has hydrophilic and charged amino acid residues while the smaller arm has hydrophobic branched residues exclusively. Side chain of most of the hydrophobic residues are oriented inwards with the exception of Leu 34, Val 36, Val 40 and Ala 42 which project outwards, that may lead to the association of one A with another A through hydrophobic interactions involving the C-terminal stretch. Analysis of the interactions stabilizing the U-like structure of Aβ_{15-42} , and identification

Figure 1. A molecular view of optimized (Aβ_{15-42}) and (Aβ_{1-16})peptide fragment in aqueous medium, Fig A and Fig B respectively

of the possible interactive regions in Abeta peptides will be utilized for construction of Aβ dimers and high-mers.

Molecular view suggests that optimized geometry for the Aβ_{1-16} monomer is stabilized by hydrogen bonds are formed primarily between the carbonyl oxygen and the amino group of the backbone as displayed in Figure B. At 300 K, the peptide is seen to exist 25% in a helical conformation. The phi, psi values of the various residues indicate the presence of gamma turn around the residue Glutamine 7 and Glycine 9. Though residues with polar side chains are present only a few of them (ARG5, GLu3, ASP1,TYR10 and Glu15) participate in hydrogen bond formation. From the graphical display of the molecule, it is apparent that the carbonyl oxygen is of His 6 backbone is near to the OH group of the aromatic ring of TYR 10 as the distance of the oxygen atom of the carbonyl group and OH moiety is 1.8 A . This may be interpreted as though the carbonyl moiety is involved in pi-pi interaction with the aromatic ring.

Acknowledgments

PA would like to thank Prof. Dr Fateh Singh Nandel, Department of Biophysics, Panjab University, Chandigarh, India for his kind guidance, and constructive criticism, extended to complete this project as part of my Ph.D program.

References

1. Barrow CJ, Yasuda A, Kenny PT, and Zagorsk MG, *Solution conformations and aggregational properties of synthetic Amyloid beta peptides of AD. Analysis of circular dichrosim spectra* J. Mol. Biol. **825 (4)**, 1075–93, 1992.
2. Benzinger T.L.S., Gregory DM Burkoth T.S., Miller Auer, H. Lynn D.G. Botto, R.E. and Meredith S.C. (2000) Biochemistry 39, 3941-3499.

Global Persistence Exponent of the Helix-Coil Transition in Polypeptides

Everaldo Arashiro[1], **José Roberto Drugowich de Felício**[2], **and Ulrich H. E. Hansmann**[1]

[1] John von Neumann Institute for Computing,
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: {e.arashiro, u.hansmann}@fz-juelich.de

[2] Departamento de Física e Matemática, FFCLRP, Universidade de São Paulo,
Ribeirão Preto, São Paulo CEP 14040-901, Brazil
E-mail: drugo@usp.br

The dynamical exponent θ_g that governs the behavior of the global persistence probability is obtained through short-time Monte Carlo simulations of the helix-coil transition. Our simulations are based on a detailed, all-atom representation of the molecules and an implicit solvation model to approximate the interaction with the surrounding solvent. Our results obtained for the polyalanine and the 34-residue human parathyroid fragment PTH(1-34) are in good agreement each other, and indicate universality of the helix-coil transition in proteinlike molecules.

1 Introduction

After the works by Janssen, Schaub, and Schmittmann[1] and Huse[2], much attention has been devoted to study of phase transitions and critical phenomena of dynamical systems at the early stage of its time evolution. By using renormalization group techniques and numerical calculations, respectively, they showed that universality and scaling behavior are already present in this stage (henceforth named as short-time stage) after quenching from high temperatures to the critical one.

The values of the critical indices obtained through this technique[3,4], have shown that short-time Monte Carlo simulations are very efficient and can be used to estimate, with good precision, the statical exponents and the dynamical exponent z.

Far from equilibrium, another dynamic critical exponent was proposed by Majumdar et al.[5] studying the behavior of the global persistence probability $P(t)$ that the order parameter has not changed its sign up to the time t. The global persistence probability $P(t)$ can be defined as

$$P(t) = 1 - \sum_{t'=1}^{t} \rho(t'), \tag{1}$$

where $\rho(t')$ is the fraction of the samples that have changed their state for the first time at the instant t'.

At criticality, $P(t)$ is expected to decay algebraically as

$$P(t) \tilde{} t^{-\theta_g} \tag{2}$$

where θ_g is the global persistence exponent. Since then, the study of the persistence behavior have attracted much interest, playing an important role in the study of systems far from equilibrium[6–8] .

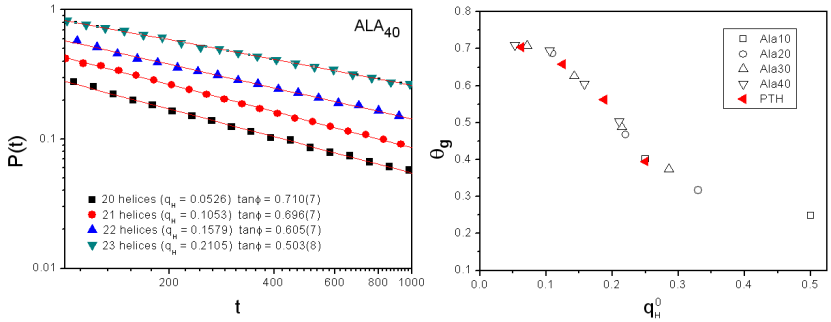


Figure 1. Left: Log-log plot of the time series of the global persistency $P(t)$ for Ala_{40} and various initial order parameter values q_H^0. Right: Persistence exponent as a function q_H^0. Open symbols are the values for polyalanine and the filled symbols are the values for PTH(1-34).

In this work, we have studied the dynamic critical behavior of helical proteins via short-time Monte Carlo simulations. We evaluated the dynamic critical exponent θ_g for $(\mathrm{Ala})_N$ chains (N=10, 20, 30 and 40) to verify the effects of finite size. Our investigation is later extended toward the 34-residue human parathyroid fragment PTH(1-34).

2 Model

Our short-time MC simulations of the helix-coil transition are based on a detailed, all-atom representation of proteins. The interaction between the atoms is described by a standard force field, ECEPP/2 (Empirical Conformational Energy Program for Peptides) as implemented in the program package SMMP (Simple Molecular Mechanics for Proteins)[9] . The interactions between our polypeptides and surrounding water are approximated by means of an implicit water model, which assumes that the solvation (free) energy is proportional to the solvent accessible surface area and utilizes the parameter set of Ref. 10 that is often used in conjunction with the ECEPP force field.

Crucial for our analysis is the definition of an order parameter. Our analog to the magnetization in spin systems is the number of helical residues $q_H = 2 * \langle n_H(T) \rangle / (N - 2) - 1$. Here we define a residue as helical if its backbone dihedral angles (ϕ, ψ) , take values in the range $(-70^\circ \pm 30^\circ, -37^\circ \pm 30^\circ)$ (their common values in an α helix), and the residue exhibits the hydrogen bonding pattern observed in α helices. The normalization factor $N-2$ (N the number of residues) is chosen because the flexible terminal residues are usually not part of an α helix. Our definition ensures that $-1 \leq q_H \leq 1$ and $q_H(T_c) = 0$.

3 Results

We start our investigation of dynamic critical exponent θ_g in the helix-coil transition by simulating polyalanine chains of lengths N=10, 20, 30, and 40 in your respectively critical temperatura $\mathrm{T}_c = 315$, 415, 450 and 470 which were found in our earlier works[11,12]. Our results are averaged over 5000 independent runs . Errors are estimated by dividing these

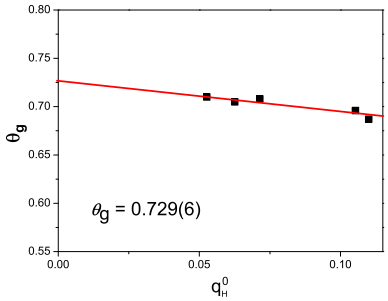


Figure 2. The scaling obtained for the 5 smaller q_H^0. The persistence exponent found is $\theta_g = 0.729(6)$.

5000 runs in bins of 100 runs and calculating the fluctuation of the averages calculated for each bin.

In order to estimate the exponent θ_g one needs to prepare the initial states carefully to obtain precise values of the initial $q_H^0 = q_H(t = 0)$. The exponent q_H^0 is obtained from Eq. 2. Figure 1 (Left) shows the corresponding log-log plots of $P(t)$ for Ala_{40} and various initial values of q_H^0 . This value, and the ones corresponding to polyalanine chains of lenght N= 10, 20 and 30 are listed in Tab. 1.

q_H^0	Ala_{10}	q_H^0	Ala_{20}	q_H^0	Ala_{30}	q_H^0	Ala_{40}
0.25	0.402(6)	0.11	0.687(7)	0.0714	0.708(7)	0.0526	0.710(8)
0.50	0.248(7)	0.22	0.469(7)	0.1429	0.626(6)	0.1053	0.696(7)
		0.33	0.317(8)	0.2143	0.488(7)	0.1579	0.605(7)
				0.2857	0.374(7)	0.2105	0.503(7)

Table 1. Global persistence exponent θ_g for polyalanine chains of length N = 10, 20, 30 and 40.

q_H^0	$PTH(1-34)$
0.0625	0.705(6)
0.125	0.658(7)
0.1875	0.562(6)
0.25	0.395(7)

Table 2. Global persistence exponent θ_g for the 34-residue human parathyroid fragment PTH(1-34).

The general behavior of the global persistence for each polyalanina indicate that when q_H^0 decreases, the value for the persistence exponent θ_g increases. In Fig 1 (Right) we show the θ_g of all polyalanine as a function of the q_H^0. We can observe that the value of

θ_g increase quickly, with the decrease of the q_H^0, but tends to stabilize by θ_g = 0.7. We also estimate the persistence exponent for the PTH(1-34), T_c = 545 (Tab. 2). In Fig 2 we show that the values de θ_g for PTH(1-34) (closed simbols) have the same behavior those the polyalanine (open simbols). From the 5 smaller q_H^0 the persistence exponents $\theta_g = 0.729(6)$ was estimated for this helical protein.

4 Conclusion

The behavior of exponent $\theta_g = 0.729(6)$ for polyalanine and the 34-residue human parathyroid fragment PTH(1-34) are in good agreement each other, and indicate universality of the helix-coil transition in proteinlike molecules.

Acknowledgments

Two of the authors (E.A. and J.R.D.F.) acknowledge support by Capes and CNPq (Brazil), and another author (U.H.E.H.) was supported by a research grant from the National Science Foundation (CHE-0313618).

References

1. H. K. Janssen, B. Schaub, and B. Schmittmann, Z. Phys. B: Condens. Matter **73**, 539 (1989).
2. D.A. Huse, Phys. Rev. B **40**, 304 (1989).
3. B. Zheng, Int. J. Mod. Phys. B **12**, 1419 (1998).
4. E. Arashiro, J.R. Drugowich de Felicio, Phys. Rev. E 67 046123 (2003).
5. S. N. Majumdar, A. J. Bray, S. J. Cornell and C. Sire, Phys. Rev. Lett. **77** 3704 (1996).
6. S. N. Majumdar and A. J. Bray, Phys. Rev. Lett. **91**, 030602 (2003).
7. H. A. Fernandes, E. Arashiro, J. R. Drugowich de Felicio, A. A. Caparica, Physica A **366**, 255 (2006).
8. R. da Silva, N. A. Alves, J. R. Drugowich de Felicio, Phys. Rev. E **67**, 057102 (2003).
9. F. Eisenmenger, U. H. E. Hansmann, Sh. Hayryan, and C.-K. Hu, Comput.Phys. Commun. **138**, 192 (2001).
10. T. Ooi, M. Obatake, G. Nemethy, and H. A. Scheraga, Proc. Natl. Acad., Sci. U.S.A. **8**, 3086 (1987).
11. E. Arashiro, J. R. Drugowich de Felício, and U. H. E. Hansmann, Phys. Rev. E **73**, 040902 (2006).
12. E. Arashiro, J. R. Drugowich de Felício, and U. H. E. Hansmann, J. Chem. Phys. **126**, 045107 (2007).

Coarse-Grained Lattice Model for Molecular Recognition

Hans Behringer, Andreas Degenhard, and Friederike Schmid

Fakultät für Physik, Universität Bielefeld, D-33615 Bielefeld, Germany
E-mail: {behringer, degenhard, schmid}@physik.uni-bielefeld.de

Equilibrium aspects of molecular recognition are investigated using coarse-grained models for the recognition process of two rigid biomolecules. To this end, a two-stage approach consisting of a design and testing step is adopted. Particular attention is paid to the influence of cooperative effects accompanying the association of biomolecules. Cooperativity is found to enhance selectivity. In addition it is discussed that a small number of strong bonds is favoured in flexible complexes compared to a situation with many but weak bonds.

1 Introduction

An understanding of the basic principles of biomolecular recognition, i. e. the ability of a biomolecule to interact specifically with another molecule in the presence of structurally similar rival molecules, is not only important from a scientific point of view but also opens up a wide field of potential biotechnological applications. The recognition process itself is governed by a complex interplay of non-covalent interactions of strengths comparable with the thermal energy thus leading to a complex problem[1,3]. In this context the study of idealised models with methods from statistical physics seems to be particularly adequate.

2 Model and General Approach

In this work we consider protein-protein recognition from a coarse-grained point of view on the level of both the structure of the biomolecules at the mutual interface and the interactions stabilising the complex. The biomolecules are assumed to undergo no refolding during the association process. This is a justified assumption for most protein-protein recognition processes, although notable exceptions do exist[1]. Motivated by the observation that hydrophobicity is the major driving force in molecular recognition[1] we describe the type of the residue at the position $i = 1, \ldots, N$ of the interface by a binary variable[2,3]. Denoting the structure of the target molecule by $\sigma_i \in \{\pm 1\}$ and that of the interaction partner by $\theta_i \in \{\pm 1\}$ we model the energetics at the interface by

$$\mathcal{H}(\sigma, \theta; S) = -\varepsilon \sum_{i=1}^{N} \frac{1 + S_i}{2} \sigma_i \theta_i - J \sum_{\langle ij \rangle} S_i S_j . \tag{1}$$

The variable S_i takes on the two discrete values ± 1 and describes the fit of the shape of the molecules at position i of the interface (on a microscopic level resulting from a rearrangement of the amino acid side chains when the complex is formed[1]). Apart from the direct contact energy with strength ε the model Hamiltonian contains an additional cooperative interaction term where the quality of a residue-residue contact couples onto the structure in its neighbourhood. To study the recognition process between two rigid

biomolecules we adopt a two-stage approach. For a fixed target structure $\sigma^{(0)}$ we first design an ensemble of probe molecules θ at a design temperature $1/\beta_{\mathrm{D}}$ leading to the distribution $P(\theta|\sigma^{(0)}) = \frac{1}{Z_{\mathrm{D}}} \sum_S \exp\left(-\beta_{\mathrm{D}} \mathcal{H}(\sigma^{(0)}, \theta; S)\right)$. In a second step the free energy difference of association at temperature $1/\beta$ is calculated for the interaction of the probe ensemble with the target molecule $\sigma^{(0)}$ and a structurally different rival molecule $\sigma^{(1)}$. In this step the free energy of the interaction of the molecule $\sigma^{(\alpha)}$ with a particular probe structure θ has to be averaged with respect to the distribution $P(\theta|\sigma^{(0)})$. This gives finally the free energy difference $\Delta F = F_{\mathrm{target}} - F_{\mathrm{rival}}$ as a function of the similarity Q between these two molecules, where Q is the number of residues N at the interface minus twice the number of point mutations that have to be carried out to convert the target into the rival. A negative ΔF then signals recognition of the target.

3 Results

It has been argued in the literature for the importance of cooperativity for molecular recognition[4]. In our coarse-grained Hamiltonian (1) cooperativity is taken into account by the second interaction term. The cooperative term rewards additional contacts in the neighbourhood of an already established one. As a consequence the fit of the two biomolecules at the interface is optimised and therefore one can expect an improved recognition ability. An investigation of the influence of this second term using our two-stage approach indeed reveals as shown in figure 1 that an increase of the cooperative interaction constant J significantly increases the recognition ability, i. e. the free energy difference. A value of J comparable to the value of ε already leads to the maximum effect of cooperativity (up to minor finite-size effects).

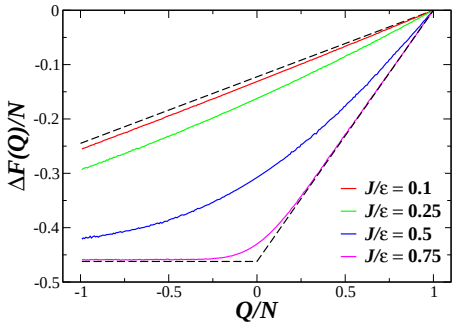


Figure 1. Influence of the cooperativity J on the free energy difference for the association of the probe molecules with the target and the rival molecule as a function of the similarity Q between these two molecules. The upper dashed line corresponds to $J = 0$, the lower one to $J \to \infty$ (in the limit $N \to \infty$).

Investigations of highly flexible antibody-antigen complexes showed that only approximately one quarter of the residue contacts at the interface contribute (significantly) to the binding energy[5] suggesting that in flexible complexes interfaces with a few strong bonds are favoured compared to a situation with many but weak bonds. We address this question of the role of varying bond strengths within our approach by considering a model which

distinguishes only between active residues, i. e. residues that contribute to the binding energy, and inactive ones. On the coarse-grained level this amounts to attributing the values 1 (active) and 0 (inactive) to the structural variables σ_i and θ_i in the Hamiltonian (1). In the following the uncooperative model with $J = 0$ is considered. In order to ensure the stability of the complex the interaction energy has to overcome the thermal energy barrier. On the other hand, however, the interaction energy has to be "small" enough to ensure the required flexibility of the complex. This can be incorporated into our approach by including the constraint that the interaction energy has to be fixed to some (suitable) value. Fixing the number of active residues A by a Lagrange multiplier, the free energy difference ΔF can be calculated as a function of the fraction A/N of active residues. Figure 2 demonstrates that the free energy difference indeed has a minimum at small fractions A/N (fairly insensitive to a variation of the interaction parameters). Our simple coarse-grained model hence predicts that recognition processes which require a certain amount of flexibility are most efficient if only a small number of fairly strong bonds is established across the contact interface as observed in antibody-antigen complexes.

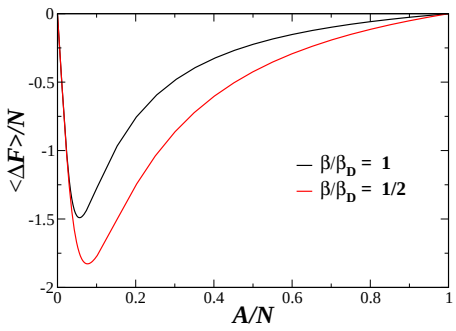


Figure 2. Averaged free energy per site as a function of the fraction A/N of active residues.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft (SFB 613) is gratefully acknowledged.

References

1. C. Kleanthous, ed., *Protein-Protein Recognition* (Oxford University Press, Oxford, 2000).
2. H. Behringer, A. Degenhard, and F. Schmid, Phys. Rev. Lett. **97**, 128101, 2006.
3. H. Behringer, T. Bogner, A. Polotsky, A. Degenhard, and F. Schmid, J. Biotechnol. **129**, 268, 2007.
4. E. di Cera, Chem. Rev. **98**, 1563, 1998.
5. B. C. Cunningham and J. A. Wells, J. Mol. Biol. **234**, 554, 1993.

Residual Entropy of Ice I from Multicanonical Simulations

Bernd A. Berg[1,2]

[1] Department of Physics, Florida State University,
Tallahassee, FL 32306-4350, USA

[2] School of Computational Science, Florida State University,
Tallahassee, FL 32306-4120, USA
E-mail: berg@scs.fsu.edu

We introduce two simple models with nearest neighbor interactions on 3D hexagonal lattices. Each model allows one to calculate the residual entropy of ice I (ordinary ice) by means of multicanonical simulations. This gives the correction to the residual entropy derived by Linus Pauling in 1935. Our estimate is found to be within less than 0.1% of an analytical approximation by Nagle which improved on Pauling's result. In biological applications at room temperature small, ice-like clusters are of importance. Their entropy can be computed by the same method.

1 Introduction

In this talk I report on a novel calculation[1] of the residual entropy of ice I. The approach can as well be used to calculate the entropy of arbitrary cluster of hydrogen bonds of known geometry. This is of importance for the interaction of water with proteins and other biomolecules, because the liquid phase of water differs from simple fluids in that there is a large qualitative remnant of ice structure in the form of local tetrahedral ordering[2]. The next section reviews briefly the residual entropy of ice and in section 3 our calculation is sketched.

2 Residual Entropy of Ice

In contrast to liquid water the properties of ice are relatively well understood. Most of them have been interpreted in terms of crystal structures, the forces between its constituent molecules, and the energy levels of the molecules themselves[3]. A two-dimensional projection of the hexagonal crystal structure of ordinary ice (ice I) is depicted in Fig. 1 (other forms of ice occur in particular at high pressures). Each oxygen atom is located at the center of a tetrahedron and straight lines (bonds) through the sites of the tetrahedron point towards four nearest-neighbor oxygen atoms. Hydrogen atoms are distributed according to the ice rules: (A) There is one hydrogen atom on each bond (then called hydrogen bond). (B) There are two hydrogen atoms near each oxygen atom (these three atoms constitute a water molecule).

By experimental discovery it was found that ice has a residual entropy[4] $S_0 = k\,\ln(W) > 0$ for temperature $T \to 0$ and moderate relaxation times. Here W is the number of configurations for N molecules. Subsequently Linus Pauling[5] derived estimates of $W = (W_1)^N$ by two approximate methods based on the ice rules, obtaining

$$W_1^{\rm Pauling} = 3/2 \tag{1}$$

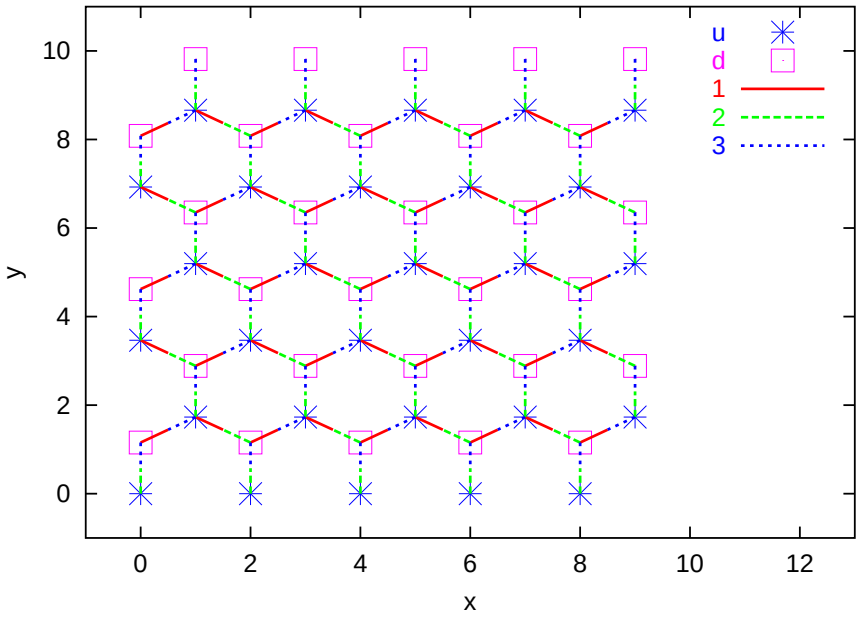


Figure 1. Lattice structure of one layer of ice I. The up (u) sites are at $z = 1/\sqrt{24}$ and the down (d) sites at $z = -1/\sqrt{24}$. The dimensions are explained in Ref. [1]. For each site three of its four pointers to nearest neighbor sites are shown.

in each case. Twenty five years later it was shown by Onsager and Dupuis that $W_1 = 1.5$ is in fact only a lower bound. Onsager's student Nagle used a series expansion method to derive the estimate[6]

$$W_1^{\mathrm{Nagle}} = 1.50685\,(15)\,. \tag{2}$$

The error bar is not statistical but reflects higher order corrections of the expansion, which are not entirely under control. The only independent theoretical value appears to be one for cubic ice, which is obtained by numerical integration of Monte Carlo data[7] and in good agreement with Nagle.

3 Multicanonical Calculation

Our calculations are based on two simple statistical models. Each model is defined on the hexagonal lattice of Fig. 1. In the first model, called 6-state H_2O molecule model, ice rule (B) is always enforced and we allow for six distinct orientations of each H_2O molecule. Its energy is defined by

$$E = -\sum_b h(b, s_b^1, s_b^2)\,. \tag{3}$$

Here, the sum is over all bonds b of the lattice and (s_b^1 and s_b^2 indicate the dependence on the states of the two H_2O molecules, which are connected by the bond)

$$h(b, s_b^1, s_b^2) = \begin{cases} 1 \text{ for a hydrogen bond,} \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

In the second model, called 2-state H-bond model, ice rule (A) is always enforced and we allow for two positions of each hydrogen nucleus on a bond. The energy is defined by

$$E = -\sum_s f(s, b_s^1, b_s^2, b_s^3, b_s^4)\,, \tag{5}$$

where the sum is over all sites (oxygen atoms) of the lattice and f is given by

$$f(s, b_s^1, b_s^2, b_s^3, b_s^4) = \begin{cases} 2 \text{ for two hydrogen nuclei close to } s, \\ 1 \text{ for one or three hydrogen nuclei close to } s, \\ 0 \text{ for zero or four hydrogen nuclei close to } s. \end{cases} \tag{6}$$

The groundstates of each model fulfill the ice rules. At $\beta = 0$ the number of configurations is 6^N for the 6-state model and 2^{2N} for the 2-state model. Because the normalizations at $\beta = 0$ are known, multicanonical (MUCA) simulations allow us in either case to estimate the number of groundstate configurations[8]. Using periodic boundary conditions (BCs), our simulations are based on a lattice construction similar to that set up in Ref. [9] for Potts model simulations, while the thermodynamic properties found are entirely different from those of Potts models.

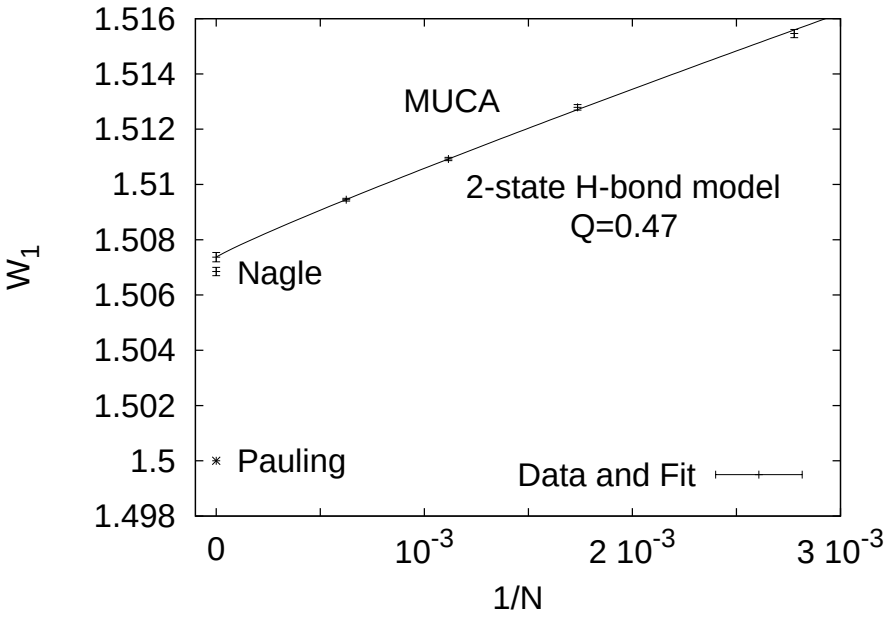


Figure 2. Fit and estimates of W_1.

The results from the 2-state model are more accurate than those from the 6-state model. Apparently, the reason is that the $\beta = 0$ entropy of the 2-state model is smaller. In Fig. 2 a fit for the data of the 2-state H-bond model to the form

$$W_1(x) = W_1(0) + a_1\, x^\theta\,, \quad x = 1/N \tag{7}$$

with $\theta = 0.920\,(25)$ is shown together with the estimates by Nagle and Pauling. The goodness of fit (chapter 2.8 of Ref. [9]) is $Q = 0.47$ as given in the figure. That we have $\theta \neq 1$ reflects bond correlations in the groundstate ensemble.

Combining this fit with that from the 6-state model leads to our final estimate

$$W_1^{\rm MUCA} = 1.50738\,(16)\,. \tag{8}$$

The difference with the estimate of Nagle is 0.035% of W_1. However, the Gaussian difference test[9] between the two estimates yields $Q = 0.016$. As the error bar in (8) covers only statistical and no systematic errors due to finite size corrections from larger lattices, the small discrepancy may be explained in this way. The precision of experimental work[10] has remained insufficient to resolve the difference with Pauling's original estimates and its theoretical improvement. Experimental verification of the difference would be an outstanding confirmation of structures imposed by the ice rules.

References

1. B. A. Berg, C. Muguruma, and Y. Okamoto, *Residual Entropy of Ordinary Ice from Multicanonical Simulations*, Phys. Rev. B **75**, 092202, 2007.
2. Literature is given in Ref. [1].
3. D. Eisenberg and W. Kauzmann, *The Structure and Properties of Water*, Oxford University Press, Oxford 1969; V. F. Petrenko and R.W. Whitworth, *Physics of Ice*, Oxford University Press, Oxford 1999.
4. W. F. Giauque and M. Ashley, *Molecular Rotation in Ice at* $10°K$*. Free Energy of Formation and Entropy of Water*, Phys. Rev. **43**, 81, 1933.
5. L. Pauling, *The Structure and Entropy of Ice and of Other Crystals with Some Randomness of Atomic Arrangement*, J. Am. Chem. Soc. **57**, 2680, 1935.
6. J. F. Nagle, *Lattice Statistics of Hydrogen Bond Crystals I. The Residual Entropy of Ice*, J. Math. Phys. **7**, 1484, 1966.
7. S. V. Isakov, K. S. Raman, R. Moessner, and S. L. Sondhi, *Magnetization Curve of Spin Ice in a [111] Magnetic Field*, Phys. Rev. B **70**, 104418, 2004.
8. B.A. Berg and T. Celik, *New Approach to Spin-Glass Simulations*, Phys. Rev. Lett. **69**, 2292, 1992.
9. B.A. Berg, *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*, World Scientific, Singapore, 2004.
10. W.F. Giauque and J.W. Stout, *The Entropy of Water and the Third Law of Thermodynamics. The Heat Capacity of Ice from 15 to* $273°K$, J. Am. Chem. Soc. **58**, 1144, 1936; O. Haida, T. Matsuo, H. Suga, and S. Seki, *Calometric Study of the Glassy State X. Enthalpy Relaxation at the Glass-Transition Temperature of Hexagonal Ice*, J. Chem. Thermodynamics **6**, 815, 1974.

REMD Simulations of Aβ_{16-22} Peptide Aggregation in Explicit Solvent

Raghunadha Reddy Burri*, Dietmar Paschek, and Alfons Geiger

Department of Physical Chemistry, University of Dortmund,
Otto-Hahn-Strasse 6, 44227 Dortmund, Germany
**E-mail: reddy.burri@udo.edu*

Experimental studies show that the short peptide fragments Aβ_{16-22} form fibrils as it is known from the full length β-amyloid peptide. This fibril growth is strongly temperature dependent. We report here a simulation study of the temperature dependent Aβ_{16-22} aggregation in explicit water. We simulated a system of ten Aβ_{16-22} peptides with 5900 SPC/E water in a cubic box and used 76 replicas (with 20 ns simulation time per replica) distributed over a temperature range from 285.0 to 606.3 K. Replica Exchange Molecular Dynamics (REMD) simulation is an efficient way for equilibration and simulation of complex molecular systems at different temperatures. The temperature dependence of radius of gyration R_G and the solvent accessible surface area(SASA) of the aggregates, as well as structural properties like mutual orientation and number of peptide-peptide hydrogen bonds can be understood by the different temperature dependence of hydrogen bond strength, electrostatic, and hydrophobic interactions.

The Aβ_{16-22} fragment is highly prone to aggregation, and a prototype molecule for the study of processes of amyloidosis. It contains a central hydrophobic core, a positive charge at the N-terminus (Lys16), and a negative charge at the C-terminus (Glu22). Solid-state NMR showed that this segment of the full amyloid-β peptide can form fibrils with an antiparallel strand organization.[1] Experimental studies of Aβ fibril formation also revealed a strong temperature dependence.[2] This could also be obtained in constant pressure MD simulations of Aβ_{16-22} peptides aggregation.[3] Here we present first results of an extension of this study, where we apply Replica Exchange Molecular Dynamics (REMD) simulations. This is an efficient way to simulate complex systems at different temperatures and is the simplest and most general form of simulated tempering.[4] It offers a much-improved approach for determining oligomer distributions relevant to aggregation.[5] The basic idea of REMD is to simulate different copies (replicas) of the system at the same time but at different temperatures. After a certain time, conformations are exchanged with a Metropolis probability, therefore permitting random walks in the temperature space and escape from local energy traps.[6] Recently, Paschek et al.[7] published such simulations, providing the first unbiased folding of the Trp-cage in explicit solvent using 40 replicas (100 ns per replica).[7]

Taking Paschek et al. work as a reference,[7] we used REMD to study Aβ_{16-22} peptides aggregation at atomic level in explicit aqueous solution.[8] The GROMACS 3.2.1[9] simulation package was used in both simulation and data processing. The OPLS-All Atom force field was chosen to represent the peptide in GROMACS. The system is coupled to an external heat bath (Nose-Hoover-thermostat) with a relaxation time of 1.5 ps. The electrostatic interactions are treated by the smooth particle mesh Ewald summation with a real space cutoff of 0.9 nm. A 2.0 fs timestep was used for all simulations. Solvent constraints were solved using the SETTLE procedure, while the SHAKE-algorithm was used for the polymer constraints.

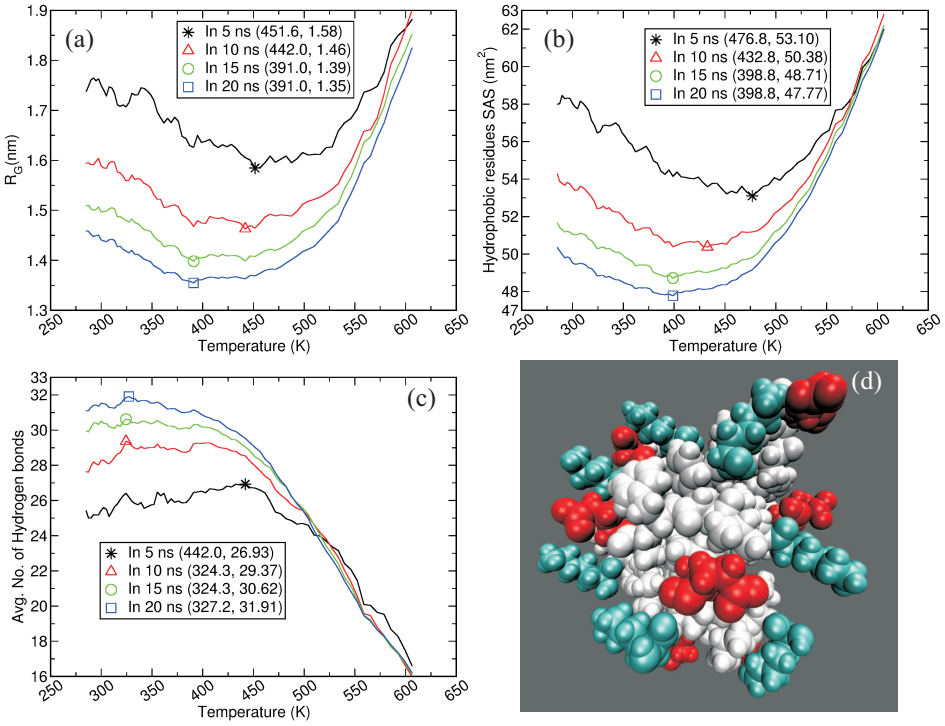


Figure 1. Properties of the peptide aggregates, averaged over different lengths of the simulation runs as a function of temperature. Lowest/highest values are marked. (a) Radius of gyration R_G of the peptide back-bone atoms. (b) SASA of hydrophobic residue atoms. (c) Average number of hydrogen bonds between the peptides. (d) Final snapshot at 398.8 K, hydrophobic residues in white, Lys(+) in cyan and Glu(-) in red.

In the starting configuration of this study, as in the previous constant pressure simulation series[3], six monomeric peptides (Capped Aβ_{16-22} with the sequence of Ace-KLVFFAE-NH_2) were placed uniformly in a distance of about 1.5 nm around the center of an ordered tetramer which was considered to serve as nucleus for further growth. It was obtained in an initial constant pressure simulation of four peptides at 360 K after 20 ns.[3] These 10 peptides are immersed in 5900 SPC/E water molecules in a $5.8 \times 5.8 \times 5.8(\mathrm{nm})^3$ cubic box and periodic boundary conditions were applied. For REMD we used 76 replicas (20 ns per replica) distributed over a temperature range from 285.0 to 606.3 K, where multiple copies (or replicas) of identical systems are simulated in parallel at different temperatures. The temperature spacing between each of the replicas was chosen such that the energy distributions overlap sufficiently and state exchange attempts are (on average) accepted with a 20 % probability.

The results from our constant pressure simulations (at seven temperatures from 280 to 460 K) show that the Aβ_{16-22} monomers first form anti-parallel hydrogen-bonded dimers at the lower T range of 280–340 K. These aggregate at middle T range from 340 to 400 K, to large structures, which show two major features of the amyloid fibrils: twisted β-sheets

formed from antiparallely oriented peptides and the onset of formation of a second layer. In the higher temperature range (from 400 to 460 K) the twist angle between the monomers increased probably to protect hydrophobic residues from water.[3]

From the REMD simulations, properties of interest have been extracted at 76 temperatures. Below 391 K, the radius of gyration R_G of peptide cluster (calculated from peptide backbone atoms) decreases with increasing temperature, and started to increase with T below 391 K [Figure 1(a)]. Consequently, also the SASA calculated from the hydrophobic residues, reached a lowest point at a temperature 398 K [Figure 1(b)]. Figures 1(a and b) show that the aggregated structures at intermediate temperatures extend and disintegrate at high temperatures. The maximum number of peptide–peptide hydrogen bonds is observed at around ~ 330 K and decreases at higher temperatures [Figure 1(c)]. The shift of the positions of the minima of R_G and SASA compared to the maximum of the number of H-bonds can be explained by the fact, that with increasing temperature the H-bonds are weakened, whereas the hydrophobic interaction strength increases. While the H-bonds tend to build a planar β-sheet structure, the increasing hydrophobic interaction produces more compact structures [Figure 1(d)]. This may be obtained by twisting the β-sheet or by building up a second sheet, as observed in the constant pressure simulation study.[3] Interestingly, Meinke and Hansmann[10] also observe for a system of six β-amyloid fragment peptides (without explicit water) above 400 K a strong increase of R_G, but do not observe a temperature minimum. This is probably due to the lack of water (and the water mediated hydrophobic interaction) in their simulations. To demonstrate the convergence of the REMD simulations, in figures 1(a, b and c) averages over different lengths of the simulation runs (5, 10, 15 and 20 ns) are shown and the minima and maxima are marked. From 15 to 20 ns the positions of there extrema on the temperature axis stay a constant, revealing that 20 ns time for each replica is reasonable to study early aggregation.

Our results are in good agreement with the previous work done by Meinke et al.[10] and Gnanakaran et al.[5]. We find that at low temperatures the structure of the aggregates is largely determined by hydrogen bonding and electrostatic interactions. This leads to the formation of well ordered antiparallel β-sheet structures. With increasing temperature, hydrophobic interactions become more important, as indicated by the formation of stacked β-sheets, as well as less regular ordered collapsed clusters. At highest temperatures the aggregates are found to disintegrate due to the strong thermal motions.

Acknowledgments

Support by a PhD fellowship from the International Max Planck Research School in Chemical Biology (IMPRS-CB) to RRB is acknowledged. We appreciate generous computing time from LiDO, Dortmund, Germany.

References

1. Balbach J. J., Ishii Y., Antzutkin O. N., Leapman R. D., Rizzo N. W., Dyda F., Reed J., and Tycko R. Amyloid fibril formation by Aβ_{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.

2. Kusumoto Y., Lomakin A., Teplow D. B., and Benedek G. B. Temperature dependence of amyloid β-protein fibrillization. *Proc. Natl. Acad. Sci. USA*, 95:12277–12282, 1998.
3. Burri R. R., Paschek D., and Geiger A. Temperature dependence of Aβ_{16-22} peptide aggregation and fibril growth in aqueous solution. *in preparation*.
4. Hansmann U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–140, 1997.
5. Gnanakaran. S., Nussinov. R., and García A. E. Atomic-level description of amyloid β-dimer formation. *J. Am. Chem. Soc.*, 128:2158–2159, 2006.
6. Sugita Y., Kitao A., and Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. *Chem. Phys.*, 113:6042–6051, 2000.
7. Paschek D., Nymeyer H., and García A. E. Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *J. Struct. Biol.*, 157:524–533, 2007.
8. Burri R. R., Paschek D., and Geiger A. Replica exchange simulations of Aβ_{16-22} peptides aggregation in explicit solvent. *in preparation*.
9. Lindahl E., Hess B., and van der Spoel D. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7:306–317, 2001.
10. Meinke J. H. and Hansmann U. H. E. Aggregation of β-amyloid fragments. *J. Chem. Phys.*, 126:014706–014711, 2007.

A Molecular Dynamic Study of the Basic Fibroblast Growth Factor - Fibroblast Growth Factor Receptor Complex

Léo Degrève* **and Gustavo Henrique Brancaleoni**

Departamento de Quimica, FFCLRP,
Universidade de São Paulo 14040-901, Ribeirão Preto, Brazil
**E-mail: leo@ffclrp.usp.br*

The basic fibroblast growth factor is a 155 amino acid residues peptide that plays versatile functions throughout the human life cycle. It binds to its tyrosine kinase-containing FGF receptor, FGFR1, and to heparan sulfate proteoglycans. The presence of ethanol interferes on the bFGF biologic activities resulting in numerous anomalies like the fetal alcohol syndrome. Experimental studies suggest that the interaction sites between bFGF and FGFR1 are the target for ethanol molecules. In order to better understand the interaction between bFGF, FGFR1 and ethanol, molecular dynamics simulations of the complex bFGF-FGFR1-heparin in water and in water-ethanol solutions were carried. Hydrogen bond structures, root mean square deviations and specific radial distribution functions were employed to identify the role of each component on the complex stability. Some residues (SER73, TYR112, THR114, TYR115 and SER117) were identified as extremely important for the perfect linking between the bFGF and FGFR1. There are exactly these residues that are the target of ethanol molecules resulting in the impossibility to occur the docking of bFGF on its receptor.

1 Introduction

The cell growth is a tightly regulated process controlled by highly specific proteins, namely the growth factors[1]. The bFGF, a mammalian fibroblast growth factor, is a sequence of 155 amino acid residues that plays versatile functions throughout the human life cycle. Two types of molecules bind bFGF to mediate its response: (1) tyrosine kinase-containing FGF receptor, FGFR1, and (2) heparan sulfate proteoglycans (heparin and heparan sulfates, HS). FGFR1 is an integral membrane protein composed by an extracellular ligand binding segment consisting of three immunoglobulin-like domains (D1, D2 and D3), a single transmembrane helix and a cytoplasmic portion that contains a largely split tyrosine kinase activity[2]. The ligand binding and the specificity residues of the FGFR1 are located in the D2 and D3 domains and in the linker that connects these domains. The bFGF:FGFR1 complex contains a 2:2 dimeric bFGF-FGFR1-HS assemblage. D2 and D3 are sufficient for the bFGF binding and to determine its specificity[3]. Experimental studies on bFGF biologic activities revealed that the presence of ethanol interferes in its activity apparently because the interaction sites between bFGF and FGFR are the targets of the ethanol molecules[2]. The objectives of the present study are (1) to describe the structure of bFGF-FGFR1 interactions and (2) to contribute to better understand the structural factors that determine the loss of bFGF activity in the presence of ethanol molecules. Molecular dynamics simulations[4] conducted with the Gromos96 force field and the GROMACS 3.2 package[5] will be used. The simulated systems consisted of one dimer bFGF-FGFR1 (initial guess: pdb code 1FQ9) immersed in water or in a mixed water-ethanol medium.The solvent concentrations

were adjusted to match, in the pure solvent regions of the simulation systems, their experimental densities. Na^+ and Cl^- ions were added to preserve the local electroneutrality. After an initial phase, the 10.0 ns long simulations were conducted at 298 K and 1 atm in a 13.5 x 13.5 x 13.5 nm^3 simulation cell, implementing the PME method with a cut-off applied at 1.4 nm.

2 Results and Discussion

The structure of the dimer bFGF2-FGFR1 consists of two bFGF molecules (residues 16-144), the dimerized receptor formed by two D2 domains (residues 149-249) linked to two D3 domains (residues 251-359) and two heparin molecules. The structure of the dimer in aqueous medium undergoes only few changes if compared with the starting X-ray structure (RMSD about 0.25nm) while, in ethanol solution, RMSDs with a mean value of 0.60 nm are observed when the dimer structure undergoes significant changes suggesting that some regions are loosing their structure. The characterization, by molecular dynamics simulations, of the peptide structures was carried out by identifying, in a first step, the intramolecular hydrogen bonds (iHB) and, a second step, the intermolecular hydrogen bonds (HB). The dimeric bFGF-FGFR1 structure is stabilized by 20 HB between the dimer and the solvent molecules, 20 between bFGF and FGFR1, 4 between the FGFR1 units, 32 between heparin and the bFGF's and 22 between heparin and the FGFR1's.

The hydration changes in the bFGF-FGFR1 complex comparing the data in aqueous solution and in ethanolic medium reveal decreases of 20% for bFGF and 9% for FGFR1. All the affected residues contain a hydroxyl group in the side-chain and are located in the neighborhood of other residues displaying a hydrophobic character[6](the bFGF residues solvated by ethanol are always the same in the presence or not of the receptor) suggesting that an energetically favorable situation takes place when the ethanol molecules are hydrogen bonded to hydroxyl groups while their methyl groups are accommodated into a hydrophobic environment. In the case of bFGF, the ethanol molecules form intermolecular hydrogen bonds with TYR33 (β1), SER73 (β5), TYR82 (β6), SER109, TYR112(β9), SER114(β9), TYR115(β9), and SER117(β9) so that two important HB between bFGF and FGFR1 D2 domain are lost: the links between SER109 of bFGF and ASP209 of FGFR1 and between TYR112 (β9) of bFGF and ALA179 located in the FGFR1 D2 domain.

The solvation by the ethanol molecules shows a more active interference of the ethanol molecules with the residues belonging to the bFGF β1, β5, β6, β9, β10, and β12 strands. The most important interference (*i.e.,* the longer residence times) is observed on the β1 (TYR33), β6 (TYR82), β9 (SER109, TYR112, TYR115 and SER117) residues. The consequence is the lost of four bFGF-FGFR1 HB, three between each one of the bFGF's and the D2 domains and one HB between each bFGF and the D3 domains. The losses of local structures are confirmed by the experimental results that also suggest that the interaction between bFGF and FGFR is the target for ethanol in its inhibition process. On the basis of the detailed description of the interactions of bFGF-FGFR1 complex with the ethanol molecules, it can be establish that the residues belonging to the β1, β6, and β9 strands, especially TYR33, TYR82, SER109, and TYR112 are extremely important for the perfect linking between the bFGF and FGFR1.

3 Concluding Remarks

The studies by molecular dynamics of the bFGF-FGFR1 complex structure conclude that the ligand binding and specificity require both D2 and D3 domains of FGFR1, that the heparin molecules are a fundamental binding element between bFGF and the receptor, that the interaction of bFGF with its membrane receptor must be described by a HB complex network where all HB are essential for the perfect bFGF binding to FGFR1 (the perfect binding that is an unavoidable step to bFGF exhibits all its biological activities). The simulation results confirm the experimental data concerning the importance of the TYR-112 and ASN-113 in the bFGF activity but also are able to identify the SER73, TYR112, THR114, TYR115 and SER117 residues as being probably important to understand the activity of the bFGF growth factor.

Acknowledgments

The Fundação de Amparo à Pesquisa do Estado de São Paulo and the Conselho Nacional de Desenvolvimento Científico e Tecnológico are acknowledged for supporting the research.

References

1. L. Strayer, *Bioquímica*, Guanabara Koogan, São Paulo, (1995).
2. J. Schlessinger, A. N. Plotnikov, O. A. Ibrahimi, A. V. Eliseenkova, B. K. Yeh, A. Yayon, R. J. Linhardt, and M. Mohammadi, Molecular Cell, 6 , 743 (2000).
3. M. Mohammadi , S. K. Olsen, O. A. Ibrahim, Cytokine & Growth Factor Reviews, 16, 107(2005).
4. M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids, Clarendon Press, Oxford, (1987).
5. D. Van Der Spoel,; E. Lindahl,; B. Hess; A.R. Van Buuren, E. Apol, P. J. Eulenhoff, D. P. Tieleman; A. L. T. M. Sijbers, K. A. Feenstra, R. Van Drunen, H. J. C. Berendsen. Gromacs User Manual, Version 3.2: `http://www.gromacs.org`.
6. G.H. Brancaleoni, M.R. Lourenzoni and L. Degrève, Genet. Mol. Res. 5 (2): 350 (2006).

Dimensionality Reduction Techniques for Protein Folding Trajectories

T. Eitrich, S. Mohanty, X. Xiao, and U. H. E. Hansmann

John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {*t.eitrich, s.mohanty, x.xiao, u.hansmann*}*@fz-juelich.de*

In our work we analyze large and high dimensional data from protein folding simulations. The main goals are to extract the underlying dimensionality, to find a small number of features that describe the data with high accuracy and to find interesting clusters in the data: in this work we treat this as a problem of dimensionality reduction. Dimensionality reduction aims to find a mapping of the original space into a space of a few interesting dimensions, which the user then can use for interpretation and analysis. We study modern dimensionality reduction techniques and combine them with promising distance measures, suitable for the description of dissimilarities between the data points generated by the package ProFASi - a Protein Folding and Aggregation Simulator.

1 Introduction

Understanding the folding of proteins is one of the most challenging problems in computational biology. We refer to protein folding as the process by which a protein assumes its native state. Protein folding trajectory data is high dimensional and thus hard to interpret after the simulation. Dimensionality reduction methods, that map the data into a new space of fewer dimensions while preserving as much relevant information as possible, can be used to find meaningful low dimensional structures in the original high dimensional datasets.

Several techniques do exist, which we may divide into

- linear methods, like principal component analysis (PCA)[6] and multi dimensional scaling (MDS)[2], and
- nonlinear methods, like locally linear embedding (LLE)[8], kernel PCA[9] and Isomap[10].

We refer to this methods as unsupervised embedding algorithms. Supervised embedding methods, which take additional properties into account, are discussed in ref 5.

2 Material and Methods

Our dataset for this work consists of a folding trajectory of a 49 residue alpha-beta protein with PDB id 2GJH, generated using the program package ProFASi: a Protein Folding and Aggregation Simulator[4]. Each data point represents the conformation of the molecule at a certain Monte Carlo time. Each successive pair of data points are separated by 1000 Monte Carlo sweeps. For this analysis, we have used 1000 such points, which were selected to span one particular folding event in the simulations.

We have used two dimensionality reduction methods, MDS and Isomap. The following are the main steps in the Isomap[10] method.

- Compute dissimilarities between all points.
- Construct a neighborhood graph according to the number of neighbors k_n
- Based on the neighborhood information, compute shortest paths between all points.
- Embed the data into a new d-dimensional space using an eigenvalue method.

In the case of MDS, only the first and the last step need to be done (no neighborhood parameter k_n), because the dissimilarity measure is assumed to be uniformly good for all pairs of points.

The success of dimensionality reduction methods highly depends on a reasonable choice of the dissimilarity measure, since these methods use distances between objects instead of the objects themselves. Thus, any information we want to preserve must be represented by the distance measure. In our study we have used and compared three different measures, i.e.

- the minimized root mean square deviation (RMSD) between atomic coordinates, as it was used in ref 3,
- the RMSD of the dihedral angels, as introduced in ref 1, and
- a two dimensional structural similarity measure based on dihedral angle distributions and a Fourier transformation process, described in ref 7.

3 Results and Discussion

In our tests we have applied different dimensionality reduction techniques as well as various distance measures for our folding trajectory dataset. We observed, that both, MDS and Isomap (using small neighborhood sizes) lead to interesting embedding dimensions. We observed, that all three distance measures give results which fit to some of the characteristics of the simulation, like energy or RMSD to the native state.

In Fig. 1, we show results using Isomap with 10 neighbors and the atomic RMSD distance measure. Plotted is the RMSD to the native state against the first embedding dimension - a descriptor for how well the protein has folded in the simulation. Please note, that the colors in all pictures express the time steps. In Fig. 2, results using MDS with the dihedral RMSD distance measure are given. Plotted is the total energy value against the first embedding dimension. In Fig. 3, MDS together with the Fourier measure was applied. Given is the helix content against the first embedding dimension.

So far we could not observe huge differences between the linear MDS and the nonlinear Isomap, which leads us to the conclusion, that for the case of protein folding trajectories, we have examined, the choice of a reasonable distance measure also leads to adequate results for linear embedding methods.

4 Summary and Future Work

In our work we studied linear and nonlinear dimensionality reduction techniques. Our application field is protein folding, for which we try to find and analyze new embedding coordinates. Future work will concentrate on the analysis of embedding data. Especially, we will use supervised methods to cluster and classify embedded data.

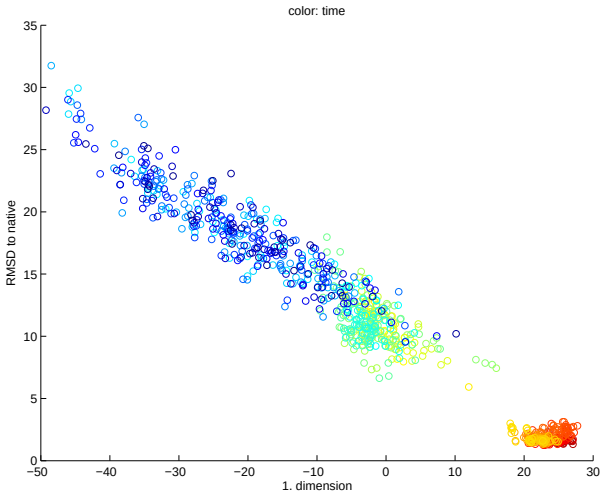


Figure 1. First embedding coordinate is correlated with the RMSD to native value.

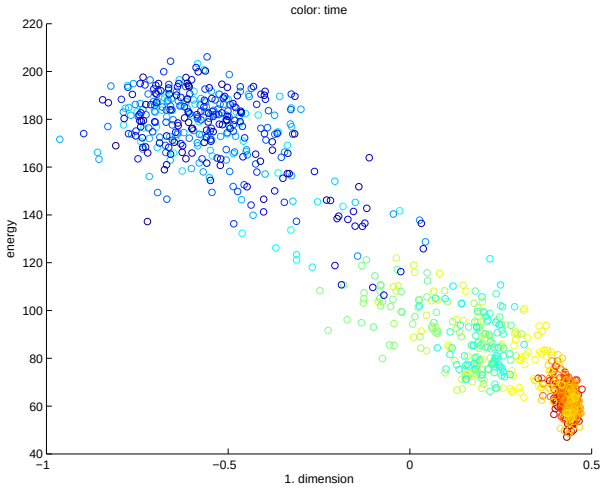


Figure 2. First embedding coordinate is correlated with the energy.

Acknowledgments

Part of this work are supported by research grants of the National Science Foundation, USA (CHE-0313618) and the National Institutes of Health, USA (GM 62838).

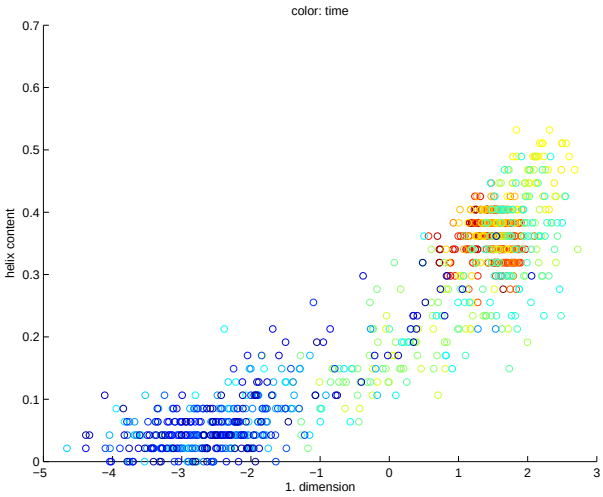


Figure 3. First embedding coordinate is correlated with the helix content.

References

1. A. Collins, R.I. Cooper, and D.J. Watkin *Structure matching: measures of similarity and pseudosymmetry. J. Appl. Cryst.* **39**, *842–849, 2006.*
2. *T.F. Cox and M.A.A. Cox.* Multidimensional scaling. *Chapman & Hall, London, 1994.*
3. *P. Das, M. Moll, H. Stamati, L.E. Kavraki, and C. Clementi* Low-dimensional free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. PNAS **103**(26), 9885–9890, 2006.
4. A. Irbäck and S. Mohanty *ProFASi: a Monte Carlo simulation package for protein folding and aggregation. J. Comput. Chem.* **27**, *1548–1555, 2006.*
5. *T. Iwata, K. Saito, N. Ueda, S. Stromsten, T.L. Griffiths, and J.B. Tenenbaum* Parametric embedding for class visualization. Advances in Neural Information Processing Systems **17** (NIPS2004), 617–624, 2005.
6. I.T. Jolliffe *Discarding variables in a principal component analysis. I: artificial data. Applied Statistics* **21**, *160–173, 1972.*
7. *N. Kandiraju, S. Dua, and S. Conrad* Dihedral angle based dimensionality reduction for protein structural comparison. Proc. Conf. ITCC2005, 14–19, 2005.
8. S.T. Roweis and L.K. Saul. *Nonlinear dimensionality reduction by locally linear embedding. Science* **290***(5500), 2323–2326, 2000.*
9. *B. Schölkopf, A.J. Smola, and K.-R. Müller.* Kernel principal component analysis. In Advances in kernel methods: support vector learning, MIT Press, 327–352, 1999.
10. J.B. Tenenbaum, V. deSilva, and J.C. Langford. *A global geometric framework for nonlinear dimensionality reduction. Science* **290***(5500), 2319–2323, 2000.*

Chemical Space of Auxins, their Multi-Phenomenology and Multiple Protein Interaction

Noel Ferro[1,2], Patrick Bultinck[3], Thomas Bredow[1], and Thomas Reinard[2]

[1] Institute for Physical and Theoretical Chemistry,
University of Bonn, Wegelerstr. 12, D-53115 Bonn, Germany
E-mail: Noel.ferro@genetik.uni-hannover.de

[2] Institute of Plant Genetic,
University of Hannover, Herrenhauser Str.2, D-30419, Germany

[3] Department of Physical Chemistry,
University of Ghent, Krijgslaan 281(S3) B-9000 Gent, Belgium

The irregular net of physiological processes in plants has been the medium for intuitive theories (hypotheses) of structure-activity relationships of auxin molecules, but even now their chemical definition is unclear. Molecular Quantum Similarity Measures (MQSM) scored a conceptual framework to uncover such phenomenon. The quantum objects (auxins) were classified by cluster analysis methods based on a feed-back with a biological consensus variable. Next, standardized multi-screening bioassays of different auxins were carried out and the data analysis was implemented at multidimensional level. In these scenarios, on the one side, a new active auxin-like substance without COOH in the side chain (2, 6-dibromo-phenol) was found, and hardness (η) was shown to be a common variable to predict the auxins biological activities. On the other side, two sets of molecular properties are postulated to explain different physiological processes such as growth and morphogenesis. These dual properties of auxins are structurally consistent with the two different binding-pocket systems disclosed until now in ABP1 and TIR1, respectively.

1 Introduction

The underlying concept of the structure-activity rule, "auxins act as a kind of co-enzyme or ergon at the growth center, which is a protein or enzyme surface of highly specific shape"[1] is not consistent with experimental facts anymore.
Due to their pleiotropic effects (e.g. organogenesis, cell elongation, cell division) auxin activity depends on more than one "receptor" and on a new biochemical mechanism of interaction of protein – small molecule – protein[2]. These statistical regularities have not been taken into account previously for the analysis of the detailed molecular causal properties of their biological activities. The physiological role of the Auxin-Binding-Protein 1 (ABP1) is still debated. But definitively, it is not involved in all the different physiological auxin effects[3]. Furthermore, the speculations about specialized receptor functions for specific transporters like PIN (PIN-FORMED) proteins should be considered as well[4,5]. Recently a new complex of three proteins SCF^{TIR1} out of which the transport inhibitor response 1 (TIR1) has been described as an auxin receptor[2]. The objective of the present article is to summarise some of our recent results and search for interactions between some methods of molecular biology and computational chemistry in order to analyse the complex physiological processes related to the auxin phenomena.

2 Structure-activity

Different approaches have been postulated: **chemical**[6] **, physico-chemical**[7] **binding site models**[8],[9]. They can not overcome the structure-activity impasse of auxins. The unconnected comprehensions of the chemical and biological view-points have failed to describe the auxin phenomena: first, suppressing ideas of statistical regularities and evaluating the assumption, "structure generates properties", as a dynamic regularity of the hormone-receptor interaction exported from the animal model. Second, distinctiveness of the molecular biology and biochemical outcomes in the last twenty years in plants has not been incorporated into the analysis.

3 Problem Solution

Our chemical computational-biostatistical approach focused on the auxin chemical space in the biological context. The analysis of the structure of the auxin-like molecules is treated as the invariant part, to which the the biological tests are relative, but not vice versa[10]. That does not mean, that the phytohormone phenomenon depends exclusively on the ligand structure, but the ligand structure analysis is the point to define the degrees of freedom of the phenomenon. The analysis presented focus on:

1. searching Molecular Quantum Similarities associated with the biological activities;
2. classifying auxin molecules based on the connections of the QMSM and the Biological Activity;
3. developing standardized bioassay for the screening of selected molecules to confirm the hypothesis of relation quantum similarity – biological activity;
4. identifying quantum molecular regions responsible for the biological activity.

The statistical treatment of similarity matrices of both, Coulomb and Overlap operators, was preceded by a Principal Component Analysis, minimizing repetitive information. Next, cluster methods analysis were applied. Relationship between clusters and biological activity was confirmed by statistic methods with probability *a posteriori*. The resulting cluster allowed to recognize different quantum similarity classes related to the biological activity[10]. Following, the biological activity of 2,6-Br-Phe (without COOH group), active as auxin by QMSM predictions[10], was experimentally confirmed[11].

Functionally, the multi-scaling analysis of the experimental data allowed us to discriminate two sets of chemical predictor variables (chemical regions), which distinguish two different kind of auxin biological activities: growth, where ABP1-binding activities inferred a physiological relation statistically; and morphogenesis events where TIR1 is involved (Fig. 1). Structurally, the two sets of accessory binding areas proposed M and G (Fig. 1 A) are closely related to the differences of the binding site requirements of TIR1 and ABP1 respectively (the two proposals of auxin receptors) (Fig. 1, B). In TIR1 – crystal is well represented the binding by the N-indole, while a complex binding system appear in the complex with Zn in ABP1. The results of both crystals suggest a duality of auxin properties: able to bind in two totally different systems. On the one side (ABP1), the

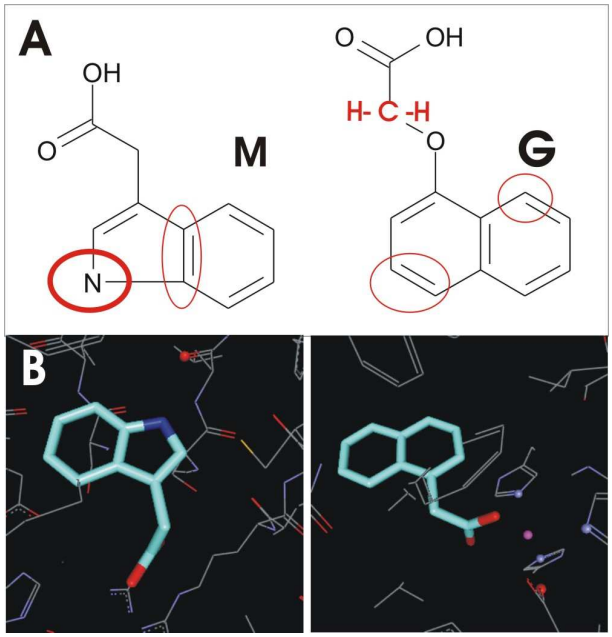


Figure 1. Accessory binding areas. (M) morphogenesis, (G) growth, (H-C-H) region buffer important for growth regulation.

molecule-protein interaction via auxin side chain evoke totally different chemical binding conditions, to which the existence of the sp^2 carbon, statistically inferred, could be important. On the other side (TIR1), the interaction via the N-indole in TIR1 was predicted before as an important point for morphogenetic events[11]. Likewise, both crystal auxin – protein do not show any rearrangement of the atoms of the protein, due to the binding. The structural information rather suggests electron arrangement, like predicted by our calculations with a common variable influences the biological activity of auxins at every level: the chemical hardness[11]. This is also able to classify the 240 auxins together with the two principal components obtained by QMSM (Fig. 2). Soft molecules are more active than hard molecules if electron transfer or rearrangement is necessary for the reaction[12]. This concept is exactly confirmed by our results. Functional dependence with hardness having natural auxins as outliers[11], suggests a relation with the limits of reactivity in biological systems; it means toxicity. Contrary to the natural auxins, synthetic auxins like 2,4-D are potent herbicides and their hardness is very high. Consequently, other sites and/or inespecific interactions in the cellular systems (proteins) are expected, it dependent on the molecular structures. These results, the broad quantity of active molecules as auxins and many of them being almost as active as the standard auxin IAA, indicate controversy between a typical specificity of a hormonal mechanism and the auxin concept in plants. The strategies described above are the key of chemical design, which will result in the next generation of new synthetic auxins and/or comprehensive bases of the action mechanism. This may be also extended to the phytohormone context.

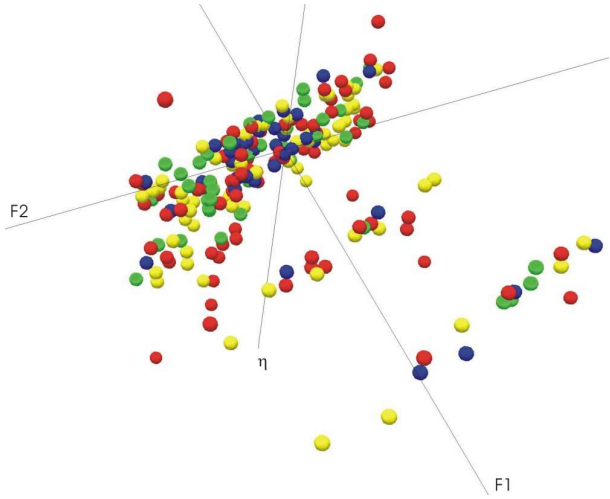


Figure 2. Influence of hardness in the classification of 241 auxin like molecules with QMSM components.

4 Conclusion

It was demonstrated that a combination of both electronic structure and intermolecular interaction descriptors was able to describe this multi-dimensional biological phenomena. The Coulomb matrix, as an electrostatic potential descriptor, was the most important differentiating auxin activity measure as evidenced by discriminant analysis. A phenol derivative, predicted as active substance, was confirmed to be active in different auxin tests. The molecular descriptors and regions analyzed functionally, from phenomenological point of view, are structurally related with the differences of both binding sites reported for ABP1 and TIR1. The electronic rearrangement in the interaction auxin - receptor predicted by hardness (η) is confirmed by non atoms rearrangements in the both auxin-ABP1 and auxin-TIR1 interactions. The existence of sp^2 carbon in the side chain turns the induction of activity of such substances at very low concentrations.

References

1. L. J. Audus, *Metabolism and mode of action*. (In: Ruhland W, ed. Encyclopediea of Plant Physiology Berlin: Springer, 1961).
2. X. Tan, L A. Calderon Villalobos, M. Sharon et al. *Mechanism of auxin perception by the TIR1 ubiquitin ligase*, Nature **446**, 640, 2007.
3. R. M. Napier, K. M. David, C. Perrot-Rechenmann, *A short history of auxin-binding proteins*, Plant Molecular Biology **49**, 339, 2002.
4. J. Friml, J. Wisniewska, E. Benkova, et al. *Lateral relocation of auxin efflux regulator PIN3 mediates tropism in Arabidopsis* Nature **415**, 806, 2002.
5. J. J. Blakeslee, A. Bandyopadhyay, W. A. Peeret, al. *Relocalization of the PIN1 auxin efflux facilitator plays a role in phototropic responses*, Plant Physiology **134**, 28, 2004.

6. C. Hansch, R. M. Muir, *The Ortho Effect in Plant Growth-Regulators*, Plant Physiology **25**, 389, 1950.
7. H. Veldstra, *The Relation of Chemical Structure to Biological Activity in Growth Substances*, Annual Review of Plant Physiology and Plant Molecular Biology **4**, 151, 1953.
8. G. F. Katekar, *Auxins - Nature of the Receptor-Site and Molecular Requirements for Auxin Activity*, Phytochemistry **18**, 223, 1979.
9. S. Tomic, R. R. Gabdoulline, B. Kojic-Prodic, R. C. Wade, *Classification of auxin plant hormones by interaction property similarity indices*, Journal of Computer-Aided Molecular Design **12**, 63, 1998.
10. N. Ferro, A. Gallegos, P. Bultinck et al. *Coulomb and Overlap Self-Similarities: A comparative selectivity analysis of structure - function relationships for auxin - like molecules*, Journal of Chemical Information and Modeling **46**, 1751, 2006.
11. N. Ferro, P.Bultinck, A. Gallegos, H.-J. Jacobsen, R. Carbo-Dorca, T. Reinard, *Unrevealed structural requirements for auxin-like molecules by theoretical and experimental evidences*, Phytochemistry **68**, 237, 2007.
12. R. G. Pearson, *Absolute Electronegativity and Hardness Correlated with Molecular-Orbital Theory*, PNAS **83**, 8440, 1986.

Crystal Water Molecules and Solvation Effects on Protein-Ligand Docking

B. Fischer and W. Wenzel

Institute for Nanotechnology,
Research Center Karlsruhe, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: {fischer-b, wenzel}@int.fzk.de

In this report, we briefly present our recent study which investigates the influence of solvation energies and of crystal water molecules on the ligand binding orientation and affinity. Recently, we integrated fully movable water molecules into our docking program *FlexScreen*. These water molecules have translational and rotational degrees of freedom and if their presence is very unfavorable at a protein-ligand conformation, they are not considered at all. Additionally, we present our progress in improving our scoring function by incorporating into the function the solvation energies of the protein-ligand complex. The solvation energies are also calculated during the docking simulation itself.

1 Methods

A high-throughput in-silico screening method like *FlexScreen* consists of three major elements: (1) a *suitable virtual representation* of existing protein-ligand systems, (2) a *scoring function* that approximates the binding energy (ideally the affinity) of the protein-ligand complex as a function of the conformation of this complex[1,2] and (3) an efficient *optimization method* that is able to locate the best protein-ligand conformation on the potential energy surface[3,4]. In this report, we present our progress in points (1) and (2). So far, we have already developed a simulation approach that allows the protein structure to adapt to the docking ligand by conformational changes in the side chains. This approach has proved to be successful and results in more accurate and reliable database screens than an approach using only one rigid protein structure for the simulations[2].

1.1 Crystal Water Molecules

An additional important aspect is the influence of water molecules on the ligand binding affinity and on the ligand binding orientation. Firstly, we focus on the influence of crystal water molecules. These water molecules are strongly stabilized in the hydrogen bond network of the protein and therefore are well preserved in the crystal structures obtained by X-ray diffraction patterns. Secondly, solvation effects are considered in sec. 1.2.
When a ligand binds to a protein, it pushes aside water molecules in the protein cavity. As a consequence, a part of the water molecules are removed into the bulk solvent whereas others are involved in the stabilization of the ligand binding orientation. The spatial distribution of crystal water molecules in the protein cavity differs for bound ligands. Using one of those distributions as a fixed element in docking simulations, prevents many other high-affinity ligands from binding. Because of this and the already large conformational protein-ligand space, changeable water molecules are usually neglected for simulations.
In the following, we report our progress and our methodology for docking simulations

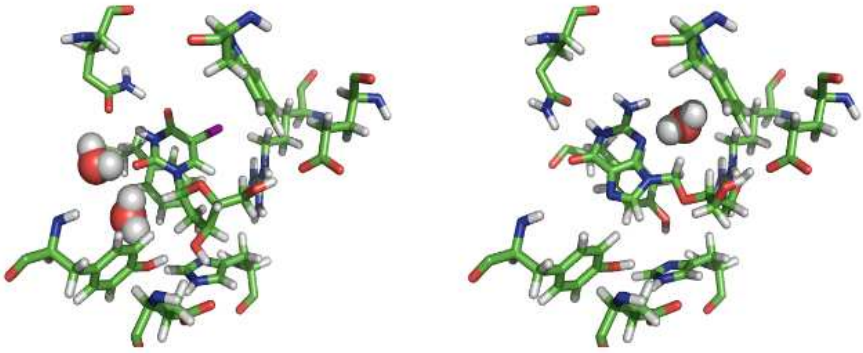


Figure 1. Illustration of two docking simulations with three additional flexible water molecules. Left picture: binding conformation of the ligand idu with the protein thymidine kinase (TK). Right picture: binding conformation of the ligand gcv with TK.

with flexible and movable water molecules in the protein cavity. Before docking, we select flexible water water molecules which have the degrees of freedom of translation, rotation and of moving into the solvent, if unfavorable. We performed docking simulations with the ligand gcv and idu to the protein thymidine kinase (TK). Figure 1 illustrates both resulting binding conformations with our new methodology. In the left picture of figure 1, the ligand idu is stabilized by two water molecules, whereas in the right picture the ligand gcv is bound to only one crystal water molecule. Our new methodology has the advantage that water molecules may actively stabilize a ligand conformation with their updated orientation. At the same time ligands are not hindered from binding.

1.2 Solvation Effects

The hydrophilic and hydrophobic properties of a ligand do not only influence the binding affinity, but also the binding orientation in protein pockets accessible to the bulk solvent. In the *FlexScreen* scoring-function we approximate the solvation energies through an implicit solvation model by calculating the solvent accessible surface areas of each atom (SASA)[5]

$$E_{Solvation} = \sum_i \gamma_i A_i^{\mathrm{SASA}}, \tag{1}$$

A_i^{SASA} being the solvent accessible surface area of atom i and γ_i the proportional factor of the atom type of atom i that relates the surface area to solvation energies. Considering these contributions during a simulation enables us to also reliably predict the binding orientations in open protein pockets. Figure 2 illustrates the consequences of neglecting and of considering the solvation effects, described with eq. 1. Docking conformations close to the native conformations can only be stabilized by taking solvation energies into account. As a consequence, methods like the thermodynamic cycle approach[6] can not be applied here. This method is only applicable, if the final protein-ligand conformation is independent of simulations being in vacuum or in solution.

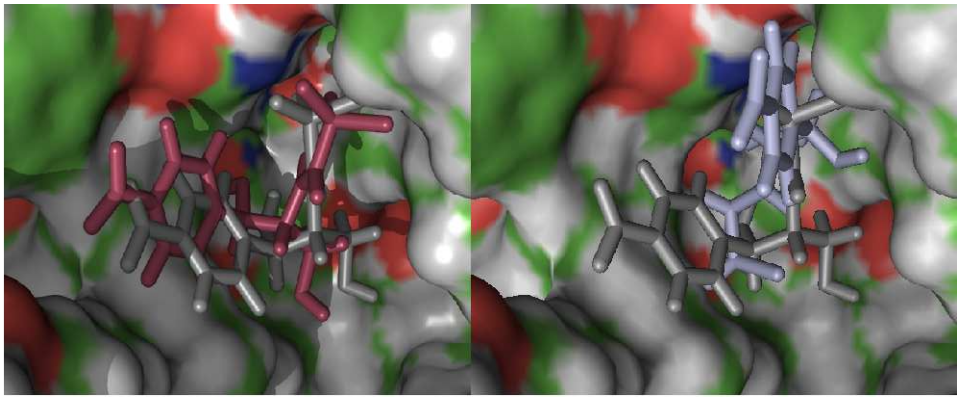


Figure 2. Comparing *FlexScreen* docking results with the experimental binding mode of the protein-ligand complex 3cla (pdb-code). The native ligand conformation is colored grey. Left picture: solvation effects are not considered and the simulation fails to predict the native binding mode. Right picture: solvation effects are included and the experimental binding conformation can be reproduced.

2 Conclusion

Crystal water molecules and water molecules from the bulk solvent sometimes have to be considered during docking simulations itself. With our approach we have started to study the effects these water molecules have. Larger quantitative studies are still missing and further work has also to be done to refine the parameters for the SASA-solvation energies.

Acknowledgments

We thank the Fond der Chemischen Industrie, the BMBF, the Deutsche Forschungsgemeinschaft (grant WE 1863/13-1) and the Kurt Eberhard Bode Stiftung for financial support.

References

1. H. Merlitz and W. Wenzel, *Comparison of stochastic optimization methods for receptor-ligand docking*, Chem. Phys. Lett. **362**, 271–7, 2002.
2. H. Merlitz and W. Wenzel, *Impact of receptor flexibility on in-silico screening performance*, Chem. Phys. Lett. **390**, 500, 2004.
3. W. Wenzel, K. Hamacher, *Stochastic Tunneling Approach for Global Minimization of Complex Potential Energy Landscapes*, Phy. Rev. Lett. **82**, 3003–07, 1999.
4. A. Schug, B. Fischer, A. Verma, H. Merlitz, W. Wenzel, G. Schön, *Biomolecular structure prediction stochastic optimization methods*, Adv. Eng. Mater. **7**, 1005–9, 2005.
5. D. Eisenberg, A.D. McLachlan, *Solvation energy in protein folding and binding*, Nature **319**, 199–203, 1986.
6. K.A. Sharp, *Virtual Screening in Drug Discovery* (Taylor & Francis, 2005).

Semiautomatic Workflow for Fold Recognition – Results from the CASP 2006 Competition

Fabian Fredel[1], **Jan Meinke**[1], **Sandipan Mohanty**[1], **Olav Zimmermann**[1], **and Ulrich H. E. Hansmann**[1,2]

[1] John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {*f.fredel, j.meinke, s.mohanty, olav.zimmermann, u.hansmann*}*@fz-juelich.de*

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

We outline a semi-automatic procedure for structure prediction of proteins. A first analysis of the performance of this procedure in the CASP 2006 competition is presented.

1 Introduction

In order to test combinations of physics-based simulation techniques and sequence-based prediction methods, our group participated in the "Critical Assessment of Techniques for Protein Structure Prediction" (CASP) competition in the summer of 2006. As a first-time participant our goal was to establish a semi-automatic workflow. We combined existing methods for fold recognition with our refinement algorithms testing heuristics for the selection at each step. In this article, we give an overview of the workflow and the results of an in-depth statistical analysis of our results. In particular, we assess the significance of measured performance differences between the prediction methods. Analyzing our workflow, we try to find the critical points where alternative decisions lead to a significant change in the results. Our aim is to obtain rules that guide the decision process in the workflow to optimize our predictions.

2 Workflow

The first step in our workflow is the manual selection of templates from 3D-Jury[1] predictions. Preference is given to high 3D-Jury-scores and agreement between the secondary structure of the template and the predicted secondary structure of the target sequence. For targets that were obviously not comparative modeling targets, 3D-Jury predictions from fold recognition servers are preferred.

We search the fold space[2] using CABS[3]. This parallel tempering Monte Carlo program uses constraints from the respective 3D-Jury templates and secondary structure prediction by PSIpred 2.5[4]. We use 32 replicas for sequences with less than 200 residues and 64 replicas for proteins with longer sequences. The statistics was between 15,000 sweeps for long sequences and 100,000 sweeps for short sequences.

Clustering is performed using hierarchical clustering with HPCM[5] using a fixed RMSD of 2.5 Å as clustering radius. Structure clusters are selected based on cluster averages of CABS energy and structure similarity (TM-score[6]) to the PDB structure on which the 3D-Jury template was based.

Averaged structures from the selected clusters are subject to regularization by SMMP[7]. Regularized structures are ranked according to the total and partial energies of the structures in SMMP. In ambiguous cases, the consistency of this ranking with a similar ranking based on energy terms of PROFASI[8] is checked.

The 5 to 10 structures ranked best are selected for refinement. For most structures, refinement involves a set of constrained simulated annealing runs with SMMP, starting from very high temperatures. Most structures dissolve and re-form into local minima of the potential that are close to the input structures of the refinement procedure. The final structures from different annealing trajectories are once again ranked following a similar procedure as for the initial selection for the refinement runs. Final selection and ranking is based on several energy terms, secondary structure content and visual inspection.

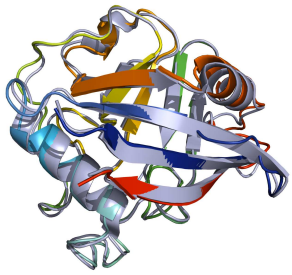


Figure 1. Grey is the experimental structure (2HE9). Colored is our best structure for T0346.

3 Comparing Prediction Methods

Not all of our predictions are as good as the one in figure 1. In order to assess the significance of measured performance differences between us and other prediction methods we use a nonparametric statistical test, the Friedman test[9]. It has a simple two-way layout for k treatments (groups) and n blocks (target).

TARGET	GROUP A_1 A_2 A_3 ... A_k	$R_i = \sum_{j=1}^{k} r_{ij}$
#1	r_{11} r_{12} r_{13} ... r_{1k}	R_1
#2	r_{21} r_{22} r_{23} ... r_{2k}	R_2
#3	r_{31} r_{32} r_{33} ... r_{3k}	R_3
... #n	r_{n1} r_{n2} r_{n3} ... r_{nk}	R_n
$R_j = \sum_{i=1}^{n} r_{ij}$	R_1 R_2 R_3 ... R_n	$\sum_{j=1}^{k} R_j = \sum_{i=1}^{n} R_i$

For each of the n experiments the k results are ranked from 1 for the best to k for the worst result. The ranked results are based on TM-score and RMSD for the predictions to the experimental structure. The Null-hypothesis of the test is: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ (no treatment differs). The rank of group j at experiment i is given by r_{ij}. R_j is the sum of the n ranks of group j. $\bar{R_j}$ is the mean of the n ranks of group j and $R.. = \frac{k+1}{2}$.
Compute

$$S = \frac{12\,n}{k\,(k+1)} \sum_{j=1}^{k} (\bar{R_j} - R..)^2.$$

S is an approximation for the χ^2 distribution.

The Nullhypothesis is rejected in all test cases. Across the field we find significant differences between all servers. The best servers are *Zhang-Server*, *MetaTasser*, *Pmodeller 6* and *BayesHH* outperforming our procedure and demonstrating the need for further improvement of our workflow.

4 Workflow Analysis

For this reason, we have decided to search the workflow for critical points where alternative decisions lead to significant changes and improvements in our results. We asked ourselves the following questions:

- Do we select the best template?
- Do we trust PSIpred for the secondary structure prediction?
- Which structures should be used for clustering?
- Which clusters are the best?
- Is the final ranking of energy terms ideal to find the best structure?

As an example, we show the analysis of the workflow for target T0354 (130 residues):

best template selected ?		best template	our template
	RMSD	3.98	3.89
	TM-score	0.515	0.395
do we trust PSIpred ?	good secondary structure prediction by PSIpred		
which structures for clustering ?	good predictions often at the lower replica numbers		
best clusters are?	average of CABS energy not always the best		
ranking of energy terms ideal ?	structure ranked 1st	TM-score 0.3323	
	best structure ranked 12th	TM-score 0.5085	

5 Conclusion

We have described a method for structure prediction of proteins. While currently not competitive with other approaches, we have shown a way to analyze its performance and to explore possible improvements.

Acknowledgment

The calculations were done on computers of the John von Neumann Institute for Computing, Forschungszentrum Jülich, Jülich, Germany.

Ulrich H. E. Hansmann is supported by a research grant (GM62838) of the National Institute of Health (USA).

References

1. K. Ginalski, A. Elofsson, D. Fischer and L. Rychlewski, Bioinformatics **19**, 1015-1018, 2003.
2. U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140, 1997.
3. A. Kolinski and J. M. Bojnicki, Proteins **61 Suppl. 7**, 84-90, 2005.
4. D. T. Jones, J. Mol. Biol. **292**, 195-202, 1999.
5. D. Gront and A. Kolinski, Bioinformatics **21**, 3179-3180, 2005.
6. Y. Zhang and J. Skolnick, Proteins **57**, 702-710, 2004.
7. F. Eisenmenger, U. H. E. Hansmann, S. Hayryan and C. K. Hu, Comp. Phys. Comm. **174**, 422, 2006.
8. A. Irbaeck and S. Mohanty, J. Comp. Chem. **27**, 1548, 2006.
9. Myles Hollander, Douglas A. Wolfe (1973), *Nonparmetric Statistical Methods*, John Wiley & Sons (ISBN 0-471-40635-X).

Interaction of Biological Matter with Nanomaterials: A First-Principles Approach

S. Gowtham[1], Ralph H. Scheicher[1], Rajeev Ahuja[2], and Ravindra Pandey[1]

[1] Department of Physics, Michigan Technological University, Houghton, MI USA
E-mail: sgowtham@mtu.edu

[2] Condensed Matter Theory Group, Department of Physics, Uppsala University, Uppsala, Sweden
Applied Materials Physics, Department of Materials and Engineering,
Royal Institute of Technology, Stockholm, Sweden

1 Introduction

DNA-coated carbon nanotubes represent a hybrid system which unites the biological regime and the nanomaterials world. They possess features which make them attractive for a broad range of applications, e.g., as an efficient method to separate carbon nanotubes (CNTs) according to their electronic properties[1–3], as highly specific nanosensors, or as an in vivo optical detector for ions. It has also been experimentally demonstrated that ssDNA can be inserted into a CNT[5], further enhancing the potential applications of this nano-bio system. Our specific interest is to assess the subtle differences in the adsorption strength of these nucleobases on graphene, which in turn will allow us to draw conclusions for the interaction of DNA and RNA with CNTs as well.

2 Computational Method

Calculations were performed using the plane-wave pseudopotential approach within the local density approximation (LDA)[6] of density functional theory (DFT)[7], as implemented in the Vienna Ab-initio Simulation Package (VASP)[8].

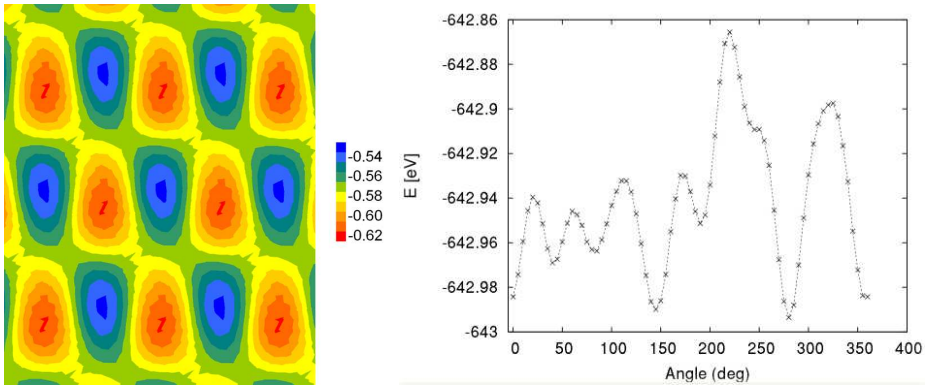
A 5×5 array of the graphene unit cell in the x-y plane and a separation of 15 Å between adjacent graphene sheets in the z-direction was found to be a suitable choice to represent the supercell. The base molecules were terminated at the cut bond to the sugar ring with a methyl group in order to generate an electronic environment in the nucleobase more closely resembling the situation in DNA and RNA rather than that of just individual isolated bases by themselves.

For each of the five nucleobases, an *initial force relaxation* calculation step determined the preferred orientation and optimum height of the planar base molecule relative to the graphene sheet. The potential energy surface was then explored by translating the relaxed base molecules parallel to the graphene plane in steps of 0.246 Å along the lattice vectors of graphene, covering its entire unit cell by a mesh of 10×10 scan points. The determination of the minimum total energy configuration was then followed by a 360° rotation of the base molecules in steps of 5° to probe the dependence of the energy on the orientation of the base molecules with respect to the underlying 2-D graphene sheet. The configuration yielding the minimum total energy was used in the final optimization step in which all atoms in the system were free to relax.

3 Results and Discussion

From the optimization steps involving the translational and the rotational scan of the energy surface, it is apparent that the energy barriers to lateral movement and to rotation of a given base can range from 0.04 to 0.10 eV (Fig. 1(a)), thus considerably affecting the mobility of the adsorbed nucleobases on the graphene sheet at room temperature, and constricting their movement to certain directions.

In their equilibrium configuration, all five bases tend to position themselves on graphene in a configuration reminiscent of the Bernal's AB stacking of two adjacent graphene layers in graphite (Fig. 2). Virtually no changes in the interatomic structure of the nucleobases were found in their equilibrium configurations with respect to the corresponding gas-phase geometries, as it could be expected for a weakly interacting system such as the one studied here. A notable exception is the R_{C-O} in guanine which shows a 10% contraction upon physisorption of the molecule on graphene.



(a) Surface Scan (b) Rotation Scan

Figure 1. Potential energy surface (PES) plot (in eV) for guanine on graphene. Qualitatively similar PES plots were obtained for the other four nucleobases.

The binding energy of the system consisting of the nucleobase and the graphene sheet is taken as the energy of the equilibrium configuration with reference to the asymptotic limit obtained by varying the distance between the base and the graphene sheet in the z-direction (Table 1). This table also includes the polarizabilities of the nucleobases calculated at the MP2 level of theory. The polarizability of the nucleobase[9], which represents the deformability of the electronic charge distribution, is known to arise from the regions associated with the aromatic rings, lone pairs of nitrogen and oxygen atoms. Accordingly, the purine base guanine appears to have the largest value, whereas the pyrimidine base uracil has the smallest value among the five nucleobases. Our calculations confirm this behavior.

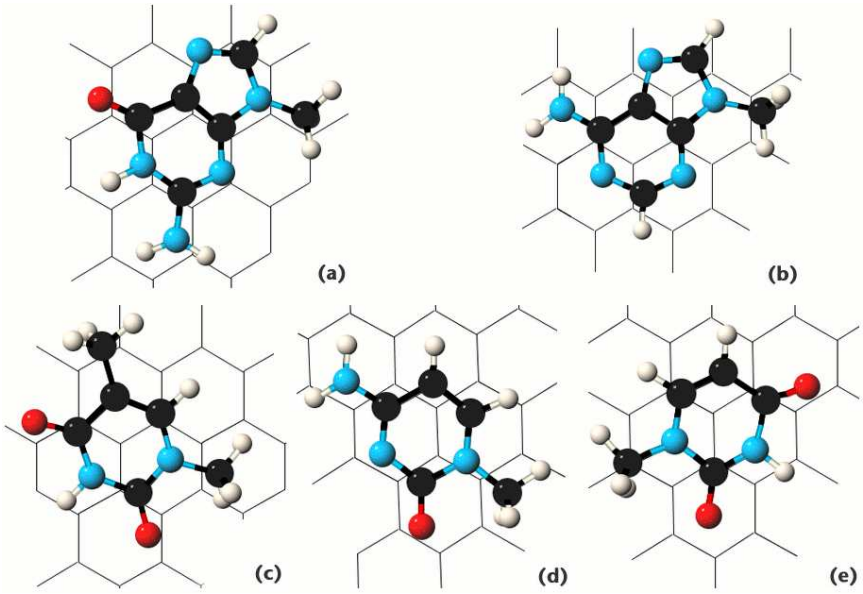


Figure 2. Equilibrium geometry of nucleobases on top of graphene: (a) guanine, (b) adenine, (c) thymine, (d) cytosine, and (e) uracil.

Base	E_b(LDA) [eV]	E_b(MP2) [eV]	α $[e^2 a_0^2 E_h^{-1}]$
G	0.61	1.07	131.2
A	0.49	0.94	123.7
T	0.49	0.83	111.4
C	0.49	0.80	108.5
U	0.44	0.74	97.6

Table 1. Binding energy E_b of the DNA/RNA nucleobases with graphene as calculated within LDA are compared with binding energy and polarizability α from MP2 calculations.

4 Conclusions

Our first-principles results clearly demonstrate that the nucleobases exhibit significantly different interaction strengths when physisorbed on graphene. This finding represents an important step towards a better understanding of experimentally observed sequence-dependent interaction of DNA with CNTs[3,4]. The calculated trend in the binding energies strongly suggests that the polarizability of the base molecules determines the interaction strength of the nucleobases with graphene. Further studies involving the investigation of nucleobases interacting with small-diameter CNTs are currently underway.

Acknowledgments

S.G., R.H.S., and R.P. would like to thank DARPA for funding. Authors are grateful to Center for Experimental Computation at MTU and the Swedish National Infrastructure (SNIC) for computing time.

References

1. N. Nakashima *et al.*, Chem. Lett. **32**, 456 (2003).
2. M. Zheng *et al.*, Nature Mater. **2**, 338 (2003).
3. M. Zheng *et al.*, Science **302**, 1545 (2003).
4. C. Staii *et al.*, Nano Lett. **5**, 1774 (2005).
5. T. Okada *et al.*, Chem. Phys. Lett. **417**, 288 (2006).
6. J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
7. P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, *ibid.* **140**, A1133 (1965).
8. G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996); G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
9. Frank Seela, Anup M. Jawalekar, and Ingo Münster, Helevetica Chimica Acta **88**, 751 (2005).

Computational Reconstruction of Macromolecular Assemblies

Axel Griewel and Matthias Rarey

Center for Bioinformatics
University of Hamburg
Bundesstr. 43
20146 Hamburg, Germany
E-mail: {griewel, rarey}@zbh.uni-hamburg.de

Cryo electron microscopy (cryo EM) has proven to be an effective technique to gather low resolution information of macromolecular structures. Methods like X-ray crystallography and NMR spectroscopy on the other hand supply atomic detail structures but tend to fail in the investigation of large specimen. We present a method for elucidating macromolecular structures by combining atomic detail structures of subunits with overall shape information gathered by cryo EM – a problem for which only few computational methods have been published so far. In our approach atomic detail structures are reduced to a set of spheres by computational methods like gift-wrapping and Delaunay triangulation. These describing spheres are then matched to a hard sphere-filtered cryo EM map employing pose clustering, a technique adapted from computer vision. Tests were performed using experimental and simulated cryo EM maps. Calculations yield results within 3.5 Å rmsd accuracy or better in a runtime of less than 5 minutes.

1 Introduction

Cryo electron microscopy is as a powerful imaging technique for the elucidation of macromolecular structures. The resolution of the resulting density maps, however, does not allow for atomic detail interpretation. For this task methods like X-ray crystallography and NMR spectroscopy are more suited, while they also tend to fail in the investigation of large specimen. Combining atomic detail structures of subunits with cryo EM is a viable method for determining macromolecular structures. We developed a fast and accurate tool incorporating information from different sources for the computational reconstruction of macromolecular assemblies (CRoMA).

All published methods for this problem except one rely on a sequential scan of the six dimensional degrees of freedom given by rotation and translation of the two objects relative to each other. Only one published approach utilizes topology representing neural networks. Three approaches for a scoring function can be found in literature. One is to calculate a synthetic cryo EM map of the atomic structure at the desired resolution and to determine the cross correlation of the two maps. Another approach utilizes laplacian filtered maps yielding a score reflecting the overlap of borders. A third scoring function uses trilinear interpolation for each atom.[1,2]

2 Method

The atomic detail structure as well as the map are preprocessed yielding descriptors consisting of spheres of different radii. These descriptors are utilized to determine favourable placements by a technique adapted from computer vision.

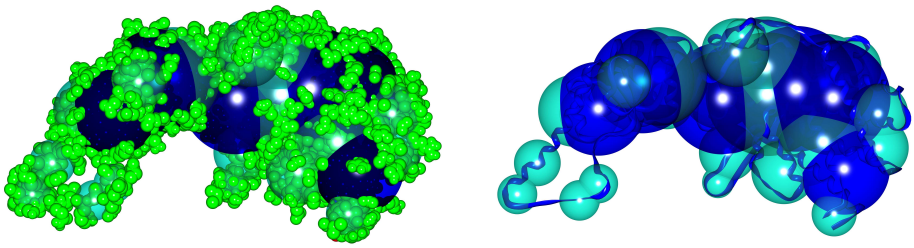


Figure 1. Atomic detail structure and ribbon model of a monomer of the thermosome from T. Acidophilium (PDB accesion code 1A6D) together with the implemented sphere-descriptor

The goal of the descriptor for structures given with atomic resolution is to determine few large spheres describing the shape of the structure in decent detail. Therefore all surface atoms of the structure are determined using a gift-wrapping algorithm. For the set of surface atoms the three-dimensional Delaunay triangulation is computed. An essential property of this triangulation is that each tetrahedron's circumsphere does not contain any other surface atom, i. e. also the largest spheres inside the structure are contained in this set of spheres. All circumspheres outside the structure are discarded while a subselection of spheres describing the interior of the structure is chosen employing a set cover algorithm. An example of a descriptor can be found in Figure 1. The cryo EM map descriptor is constructed by filtering with isodense spheres of different radii. The radii for filtering are chosen according to the size of the spheres found as structure descriptor.

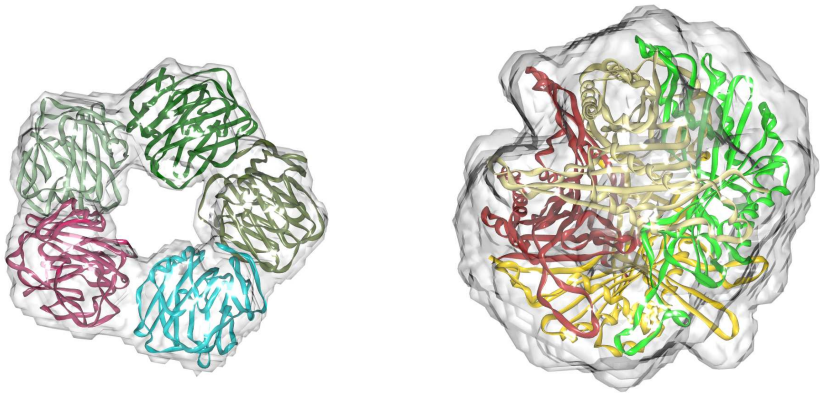
For calculating the placement of the high resolution structure in the low resolution density map, pose clustering[3] – a powerful method from pattern recognition – is used. Triangles between centers of characteristic spheres in the structure-descriptor are used to query a database of triangles build up by selected voxels of the filtered map. The pose clustering algorithm then proposes positions in which the three structure spheres generating the triangle coincide with map spheres of similar radius.

All proposed placements are subject to a quick scoring function which utilizes the sphere representation of structure and map. A certain amount of proposed placements is then retained and scored according to a fine-grained scoring function based on trilinear interpolation. The resulting placements can be clustered given a user defined rmsd threshold.

3 Results

A first test was performed by docking monomers taken from the molecular complex GroEL deposited in the PDB with accession code 1OEL to map 1081 taken from the EM-Search database[4]. The rmsd between the calculated and the correct position was 1.7 Å and the computing time was less than four minutes.

Further tests were conducted on simulated maps using various structures. The result of docking structures to urate oxidase (2FUB) is shown in Figure 2. Within ten minutes runtime placements with 3.4 Å rmsd to the original placement were found. Another example



Human c-reactive protein (1B09) Urate oxidase (2FUB)

Figure 2. Solution structures

is docking monomers to a map generated from human c-reactive protein (1B09). Here a rmsd of 2.3 Å was achieved in a runtime below one minute. The resulting placements for the latter two examples can be found in Figure 2.

4 Perspectives

Besides the success in several cases, the algorithm fails in a few docking scenarios. This is the case, if the molecular mass of the structure to be docked makes up only a very small fraction of the overall mass. Further improvements to the scoring function will help tackling these cases. Although not yet fully developed, first applications of this approach suggest that the algorithm is going to be highly efficient and effective.

References

1. J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies* (Oxford University Press, New York, 2006).
2. M. G. Rossmann, Marc C. Morais, Petr G. Leiman, and Wei Zhang, *Combining X-ray Crystallography and Electron Microscopy*, Structure **13**, 355–362, 2005.
3. S. Linnainmaa, D. Harwood, L. S. Davis, *Pose determination of a three-dimensional object using triangle pairs*, IEEE trans. pattern anal. mach. intell. **10**, 634–647, 1988.
4. *A database for EM*, Nat Struct Biol **10**, 313, 2003.

Verification of Protein-Protein Interactions by Use of Docking Techniques

Florian Fink[1], **Rainer Merkl**[2], **and Wolfram Gronwald**[1]

[1] Institute of Functional Genomics,
University of Regensburg, Josef-Engert-Strasse 9,
93053 Regensburg, Germany
E-mail: florian.fink@klinik.uni-regensburg.de,
wolfram.gronwald@biologie.uni-regensburg.de

[2] Institute for Biophysics and Physical Biochemistry,
Postfach, 93040 Regensburg, Germany
E-mail: rainer.merkl@biologie.uni-regensburg.de

For the understanding of large macromolecular complexes such as ribosomes the analysis of protein-protein interactions is essential. These intermolecular interactions are strongly dependent on the three-dimensional structures of the corresponding molecules. In case that the structures are known they can be directly used while in many other cases homology modeling techniques can be applied. We have developed a novel algorithm for this purpose that allows the combination with additional experimental data to further improve the structural models. Currently we are developing tools based on a data driven docking approach and the 3D structures of the individual molecules to investigate whether proposed intermolecular interactions can be verified or falsified. In this contribution we will show first results to demonstrate the principal applicability of our approach.

1 Introduction

Most of the various functions in a cell are mediated by large protein-protein interaction networks. For a detailed understanding of these interactions knowledge of the corresponding three-dimensional complex structures is required. However, a significant amount of these complex structures will be extremely difficult to study by conventional experimental structure determination methods. One avenue to circumvent this problem is to use the structures of the individual molecules in combination with computational docking techniques. It has been shown that reliable results can be obtained when data driven docking techniques are applied[1] and only moderate structural changes occur during complex formation. As mentioned above three-dimensional structures of the individual molecules are required for the application of docking techniques. If these structures are already available they can be directly used. In many other cases the use of homology models is applicable. We have developed for this purpose the homology modeling program PERMOL[2] that is based on restrained molecular dynamics in torsion angle space. For the case that sufficient experimental data is available to obtain low resolution structures of the single molecules we developed the ISIC[3] algorithm to improve the structural quality by combining information from different sources. The key question here is how to combine the available information ensuring that no wrong structural bias is introduced.

Next the 3D-structures of the single molecules are used for *in silico* complex formation. In this contribution we focus on the question if data driven docking techniques provide

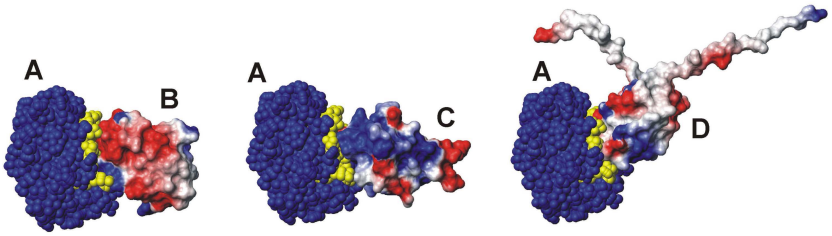


Figure 1. Example on the use of docking to query proposed intermolecular interactions. Molecules B, C and D are forced to interact with the target molecule A for which the interface (shown in yellow) is known.

Interaction A-B	Interaction A-C	Interaction A-D
E_{inter} -2512,2 kJ/mol	E_{inter} -1991,4 kJ/mol	E_{inter} -1695,6 kJ/mol

Table 1. Average interactions energies for three different trial molecules (B, C, and D) docked to the cytoplasmatic A domain (IIA(MTL)) of the mannitol transporter II (A). (B) histidine containing phospocarrier protein (HPr), (C) human cyclin dependent kinase subunit type I (CKSHS1), and (D) apo form of HMA domain of copper chaperone for superoxide dismutase.

additional information to query proposed binary protein-protein interactions. More specifically we assume that the binding-interface of one molecule is known and it is investigated which of several proposed partner molecules is the correct interaction partner.

2 Motivation

Over the last few years several high-throughput protein-protein interaction detection methods have been developed. However, as shown in the paper by von Mering et al[4] the proposed interactions of these methods usually contain many false positives. And although substantially improvements can be obtained by combing several methods it is clear that additional work is required to reduce the amount of false positives. Since intermolecular interactions are strongly dependent on the three-dimensional structures of the corresponding molecules, docking techniques should provide additional information in this regard.

3 Materials and Methods

To query proposed protein-protein interactions different trial molecules are forced to bind to the known interface of one target molecule. For this purpose the data-driven docking algorithm HADDOCK[1] that is based on the use of ambiguous interaction restraints is used. The known interface information of the target protein is provided as restraint information to the docking algorithm. This information can be obtained for example from NMR chemical shift perturbation data, mutagenesis data etc. For the proposed trial molecules

we assume that nothing is known about their interface and the complete surface of these molecules is defined as potential binding interface. Separate docking runs are performed for each of the trial molecules. Results are ordered based on the obtained interaction energies. Interaction energies are calculated based on the intermolecular van der Waals and electrostatic interface energies.

4 Results and Discussion

For testing, two molecules A and B that are known to interact are selected in the example shown in Fig. 1. Then in addition two molecules C and D, randomly selected from the PDB database are used, to perform docking runs with molecule A. For the target the binding interface shown in yellow in Fig. 1 is assumed to be known while for molecules B, C, and D the whole molecule is defined as possible interaction site. Results for this example demonstrate that the lowest interaction energies shown in Tab. 1 are obtained for the correctly interacting pair A-B while considerably higher energies are obtained for the non-interacting pairs A-C and A-D allowing to correctly discriminate between interacting and non-interacting proteins. These tests were repeated for several different test cases to investigate the general applicability of the method, where the so far obtained data show similar results (data not shown).

Therefore, in summary one can say that docking techniques can provide useful additional information to interrogate proposed protein-protein interactions.

Acknowledgments

The authors thank the bavarian genomic network for financial support.

References

1. C. Dominguez, R. Boelens and A. M. J. J. Bonvin, *HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information*, J. Am. Chem. Soc. **125**, 1731–1737, 2003.
2. A. Möglich, D. Weinfurtner, T. Maurer, W. Gronwald and H. R. Kalbitzer, *A Restraint Molecular Dynamics and Simulated Annealing Approach for Protein Homology Modeling Utilizing Mean angles*, BMC-Bioinformatics **6**, 91, 2005.
3. K. Brunner, W. Gronwald, J. M. Trenner, K.-P. Neidig and H. R. Kalbitzer, *A general Method for the Unbiased Improvement of Solution NMR Structures By the Use of Related X-Ray Data, the AUREMOL-ISIC Algorithm*, BMC Struc.Biol. **6**, 14, 2006.
4. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, *Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions*, Nature **417**, 399–403, 2002.

Boundary Element Method (BEM) with Parametric Surfaces

Sergei Grudinin[1,2] **and Artur Baumgaertner**[1]

[1] Institute of Solid State Research (IFF),
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {S.Grudinin, A.Baumgaertner}@fz-juelich.de

[2] Institute of Neurosciences and Biophysics (INB-2),
Research Centre Jülich, 52425 Jülich, Germany

This work describes a new Boundary Element Method (BEM) implementation for biomolecular solvation with parametric surfaces. First, multi-scale volumetric synthetic electron density maps are constructed from parsed atomic location data of biomolecules using Gaussian isotropic kernels. Next, three different methods are used to extract triangular meshes for the molecular surface. They are: marching cubes, marching tetrahedra and marching tetrahedra with dual contouring. Then generated meshes are used in BEM electrostatic calculations. In this work we study: 1) calculation time and accuracy for multipole and direct electrostatic solvers; 2) energy convergence and calculation time with the density of boundary points different meshing algorithms; 3) energy convergence for different iterative linear solvers.

1 Introduction

Electrostatic interactions are known to play a key role in determining the structure and activity of biomolecules. Interactions with solvent are very important for biomolecular functioning. Here we present a new method for the implicit representation of solvent along with algorithms for calculation of electrostatic interactions.

2 Molecular Surface Representation

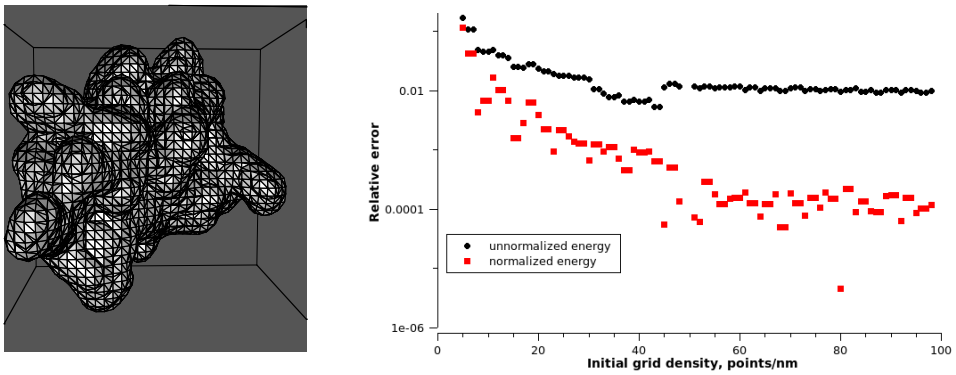
For our model we have chosen an implicit surface representation. The 3-dimensional scalar function $D(\vec{r})$ (Eq. 1), which is a sum of Gaussians centered at atomic nuclei, is similar to the electron density function. Then the isosurface $D = 0$ represents a molecular solvent-accessible surface with R_i parameters equal to WdV radii of corresponding atoms and a_i parameters responsible for surface smoothing[1,2].

$$D = \sum_i \exp^{-a_i(\frac{|\vec{r}-\vec{r_i}|}{R_i}-1)^2} \tag{1}$$

The Gauss map $\mathbf{S}$ of such a function is given in Eq. 2 and Gaussian K and mean H curvatures[6] can be calculated as shown in Eq. 3.

$$\mathbf{S} = \frac{1}{\|\nabla \mathrm{D}\|^3} \begin{pmatrix} d_{xx}(d_y^2+d_z^2) & -d_x d_y d_{yy} & -d_x d_z d_{zz} \\ -d_y d_x d_{xx} & d_{yy}(d_x^2+d_z^2) & -d_y d_z d_{zz} \\ -d_z d_x d_{xx} & -d_z d_y d_{yy} & d_{zz}(d_x^2+d_y^2) \end{pmatrix} \tag{2}$$

$$\begin{aligned} K &= det(\mathbf{S}) \\ H &= 1/2\, Tr(\mathbf{S}) \end{aligned} \tag{3}$$



(a) (b)

Figure 1. (a) Molecular surface of the BPTI protein. Marching cubes algorithm for the surface extraction has been used. (b) Error in solvation energy versus density of surface point for the marching cubes algorithm.

Marching algorithms have been chosen to extract the isovalue surface. We have used marching cubes and marching tetrahedra methods in our model. Difference between them will be discussed below. BPTI protein molecular surface, evaluated with marching cubes, is shown in Fig. 1(a).

3 Boundary Element Method

A well-known result for the induced charges σ on a surface between two dielectric media, which is a solution of Poisson's Equation, is given by:

$$\sigma = \left(\frac{\epsilon_{in} - \epsilon_{out}}{4\pi\epsilon_{in}}\right) \sum_i \frac{q_i(\mathbf{r} - \mathbf{r_i}) \cdot \mathbf{n}}{\epsilon_{in}|\mathbf{r} - \mathbf{r_i}|^3} + \left(\frac{\epsilon_{in} - \epsilon_{out}}{2\epsilon_{in}}\right) \sigma$$
$$+ \left(\frac{\epsilon_{in} - \epsilon_{out}}{4\pi\epsilon_{in}}\right) \oint \frac{(\mathbf{r} - \mathbf{r_s}) \cdot \mathbf{n}}{|\mathbf{r} - \mathbf{r_s}|^3} \sigma_s ds \qquad (4)$$

This equation can be solved in a matrix form $\mathbf{A}\sigma = \mathrm{b}$ using direct or iterative methods[4]. As the matrix size grows the iterative methods become more preferable. We have solved this equation with a number of modern iterative algorithms CGS, BiCGSTAB, GMRES, BiCG, QMR and Chebyshev iteration. All the algorithms are Krylov-subspace methods and provide well convergence. GMRES, unlike the other methods, needs only one matrix-vector multiplication per cycle and is a method of choice in many applications. QMR requires one matrix-vector and one transpose matrix-vector multiplications per cycle, but these two operations can be performed in a single function call in our modified DPMTA scheme[3]. So our implementation of QMR iteration is almost as fast as GMRES, but does not require a restart and uses less memory. All other algorithms, except Chebyshev, need two matrix-vector multiplications, which slow down their performance. Chebyshev iteration, although being very simple, needs some additional information about the spectrum of the matrix $\mathbf{A}$ and has been used here only for test purposes.

Since BiCG and QMR methods require a matrix transpose, we have extended DPMTA algorithm[3] to calculate the transpose of matrix $\mathbf{A}$. The multipole expansion $\mathbf{M}^{\mathrm{T}}_{\mathrm{n,m}}$ has a vector form and contains three new components:

$$\vec{\mathbf{M}}^{T}_{n,m}(\vec{r}) = \sum_{i=1}^{k} q_i s_i \vec{n}_i \mathbf{F}^{*}_{n,m}(\vec{r}_i) \tag{5}$$

Forces then calculated as:

$$q_i \nabla \Phi^{T}(\vec{r}) = Tr\left[q_i \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \vec{\mathbf{M}}^{T}_{n,m} \nabla \mathbf{G}_{n,m}(\vec{r}) \right] \tag{6}$$

Once the polarization charges σ are calculated, the solvation energy is evaluated as:

$$U = \frac{1}{\epsilon_{in}} \sum_{i} \frac{q_i}{|\mathbf{r} - \mathbf{r_i}|} + \sum_{k} \frac{\sigma_k}{|\mathbf{r} - \mathbf{r_k}|} \tag{7}$$

Finally the surface charges can be normalized according to Gauss's law:

$$\sum_{k} \sigma_k s_k = \frac{(\epsilon_{in} - \epsilon_{out})}{(\epsilon_{in} * \epsilon_{out})} \sum_{i} q_i \tag{8}$$

4 Code

The presented algorithm was written as a C/C++ extension to MMTK molecular modelling toolkit[5].

5 Numerical Tests

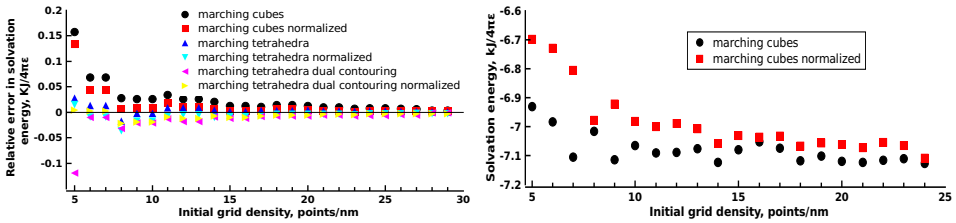
As a first example we have calculated solvation energies for a spherical cavity and compared them with analytical results given by Born equation. Results for this comparison are given in Fig. 1(b). Starting at point density $5\ points/nm$ the normalized energy error is below 1%.

Comparison for different iterative algorithms for the BPTI protein with 2500 surface elements is given in Tab. 1. GMRES and QMR methods perform the best, while Chebyshev did not converge due to a poor matrix $\mathbf{A}$ eigenvalue estimation.

	CGS	BiCGSTAB	GMRES	BiCG	QMR	Cheby
steps	24	23	30	36	37	-

Table 1. Convergence of iterative methods for the BPTI protein with point density 5 points/nm. Tolerance = 1e-6, 2500 surface elements, no preconditioner was used.

Numerical results for marching cubes, marching tetrahedra and marching tetrahedra dual contouring are given in Fig.2(a). All the methods show similar energy errors. Marching cubes provide a better mesh spacing and triangular quality, while marching tetrahedra



(a) (b)

Figure 2. (a) Three different marching schemes for a spherical cavity are compared. (b) Solvation energy for the BPTI protein at different grid point densities.

algorithm is simpler to implement. Convergence of the marching cubes scheme for the BPTI solvation energy is shown in Fig. 2(b).

6 Concluding Remarks

In the current work we have shown how to use a parametric solvation model together with BEM Poisson's solver. Three different marching schemes have been implemented and tested. We have also extended the DPMTA multipole algorithm[3] to deal with transpose matrices, which allowed us to use new iterative schemes[4]. The method has been coded as an extension to the MMTK molecular modeling toolkit[5] and will be soon available for download.

Acknowledgments

S.G. is thankful to Prof. Gerhard Gompper for the support in his postdoctoral studies.

References

1. The interpretation of protein structures: estimation of static accessibility. B Lee, FM Richards, J. Mol. Biol. **55**, 379-400, 1971.
2. Quality Meshing of Implicit Solvation Models of Biomolecular Structures. Y Zhang, G Xu and C Bajaj, Computer Aided Geometric Design **23**, 510-530, 2006.
3. DPMTA - A Distributed Implementation of the Parallel Multipole Tree Algorithm - Version 3.0. William T. Rankin `http://www.ee.duke.edu/Research/SciComp/Docs/Dpmta`.
4. A Historical Overview of Iterative Methods, D. M. Young, Comp. Phys. Comm. **53**, 1-17, 1989.
5. The molecular modeling toolkit: A new approach to molecular simulations. K. Hinsen, J. Comp. Chem **21(2)**, 79-85, 2000.
6. Modern Differential Geometry of Curves and Surfaces with Mathematica, A. Gray, 2nd ed. Boca Raton, FL: CRC Press, 1997.

Parameterization of the Potential Energy Surface of the Double Proton Transfer in Porphyrins

Marta Hallay-Suszek[1,3], **Pawel Grochowski**[2,3], **and Bogdan Lesyng**[1,3]

[1] Department of Biophysics, Institute of Experimental Physics, Warsaw University, Zwirki i Wigury 93, 02-089 Warsaw, Poland
E-mail: m.hallay@icm.edu.pl

[2] Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Pawinskiego 5A, 02-106 Warsaw, Poland

[3] CoE BioExploratorium, University of Warsaw, Zwirki i Wigury 93, 02-089 Warsaw, Poland

We present a fast and precise generator of the potential energy surface (PES) for the double proton transfer occuring in free-base porphyrin. We propose novel analytical formulae of the PES along with their parameterizations to describe correlated motions of the protons.

1 Introduction

Double proton transfer processes (DPTPs) are commonly observed, from small systems to large ones, including proteins. For a better understanding of DPTPs and coupled, correlated atomic motions, we plan to use quantum-classical molecular dynamics (QCMD) simulations. The main problem in the QCMD simulations is the lack of reliable and fast PES generators for the atomic motions. Based on effective quantum-classical models[1,2] and their numerical implementations, we decided to construct PES for model molecular systems with DPTPs. Drawing experience from our previous research on DPTP in formic acid dimer (a system with two hydrogen bonds), we focused our research on larger and experimentally well-studied molecules with the intramolecular proton transfer: porphyrin and porphycene. Free-base porphyrin and porphycene are built of four pyrrole rings connected by carbon bridges and have two hydrogen atoms (protons) in the inner part of the skeleton. The two inner protons migrate in the molecular scaffold containing four nitrogen atoms. The double proton transfer could - in principle - proceed either along a synchronous pathway or *via* a two-step mechanism involving a metastable *cis* intermediate. All the stationary states presented in Fig. 1. were optimized using GAUSSIAN 03[3] with

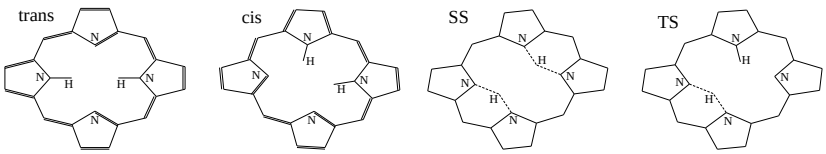


Figure 1. The stationary structural states of the porphyrin molecule obtained with the DFT/B3LYP method.

the spin-restricted B3LYP density functional method (DFT) and the 6-31G(d,p) basis set.

The obtained energy values are presented in Tab. 1, they are in good agreement with the B3LYP/TZ2P values given by Baker *et al.*[4].

structure	energy (V^G)	energy[4]	structure	energy (V^G)	energy[4]
trans (D_{2h})	0.0	0.0	cis (C_{2v})	34.3	34.8
TS (C_s)	64.0	67.8	SS (D_{2h})	94.2	102.2

Table 1. Relative energies for the stationary states of the porphyrin PES (energies are given in [kJ/mol], the *trans* porphyrin is taken as the reference state).

2 Potential Energy Functions

Our aim is to construct a PES generator for the motions of the two inner protons and vibrations of the molecular skeleton, occurring in the ground Born-Oppenheimer state. Such a generator can be parametrized based on the *ab initio* or DFT calculations for the energies and the geometries of the stationary states (including the minima and the saddle points of the PES), and the corresponding Hessians, as well as on the PES scans for proton motions inside the molecular scaffold. Several models are considered; in one of them, the PES for the protons and for the skeleton is obtained with the Approximate Valence Bond (AVB) method[1] as the lowest eigenstate of the AVB Hamiltonian; in two other approaches, the PES for the molecular skeleton is described by a classical force field or by an average harmonic potential, and a more rigorous analytical representation of the PES for the proton motions is used, allowing for better approximation of the energy scans obtained with GAUSSIAN 03.

3 Mechanical PES Model

The following analytical formulae are used to describe the PES for the inner protons:

$$V_M := \sum_H \Big(\sum_N V_{NH} + \sum_C V_{CH} + \sum_{pairsNC} V_{HNC} + \sum_{pairsN} V_{NHN} \Big) + V_{HH} + \sum_N V_{HNH} + V_{const}(R_{skeleton}),$$

where the first sum runs over the two protons and includes the following terms for each proton:

- two-body interactions with C and N atoms described by the Morse function for the bond length, CH and NH,

$$V_{CH} := P_{L1}(\exp(-P_{N1} r_{CH}) + P_{N2} \exp(-2P_{N1} r_{CH})),$$

$$V_{NH} := P_{L2}(\exp(-P_{N3} r_{NH}) + P_{N4} \exp(-2P_{N3} r_{NH})),$$

- three-body interactions with N and C atoms, ensuring proper angles between the NC and NH bonds,

$$V_{HNC} := P_{L3}\exp(-P_{N5}r_{NH}^2)\Big(\frac{\vec{r}_{NC}\cdot\vec{r}_{NH}}{|\vec{r}_{NC}||\vec{r}_{NH}|}+P_{N6}\Big)^2,$$

- three-body interactions with pairs of N atoms (computed for each pair),

$$V_{NHN} := P_{L4}\exp(-P_{N7}r_{N_1H})\exp(-P_{N7}r_{N_2H}).$$

The other terms are:

- a repulsive HH interaction,

$$V_{HH} := \frac{P_{L5}+P_{L6}\exp(-P_{N8}r_{HH}^2)}{r_{HH}},$$

- a three-body interaction between the two protons and a nitrogen atom (computed for each nitrogen),

$$V_{HNH} := P_{L7}\exp(-P_{N9}r_{NH_1})\exp(-P_{N9}r_{NH_2}).$$

3.1 Fitting the Parameters

The mechanical PES model contains a set of parameters: $\{P_{Li}\}_{i=1}^{7}$,$\{P_{Ni}\}_{i=1}^{9}$, which have to be optimized in order to fit the model to the DFT results. The optimization is based on minimization of the root mean square deviation (RMSD) function:

$$RMSD := \sqrt{\frac{\sum_{\{geom\}}\Big(\sum_{\{map\}}\mathcal{I}(V^G)(V_M-V^G)^2\Big)}{\sum_{\{geom\}}\Big(\sum_{\{map\}}\mathcal{I}(V^G)\Big)}},$$

where

$$\mathcal{I} = \begin{cases} 1 & \text{if } V^G \le V^S \\ \exp(-V^G/V^S) & \text{if } V^G > V^S \end{cases},$$

and $\{map\}$ represents the set of positions of the protons, used to scan the PES at a fixed geometry of the skeleton, $\{geom\}$ represents the set of all the geometries of the skeleton (minima and transition states) for which the scans have been performed, V^S is an energy cutoff (e.g. V^S=100 kJ/mol for porphyrin). It should be emphasized that optimization is carried out simultaneously for all geometries that represent stationary states of the molecule.

4 Results

The mechanical PES model was fitted to the DFT calculations with a satisfactory accuracy. In the most important, low energy regions, up to 100 kJ/mol, the discrepancy between the model and the DFT results does not exceed 8 kJ/mol. In order to visualize the results of the parameterization, the protonic scans for the porphyrin molecule in the *trans* minimum state are presented in Fig 2. From the left to the right are shown: the reference PES scans obtained with GAUSSIAN 03 by moving the right inner proton in the plane of the

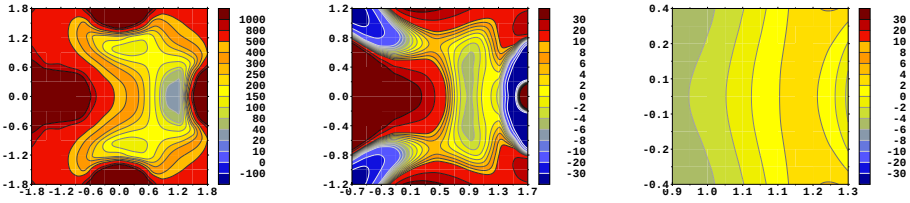


Figure 2. An *ab initio* scan of the PES (left), and the differential energy maps (middle and right) - for details, see the text.

molecule (x, y) in the region $(x, y) \in (-1.8; 1.8)$, while all other atoms were fixed in space; the differences between the protonic scans obtained with GAUSSIAN 03 and with the analytical model in the region limited to $x \in (-0.7; 1.7), y \in (-1.2; 1.2)$; and the right one to $x \in (0.9; 1.3), y \in (-0.4; 0.4)$. Note that the (x,y) coordinates of the nitrogen atoms are: (0.00;2.03), (0.00;-2.03), (2.12;0.00), (-2.12;0.00). The NH bond lengths in the local minima, the transition state and the saddle point of GAUSSIAN 03 as well as the parametrized PES (V_M) are summarized in Tab. 2. In the case of the transition states, the distances between the proton and the two closest nitrogen atoms are given.

structure	R_1	R_1^G	R_2	R_2^G
trans (D_{2h})	1.0074	1.0145	1.0074	1.0145
cis (C_{2v})	1.0204	1.0277	1.0204	1.0277
TS (C_s)	1.0225	1.0273	1.9020 ,1.2982	1.8772, 1.2833
SS (D_{2h})	1.2927 ,1.2927	1.2943,1.2943	1.2927 ,1.2927	1.2943, 1.2943

Table 2. NH bond lengths (given in Å) of the stationary states in porphyrin.

Acknowledgments

These studies were partially supported by BST funds and CoE BioExploratorium, University of Warsaw.

References

1. P. Bala, P. Grochowski, K. Nowinski, B. Lesyng and J. McCammon, *Quantum-Dynamical Picture of a Multi-Step Enzymatic Process Reaction Catalyzed by Phospholipase A2*, Biophys. J. **79**, 1253–1262, 2000.
2. P. Grochowski and B. Lesyng, *Extended Hellmann-Feynman Forces, Canonical Representations, and Exponential Propagators in the Mixed Quantum-Classical Molecular Dynamics*, J. Chem. Phys. **119**, 11541–11555, 2003.
3. M. J. Frisch *et al.* GAUSSIAN 03, Gaussian, Inc., Wallingford, CT, 2004.
4. J. Baker, P. M. Kozlowski, A. A. Jarzecki and P. Pulay, *The inner hydrogen migration in free base porphyrin*, Theor. Chem. Acc. **97**, 59–66, 1997.

A Statistical Approach to Deriving and Analyzing a Propensity Scale for Predicting Exposed Transmembrane Beta Barrel Residues from Protein Sequence

Sikander Hayat, Yungki Park, and Volkhard Helms

Center for Bioinformatics, Saarland University, 66041 Saarbrücken, Germany
E-mail: {s.hayat, volkhard.helms}@bioinformatik.uni-saarland.de

In the current study, we implement an algorithm to analytically derive a novel propensity scale for the Transmembrane Beta barrel (TMB) residues to be exposed to the lipid bilayer. Since it is very difficult to experimentally determine their 3D structures and given the fact that they perform several important functions in the cell proteome of both gram-negative bacteria and eukaryotes, it is imperative to develop *in silico* methods for the modeling of their 3D structure. The algorithm previously described by us for Transmembrane Alpha helical proteins takes into account the evolutionary conservation and frequency profile to derive a positional score for a given Transmembrane residue. The scale, based on ridge regression, is derived such that the positional score for a given residue is maximally correlated with its relative solvent-accessible surface area (rSASA) value. A leave-one-out test with known structures demonstrates the correlation coefficient between the observed and predicted rSASA values to be around 0.52. Analysis of scales derived for both the interface and the hydrophobic core of the TMBs provides interesting insights into structural aspects of TMB residues.

Keywords: ridge regression, TM Beta Barrel proteins, frequency profile, propensity scale

1 Introduction

TMBs, composed of antiparallel Transmembrane beta strands and connected by soluble loop regions are inserted into the outer membrane (OM) of gram-negative bacteria, mitochondria and chloroplasts, where they perform a variety of functions, including passive transport of ions and small hydrophilic molecules, membrane anchoring and a role in bacterial pathogenicity[1–4]. The lipid-facing surfaces of the barrels are composed of hydrophobic residues and the residues facing the interior of the barrel are mostly polar residues[5] to facilitate transfer of small solute molecules. Accurate structure prediction of TMBs still remains a challenge, as Transmembrane protein structure can not be easily studied using X-ray crystallographic and NMR methods. As of now, no propensity scale exists that can account for the propensities of TMB residues to be exposed to the lipid bilayer. In this study, we implement an algorithm previously established by us[7] for deriving a propensity scale for Transmembrane Alpha helical residues to formulate two propensity scales for the Beta barrel residues to be exposed to the lipid bilayer at the hydrophobic core and at the interface region, respectively. Briefly, the algorithm tries to maximize the correlation coefficient between the observed and the predicted rSASA values for a given residue. The rSASA value of a residue describes the extent to which that particular residue is exposed to the lipid bilayer. Previously, it has been observed that polar residues tend to be buried inside and hence are less exposed to the bilayer in the hydrophobic core region[5]. The same is not true for residues at the interfaces, where the distinction between the exposure patterns is influenced by the changing nature of the lipid environment. It is also known that

the more conserved a residue is, the more it is structurally and/or functionally important for the protein. Consequently, the frequency profile of the 20 naturally occurring amino acids and their positional conservation Index is used as the input vector.

2 Materials and Methods

A non redundant data set of known TMB structures was compiled primarily based on the list provided by Tusnady *et al.*[6]. The final dataset comprises of 25 protein chains with 2305 and 1195 TM residues in the hydrophobic core and interface region, respectively. The hydrophobic region was derived from the OPM database[10]. The classification of each residue as being buried or exposed to the lipid bilayer was based on its rSASA value. As previously described by us[7], the SASA values were calculated with the VOLBL program suite employing a probe radius of 2.2 Å. To prevent the residues lining the internal cavity of the protein from being classified as exposed, capping with dummy residues was performed for both the interfaces and the hydrophobic core. SASA values were normalized to generate rSASA values by considering the SASA values for each amino acid X in the context of the tri-peptide G-X-G. The positional frequency for a given residue was calculated using AL2CO program suite. The complexity parameter for ridge regression was found by employing 10 fold cross validation. Calculations performed using R yielded 0.691 and 0.301 as complexity parameters for hydrophobic and interface regions, respectively.

3 Results and Discussion

3.1 Implementation of an Optimum Propensity Scale for TMBs

	PHE	MET	TRP	ILE	VAL	LEU	ALA	PRO	ASP	GLU
HMOB	0.086	-0.086	0.030	0.023	0.077	0.246	0.056	0.068	-0.062	-0.055
IMOB	0.096	-0.040	0.094	0.044	0.019	0.100	-0.017	0.030	-0.057	-0.059
MO	-0.010	-0.230	-0.030	0.050	0.020	0.020	-0.090	-0.100	-0.270	-0.200
	CYS	ASN	GLN	THR	TYR	SER	GLY	HIS	ARG	LYS
HMOB	0.047	-0.061	-0.060	-0.035	-0.020	-0.075	-0.026	-0.026	-0.062	-0.063
IMOB	-0.050	-0.043	-0.031	-0.006	0.078	-0.063	-0.038	0.020	-0.038	-0.041
MO	-0.160	-0.230	-0.220	-0.180	-0.150	-0.190	-0.180	-0.240	-0.210	-0.100

Table 1. The propensity scales for transmembrane Beta-Barrel proteins. HMOB = Hydrophobic MO Beta, IMOB = Interface MO Beta, MO = MO scale for transmembrane alpha helical proteins. The standard deviation of individual propensity values with leave-one-out test was found to be less than 0.008 in all cases.

As can be seen in Table 1, the propensity values for TMB residues to be exposed to the bilayer are smaller in magnitude than their counterparts in Transmembrane Alpha helices. This could be due to the less hydrophobic exterior of TMBs, which is necessary for their translocation via the inner membrane[8]. The affinity for hydrophobic residues to be exposed to the hydrophobic environment of the lipid bilayer is in concert with the experimental results[4].

	0.00^a $(0.44)^b$	0.01 (0.37)	0.02 (0.35)	0.03 (0.34)	0.04 (0.33)	0.05 (0.33)
HMOB CC(0.51)	67.2	73.1	75.0	75.8	76.0	77.4
+CI CC(0.52)	69.9	76.3	78.3	78.7	79.1	79.7
	0.00^a $(0.49)^b$	0.01 (0.45)	0.02 (0.44)	0.03 (0.42)	0.04 (0.41)	0.05 (0.39)
IMOB CC(0.48)	76.6	76.6	76.0	76.6	77.0	77.1
+CI CC(0.49)	77.2	78.3	77.7	78.0	78.8	78.7

Table 2. Performance comparison of the derived scales at different rSASA values in the hydrophobic core region. Entries reflect the accuracy of prediction, which improves with the inclusion of conservation index as an input parameter. a = Threshold rSASA, b = fraction of exposed residues, CC =Absolute magnitude of Correlation coefficient between the observed rSASA and the computed positional score, ACC = Accuracy of correct prediction in percentage, CI = Conservation index.

3.2 Performance Comparison of Propensity Scales

A leave one out test was conducted to measure the performance of the propensity scales. The performance of the particular scale was assessed by analyzing the correlation between positional score and the observed rSASA values and by the corresponding prediction accuracy. A support vector machine in R was used to implement the test. As depicted in Table 2, the choice of the cutoff value for the rSASA value is a major factor when it comes to predicting the burial status. In the current study, this cutoff was objectively chosen based on the SVM. The table also shows that inclusion of conservation indices as an input parameter enhances the performance of the derived scales. The overall weak correlation between the observed rSASA and the computed positional score suggests that the lipid protein interaction might not be the only factor involved in protein insertion and folding mechanism[8].

3.3 Comparative Analysis and Correlation with Other Scales

	Scale (Reference)	HMOB	IMOB
Hydrophobicity	KD (Kyte *et al.*, 1982)	0.68	0.46
	EIS (Eisenberg *et al.*, 1984)	0.63	0.61
	GES (Engelman *et al.*, 1986)	0.54	0.51
	WW (Wimbley *et al.*, 1996)	0.56	-
	Hessa (Hessa *et al.*, 2005)	-0.58	-0.52
Size	Bulkiness (Zimmerman *et al.*, 1968)	0.53	0.74
Packing	Partial specific volume (Cohn *et al.*, 1943)	0.60	0.59
Others	KPROT (Pilpel *et al.*, 1999)	0.61	0.30

Table 3. Correlational analysis with other scales. The propensity scales derived here shows weak correlation with other scales.

As shown in Table 3, one method to discover the other factors involved in TMB folding could be to find other scales that strongly correlate with the scales derived here. As expected, the hydrophobicity scales show a weak correlation with the propensity scales derived here, which can be atributed to the less hydrophobic exterior of TMBs.

4 Conclusion

The current study successfully implements the MO algorithm for Transmembrane Beta barrel proteins. The propensity scales for both the hydrophobic core and the interface region are presented. The derived scales confirm the less hydrophobic exterior of the TMBs and are weakly correlated with hydrophobicity scales. A further analysis of the scale in terms of principal components, distance of the residue from the lipid bilayer core and more advanced statistical methods need to be employed to fully understand the insertion and the folding mechanism of the TMBs in an analytical way. Development of a reliable predictor for the burial status of TMBs can be used to impose additional constraints on starting template models while performing *ab initio* structure prediction[9]. Further analysis of TMB residue propensities might provide important insights into the evolution of the mitochondrial OM and the development of its protein biogenesis system[8].

References

1. R. Koebnik, K. P. Locher, and Van Gelder P., *Structure and function of bacterial outer membrane proteins: barrels in a nutshell*, Mol. Microbiol. **60**, 1-4, 2000.
2. G. E. Schulz, *Porins: general to specific, native to engineered passive pores*, Curr. Opin. Struct. Biol. Mol. **6**, 485-490, 1996.
3. A. Pautsch, and G. E. Schulz, *Structure of the outer membrane protein A transmembrane domain*, Nat. Struct. Biol. **5**, 1013-1017, 1998.
4. G. E. Schulz, *The structure of bacterial outer membrane proteins*, Biochim. Biophys. Acta. **1565**, 308-317, 2002.
5. L. K. Tamm, H. Hong, and B. Liang, *Folding and assemble of beta barrel membrane proteins*, Biochim. Biophys. Acta. **1666**, 250-263, 2004.
6. `http://pdbtm.enzim.hu/` G. E. Tusnady, Z. Dosztanyi, and I. Simon, *Transmembrane proteins in the Protein Data Bank: identification and classification*, Bioinf. **20**, 2964-2972, 2004.
7. Y. Park, and V. Helms, *On the Derivation of Propensity Scales for Predicting Exposed Transmembrane Residues of Helical Membrane Proteins*, Bioinf. **23**, 701-708, 2007.
8. N. Pfanner, N. Wiedemann, C. Meisinger, and T. Lithgow, *Assembling the mitochondrial outer membrane*, Nat. Struct. and Mol. Biol. **11**, 1044-1048, 2004.
9. L. Adamian, and J. Liang, *Prediction of transmembrane helix orientation in polytopic membrane proteins*, BMC Struct. Biol. **6**, 2006.
10. M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, *OPM: Orientations of Proteins in Membranes database*, Bioinf. **22**, 623-625, 2006.

Simulation of Small Peptide Using Combined Wang-Landau-Transition Matrix Monte Carlo Algorithm

Sh. Hayryan[1], **R. Ghulghazaryan**[1,2], **N. Poklar Ulrih**[3], **V. Čurin Šerbec**[4], **and Chin-Kun Hu**[5]

[1] Institute of Physics, Academia Sinica,
Nankang 11529, Taipei, Taiwan
E-mail: {shura, huck}@phys.sinica.edu.tw

[2] Yerevan Physics Institute, Alikhanian Brothers 2, 375036 Yerevan, Armenia
E-mail: ghulr@yahoo.com

[3] Department of Food Science and Technology, Biotechnical Faculty,
University of Ljubljana, Jamnikarjeva 101, 1000 Ljubljana, Slovenia

[4] Blood Transfusion Center of Slovenia,
Šlajmerjeva 6, School of Medicine, 1000 Ljubljana, Slovenia

[5] Department of Physics, Chung-Yuan Christian University,
Chungli 320, Taiwan

We applied our recently suggested modification of the Wang-Landau algorithm to study the folding thermodynamics of 13-residue peptide P1 from C-terminal α-helix part of the human Prion Protein (PrP). Temperature dependencies of the average energy, the specific heat, the relative free energy and entropy have been evaluated for the single peptide in gas phase as well as in the implicit water environment. Simulations of single molecule show that this peptide is a good folder and folds into a fully helical conformation. This is consistent with the conformation of this part in the native structure of the mother protein. However, CD spectrum of the water solution of P1 peptide shows only very small amount of α-helix. This may indicate that in water solution the peptide forms aggregates. We used the simulated annealing algorithm to determine the ground state of the system of two peptides. The results show that the ground state conformation indeed contains a small amount of α-helix.

1 Introduction

The Monte Carlo method has been an effective tool in protein simulations for a long time. Due to the very complicated energy landscape of the proteins with high potential barriers, the conventional Metropolis algorithm does not allow to sample the energy space more or less uniformly, especially in simulations with all-atom potentials. The situation has improved dramatically when several non-conventional (so called generalized-ensemble) algorithms have been introduced[1–4]. However, even the improved algorithms do not meet all challenges in protein simulations, and there is a constant quest for new, more sophisticated and efficient algorithms. Recently we have suggested a new approach which is based on the combination of the Wang-Landau and Transition Matrix Monte Carlo methods[5]. It is well known that the Wang-Landau algorithm allows to easily jump over the energy barriers at the initial stage of simulation and to sample all energy levels uniformly. Nonetheless, at the later stages the statistical errors do not decrease with increasing samples. This is considered as a drawback of the algorithm. At the same time, the transition matrix method

gives a good estimate of the density of states with small statistical errors but it requires a proper initial guess for the density of states which is usually not available. In our algorithm we use the advantages of the two methods within a combined approach whereby the algorithm behaves like the Wang-Landau method in initial stages and uses the transition matrix method to reduce the statistical errors at the later stage. Here we apply our algorithm to simulate a small peptide P1 from the C-terminus of human prion protein (PrP). The biological importance of this peptide is that it has been used as an antigen to produce a monoclonal antibody which recognizes an epitope specific to pathological isoform of PrP from brain samples of Creutzfeld-Jakob disease patients. As a part of the mother protein in the native state this peptide has a fully α-helical structure whereas the CD spectrum of the water solution shows only a tiny amount of α-helix. We use the standard geometry model with ECEPP/3 forcefield as implemented in protein simulation package SMMP[6].

2 Calculation of the Thermodynamical Quantities

The advantage of knowing the density of states is that one can calculate directly the partition function and all related quantities

$$Z = \sum_E n(E) e^{-\beta E}, \tag{1}$$

where Z is the partition function and $n(E)$ are the density of states. We performed our algorithm to calculate the partition function of P1 peptide both in gas phase and with implicit water environment treated by solvent accessible area method. After calculating the partition function we obtained the temperature dependencies of the potential energy and the specific heat:

$$\begin{aligned} E(T) = \langle E \rangle_T &= Z^{-1} \sum_E E n(E) e^{E/k_B T} \\ C(T) &= \frac{\left\langle E^2 \right\rangle_T - \langle E \rangle_T^2}{k_B T^2}. \end{aligned} \tag{2}$$

These quantities are plotted on Fig. (1). One can see that the specific heat has a well defined maximum indicating cooperative transition.

3 Modification of the SMMP Code

Since the experiment shows a little amount of the α-helix the suggestion was made that the molecules of P1 in solution don't behave similar to the single molecule. The most straightforward explanation may be that they aggregate and do not allow each other to undergo helix-coil transition. We have addressed this problem by finding the ground state conformation of the system of two peptides. For this purpose a minor modification of the code was necessary because normally it is impossible to simulate more than one molecule in SMMP. We introduced a "fake" residue which consists only of the backbone bonds and contains atoms which do not interact neither with each other nor with other atoms. Thus they don't contribute into the energy function. The rotations around all bonds of

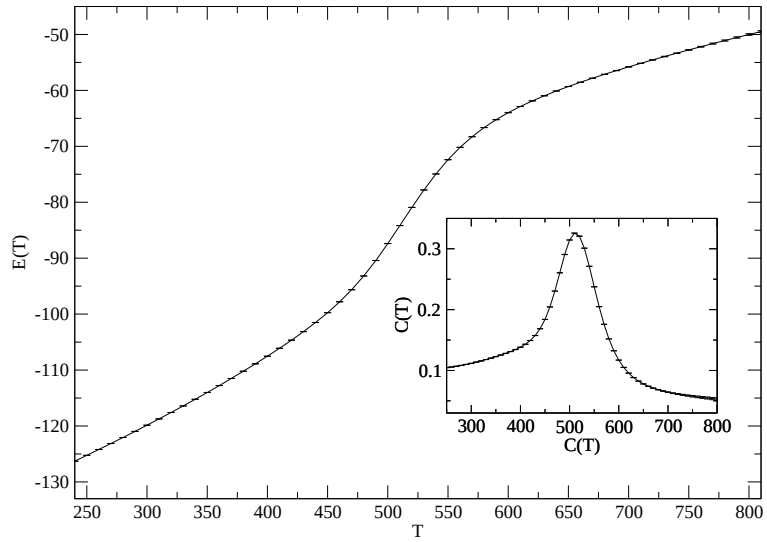


Figure 1. The temperature dependence of the potential energy and the specific heat (inlet) of the peptide P1 in implicit water environment.

virtual residues are completely free. Connecting two peptides with sufficiently long chain of virtual residues allows the two molecules to accept any positions with respect to each other. We have performed a large number of simulated annealing for such a system and obtained that in more that 90% cases the ground state of the system contains only very small amount of α-helix. In all ground states the two peptides are positioned parallel or antiparallel, creating intermolecular hydrogen bonds instead of itramolecular ones.

Acknowledgments

This work was suported by the National Science Council of Taiwan under the grant No. NSC 95-2112-M 001-008, National Center for Theoretical Sciences in Taiwan, and Academia Sinica (Taiwan) under Grant Nos. AS-92-TP-A09 and AS-095-TP-A07.

References

1. U. H. E. Hansmann, Y. Okamoto, J Comp Chem **14**, 1333 (1993).
2. Y. Okamoto, U. H. E. Hansmann, J Phys Chem **99**, 11276 (1995).
3. A. Berg, U. H. E. Hansmann, Eur Phys J B **6**, 395(1998).
4. F. G. Wang, D. P. Landau, Phys Rev Lett **86**, 2050 (2001).
5. R. Ghulghazaryan, Sh. Hayryan, C.-K. Hu, J Comp Chem **28**, 715 (2007).
6. F. Eisenmenger, U. H. E. Hansmann, Sh. Hayryan, C.-K. Hu, Comp Phys Comm **138**, 192 (2001).

Conformational Studies of UDP-GlcNAc in Environments of Increasing Complexity

Martin Held[1], **Eike Meerbach**[1,3], **Stephan Hinderlich**[2], **Werner Reutter**[2], **and Christof Schütte**[1]

[1] Fachbereich Mathematik und Informatik,
Freie Universität Berlin
Arnimallee 2-6, 14195 Berlin
E-mail: {held, meerbach, schuette}@mi.fu-berlin.de

[2] Institut für Biochemie und Molekularbiologie
Campus Benjamin Franklin
Charité Universitätsmedizin Berlin
Arnimallee 22, 14195 Berlin
E-mail: {Stephan.Hinderlich, Werner.Reutter}@Charite.de

[3] Supported by the DFG research center "Mathematics for key technologies".

The "effective" dynamics of a biomolecular system can often be described by means of a Markov chain describing "flipping dynamics" between metastable (geometrical large scale) molecule conformations on "long" time scales. In this work we present our methods for the identification of metastable conformations by a study of UDP-GlcNAc - a key substrate in sialic acid synthesis. We investigate the system in different environments of increasing complexity (vacuum, water, peptide) by performing force field based molecular dynamic simulations. By applying our analysis techniques to the obtained data we get insights into the conformational dynamics of the system. A comparison of the results reveals interesting environment-dependent properties of UDP-GlcNAc.

1 Introduction

Sialic acids play an important role in various biological processes, e.g., immune response, tumor metastasis or inflammatory reactions. In mammalian cells the key reaction of the sialic acid synthesis pathway is carried out by the bifunctiononal enzyme UDP-GlcNAc-2-Epimerase/ManNAc-Kinase[1]. In cooperation with researchers at the Charité we investigate UDP-N-Acetylglucosamine (UDP-GlcNAc) (see Fig. 1), the ligand of this key reaction by means of molecular dynamic simulations. Thereby, we expect to gain a better understanding of the actual reaction and, in the long term, insights that assist in the design of potential inhibitors.

2 Simulation

By a systematic conformational search in dihedral space we generated 20 initial conformers of UDP-GlcNAc. For each conformer we run molecular dynamic simulations *in vacuo* and in explicit water using AMBER 8. Additionally, we simulated a set of possible binding conformations of UDP-GlcNAc. To derive these, crystallographic data of UDP-GlcNAc-2-Epimerase in complex with UDP was utilized. To compensate for the missing information

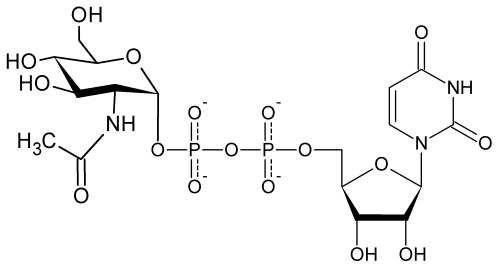


Figure 1. UDP-N-acetylglucosamine

of the GlcNAc orientation, we generated conformers of UDP-GlcNAc by systematically turning the three dihedrals that qualify the orientation of the GlcNAc moiety . The force field parameters needed for this non-standard residue were obtained from the literature[12, 14].

	Atoms	Sim. Time	Step Size	Temperature	Pressure
In Vacuo	64	100 ns (2 μs)	1 fs	300 K	-
Water	3964	10 ns (200 ns)	2 fs	300 K	1 atm
Protein	46646	100 ps	2 fs	300 K	1 atm

Table 1. Simulation Parameters

3 Analysis Methods

Our aim is to identify so-called *metastable* conformations of UDP-GlcNAc, i.e., geometries which are persistent for long periods of time, from the generated trajectory data. Such identification corresponds to the set up of a coarse-grained model for the dynamics, as on long time scales the typical dynamics of biomolecular systems can often be described as a (Markovian) flipping process between such conformations[6, 8, 13]. The challenge of identification lies in the rich temporal multiscale structure of the data, induced by flexibility within the metastable conformations.

The groups of P. Deuflhard and C. Schütte developed the so-called *Perron Cluster Cluster Analysis* (PCCA) for the identification of metastable conformations[2, 3, 5]. The approach is based on assuming a Markov dynamics of the observed system on a certain time scale. Discretization of the state space, e.g., the dihedral angle space of some molecule, allows to set up a stochastic matrix, w.r.t. the chosen time scale, by counting transitions between discrete states in the trajectory. The information contained in the spectrum of this matrix allows to aggregate parts of the state space to metastable states, as the number of metastable states corresponds to the number of eigenvalues close to unity, while information about an appropriate clustering is encoded in the eigenvectors.

In addition a variety of methods based on Hidden Markov Models (HMMs) were recently developed in the group of C. Schütte[7, 9–11]. An intriguing feature of HMM analysis is that the chosen set of observables need not to enable a geometric separation of the

metastable conformations. Instead, the *observed* time series is used to fit a model for an *unobserved* Markov chain, by extracting geometric as well as dynamical properties out of the given observation sequences, and thus allowing for separation of overlapping conformations.

For our analysis we used a combination of these approaches. First each chosen torsion angle is analyzed separately by using an HMM with gaussian output distributions, resulting in a discrete time series corresponding to the hidden states for each angle. Then the obtained information is aggregated by superposition of the discrete one dimensional time series. The (combined) discrete time series is then further analyzed by means of PCCA. For more details refer to, e.g.,[11].

4 Results

Based on the *in vacuo* simulation data we identified 5 metastable conformations via an analysis on the dihedral space. All found conformations are stabilized by intramolecular hydrogen bonds, where the most compact structure corresponds to the metastable set with the highest weight (66.6%). In the water trajectory data we found 6 metastable sets. Here, the related structures appear to be less stiff, as the effect of inner hydrogen bonds is weakened, due to the shielding effect of the present water molecules.

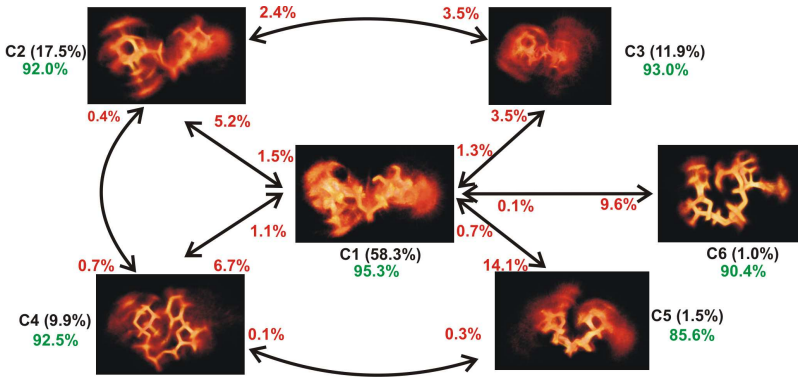


Figure 2. Transition network of UDP-GlcNAc in explicit water with density plots showing the flexibility within each metastable set. (in brackets – relative weighting, below brackets – persistence probability, next to pictures – exit probability) – Visualizations using Amira[15]

Screening the water simulation data for the generated potential protein binding conformations of UDP-GlcNAc did not result in any match. This leads to the conjecture that the binding of the ligand must be accompanied by an induced-fit effect. The protein has to alter the conformation of the ligand in order to bind it. In contrast to a conformational selection scenario, where an available conformation would be ”selected” to bind.

References

1. S. Hinderlich, R. Stäsche, R. Zeitler, and W. Reutter A bifunctional enzyme catalyzes the first two steps in N-acetylneuraminic acid biosynthesis of rat liver. Purification and characterization of UDP-N-acetylglucosamine2-epimerase/ N-acetylmannosamine kinase. *J Biol Chem 272*, pages 24313-24318 (1997).
2. F. Cordes, M. Weber, and J. Schmidt-Ehrenberg. Metastable conformations via successive Perron cluster analysis of dihedrals. ZIB-Report 02-40, Zuse Institute Berlin, 2002.
3. P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.
4. P. Deuflhard and Ch. Schütte. Molecular conformation dynamics and computational drug design. In J. M. Hill and R. Moore, editors, *Applied Mathematics Entering the 21st Century*, pages 91–119. SIAM, 2004.
5. P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. ZIB-Report 03-19, Zuse Institute Berlin, 2003.
6. R. Elber and M. Karplus. Multiple conformational states of proteins: A molecular dynamics analysis of Myoglobin. *Science*, 235:318–321, 1987.
7. A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and Ch. Schütte Identification of Biomolecular Conformations from Incomplete Torsion Angle Observations by Hidden Markov Models *J. Comp. Chem.*, accepted (2006).
8. C. Schütte and W. Huisinga. Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis X*, Special Volume Computational Chemistry, pages 699–744, 2003.
9. I. Horenko, E. Dittmer, A. Fischer, and Ch. Schütte Automated Model Reduction for Complex Systems exhibiting Metastability, *Mult. Mod. Sim. 5 (3)*, pages 475–491 (2006).
10. I. Horenko, E. Dittmer, F. Lankas, J. Maddocks, Ph. Metzner, and Ch. Schütte Macroscopic Dynamics of Complex Metastable Systems: Theory, Algorithms, and Application to B-DNA Submitted to *J. Appl. Dyn. Syst.* (2005).
11. E. Meerbach, E. Dittmer, I. Horenko, and Ch. Schütte. *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology*, volume 1 of *Lecture Notes in Physics*, chapter Multiscale Modelling in Molecular Dynamics: Biomolecular Conformations as Metastable States, pages 475–497. Springer, 2006.
12. P. Petrova, J. Koca, and A. Imberty. Potential energy hypersurfaces of nucleotide sugars: Ab initio calculations, force field parametrization and exploration of the flexibility. *J. Am. Chem. Soc.*, 121:5535–5547, 1999.
13. C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.
14. Woods Group. GLYCAM Web . 15 January 2007. Complex Carbohydrate Research Center , The University of Georgia , Athens , GA. `http://www.glycam.com`.
15. Amira—advanced visualization, data analysis and geometry reconstruction, user's guide and reference manual. Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Mercury Computer Systems GmbH and TGS Template Graphics Software Inc., 2000.

Photosensory Proteins as a Tool in Synthetic Biology: Bridging Computational Biophysics and Systems Biology

Johnny Hendriks[1], **Marcela Avila Perez**[1], **Frank Bruggeman**[2], **and Klaas J. Hellingwerf**[1]

[1] Laboratory for Microbiology, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, The Netherlands
E-mail: khelling@science.uva.nl

[2] Molecular Cell Physiology, Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1085, NL-1081 HV Amsterdam, The Netherlands

1 Introduction

The systems biology approach to understand basic questions related to health, food and the environment can be applied using different levels of molecular detail. Often it is claimed that systems-biologal descriptions are based on 'first principles'. Actually it will still take several decades before such an approach may become feasible (that is if 'first principles' is assumed to have its regular meaning in chemistry). Currently, at best, the most detailed elements of system descriptions are in terms of enzyme-kinetic characteristics (i.e. K_M and V_{max}). This approach, useful as it may be, is poorly suited for use in a synthetic biology approach. The latter would ideally integrate molecular dynamics modeling and molecular engineering in a systems description. Inclusion of this level of detail is important for applications in molecular medicine, e.g. when new enzyme inhibitors are to be designed.

When it comes to analyzing functional sub-states during enzymatic activity, photosensory proteins are the "star actors of modern times"[1]. For important classes of photoreceptors both experimental measurements and molecular dynamics simulations of both the receptor state and the functional signaling state are available and mutually reinforcing. 'Proof of principle' for this approach has been obtained with e.g. Photoactive Yellow Protein[2,3]. The predicted structure of the signaling state of this protein, calculated with molecular dynamics simulations, including the parallel tempering approach, turned out to correlate well with the results obtained with NMR measurements of the solution structure of this (sub) state. In ongoing work, transition-path sampling is used to analyze the key factors that govern the transitions between the relevant sub-states of this photoreceptor.

Light is a very convenient substrate, not only in photoreceptor studies, but also in experimental work on cellular biological systems. It is therefore surprising that not more systems biology modeling work has been done on biological systems containing photosensory proteins. As a result very little progress has been made in method development to incorporate light in a systems biology approach. Here we have developed a method based on the Lambert-Beer law. To test this method we used light-regulated gene expression via the BLUF-domain of the AppA protein from *Rhodobacter sphaeroides*. The AppA protein is involved in the blue-light repression of photosynthesis genes[4–6] . *In vitro* experimentation has shown the photocycle of AppA can be summarized in a simple photocycle scheme with only 2 main components[7,8], the receptor state R and the signaling state S (see Fig. 1a).

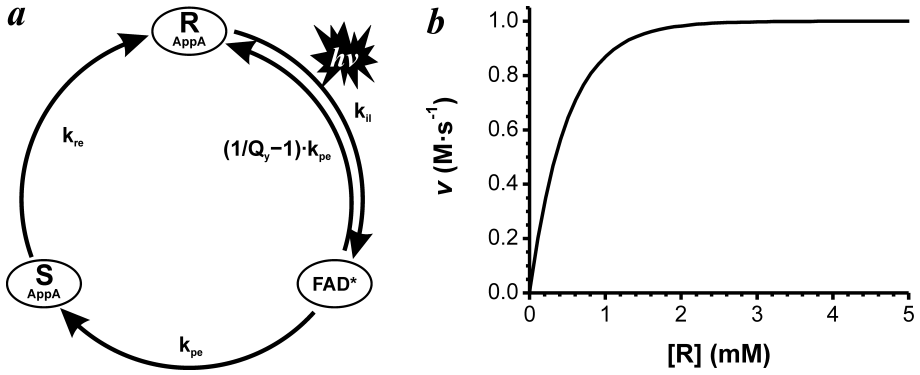


Figure 1. Panel a: Representation of the AppA photocycle model, with R the receptor state, FAD* the excited state upon light absorption, and S the signaling state. Panel b: Rate of excited state formation as function of the R state concentration. This plot was generated using equation 3, where $V = 1$ cm^3; $l = 1$ cm; $I_0 = 1$ mmol·cm^{-2}·s^{-1}; c is the concentration of R in M. Note that [R] is plotted in mM and that k_{il} ranges from $2{\cdot}10^3$ to $2{\cdot}10^2$ s^{-1} for [R] is 0.01 to 5 mM.

The signaling state is formed on a nanosecond timescale, whereas the recovery from S to R takes place on a timescale of minutes.

2 Methods

2.1 Kinetic Model of the AppA Photocycle

The model depicted in Fig. 1a was used to test the incorporation of light in the AppA model. Here k_{pe} is the photocycle entry rate or the rate with which the signaling state is formed from the excited state. Q_y is the photocycle quantum yield or the fraction of excited states that go on to form the signaling state. k_{re} is the recovery rate or the rate with which the receptor state is formed from the signaling state. k_{il} is the illumination rate constant. This rate constant is dependent on both the light intensity and the concentration of the sample. Derivation of this rate equation is shown below. For the other reactions first-order rates were assumed, in accordance with the results of the time-resolved spectroscopy experiments. Calculations with the differential equations that describe the model depicted in Fig. 1a, were performed in MatLab 7.1.

2.2 Derivation of Illumination Rate Constant (k_{il})

For the derivation of the illumination rate constant we assumed that for each photon that is absorbed, a receptor state protein is converted to the excited state. By subtracting the intensity of light used to illuminate the sample (I_0) by the intensity of light after it has passed through the sample (I) we can determine the amount of light that was absorbed. By choosing the unit of light intensity as mmol·cm^{-2}·s^{-1} (= 10^{-7} μmol·m^{-2}·s^{-1}) dividing it by the illuminated sample volume (V) and subsequently multiplying it with the area (a) of the sample that is illuminated, the velocity (v) is obtained with which the excited state of the

photoactive protein is formed. From this velocity an illumination rate constant (k_{il}) can be determined.

$$A = \log\frac{I_0}{I} = \varepsilon \cdot l \cdot c \tag{1}$$

$$(I_0 - I) = I_0 - \frac{I_0}{10^{\varepsilon \cdot l \cdot c}} = I_0 \cdot \left(1 - \frac{1}{10^{\varepsilon \cdot l \cdot c}}\right) \tag{2}$$

$$v = c \cdot k_{il} = \frac{a}{V} \cdot I_0 \cdot \left(1 - \frac{1}{10^{\varepsilon \cdot l \cdot c}}\right) \tag{3}$$

$$k_{il} = \frac{I_0 \cdot \left(1 - \frac{1}{10^{\varepsilon \cdot l \cdot c}}\right)}{l \cdot c} \tag{4}$$

Equations 1-4 depict the different steps of the derivation, where A is absorption; I_0 and I are light intensities before and after absorption by sample (mmol·cm^{-2}·s^{-1}); c is the sample concentration (M); ε is the Molar extinction coefficient of the sample (M^{-1}·cm^{-1}); l is the path length of the light through the sample (cm); a is the area of the sample that is illuminated; V is the volume of sample illuminated ($V = a \cdot l$; cm^3); v is the velocity with the excited state of the sample is formed (M·s^{-1}); k_{il} is the illumination rate constant (s^{-1}).

3 Results and Discussion

The k_{il} is linearly dependant on light intensity, however the dependence on the sample concentration is a little more complex as is illustrated in Fig. 1b, where rate of excited state formation v ($c \cdot k_{il}$) is plotted as a function of sample concentration. Note that v varies in the concentration range of 0 to 2 mM which is equivalent to a sample OD of 0 to 17 of the flavin absorption peak at 446 nm. This falls not only within the biologically relevant range but also within the range in which *in vitro* measurements are generally performed.

The model depicted in Fig. 1a was tested on data where the N-terminal 5-125 AppA fragment was placed in a spectrophotometer and bleaching of the sample by the spectrophotometers probe light was followed in time. The model was able to accurately describe the data. The values for k_{re} and I_0 were fitted to the data and, within the error of the measurements, resulted in the same values as those that were measured (Data not shown).

Having established that the model can accurately describe the photocycle of AppA, it was used to simulate the effect of k_{re} on the light sensitivity of the protein. In Fig. 2 a simulated light titration of the N-terminal 5-125 AppA fragment and the W104A mutant of the same protein fragment is depicted. The major difference here is the value for k_{re} which is almost 200 times faster for the W104A mutant[9]. The graph clearly shows that in the case of the W104A mutant much more light is needed in order to obtain similar amounts of the signaling state. This has implications for studies on the biological function of such a mutant, as one might mistakenly come to the conclusion that the mutant has lost its function (e.g. by analyzing its function only at 100 μmol·m^{-2}·s^{-1}), whereas in reality the used light intensity was insufficient to activate the mutant.

The essential part of this photocycle model has been incorporated in an overall description of light-mediated regulation of gene expression via AppA (see Fig. 3). In ongoing experiments we are analyzing the light intensity dependence of *puf* expression in intact

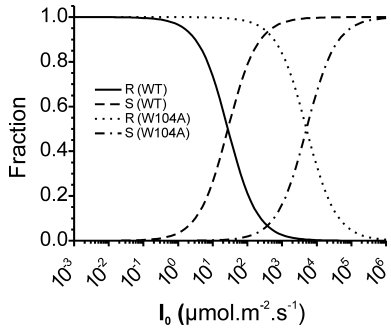


Figure 2. Simulation of the light sensitivity of AppA. Fractions of the R and S state of AppA at equilibrium are plotted as function of the used light intensity. This plot was generated with the photocycle model depicted in Fig. 1a, with a = 1 cm^2; l = 1 cm; Q_y = 0.24; k_{pe} = 10^9 s^{-1}; $\varepsilon_{400\text{-}700nm}$ = 884 M^{-1}·cm^{-1}; [AppA] = 60 μM. I_0 was varied from 10^{-3} to 10^6 μmol·m^{-2}·s^{-1}. To simulate the N-terminal 5-125 AppA Wild type fragment a k_{re} of 0.0013 s^{-1} was used. For the W104A mutant a k_{re} of 0.24 s^{-1} was used.

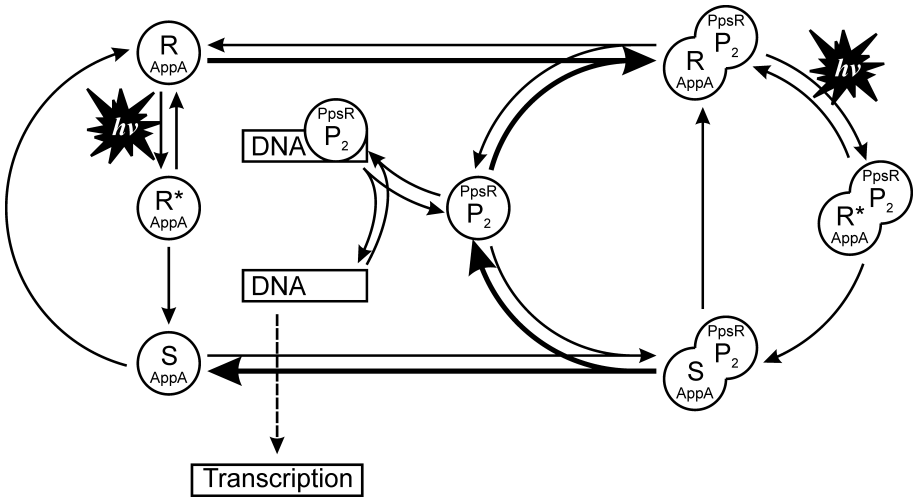


Figure 3. Model for light-mediated regulation of gene expression via AppA.

cells of *Rhodobacter sphaeroides*. In this approach we will compare the wild type protein with several of its site-directed mutants (including the W104F mutant).

We are also setting up a similar model for the YtvA mediated blue-light induced activation of the σ^B-dependent stress response of *Bacillus subtilis*[10]. This system makes use of very large ($>$ 10 MDa) signaling complexes (so-called 'stressosomes'), which generates very interesting questions regarding sensitivity, cooperativity, etc., in this signal transfer pathway. Furthermore, the light-sensitivity of YtvA provides an ideal handle to analyze the linkage of the signal transduction network that activates σ^B, with other (late growth) signal transduction networks in this organism.

References

1. M. A. Van der Horst and K. J. Hellingwerf, Accounts Chem. Res., **37**, 13-20, (2004).
2. C. Bernard, K. Houben, N. M. Derix, D. Marks, M. A. van der Horst, K. J. Hellingwerf, R. Boelens, R. Kaptein and N. A. J. van Nuland, Structure, **13**, 953-962, (2005).
3. J. Vreede, W. Crielaard, K. J. Hellingwerf and P. G. Bolhuis, Biophys. J., **88**, 3525-3535, (2005).
4. S. Braatsch, M. Gomelsky, S. Kuphal and G. Klug, Mol. Microbiol., **45**, 827-836, (2002).
5. M. Gomelsky and S. Kaplan, J. Bacteriol., **179**, 128-134, (1997).
6. S. Masuda and C. E. Bauer, Cell, **110**, 613-623, (2002).
7. M. Gauden, S. Yeremenko, W. Laan, I. H. M. van Stokkum, J. A. Ihalainen, R. van Grondelle, K. J. Hellingwerf and J. T. M. Kennis, Biochemistry, **44**, 3653-3662, (2005).
8. W. Laan, M. Gauden, S. Yeremenko, R. van Grondelle, J. T. Kennis and K. J. Hellingwerf, Biochemistry, **45**, 51-60, (2006).
9. S. Masuda, K. Hasegawa and T. A. Ono, Plant Cell Physiol., **46**, 1894-1901, (2005).
10. M. Avila-Perez, K. J. Hellingwerf and R. Kort, J. Bacteriol., **188**, 6411-6414, (2006).

Comparing Semi-Empirical versus Classic Charge Assignments in BioMolecules and their Effect on Electrostatic Potentials

Parimal Kar[1], Max Seel[1], Ulrich H. E. Hansmann[1,2], and Siegfried Höfinger[1]

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

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

The program LocalSCF is used to consider an antifungal protein at the semi-empirical QM level of theory. Model Hamiltonians include AM1, MNDO, PM3 and PM5. The biomolecule is also studied with classic charge assignment using the AMBER force field. An important aspect of classic biomolecular simulation can thus be addressed, namely to what extent the usual concept of a single set of static atomic partial charges per type of amino acid will hold in general for the entire global protein structure. Semi-empirical charges will vary with different chemical neighborhood inside the protein and the question remains how severely these alterations will affect global electrostatic properties of the protein. In order to probe this effect we use grid maps of electrostatic potentials obtained from solutions to the Poisson-Boltzmann equation. Source charges are either the classic AMBER ones, or some set of semi-empirical charges from the list of models mentioned above. In comparing different potential maps we aim to recognize systematic trends as well as to identify a recommended way of proper charge assignment in proteins.

1 Introduction

Electrostatics plays an integral part in the study of structure and function of proteins at physiological conditions[1]. Theoretical considerations of the electrostatics in proteins are usually based on solutions to the Poisson-Boltzmann (PB) equation[2,3]. All these theoretical descriptions will involve a certain type of charge assignment to the atoms of the protein. Since the result of the PB calculation will inevitably depend on the particular choice made for the charges, it might be of interest to study the influence and variation resulting from different charge assignments. Of particular interest will be the comparison between a set of classic charges, ie from force fields commonly employed in the simulation of biomolecules, and charges derived from ab-inito calculations performed at a certain level of Quantum Mechanical (QM) theory.

A convenient method to compare different charge assignments to each other is to study the shape and appearance of electrostatic potential (ESP) maps. These ESP maps describe the way the protein will represent itself to its environment in electrostatic terms. Since the solution to the PB equation is included, ESP maps render a reasonably complete picture of the protein in its native environment, ie at physiological conditions. Moreover, ESP maps are a useful tool with many direct applications in structural biology. For example, from ESP maps we can learn whether a protein, (i) is likely to migrate to the membrane[4],

(ii) will potentially bind RNA or DNA[5,6], (iii) belongs to a certain family[7–9], (iv) offers a chemically attractive binding site to ligands and other proteins.

In this present study we therefore compare ESP maps based on classic charge assignments using AMBER parameters[10] with ESP maps resulting from semi-empirical charges computed with program LocalSCF[11] at several levels of semi-empirical theory, ie AM1, MNDO, PM3 and PM5. The PB program POLCH[12] is used throughout.

2 Methods

After download of the protein with pdb code EAFP2 from the pdb data bank, a PB calculation is performed using program POLCH[12] and classic AMBER partial charges[10]. Inner/outer dielectric constants are set to 1 and 80 respectively. The net charge is +4 due to the four Arg residues. ESP maps are computed on the molecular surface and on a cubic grid superimposing the protein. Only ESP maps directly mapped onto the molecular surface are used for further analysis. Semi-empirical calculations are then carried out on the protein EAFP2 using LocalSCF[11] and finally computed partial charges are extracted from the output. The net charge is +2 due to different treatment of lone-pairs in the semi-empirical models. AM1, MNDO, PM3 and PM5 methods are applied. Classic AMBER partial charges are then replaced with either charge set derived from the semi-empirical calculations and PB calculations are repeated with the changed charge assignment. Resulting ESP maps are compared in the form of difference ESP maps.

3 Results and Conclusions

A structural sketch of the antifungal protein EAFP2 is shown in Table 1 (a) with corresponding representation of the molecular surface (b). Here the N-terminal end is colored in red while the C-terminus is given in blue. The ESP map based on classic AMBER charge assignment after PB calculation is represented in Table 1 (c). ESP levels are color-coded as +5 kT/e (blue), 0 kT/e (green) and -5 kT/e (red). It becomes clear that the major appearance of EAFP2 in aqueous solution is that of a macroscopic particle of largely positive ESP, hence the tendency to migrate to the membrane can be explained straightforwardly[4] (which also implies the antifungal mode of action). An initial test regarding the sensitivity to counter ions is shown in Table 1 (d). Here explicit Cl^- counter ions have been included in the PB calculation and corresponding ESP maps produced. The change in major ESP patterns introduced by counter ions is only marginal, thus the rest of the analysis is performed without consideration of counter ions. A differential ESP map representing ESP(AM1) - ESP(AMBER) is shown in Table 1 (e). Identical color-coding is used as mentioned above. It becomes clear that the AM1-based ESP map is comparable in sign, but significantly different in magnitude (individual ESP values have become less positive). Extended red patches mark off regions of most severe difference. Contrary to the change seen in the AM1-AMBER differential map, when comparing AM1 with MNDO we obtain essentially only green patches (see Table 1 (f)). Thus AM1 and MNDO deliver essentially the same ESP properties. Comparison of PM3 with AMBER is represented in the differential ESP map shown in Table 1 (g).

The trend is similar to the one seen with AM1, but the difference is less severely pronounced (ie certain extended red regions turn yellow or green). Switching further to PM5

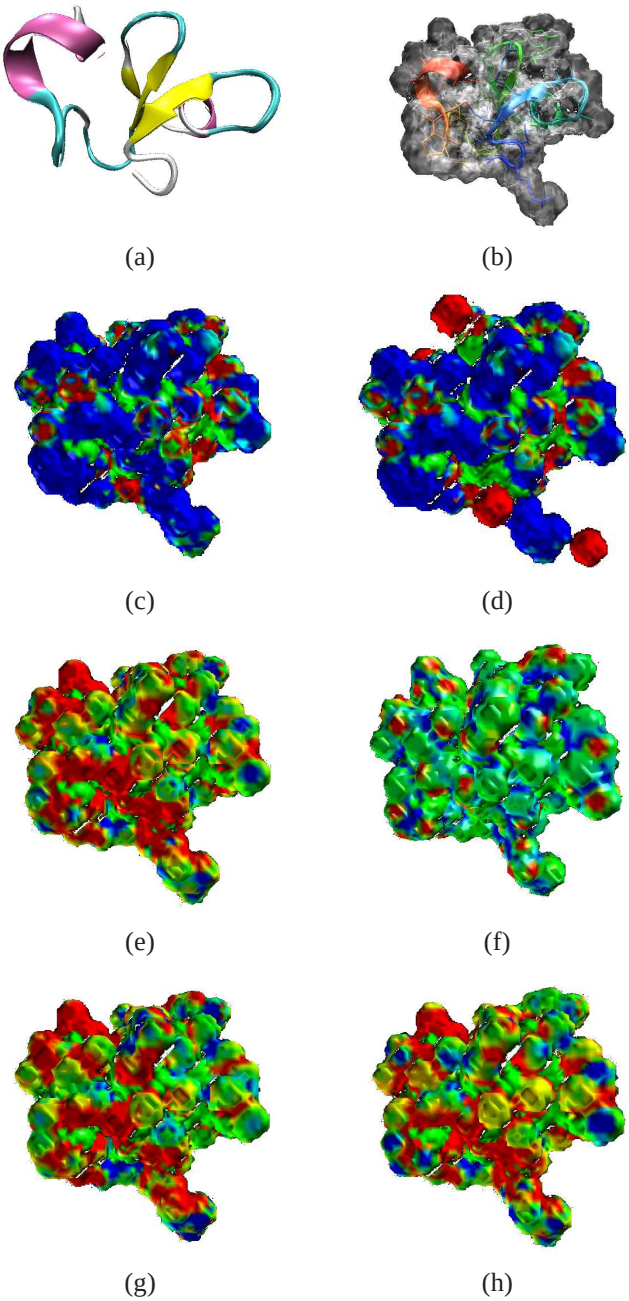


Table 1. Electrostatic Potential (ESP) maps for the antifungal protein EAFP2 (pdb code). Major structural elements are shown in (a) and a corresponding representation of the molecular surface is shown in (b), where the N-terminal helix is given in red and the C-terminus is shown in blue. The ESP mapped onto the molecular surface after solution of the PB equation based on AMBER charge assignment is shown in (c). Blue patches correspond to the +5 kT/e level, green regions represent neutral ESP and red domains indicate negative ESP of -5 kT/e. The marginal change when including 4 explicit Cl^- counter ions is shown in (d). A differential ESP map representing the difference between ESP(AM1) and ESP(AMBER) is shown in (e) with the same color-coding scheme used in (c). Further differential maps are ESP(AM1)-ESP(MNDO) (f), ESP(PM3)-ESP(AMBER) (g) and ESP(PM5)-ESP(AMBER) (h).

description is continuing the trend, ie lessening the deviation from the AMBER-based map again (see Table 1 (h)). Closer examination of the residues lying beneath the red-colored patches (indicating most severe deviation) reveals a specific role of Arg residues and the charges assigned to the N-atoms of Asn and Gln. In summary, semi-empirical charge assignments deliver a consistent picture of significant differences seen for the charged residues. However, individual semi-empirical models differ considerably amongst each other. With increasing sophistication of the semi-empirical model the deviation from the classic AMBER results becomes less severe.

Acknowledgments

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

References

1. B. Honig, A. Nicholls, *Classical electrostatics in biology and chemistry*, Science **268**, 1144–1149, 1995.
2. J. Warwicker, H. C. Watson, *Calculation of the electric potential in the active site cleft due to alpha-helix dipoles*, J. Mol. Biol. **157**, 671–679, 1982.
3. R. J. Zauhar, R. S. Morgan, *A new method for computing the macromolecular electric potential*, J. Mol. Biol. **186**, 815–820, 1985.
4. Y. Xiang, R. H. Huang, X. Z. Liu, Y. Zhang, D. C. Wang, *Crystal structure of a novel antifungal protein distinct with five disulfide bridges from Eucommia ulmoides Oliver at an atomic resolution* , J. Struct. Biol. **148**, 86–97, 2004.
5. S. S. Kim, R. g. Zhang, S. E. Braunstein, A. Joachimiak, A. Cvekl, R. S. Hegde, *Structure of the Retinal Determination Protein Dachshund Reveals a DNA Binding Motif*, Structure **10**, 787–795, 2002.
6. Q. Ye, R. M. Krug, Y. J. Tao, *The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA*, Nature **444**, 1078–1082, 2006.
7. J. W. Yu, J. M. Mendrola, A. Audhya, S. Singh, D. Keleti, D. B. deWald, D. Murray, S. D. Emr, M. A. Lemon, *Genome-Wide Analysis of Membrane Targeting by S. cerevisiae Pleckstrin Homology Domains*, Mol. Cell **13**, 677–688, 2004.
8. D. Morikis, J. D. Lambrish, *Physical methods for structure, dynamics and binding in immunological research*, Trends Immunol. **25**, 700–707, 2004.
9. K. Schleinkofer, U. Wiedemann, L. Otte, T. Wang, G. Krause, H. Oschkinat, R. C. Wade, *Comparative Structural and Energetic Analysis of WW DomainPeptide Interactions*, J. Mol. Biol. **344**, 865-881, 2004.
10. 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.
11. N. A. Anikin, V. M. Anisimov, V. L. Bugaenko, V. V. Bobrikov, A. M. Andreyev, *LocalSCF Method for Semi-empirical Quantum-Chemical Calculation of Ultra-large Bio-molecules*, J. Chem. Phys. **121**, 1266–1270, 2004.
12. S. Höfinger, *Solving the Poisson-Boltzmann Equation with the Specialized Computer Chip MD-GRAPE-2*, J. Comp. Chem. **26**, 1148–1154, 2005.

Role of Filopodia in Adhesion Formation During Migration of Epithelial Cells

Bodo Borm, Simone Born, Rudolf Merkel, and Bernd Hoffmann*

Institute of Bio- and Nanosystems 4: Biomechanics,
Research Centre Jülich, 52425 Jülich, Germany
**E-mail: b.hoffmann@fz-juelich.de*

Filopodia are a characteristic feature of many motile cells, such as keratinocytes and other epithelial cells. Filopodia are believed to act as guiding cues to direct cell migration, but how their specific functionality is implemented still remains elusive. Using live cell imaging and immunolabeling we show that filopodia are sites of adhesion formation during cell migration. Cell-substrate adhesions formed by migrating epithelial cells strongly depend on filopodia, showing a spatial and temporal correlation between filopodia dynamics and adhesion formation. We observed that: 1) the formation of adhesions depends on the persistence of filopodia. 2) Adhesion sites within a certain filopodium contain early adhesion markers such as vinculin or vasp. 3) Adhesions formed within filopodia are additionally defined by markers for mature adhesions such as tensin or zyxin. 4) In the nascent lamellipodium the filopodial adhesions grow in size whereas no obvious change in composition takes place. Our results implicate an essential role for filopodia in guiding cell migration by directing the formation of new cell-substrate adhesions.

1 Introduction

Cell adhesion is one of the most essential processes for proper cell function and movement of cells. For example, formation of multicellular organisms or the locomotion of cells of the immune system is impossible without cell adhesion and movement. The adhesion itself depends on protein complexes called focal adhesion sites[1]. These micrometer-sized adhesion sites form a connection between the extracellular matrix surrounding the cell and the actin cytoskeleton. Focal adhesions are characterized by a specific set of proteins such as integrins, spanning the plasma membrane, regulatory kinases or proteins like vinculin, zyxin or VASP, bridging the integrins to actin fibers. Latest adhesion site formation models imply for moving cells a formation of focal adhesion sites closely behind the leading edge of the cell's lamellipodia (Fig. 1)[2,3]. Over time the cell moves over the freshly formed adhesion sites transducing them to the rear site of the cell where release takes place. Formation and release times are highly dynamic in the range of seconds to minutes. In these models filopodia, i.e. finger like protrusion emanating from the lamellipodium, are mostly ignored although they are present in almost any motile cell. Instead it is assumed that filopodia mainly sense the environmental conditions without affecting adhesion site formation[4].

Our experiments show for the first time an essential role of filopodia in the life cycle of adhesion sites. Filopodia determine the location of almost all focal adhesion sites upon cell movement and are the first place of complete adhesion site formation. Such site becomes only increased in size when reach by the lamellipodium and therefore visible by classical microscopic techniques.

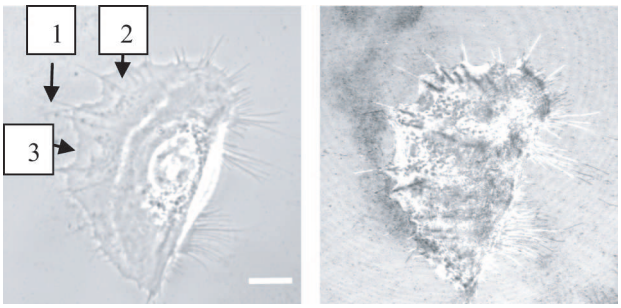


Figure 1. Keratinocyte motility was induced by EGF and cells were analyzed by phase contrast (left) and RICM (right). Note that focal adhesion sites appear as black lines in RICM and describe prolongations of filopodia. 1 = filopodium, 2 = lamellipodium, 3 = leading edge. Scale bar = 10 μm.

2 Experimental

Cell analyses were performed using primary human foreskin keratinocytes. Cell biological procedures were performed according to standard protocols. Cells were either analyzed directly by phase contrast microscopy or reflection interference contrast microscopy (RICM) or transfected with GFP-fusion proteins and analyzed by fluorescence microscopy. RICM allows the determination of distances between surface and adhered cell. The darker the signal is, the smaller the distance. Alternatively, cells were fixed and proteins were stained by immunolabelling.

3 Results and Discussion

In order to analyze filopodia function in focal adhesion site formation, keratinocytes were incubated for one day and subsequently activated by addition of epidermal growth factor (EGF). EGF stimulates cell motility of keratinocytes. Individual moving cells were analyzed by phase contrast and simultaneously by RICM. These data indicated a formation of focal adhesion sites right behind adhered filopodia (Fig. 1). Additionally, the shape of these adhesion sites resembled the one of filopodia.

The given data indicate a role of filopodia in focal adhesion site formation. Such hypothesis was analyzed in more detail. As filopodia are continuously formed ”spikes” sensing the substrate conditions in the direction of movement they can either attach to the substrate to form temporarily stable connections or become retracted if no connection could be formed. By analyzing the leading edge over time we could determine that focal adhesion sites were formed only at positions of stable adhered filopodia. Areas of the leading edge characterized by filopodia unable to attach to the substrate could not respond with the formation of focal adhesion sites (Fig. 2).

The formation of focal adhesion sites always behind stably attached filopodia could be explained by two adhesion models. First, filopodia could sense the substrate. Their stable adhesion could then function as signal in order to form focal adhesion sites right behind the filopodia. The alternative would be the formation of small adhesion sites already in the filopodia. In this case the lamellipodium would just function as enlargement factor as

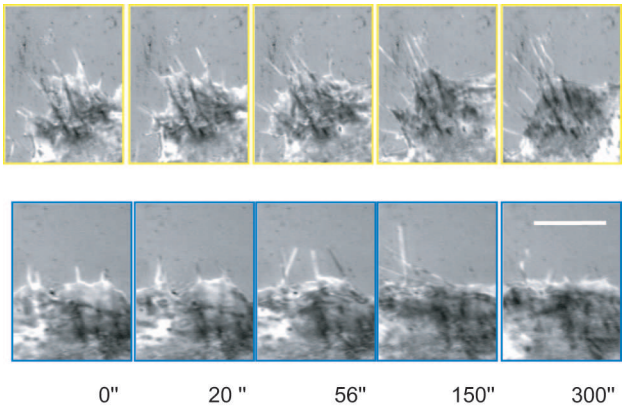


Figure 2. Motile keratinocytes were analyzed by RICM in areas of stable (top) and unstable (bottom) filopodia formation over time. Focal adhesion sites were only detected at positions of stable filopodia. Scale bar = 5 μm.

soon as it has reached the filopodia adhesion site. These models were tested by analyzing the protein localization of focal adhesion site proteins in detail as the latter model fully depends on a localization of these proteins already in the filopodia. Therefore various adhesion site specific proteins were analyzed in fixed samples by immunofluorescence experiments (Fig. 3).

The data pointed out that every focal adhesion site specific protein analyzed (in total 8) could be detected in stably adhered filopodia. Even proteins believed to be markers for old focal adhesions, as zyxin, were present.

When these proteins were analyzed in living cells using GFP-fusions over time, localizations of all proteins in filopodia could be confirmed. In addition, upon cell movement filopodia signals became overgrown by the lamellipodia leading edge. At that moment filopodia adhesion sites grew in size and became classical focal adhesion sites within a few seconds (Fig. 4). Adhesion site formation at other sites than filopodia adhesions were barely detected.

As given results identified filopodia as origin for focal adhesion site formation, we were interested how adhesion behaviour would change upon inhibition of filopodia formation. Such formation depends on the second messenger PI(4,5)P2 finally triggering filopodia outgrowth via Cdc42 and WASP. Sequestering PI(4,5)P2 by neomycin sulphate we were able to fully block filopodia formation. Such block neither affected actin filament organization nor cell polarization but induced formation of small dot like adhesion structures (Fig. 5). Additionally, these adhesion structures were clustered and displaced into the lamella compared to untreated cells. Therefore the neomycin experiments argue for a direct influence of filopodia in early adhesion site formation right at the leading edge. As soon as filopodia are blocked in their formation, localization and form of adhesion structures resemble those known from cells naturally characterized by the absence of filopodia. If such morphological switch affects migration efficiency needs to be shown.

In summary, the given data strongly argue for a new adhesion model for motile cells. Sites of first contact to newly explored substrates are filopodia adhesions. Whenever these

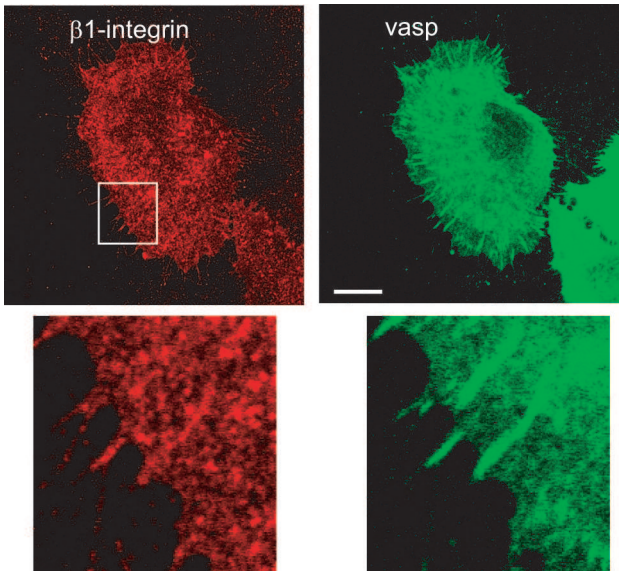


Figure 3. Motile keratinocytes were fixed and immunofluorescently labelled using antibodies against focal adhesion site specific proteins 1 integrin or vasp. Note that both proteins can be found in the filopodia extensions. Scale bar = 10 μm.

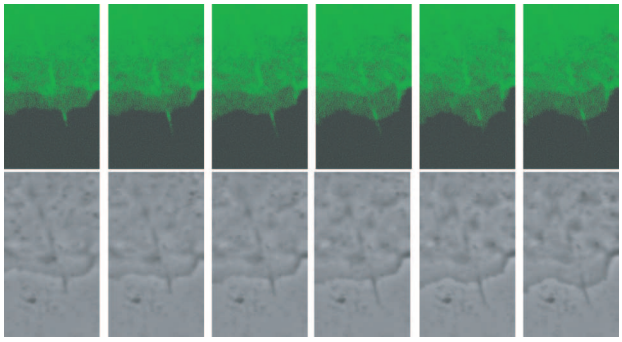


Figure 4. GFP-VASP transfected keratinocytes were simultaneously analyzed in fluorescence (upper row) and phase contrast (lower row). Note that the filopodia staining becomes a focal adhesion site over time. Image intervals = 10 s.

adhesions can not be built up under regular conditions or are unstable, no subsequent focal adhesion site will be formed and movement in that direction is blocked. In contrast, whenever a filopodia stably contacts the substrate, such a contact resembles a fully developed, yet very small adhesion site. When reached by the lamellipodium, such site is just built up in size. Therefore, filopodia are the key factors for movement, the direction of movement and finally for the formation and localization of focal adhesion sites in motile cells.

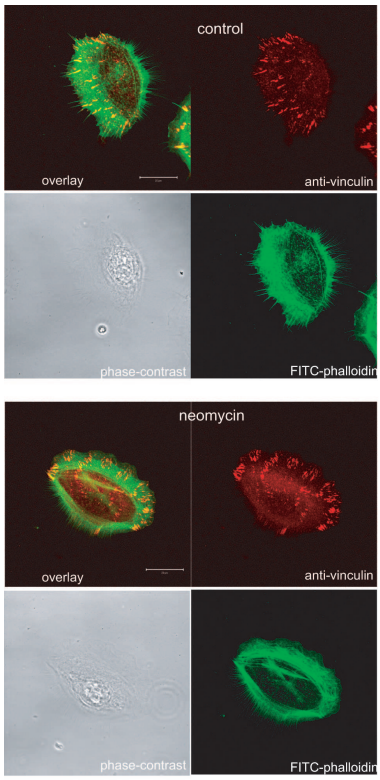


Figure 5. Keratinocytes were stimulated with eGF and analyzed for adhesion site formation in the absence (control) or presence of 10 mM Neomycin. Subsequently, cells were fixed and analyzed for vinculin (red) and actin (green). Cell morphology was additionally determined in phase contrast. Scale bar = 10 μm.

References

1. R. Zaidel-Bar, M. Cohen, L. Addadi, and B. Geiger, *Hierarchical assembly of cell-matrix adhesion complexes*, Biochem. Soc. Trans. **32**, 416–20, 2004.
2. A. D. Bershadsky, C. Ballestrem, L. Carramusa, Y. Zilberman, B. Gilquin, S. Khochbin, A. Y. Alexandrova, A. B. Verkhovsky, T. Shemesh, and M. M. Kozlov, *Assembly and mechanosensory function of focal adhesions: experiments and models* Eur. J. Cell. Biol. **85**, 165–73, 2006.
3. M. A. Wozniak, K. Modzelewska, L. Kwong, and P. J. Keely, *Focal adhesion regulation of cell behavior* Biochim. Biophys. Acta **1692**, 103–19, 2004.
4. J. Faix, and K. Rottner, *The making of filopodia* Curr. Opin. Cell. Biol. **18**, 18–25, 2006.

Study of Protein Structural Descriptors: Towards Similarity and Classification

Pooja Jain and Jonathan D. Hirst

School of Chemistry,
University of Nottingham, Nottingham,
NG7 2RD, United Kingdom
E-mail: {pcxpj1, jonathan.hirst}@nottingham.ac.uk

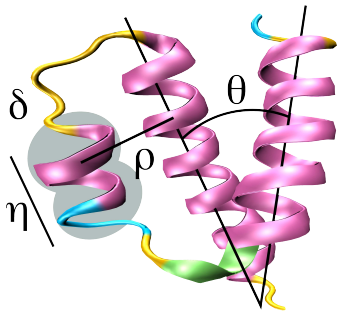
We have investigated structural descriptors for structural similarity and classification of 77 proteins extracted from SCOP. A Support Vector Machine was trained to predict structural similarity based on paired protein profiles, composed of structural descriptors derived from the geometric properties of secondary structure elements. Ten fold cross-validation, against the standard similarity measure from DALI gave a cross-validated correlation coefficient, q^2, of 0.91. A coefficient of dissimilarity was derived as the Euclidean distance among different descriptor types of two proteins. This coefficient was evaluated for the classification of protein pairs to different levels in SCOP hierarchy.

1 Introduction

Protein structure comparison can provide useful information on the biological function of a protein[1] and can imply evolutionary relationships between proteins with low sequence similarity. This information is crucial in the identification of new protein folds and understanding the organisation of the known universe of protein structures. The aim of this work is to perform protein structure analysis and comparison through the use of structural descriptors. These are numerical values that characterise the secondary structure elements of a protein. For example, they may represent the physico-chemical properties or geometric properties of the secondary structure elements derived from the 3D coordinates. In our study we make use of SCOP, DALI and USM. SCOP is a curated database which aims to provide a comprehensive description of the structural and evolutionary relationships between all protein structures[2]. The principal levels in the SCOP hierarchy are class, fold, superfamily and family. DALI is a common and popular structural alignment and comparison method[1]. It represents a protein as a matrix of contact patterns between successive hexapeptide fragments and makes comparisons with such matrices of other proteins. USM is based on the comparison of compressed protein contact maps using the principle of Kolmogorov complexity[3].

2 Structural Descriptors

A protein was defined in terms of structural descriptors derived from its secondary structure elements. A set of 77 proteins containing exactly and only three α helices from the all-α class of SCOP database was used. The descriptors, such as the pairwise separation ρ between the centre of mass COM of any two secondary structure elements i and j of lengths n and m in a protein along Cartesian coordinates A, the relative orientation cosθ,



$$\rho_{ij} = \sqrt{\sum_{A=X,Y,Z}(COM_{Ai} - COM_{Aj})^2}$$

$$cos\theta_{ij} = \frac{\vec{V_i} \cdot \vec{V_j}}{|V_i||V_j|}$$

$$\delta_i = \frac{\sum_1^n \delta_n}{n}$$

$$\eta_i = n_i$$

Figure 1. A pictorial representation of structural descriptors and their mathematical definitions.

the individual surface accessibility δ and the length η for each of them were derived from the DSSP assignments[4]. Figure 1 illustrates these descriptors in a protein structure and give their mathematical definitions.

3 The Protein Profile

The geometric profiles made up of above defined descriptors was used to pair up any two proteins. Such paired protein profiles included along with the 12 descriptors for each of the proteins, the RMS difference between respective descriptor types for that pair. For the 77 proteins, 2,926 paired protein profiles containing 28 elements each were generated.

3.1 Profile Based Structural Similarity

Using Support Vector Machines (SVMs) the paired protein profiles were subjected to the non-linear (multivariate) regression against the protein similarity values assigned by DALI and USM. The model was trained by the Sequential Minimal Optimisation algorithm for regression analysis (SMOreg)[5] from Waikato Environment for Knowledge Analysis (WEKA) software package[6]. Parameter tuning was performed to choose the best values for complexity parameter and kernel function. Finally, the model assessment was performed using 10-fold cross-validation.

3.2 Profile Based Structural Classification

For a pair of proteins x and y, a coefficient of dissimilarity Ω_{xy} that gives the Euclidean distance between them was derived from the RMS difference of different descriptor types as below:

$$\Omega_{xy} = \sqrt{\rho_{xy}^{rmsd} + \theta_{xy}^{rmsd} + \eta_{xy}^{rmsd} + \delta_{xy}^{rmsd}}$$

The higher Ω more dissimilar are the proteins.

4 Results

The outcome of multivariate regression of paired protein profiles against the similarity values assigned by USM and DALI was significant with cross-validated correlation coefficients (q^2) of 0.74 and 0.91, respectively.

The structural classification of protein pairs was based on the coefficient of dissimilarity Ω. The protein pairs belonging to the same family congregated towards a lower dissimilarity threshold, whilst those sharing the same fold were associated with higher values.

5 Concluding Remarks

The results from multivariate regression of protein profiles suggest their potential as a representation of protein structures and there further use in the protein structure comparison. Ideally, the profile based structural classification of proteins to different levels in SCOP hierarchy should be distinctive. Our results show some bias towards this. Efforts to improve this are ongoing. Analysis continues on larger datasets comprising proteins containing three or four secondary structure elements.

Acknowledgements

This work was supported by BIOPTRAIN project MEST-CT-2004-007597 under the Sixth framework program of the European Community. We thank the University of Nottingham for providing high performance computational resources.

References

1. L. Holm and C. Sander, *Protein-structure comparison by alignment of distance matrices*, J. Mol. Biol. **233**, 123–138, 1993.
2. A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *SCOP: a structural classification of proteins database for the investigation of sequences and structures*, J. Mol. Biol. **247**, 536–340, 1995.
3. N. Krasnogor and D. A. Pelta, *Measuring the similarity of protein structures by means of the universal similarity metric*, Bioinformatics **20**, 1015–1021, 2004.
4. W. Kabsch and C. Sander, *Dictionary of Protein Secondary Structure: Pattern recognition of Hydrogen-bonded and Geometrical Features*, Biopolymers **22**, 2577–2637, 1983.
5. G. W. Flake and S. R. Lawrence, *Efficient SVM regression training with SMO*, Machine Learning **46**, 271–290, 2002.
6. I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2nd Edition, 2005.

Phase Separation in Peptide Aggregation Processes – Multicanonical Study of a Mesoscopic Model

Christoph Junghans[1,2], **Michael Bachmann**[1,3], **and Wolfhard Janke**[1]

[1] Institut für Theoretische Physik and Centre for Theoretical Sciences (NTZ), Universität Leipzig, Postfach 100 920, D-04009 Leipzig, Germany
E-mail: janke@itp.uni-leipzig.de

[2] Max-Planck-Institut für Polymerforschung, Ackermannweg 10, D-55128 Mainz, Germany
E-mail: junghans@mpip-mainz.mpg.de

[3] Computational Biology & Biological Physics Group, Department of Theoretical Physics, Lunds Universitet, Sölvegatan 14A, SE-223 62 Lund, Sweden
E-mail: bachmann@thep.lu.se

We have performed multicanonical computer simulations of a small system of short protein-like heteropolymers and found that their aggregation transition possesses similarities to first-order phase separation processes. Not being a phase transition in the thermodynamic sense, the observed folding-binding behavior exhibits fascinating features leading to the conclusion that the temperature is no suitable control parameter in the transition region. More formally, for such small systems the microcanonical interpretation is more favorable than the typically used canonical picture.

1 Introduction

Folding-binding and docking processes between proteins are significant for catalysis, transport, and cell stabilization in biological systems. Also, gene replication and expression are impossible without defined binding mechanisms of molecules. However, the mutual influence of proteins on each other can also result in refolding of proteins (which often leads to the loss of their functionality and thus biological activity) or cluster formation. In the latter case, proteins self-assemble and form aggregates. The effects of plaque can be disastrous and cause heavy diseases: First, the assembled proteins lose their individual functionality and second, in the passive case, the aggregates hinder transport and signal exchange processes which are significant for the life of cells. In an active process, specific aggregates might be able to bind to cell membranes and to change the membrane morphology, e.g., by forming pores. In the amyloid hypothesis for the onset of Alzheimer's disease, for example, aggregates of Aβ proteins are believed to form pores in membranes of neuron cells, thus opening ion channels for neurotoxic calcium.[1]

We focus here on thermodynamic properties of the aggregation transition of small peptides. For this purpose, a simple hydrophobic-polar aggregation model is introduced and employed in a multicanonical study of a few short heteropolymers.[2,3]

2 Aggregation Model

For the aggregation study, we extend the AB model[4] by an additional interchain interaction between the M heteropolymers. As in the single-chain model, which has proven to be quite

useful in qualitative studies of tertiary folding behavior,[5] only two types of amino acids are considered: hydrophobic residues (A) which avoid contact with the polar environment and polar residues (B) being favorably attracted by the solvent. The single-chain energy of the μth heteropolymer ($\mu = 1, \ldots, M$) composed of N_μ monomers is given by[4]

$$E_{\mathrm{AB}}^{(\mu)} = \frac{1}{4} \sum_{i_\mu} (1 - \cos \vartheta_{i_\mu}) + \sum_{j_\mu > i_\mu + 1} \Phi(r_{i_\mu j_\mu}; \sigma_{i_\mu}, \sigma_{j_\mu}), \tag{1}$$

where $0 \leq \vartheta_{i_\mu} \leq \pi$ denotes the virtual bending angle between the monomers i_μ, $i_\mu + 1$, and $i_\mu + 2$. Not discriminating nonbonded interactions between monomers of the same or different polymers, our aggregation model reads[2]

$$E = \sum_{\mu} E_{\mathrm{AB}}^{(\mu)} + \sum_{\mu < \nu} \sum_{i_\mu, j_\nu} \Phi(r_{i_\mu j_\nu}; \sigma_{i_\mu}, \sigma_{j_\nu}), \tag{2}$$

where μ, ν label the M polymers interacting with each other, and i_μ, j_ν index the monomers of the respective μth and νth polymer. The nonbonded interresidue pair potential $\Phi(r_{i_\mu j_\nu}; \sigma_{i_\mu}, \sigma_{j_\nu}) = 4[r_{i_\mu j_\nu}^{-12} - C(\sigma_{i_\mu}, \sigma_{j_\nu}) r_{i_\mu j_\nu}^{-6}]$ depends on distance $r_{i_\mu j_\nu}$ and residue type $\sigma_{i_\mu} = A, B$. The long-range behavior is attractive for like pairs of residues [$C(A, A) = 1$, $C(B, B) = 0.5$] and repulsive otherwise [$C(A, B) = C(B, A) = -0.5$]. The length of all virtual peptide bonds is unity. In this short note, we focus on a system of two identical chains with the Fibonacci sequence $AB_2AB_2ABAB_2AB$, where the single-chain properties are known.[6] Our primary interest is devoted to the phase behavior of the system and for this purpose, the density of states $g(E)$ is a suitable quantity that we obtained by means of multicanonical computer simulations.[7]

3 Microcanonical vs. Canonical View

The Hertz definition of the microcanonical entropy is given by $\mathcal{S}(E) = k_B \ln \Gamma(E)$, where k_B is the Boltzmann constant ($k_B = 1$ in our simulations) and $\Gamma(E) = \int_{E_{\min}}^{E} dE'\, g(E')$ (where $E_{\min}$ is the ground-state energy) is the phase-space volume. In Fig. 1, $\mathcal{S}(E)$ is shown for our two-peptide system. Interestingly, in the energy region between E_{agg} and E_{frag} the entropy exhibits a convex behavior, which is a strong indication for surface effects within this small system.[8] Also shown in Fig. 1 is the corresponding concave hull $\mathcal{H}_{\mathcal{S}}(E)$, i.e., the Gibbs construction. The surface entropy, defined as $\Delta\mathcal{S}(E) = \mathcal{H}_{\mathcal{S}}(E) - \mathcal{S}(E)$ is maximal at the energy E_{sep}. The reason for the nonvanishing surface entropy is that the transition between the fragmented, i.e., separated chains, and the formation of a joint aggregate is a process with phase separation which is "delayed" due to steric surface effects reducing the entropy of the total system. Since entropy reduction is only achieved by additional energy consumption, the surprising side effect is that in the transition regime the aggregate becomes colder with increasing system energy. This is verified by considering the caloric temperature which is defined via $T^{-1}(E) = \partial\mathcal{S}(E)/\partial E$, also shown in Fig. 1. Actually, in the transition region, $T^{-1}(E)$ bends back with increasing energy.

Consequently, there is no unique mapping between temperature and energy in the transition region (or more precisely, within the bounds $T_<^{-1}$ and $T_>^{-1}$ indicated in Fig. 1). Thus the temperature should not be considered as a suitable external control parameter. From a statistical point of view this means that for transitions with phase separation in small

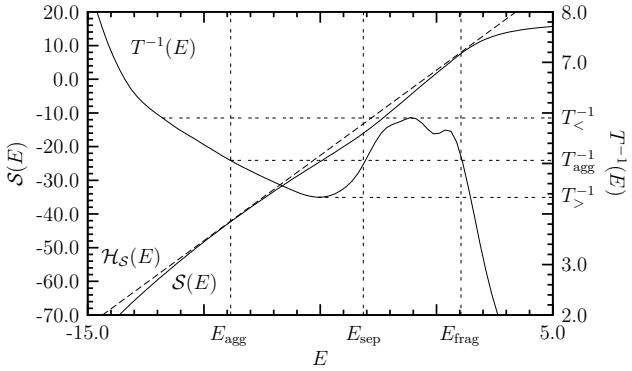


Figure 1. Microcanonical Hertz entropy $\mathcal{S}(E)$, concave Gibbs hull $\mathcal{H}_\mathcal{S}(E)$, and inverse caloric temperature $T^{-1}(E)$ as functions of energy. The phase separation regime is bounded by $E_{\rm agg}$ and $E_{\rm frag}$; the temperature region, where temperature is no suitable external control parameter and the canonical interpretation breaks down, ranges from $T_<^{-1}$ to $T_>^{-1}$. The slope of the Gibbs hull defines the aggregation temperature, $T_{\rm agg}^{-1} = \partial\mathcal{H}_\mathcal{S}(E)/\partial E = \text{const.}$

systems a microcanonical interpretation is preferred over the typically used canonical formalism. Since the backbending effect in the peptide aggregation process is a real physical effect, it should also be accessible to experimental verification, as it has indeed already been observed, for example, in experiments of sodium cluster formation processes.[9]

Acknowledgments

This work is partially supported by the DFG under Grant No. JA 483/24-1/2 and the computer time Grant No. hlz11 of NIC, Forschungszentrum Jülich. M.B. thanks the DFG and Wenner-Gren Foundation for research fellowships. We are also grateful to DAAD-STINT for a Personnel Exchange Programme with Sweden.

References

1. H. A. Lashuel and P. T. Lansbury jr., Quart. Rev. Biophys. **39**, 167 (2006).
2. C. Junghans, M. Bachmann, and W. Janke, Phys. Rev. Lett. **97**, 218103 (2006).
3. C. Junghans, M. Bachmann, and W. Janke, preprint (2007).
4. F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993).
5. S. Schnabel, M. Bachmann, and W. Janke, Phys. Rev. Lett. **98**, 048103 (2007); J. Chem. Phys. **126**, 105102 (2007).
6. M. Bachmann, H. Arkın, and W. Janke, Phys. Rev. E **71**, 031906 (2005).
7. B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).
8. D. H. E. Gross, *Microcanonical Thermodynamics* (World Scientific, Singapore, 2001).
9. M. Schmidt, R. Kusche, T. Hippler, J. Donges, W. Kronmüller, B. von Issendorff, and H. Haberland, Phys. Rev. Lett. **86**, 1191 (2001).

Algorithmic Refinements to an Enhanced Poisson-Boltzmann Approach Used in BioMolecular Simulation

Parimal Kar[1], Max Seel[1], Ulrich H. E. Hansmann[1,2], and Siegfried Höfinger[1]

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

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

In a series of recent publications we have introduced a very general description of solvation for biomolecular structure. It was shown that the enhanced continuum electrostatics approach can be decomposed into a series of individual terms, each of them representing its own portion of distinct physical interaction. Care has been taken to operate the model at conditions that guaranteed a maximum level of numerical accuracy. However, a number of internal parameters still need to be optimized further in order to speed up the procedure. Among these factors are i) the exact value of the exit criterion used to terminate the calculation, ii) the array dimension regulating the allowed number of consecutive DIIS steps, iii) the switch criterion used to move from the pre-DIIS stage to the DIIS stage, iv) the dependence on system size of the number of necessary iterations to achieve convergence, v) the dependence on renormalization factors applied to the net sum of polarization charges, vi) the influence of very small-sized boundary elements, or the introduced change when merging these very small-sized elements to larger ones from the neighborhood, and vii) the surface resolution necessary to calculate the dispersion term. Therefore in this present study we want to address these points and examine their consequences on run-time performance. A series of ten proteins of increasing size will be used as a testbed.

1 Introduction

Biological molecules typically reside in aqueous environments. Reliable consideration of the effect of water on structure and dynamics of biomolecules is among the key factors governing accurate descriptions of biological matter[1]. Here we focus on an implicit solvation model. Among other methods, e.g. SASA, GB, FDPB, the Poisson - Boltzmann (PB) approach[2] within the Boundary Element Method (BEM)[3] is frequently chosen due to its intermediate position regarding computational cost versus achievable accuracy. In our recent series of publications[4,5] we have outlined a generalization of the Polarizable Continuum Model (PCM)[6] applied to biomolecular structure. Each of the considered terms represents a separate portion of distinct physical interaction,

$$\Delta G^{sol} = \Delta G^{pol} + \Delta G^{disp,rep} + \Delta G^{cav} \tag{1}$$

which are polarization, dispersion and cavitation. The latter plays an important role in hydrophobicity related phenomena[7]. Care has been taken to operate the model at conditions

that guaranteed a maximum level of numerical accuracy. However, a number of internal parameters could still profit from further optimization.

2 Methods

2.1 Aims

In this present study, we address the following factors and examine their consequences on run-time performance with regard to a series of test proteins of increasing size that we have studied earlier[5], (i) the exact value of the exit criterion used to terminate the calculation of the polarization term, Δ G^{pol}, (ii) the array dimension regulating the allowed number of consecutive DIIS[8] steps, (iii) the switch criterion used to move from the pre-DIIS stage to the DIIS stage, (iv) the dependence on system size of the number of necessary iterations to achieve convergence, (v) the dependence on renormalization factors applied to the net sum of polarization charges, (vi) the influence of very small-sized boundary elements (BEs), or the introduced change when merging these very small-sized elements to larger ones from the neighborhood, (vii) the necessary degree of surface resolution for accurate calculation of the dispersion term, Δ G^{disp}.

2.2 Procedure

We select 10 proteins of different size (number of residues reaching from 41 to 430). Initially we run the PB/BEM program POLCH[9] at default conditions. The run time for all the 10 cases is recorded and forms a reference set. At first, we adjust the parameter MAXNIT which defines the maximum number of successive DIIS steps, hence determines the size of the DIIS matrix, and compute the run time deviation from the reference set for all the 10 proteins. Once parameter MAXNIT is optimized, we rerun the entire test set and extract net solvation free energies, Δ G^{sol}, which serve as a new reference. ACCURA is the second parameter to be optimized. It defines the threshold criterion used for termination of the iterative process when computing the polarization term, Δ G^{pol}. For optimizing ACCURA we require the deviation from the reference set not to exceed $\pm$ 0.05 kcal/mol for any of the proteins. Once ACCURA is optimized we redo the whole set of test proteins at optimized conditions for either parameter, ACCURA as well as MAXNIT. We extract the number of iterations needed for completion and use these as a new reference. In our next step we optimize the parameter DSNTRC. This parameter sets the switch criterion used to move from a pre-DIIS stage to the DIIS stage. It represents the mean square deviation of two successive sets of polarization charges. We keep changing DSNTRC and optimize for a minimum number of necessary iterations. The next point is concerned with renormalization of the polarization charges according to Gauss' Law. We study the effect this has on net solvation free energies. The solvation free energies obtained after renormalization form another reference set for our next investigation. Here, we study the influence of very small-sized BEs. We will merge these very small-sized elements to larger ones from the neighborhood. We change the parameter REQSZ (the required minium size of a BE) and compute the deviation of solvation free energies from the reference set. We again do not allow the energy to change more than by $\pm$ 0.05 kcal/mol in all test runs. Finally, we use all previously optimized parameters for a final test focusing on the

Protein PDB Code	No. of Res.	Molecular Charge (a.u.)	No. of Iter.	Δ G^{sol} Without Normalization (kcal/mol)	Δ G^{sol} Including Normalization (kcal/mol)	Δ G^{sol} Deviation (unsigned) (kcal/mol)
1P9GA	41	+3	9	-319.73	-321.54	1.81
2B97	70	+2	8	-40.22	-39.26	0.96
1LNI	96	-5	10	-534.64	-536.39	1.75
1NKI	134	+5	10	-456.20	-454.94	1.26
1EB6	177	-11	10	-1224.17	-122.75	1.42
1G66	207	-2	9	-118.26	-119.01	0.75
1P1X	250	+3	11	-636.41	-636.41	0.00
1RTQ	297	-16	11	-1998.72	-2011.23	12.51
1YQS	345	+2	11	-217.22	-217.22	0.04
1GPI	430	-12	12	-1259.54	-1271.59	12.05

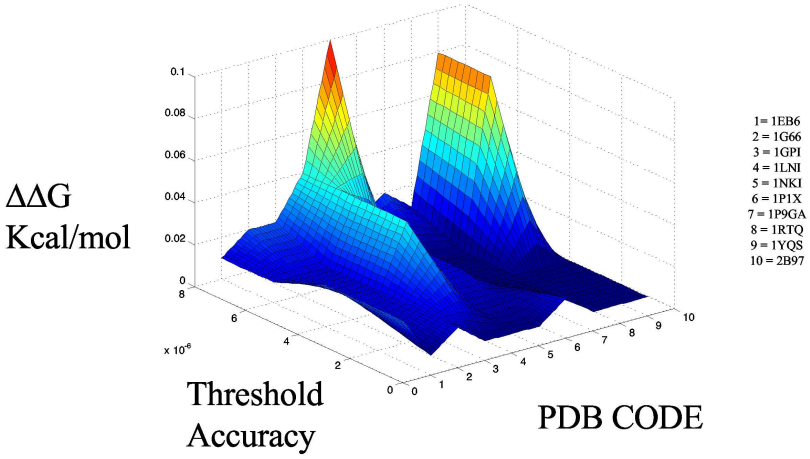


Figure 1. Numerical sensitivity of the employed enhanced Poisson-Boltzmann approach to the threshold criterion used for termination of the iterative sequence to calculate the polarization term, ΔG^{pol}. A series of 10 proteins is tested and the threshold criterion is varied between 1.0x10^{-6} and 8.0x10^{-6}. When requiring the results to be numerically accurate at least up to the first digit behind the decimal point, ie allowing fluctuations $< \pm 0.05$ (light blue patches on the surface) then the optimal value for this termination criterion is identified as 4.0x10^{-6}.

dispersion term. We change the resolution of the boundary used for calculation of Δ G^{disp} which need not be maintained at such rigorous levels as identified for the polarization term[4].

3 Results and Conclusions

Sensitivity to total system size, total charge and renormalization attempts is represented in Table 3. Variation of the termination criterion is graphically represented in Figure 1. In summary we find that the following parameters lead to a reasonable degree of numerical accuracy. (1) Best performance is achieved when the DIIS matrix is dimensioned 7x7, (2) Using a threshold criterion of 4.0x10^{-6} for termination of the iterative sequence occurring in ΔG^{pol} computation leads to stable numerical results. (3) The best switch criterion to move from the pre-DIIS stage to the DIIS stage is given when the root mean square deviation between two successive sets of polarization charges falls below 0.05 a.u. (4) The number of iterations necessary to achieve convergence does not depend on system size. (5) A renormalization process will affect the net solvation free energies, ΔG^{sol}, on the order of $\pm$ 1-2 % of their total values. Systems with large net charges are more sensitive to renormalization. (6) If we merge small sized BEs to larger ones then no significant changes will occur when this procedure is limited to elements smaller than 8 % of the mean size (0.31 Å^2). A reduction in number of BEs will lower the computational cost and foster numerical stability. (7) For calculation of the dispersion term, ΔG^{disp}, we can reduce the discretization of the boundary into BEs of average size 0.45 Å^2 without loss of accuracy.

Acknowledgments

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

References

1. B. Honig, A. Nicholls, *Classical electrostatics in biology and chemistry*, Science **268**, 1144–1149, 1995.
2. J. Warwicker, H. C. Watson, *Calculation of the electric potential in the active site cleft due to alpha-helix dipoles*, J. Mol. Biol. **157**, 671–679, 1982.
3. R. J. Zauhar, R. S. Morgan, *A new method for computing the macromolecular electric potential*, J. Mol. Biol. **186**, 815–820, 1985.
4. P. Kar, Y. Wei, U. H. E. Hansmann, S. Höfinger, *Systematic Study of the Boundary Composition in Poisson Boltzmann Calculations*, J. Comput. Chem., in press, 2007.
5. P. Kar, M. Seel, U. H. E. Hansmann, S. Höfinger, *Dispersion Terms and Analysis of Size- and Charge-Dependence in an Enhanced Poisson-Boltzmann Approach*, submitted, 2007.
6. J. Tomasi, B. Mennucci, R. Cammi, *Quantum Mechanical Continuum Solvation Models*, Chem. Rev. **105**, 2999–3094, 2005.
7. M. Makowski, A. Liwo, H. A. Scheraga, *Simple physicsbased analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water.I. Approximate expression for the free energy of hydrophobic association based on a Gaussian overlap model* J. Phys. Chem. B **111**, pp 2910–2916, 2007.
8. P. Pulay, *Convergence acceleration of iterative sequences. The case of scf iteration* , Chem. Phys. Lett. **73**, 393–398, 1980.
9. S. Höfinger, *Solving the Poisson-Boltzmann Equation with the Specialized Computer Chip MD-GRAPE-2*, J. Comp. Chem. **26**, 1148–1154, 2005.

Aggregation of the Amyloid-β Protein: Monte Carlo Optimization Study

Srinivasa M. Gopal, Konstantin V. Klenin, and Wolfgang Wenzel

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

The free-energy approach has delivered promising results for protein folding and structure prediction in recent years. The native state is found as the global minimum of an all-atom free-energy forcefield. Now, we used this approach to simulate the aggregation of Aβ fragment 16-22. This aggregation is believed to be associated with the Alzheimer's disease. The model system contained 2 polypeptide chains. The obtained structures of the aggregates consisted of either parallel or anti-parallel β-sheets, the latter were preferable.

1 Introduction

The Alzheimer's disease is associated with misfolding of the amyloid protein. In the functional native form, this small protein, of 42 amino acids, has globular structure with two α-helices. However, under certain conditions, it can form aggregates of β-sheets that, on a larger scale, are arranged as long insoluble fibrils[3]. The resulting fibrils have toxic effects in the extracellular space of the brain of patients with the Alzheimer's disease. The amyloid protein in this form is known as Aβ peptide.

This protein has been extensively studied by both experimental and computational methods[8,7]. The experimental techniques such as solid-state NMR, X-diffraction and electron microscopy were used for characterizing the structure of aggregates[8]. Most of the computational studies were focused on short polypeptide fragments of this protein, such as $LYS_{16} - LEU - VAL - PHE - PHE - ALA - GLU_{22}$[4]. This fragment is believed to play a key role in the formation of the aggregates. In the present work, we study the systems of one and two chains of Aβ_{16-22} peptide[2].

2 Methods and Results

The free-energy approach has delivered promising results for protein folding and structure prediction in recent years. Following Anfisen's hypothesis[1] the native state is postulated to be the global minimum of a all-atom free-energy function. This minimum can be found by optimization methods[9] including the Monte Carlo procedure. This approach was successfully used to fold α-, β- and mixed proteins of moderate length[10,12,5].

This approach requires an accurate, transferable protein free-energy forcefield. In this study we used Protein Force Field (PFF02)[6,11]. In our model, all the atoms are explicitly represented (the apolar group CH_n is considered as a single atom). The bond angles and the bond lengths are fixed. The degrees of freedom considered are the backbone (ψ, ϕ) and the sidechain (χ_i) dihedral angles.

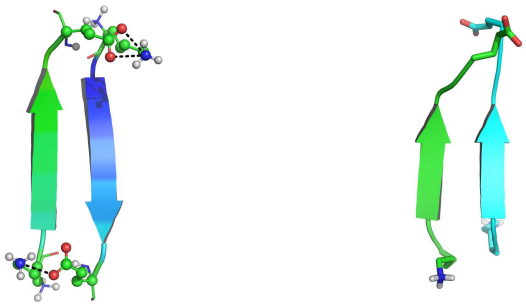


Figure 1. **Left:** The lowest energy structure is an anti-parallel β-sheet. The polar contacts between side-chain atoms N of LYS_{16} and O of GLU_{22} are shown explicitly. **Right:** The energy of the best parallel β-sheet structure is higher by 10 kcal/mol.

The energy function contains the following terms: (1) the standard Lennard-Jones potential, (2) the Coulomb energy of electrotatic interaction with group specific dielectric constants, (3) a term for the hydrogen bonding, (4) a SASA-based solvation term that implicitly takes into account the influence of the solvent. We used several optimization methods used to find the global minimum of the energy function: (1) basin hopping technique (BHT), (2) parallel tempering, and (3) evolutionary strategies[9].

Recently, our protein simulation package, POEM, has been modified to treat multiple polypeptide chains. In the present study, we simulated the systems of one and two chains of Aβ_{16-22} using a modified version of BHT[12]. In the simulations of a single chain, no specific unique structure was formed, in agreement with the earlier studies[4]. In contrast, the system of two chains formed an anti-parallel β-sheet in most of the cases. The lowest energy structure is shown in Fig. 1. The main reason for the antiparallel orientation has been much debated in the literature, the two options being the salt bridges[7] and the efficient sidechain packing[4].

Analysis of the simulations shows the overwhelming preference for the anti-parallel orientation. The free energy surface of the two-chain system is presented in Fig. 2. The lowest energy structures are the anti-parallel β-sheets. However, in the simulations the parallel orientation was also observed. The parallel β-structure with the lowest energy is shown in Fig. 1. The anti-parallel orientation is stabilized by the efficient sidechain packing as well as by the polar interactions between the sidechains of LYS_{16} and GLU_{22}.

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 parts of code that was used in our simulations. We are thankful to the Barcelona Supercomputer Center for computational resources.

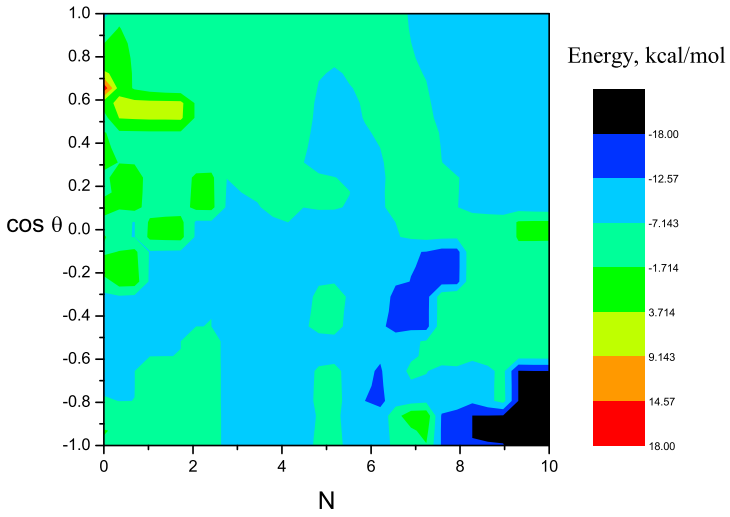


Figure 2. The free energy surface of the two-chain system. The co-ordinates are the number N of the residues in the β-conformation and the cosine of the angle θ between the two end-to-end vectors.

References

1. C. B. Anfinsen, *Principles that govern the Folding of Protein Chains*, Science **181**, 223, 1973.
2. J. J. Balbach and Y. Ishii and O. N. Antzutkin and R. D. Leapman and N. W. Rizzo and F. Dyda and J. Reed and R. Tycko, *Amyloid fibril formation by A beta 16-22, a seven-residue fragment of the Alzheimer's beta-amyloid peptide, and structural characterization by solid state NMR*, Biochemistry **39**, 13748, 2000.
3. C. M. Dobson, *The structural basis of protein folding and its links with human disease*, Phil. Trans. R. Soc. Lond. B **356**, 133, 2001.
4. G. Favrin and A. Irbäck and S. Mohanty, *Oligomerization of amyloid $A\beta_{16-22}$ peptides using hydrogen bonds and hydrophobicity forces*, Biophysical Journal **87**, 3657, 2004.
5. S. M. Gopal and W. Wenzel, *De Novo Folding of the DNA-Binding ATF-2 Zinc Finger Motif in an All-Atom Free-Energy Forcefield*, Angew. Chemie Int. **45**, 7726, 2006.
6. T. Herges and W. Wenzel, *An All-Atom Force Field for Tertiary Structure Prediction of Helical Proteins*, Biophysical Journal **87**, 3100, 2004.
7. D. K. Klimov and D. Thirumalai, *Dissecting the assemble of $A\beta_{16-22}$ amyloid peptides into antiparallel β sheets*, Structure **11**, 295, 2003.
8. T. Lührs and C. Ritter and M. Adrian and D. Riek-Loher and B. Bohrmann and H. Döbeli and D. Schubert and R. Riek, *3D structure of Alzheimer's amyloid-beta(1-42) fibrils*, Proc. Nat. Acad. Sci. **102**, 17342, 2005.
9. A. Schug, A. Verma, W. Wenzel, and G. Schoen, *Biomolecular structure prediction*

with stochastic optimization methods, Adv. Eng. Materials **7**, 1005, 2005.
10. A. Verma and S. Murthy and K. H. Lee and E. Starikov and W. Wenzel, *De novo all atom folding of helical proteins*, NIC Series **34**, 45, 2006.
11. A. Verma and W. Wenzel, *Towards a universal Free Energy Forcefield for All atom Protein Folding*, (Submited) 2007.
12. W. Wenzel, *Predictive folding of a β hairpin in an all-atom free-energy model*, Europhys. Letters **76**, 156, 2006.

Analysis and Optimization of the Flex-Screen Docking Approach Using DUD Benchmarking Database

Daria Kokh[1], **Bernhard Fischer**[2], **and Wolfgang Wenzel**[2]

[1] FB.C - Mathematik und Naturwissenschaften,
Bergische Universität Wuppertal, 42097 Wuppertal, Germany
E-mail: kokh@uni-wuppertal.de

[2] Institut für Nanotechnologie, Forschungszentrum Karlsruhe, GmbH,
Postfach 3640, 76021 Karlsruhe, Germany
E-mail: {*bernhard.fischer, wenzel*}@int.fzk.de

Screening performance of the all-atom Flex-Screen docking approach including receptor flexibility is investigated by using of a "directory of useful decoys", DUD. DUD is a bias-corrected benchmarking database based on 40 different target proteins, where each receptor is associated with a native ligand, a set of annotated ligands, and a set of decoy molecules that are unlikely to be binders. The docking performance is evaluated using two criteria: 1) geometrical fidelity of the docked poses compared to those of the experimental structures; 2) enrichment of annotated ligands among their decoys, which shows the ability of the docking calculations to distinct between true positives and nonbinder molecules with the same physical properties. Based on these results the scoring function and the receptor side-chain rearrangement procedure are optimized.

1 Method

FlexScreen[1] is an all-atom docking approach based on the stochastic tunneling method of the energy minimization and a simple atomistic scoring function that contains a sum of the Van-der-Waals, electrostatic, hydrogen-bond and salvation energies. The VdW and hydrogen-bond parameters are taken from OPLSAA[2] and AutoDock[3], respectively, the partial charges of the receptors are computed with MOE, and the atomic salvation parameters are optimized as described below. The method enables rotation up to 15 side-chain bonds of the receptor.
Scoring performance of the Flex-Screen approach is benchmarked by using of the DUD database[4] based on 40 target proteins of different classes with available ligand-bound X-ray crystal structures. For each protein the database includes: 1) crystal structures of the receptor and its native ligand; 2) a set of the annotated ligands that should in principle dock well (15-450 molecules); 3) a set of the decoys (about 36 molecules for each annotated ligand) that resemble the particular ligand in physical properties, but differ from the ligand topologically, so that they are likely to be nonbinders.

2 Results

2.1 Optimization of the Salvation Energy Parameters

Salvation energy is described as a sum of energies for the individual atoms that are assumed to be proportional to the solvent accessible surface area and atomic salvation parameter, ASP[5]. All atoms are divided into two groups: those responsible for 1) hydrophobic and

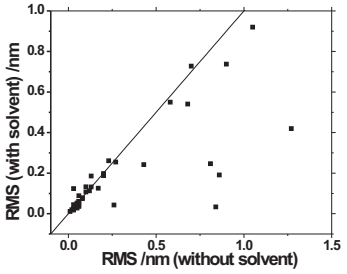


Figure 1. RMS deviation of the docked poses from the crystal ones for optimized ASPs versus the same values computed without salvation energy. 40 protein-ligand structures are included.

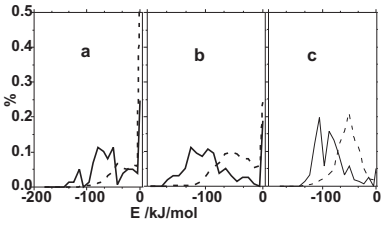


Figure 2. Distributions of annotated ligands in percentage (solid lines) and decoys (dashed lines) of the androgen receptor plotted as a function on their binding energies: (a) receptor is rigid; (b) 10 receptor side chains are flexible; (c) 15 receptor residues are shifted by 0.5 nm away from the cavity centre.

2) hydrophilic effects, so that only two ASPs have to be optimized. As an optimization criterion we use the sum of the RMS deviations of the docked conformations of the native ligands from the experimental ones.

Inclusion of the salvation energy in the scoring function in general improves docking poses for most of the ligands (see Fig.1), although some of them (10%) fail to find correct conformation regardless salvation effects and, therefore, need additional analysis.

2.2 Receptor Side Chain Rearrangement

Docking screen of the annotated ligand sets has shown that in some cases binding mode cannot be found because of high-energy clashes between protein and ligand atoms arising from the VdW term. To improve docking efficiency we enable receptor rearrangement by using of two approaches:1) rotation of up to 15 receptor side-chain bonds, 2) shift (by about 0.25-0.5 nm) of the receptor residues that are involved in clashes away from the binding pocket centre. Fig.2 demonstrates how both methods influence docking efficiency of annotated ligands and their decoys.

It is important to note, that receptor flexibility either improves or at least does not change scoring performance of the docking method. Although both approaches help to reduce a number of nondocking ligands and increase absolute value of the binding energy, the second method has been found to be usually more effective and notably less expensive.

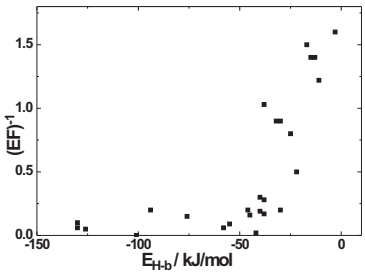


Figure 3. EF^{-1} for 28 receptors plotted as a function of the average energy of hydrogen bonds formed by docked ligands.

2.3 Enrichment of Annotated Ligands Among their Own Decoys

Efficiency of the docking method selectivity is estimated by computing of the inverse enrichment factor, EF^{-1}, for each receptor, described as a relation of the top-scoring decoys to the top-scoring annotated ligands (expressed as a percentage of the total number of the decoys and ligands, respectively). "top-scoring" molecule means that the absolute value of its binding energy is larger than at least 80% of that for the corresponding native ligand. A small value of EF^{-1} shows that annotated ligands dock notably better than their decoys for the particular receptor, whereas $EF^{-1} > 1$ indicates that the docking approach cannot distinguish between molecules with similar physical but different chemical properties.
Docking results of the preliminary calculations without salvation effects, summarized in Fig.3, indicate that: 1) the screening performance of the method is quite effective for about 70% of the targets, and 2) the docking efficiency depends strongly on the hydrogen-bond energy, e.g. essentially all molecules can find appropriate binding mode by unspecific interaction. Since most of the receptors with large value of EF^{-1} have open binding pockets, we expect that inclusion of the salvation energy in the scoring function will improve docking performance in these cases. These calculations are in progress.

References

1. H. Merlitz and W. Wenzel, *High Throughput in-silico Screening against Flexible Protein Receptors* , Lecture Notes in Computer Science **465**, 3045, 2004.
2. W. L. Jorgensen and N. A. McDonald, *Development of an all-atomic force field for heterocycles properties of liquid pyridine and diazenes* , J. Mol. Struct. **424**, 145, 1997.
3. G. M. Morris, D. S. Goodsell, R. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, *Automated docking using a Lamarckian genetic algorithm and an empirical binding energy function* , J. Comptut. Chem. **19**, 1639, 1998.
4. N. Huang, B. K. Shoichet, and J. J. Irwin *Benchmark sets for molecular docking* , J. Med. Chem. **49**, 6789-6801, 2006.
   `http://blaster.docking.org/dud/`
5. D. Eisenberg and A. D. McLachlan, *Comparison of of stochastic optimization methods for receptor-ligand docking* , Nature **319**, 199-203, 1986.

Folding and Structure Prediction of Proteins Containing Disulfide Bridges

Ivan Kondov[1], **Abhinav Verma**[2], **and Wolfgang Wenzel**[2]

[1] Institute for Scientific Computing, Forschungszentrum Karlsruhe,
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: ivan.kondov@iwr.fzk.de

[2] Institute for Nanotechnology, Forschungszentrum Karlsruhe,
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

A computational study of the proteins 1KVG and 1WQE containing disulfide bridges is presented. The simulation method bases on the protein force field PFF02 and the basin hopping technique. Both proteins were folded correctly from extended conformations with inclusion of a constraining potential.

1 Introduction

Computational prediction of tertiary structure of proteins with high accuracy on the basis of the primary structure requires development of transferable protein force fields as well as powerful optimization methods. Particularly, proteins with disulfide bridges connecting cysteine residues represent a major challenge for computational biophysics. Recently, plenty of proteins have been folded to their native conformations within experimental resolution using all-atom protein force fields and stochastic optimization methods[1–3]. However, there are only few works that study the folding behavior of proteins with disulfide bonds, e.g., by means of molecular dynamics[4,5], conformational space annealing with a united-residue force field[6], lattice models[7], topology-based approach[8], distance geometry[9], neural networks[10] and the island model[11]. In this paper, we report results of protein folding simulations for proteins containing disulfide bridges: the β-hairpin 1KVG and the potassium channel blocker 1WQE. In particular, we will focus on an approach with inclusion of additional binding potential to the free energy of the protein.

2 Methods

All-atom force fields PFF01[1,2] and PFF02[12] have been developed recently to describe the internal free energy of proteins. Besides α-helical proteins, the PFF02 enables correct description of β-sheet regions and allows unbiased comparison of results for proteins of the two types. To search for the free energy minimum, corresponding to the protein native conformation, the basin hopping technique is employed as outlined in previous work[3].

To promote formation of disulfide bridges a constraining Morse potential $V_{\mathrm{SS}}(r) = -E_0 \left[(1 - e^{-\beta(r-r_0)})^2 - 1\right]$ was considered, where r_0 is the equilibrium distance between the sulfur atoms of cysteine residues forming a disulfide bridge, E_0 is the energy corresponding to r_0, and β is the spacial extent of the potential. The choice of Morse potential is motivated by the fact that disulfide bridges are covalent bonds. In what follows, $r_0 = 2$ Å and $\beta = 1$ Å^{-1}.

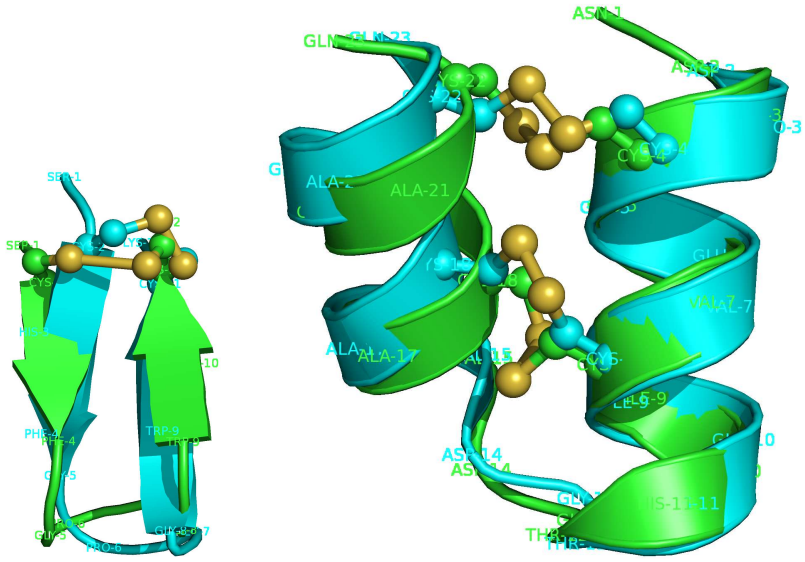


Figure 1. Cartoons of the folded proteins 1KVG (left) and 1WQE (right), shown in green, superimposed on the corresponding native structures shown in cyan. Sulfur atoms of cysteine residues forming disulfide bridges are depicted as spheres in golden color.

3 Results and Discussion

Simulations of thirty independent runs were performed for different values for E_0 starting from the extended structure, i.e. all backbone dihedral angles were set to 180°. Prediction accuracy of the results was assessed by comparing the simulated structures with natural structures determined by NMR. This comparison can be visualized overlaying the simulated structure on the reference, as shown in Fig. 1. In addition, the root mean square deviation of the backbone (RMSDb), the disulfide bond lengths and secondary structure contents are summarized in Table 1.

The α-helical 1WQE could be folded correctly to near-native conformations with and without constraining potential. However, the use of constraints increases the efficiency of the stochastic search for the global minimum significantly[13]. The other protein, 1KVG, could be folded to native conformation only in presence of the constraining potential. Turning off the constraints, the simulation of 1KVG ends up in all runs with an unstructured coil conformation. As can be seen in Table 1 the inclusion of constraining potential stabilizes the native β-sheet structure. For both proteins, the larger values for E_0 tighten the disulfide bonds while smaller values yield better overall structure.

4 Concluding Remarks

A constraining Morse potential was adopted to take into account disulfide bonds in proteins. For all proteins studied, inclusion of the constraining potential resulted in improved

E_0 [kcal/mol]	RMSDb	Res. Nr / r_{SS} [Å]		Sequence / secondary structure
1KVG		2-11		SCHFGPLGWVCK
natural	—	2.1		CEEEETTEEEEC
0	2.1	3.3		CBCBTTTBSCBC
2†	2.1	2.9		CEEEETTEEEEC
5	2.3	2.7		CEEESSSSEEEC
1WQE		4–22	8–18	NDPCEEVCIQHTGDVKACEEACQ
natural	—	2.0	2.0	CCHHHHHHHHHTCCHHHHHHHHC
0	2.1	3.3	5.5	CHHHHHHHHHHTCCHHHHHHHHC
2†	1.9	2.8	2.9	CHHHHHHHHHHTCCHHHHHHHHC
5	4.4	2.7	2.7	CHHHHHHSCSSTTTCHHHHHHHC

Table 1. Characteristics of the folded conformations in comparison with the natural conformations. In columns 3 and 4 the distances r_{SS} between the sulfur atoms of specified cysteine residues are given. Cases denoted by † are shown in Fig. 1.

RMSDb values and distances between cysteine sulfur atoms compared to constraint-free simulations. It was demonstrated that the constraining potential can be decisive for correct folding and prediction of the three-dimensional structure.

Acknowledgments

Grant of computing time by project CampusGrid at the Research Center Karlsruhe is gratefully acknowledged.

References

1. T. Herges and W. Wenzel. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.*, 87:3100, 2004.
2. T. Herges and W. Wenzel. In silico folding of a three helix protein and characterization of its free-energy landscape in an all-atom force field. *Phys. Rev. Lett.*, 94:018101, 2005.
3. A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. *J. Chem. Phys.*, 124:044515, 2006.
4. M. Qin, J. Zhang, and W. Wang. Effects of disulfide bonds on folding behavior and mechanism of the β-sheet protein tendamistat. *Biophys. J.*, 90:272, 2006.
5. M. A. Martí-Renom, R. H. Stote, E. Querol, F. X. Avilés, and M. Karplus. Refolding of potato carboxypeptidase inhibitor by molecular dynamics simulations with disulfide bond constraints. *J. Mol. Biol.*, 284:145, 1998.
6. C. Czaplewski, S. Oldziej, A. Liwo, and H. A. Scheraga. Prediction of the structures of proteins with the unres force field, including dynamic formation and breaking of disulfide bonds. *Protein Engineering, Design & Selection*, 17(1):29, 2004.
7. V. I. Abkevich and E. I. Shakhnovich. What can disulfide bonds tell us about protein energetics, function and folding: Simulations and bioninformatics analysis. *J. Mol. Biol.*, 300(4):975, 2000.

8. C. Micheletti, V. de Filippis, A. Maritan, and F. Seno. Elucidation of the disulfide-folding pathway of hirudin by a topology-based approach. *Proteins*, 53:720, 2003.
9. E. S. Huang, R. Samudrala, and J. W. Ponder. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol.*, 290(1-2):267, 1999.
10. S. M. Muskal, S. R. Holbrook, and S.-H. Kim. Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng.*, 3(8):667, 1990.
11. Y. Kobayashi, I. Sasabe, T. Akutsu, and N. Saito. Mechanism of protein folding. iv. forming and breaking of disulfide bonds in bovine pancreatic tripsin inhibitor. *Biophys. Chem.*, 44:113, 1992.
12. A. Verma and W. Wenzel. Towards a universal free-energy approach for all-atom protein folding and structure prediction. Submitted.
13. I. Kondov, A. Verma, and W. Wenzel. In preparation.

DNA Packaging and Electrostatic Interactions

Nikolay Korolev and Lars Nordenskiöld

School of Biological Sciences, Nanyang Technological University,
60 Nanyang Drive, 637551 Singapore
E-mail: korolev@ntu.edu.sg

In cell nucleus, all DNA exists as a highly packed nucleoprotein complex called chromatin. DNA is highly negatively charged polyelectrolyte and as consequence DNA-DNA repulsion must be overcome to form chromatin. The first level of DNA packaging is the nucleosome core particle (NCP), composed of 150 bp of DNA wrapped around a histone octamer core and flexible positively N-terminal domains (histone tails) protruding out from the NCP. Linear arrays of the NCP are further folded into higher level chromatin structures with the histone tails playing important roles in their formation. Theoretical models from Poisson Boltzmann (PB) approximation to all-atom molecular dynamics simulations are used to describe chromatin/nucleosome statics and dynamics. Applying the PB theory we show that free energy of NCP formation is extremely favorable that challenges established opinion about marginal stability of the NCP. To describe the influence of the histone tails on aggregation and dynamics of the NCPs, we carried out molecular dynamics simulations with coarse-grained approximation of the NCP. Our results are in good agreement with experimental neutron scattering and X-ray diffraction data. To reveal molecular details of the histone tail-DNA binding and dynamics, all-atom MD simulations were undertaken in a system comprising several DNA oligomers and fragments of the histone tails. Correlation between DNA-DNA distance and binding the histone tails to DNA is clearly observed. At the same time, binding of the tails does not restrict internal dynamics of the DNA.

In eukaryotic cell, long DNA (roughly 2 m in humans) is packed about 400,000 times inside a cell nucleus of 10-20 μm in diameter. DNA packaging is performed by nuclear proteins and this specific nucleoprotein complex called chromatin performs contradictory functions keeping DNA compact and protected while remaining selectively accessible and dynamic. About 85% of DNA in chromatin is represented by uniform units, the nucleosomes, which are the complex of 160-230 base pairs (bp) of DNA double helix with five histone proteins (H2A, H2B, H3, H4, and H1). The most regular central part of the nucleosome is called the nucleosome core particle (NCP) and consists of 147 bp DNA wrapped as a 1.75 turn superhelix around the histone octamer formed from one $(H3/H4)_2$ tetramer and two H2A/H2B dimers.[1,2] Variable length (10-80 bp) of the DNA (called linker DNA) connects the NCPs to each other and binds to the linker histone H1. Linear arrays of the NCP are further folded into higher level chromatin structures with the histone tails playing important roles in their formation. The histones are responsible for the first step of DNA compaction in chromatin and pose a major obstacle for direct access to the DNA of proteins responsible for DNA replication, transcription, repair, and recombination. Variable length double stranded linker DNA connects the NCPs with each other to form nucleosomal arrays which condense into the 30-nm chromatin fibers.[3,4] Each of the core histones has unstructured N-terminal domain called "histone tail" protruding through the DNA superhelix. The tails are able to interact with DNA of the NCP, linker DNA and variety of nuclear proteins. The structure of the histone tails largely escapes detection by X ray crystallography and by other experimental methods, implying that they are highly flexible and dynamic. The histone tails are essential for maintenance of the higher order compact folded structures of chromatin and for regulation of transcription and replication.[2,5] These

functions of the histone tails are regulated by covalent modifications of the amino acids, which may change the net charge and distribution of the charged groups in the tails.

DNA is highly negatively charged polyelectrolyte and therefore huge repulsive force between the DNA molecules must be overcome to form chromatin. Positively charged amino acids Lys^+ and Arg^+ of the "histone-fold" domains form a distinct charged surface which direct DNA wrapping in the NCP. The histone tails are highly basic. One might imagine that huge forces of electrostatic origin are involved in all transitions of chromatin. However, the role of electrostatic interactions contributing to chromatin/nucleosome structure and dynamics, has received less attention compared to the other phenomena (e.g. DNA bending properties, topological problems related to the DNA unwinding and twisting). Our work analyses and models contribution of electrostatic forces to formation of the nucleosome and ability of the nucleosomes to aggregate and mediation of the charged groups of the histone tails in close DNA-DNA contacts.

1 Estimation of the NCP Stability Using Poisson Boltzmann Polyelectrolyte Model

In the process of unwrapping DNA from histones two strongly charged polyelectrolyte entities are formed: the negatively charged DNA and positively charged histone core. Electrostatic interactions arising in such a case must have a profound influence on the course of whole process and on equilibrium between the two phases. In a single-molecule experiments,[6,7] the electrostatic forces resist the detachment of the DNA from the histones during the single-chromatin fiber stretching under the influence of a mechanical pulling force. The degree of this resistance is dependent not only on the strength of the individual histone core – DNA contacts but also on the ionic conditions in the solution as well as on the balance between the positively and negatively charged groups in the chromatin. Other important contributions to the detachment process, the mechanical bending force and specific short-range DNA-histone interactions, are not dependent on the ionic conditions. Despite the fact that the electrostatic interactions must have a significant influence on DNA unwrapping, their role is rarely discussed in the literature. Recently, we analyze the electrostatic component of the DNA-histone interactions within the NCP.[8,9] Using the mean theory (Poisson Boltzmann) approximation, the contribution of the electrostatic forces to a formation of the elementary structural unit of chromatin, the nucleosome was estimated. Results of the modeling were discussed in relation to the recent DNA stretching experiments.[6,7]

Figure 1 shows contribution of the electrostatic forces involved in gradual removal of the DNA from the histone proteins in chromatin.[8,9] Results from the PB theory challenge an established opinion that the free energy of the NCP formation is small and that the forces related to the bending of the DNA double helix plays make the major contribution to the NCP energetics. Instead, calculations within the PB model reveal that of the electrostatic free energy of the NCP formation is extremely favorable (see Fig. 1). Remarkably, simple polyelectrolyte model is in a semi-quantitative agreement with recent optical tweezers stretching experiments measuring the force necessary to unwrap DNA from the histone core.[6,7] Our analysis shows that the electrostatic interactions between the highly negatively charged polymeric DNA and the positively charged histones play a determining role in stabilizing the nucleosomes at physiological conditions.

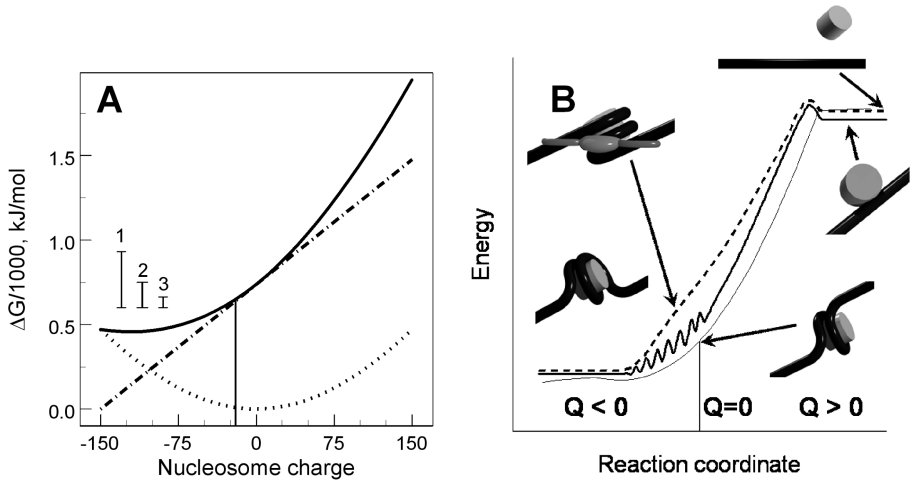


Figure 1. (A) Estimation of electrostatic free energy of the NCP dissociation. Degree of the nucleosome unwinding (abscissa) is measured as change of the total charge of the NCP (histone octamer + bound DNA). Dash-dotted line shows the free energy contribution from DNA, released from the octamer; dotted line is the electrostatic free energy of the NCP, approximated as a sphere of 50 Å radius; solid line is a sum of the two terms. Vertical bar shows the position when one full turn of DNA is wrapped around the histone core. PB cell model has been applied with conditions modeling chromatin stretching experiments. Vertical bars 1-3 are estimations of the DNA bending energy given in the literature and are drawn to highlight the importance of the electrostatic interactions relative to the DNA bending. (B). Change of the energy of the NCP in the process of the chromatin fiber unwinding. The thin line displays a hypothetical equilibrium DNA unwrapping from the histone core; solid and dashed curves are estimations of energy passes in single-fiber-pulling experiments from refs. and, respectively. Cartoons illustrate the NCP state at different stages of the DNA fiber stretching. See ref. 8 for details.

2 Coarse-Grained Molecular Dynamics for Salt-Dependent NCP-NCP Interaction

Molecular dynamics (MD) computer simulations of charged histone tail–DNA interactions in systems mimicking nucleosome core particles (NCP) have been conducted using coarse-grained model of the NCP.[10] The nucleosome is approximated as a negatively charged spherical particle with flexible polycationic histone tails attached to it in a dielectric continuum with explicit mobile counterions and added salt. The size, charge and distribution of the tails relative to the core were built mimicking real NCP. In this way, attractive ion-ion correlation effects due to fluctuations in the ion cloud and the attractive entropic and energetic tail bridging effects were incorporated. MD simulations in a dielectric continuum model containing explicit mobile counterions and various amount of added salt describe the effect of changing experimental conditions were carried out.[10] In agreement with experimental data,[11,12] increase of monovalent salt content from salt-free to physiological concentration leads to formation of NCP aggregates; and likewise in the presence of $MgCl_2$ the NCPs form condensed systems via histone tail bridging and accumulation of counterions (Fig. 2). The simulations give insight into the tail mediated bridging between core particles and are also of relevance for the mechanism of secondary and tertiary condensation of nucleosomal arrays. To our knowledge this is the first theoretical demonstration of

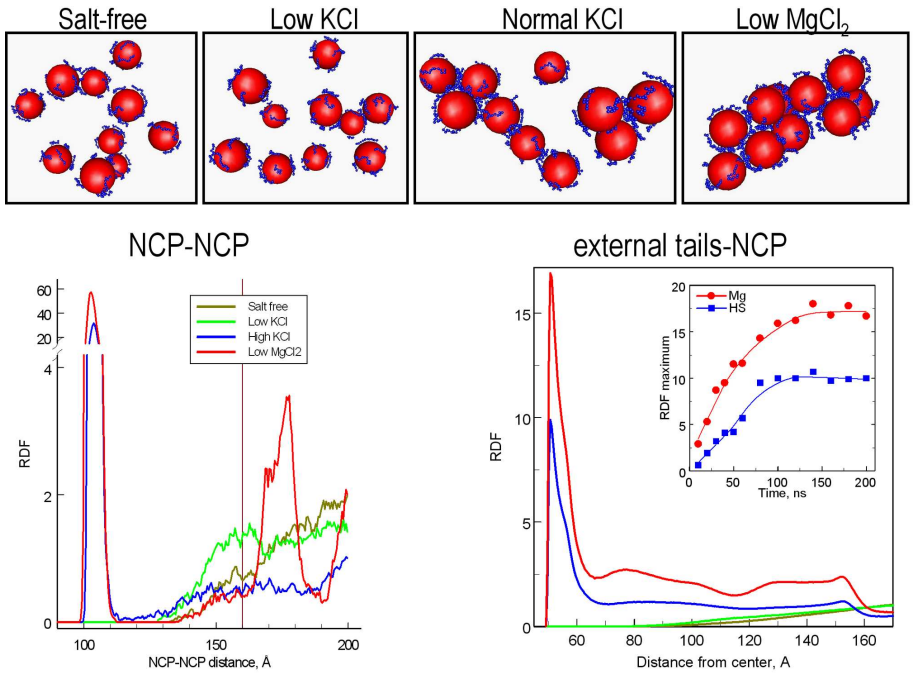


Figure 2. Top: Snapshots showing distribution of the NCP particles at the end of the four 200 ns coarse-grained MD simulations with increasing concentration of K^+ (3 cartoones to the left) and in the presence of Mg^{2+} (right). Bottom: RDFs calculated from the last 25% of the four coarse-grained MD simulations. Left: NCP-NCP correlation function. Maxima between 100-110 Å indicate NCP-NCP aggregation. Right: Intermolecular tail-NCP RDF. Appearance of maximum below 60 Å reveals crosslinking of the histone tails in the aggregated NCPs. Insert displays the dependence of the maximum of the RDF on simulation time. See ref. 11 for details.

nucleosome-nucleosome tail bridging attraction within a statistical mechanical treatment which (within the model) gives the equilibrium description (NVT ensemble simulations) for a system whose parameters model the NCP and which explicitly includes all charged mobile particles taking into account salt dependent ion-ion correlation and flexible charged tails.

3 All Atom MD Simulation of DNA-DNA Attraction Mediated by the Histone Tails

To investigate atomic details of the histone-tail mediated DNA-DNA interaction, two series of all-atom MD simulations were performed during 30-50 ns for a system of 3-4 identical DNA 22-mers, 14-20 short fragments of the charged H4 histone tail peptide fragments (amino acids 5-12, $[KGGKGLGK]^{3+}$,[10] and 13-19, $[GGAKRHRK]^{4+}$ (unpublished work)) with K^+ counterions and explicit water. The simulation setup mimics the crowded conditions of DNA in eukaryotic chromatin. To assess the influence of tail fragments on DNA structure and dynamics, 'control' MD simulations were carried for systems with the

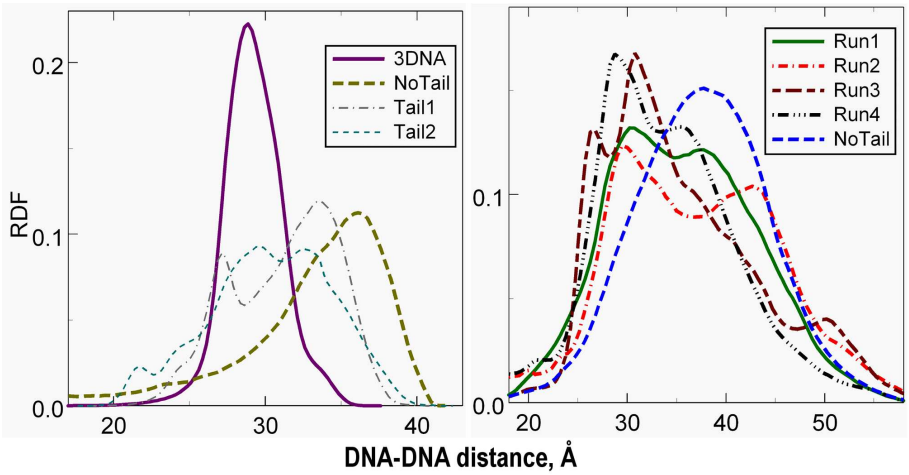


Figure 3. Distribution of the average DNA-DNA distance in MD simulations with 3 (left) and 4 (right) DNA 22-mers in the simuation cell. Short-dashed lines show DNA-DNA distance in the absence of the H4 histone tail fragments. Thin dotted lines (Tail1, Tail2) in the left graph display DNA-DNA distance in the presence of the 5-12 a.a. $[KGGKGLGK]^{3+}$ fragment. Solid line in the left graph ("3DNA") and Run1-Run4 curves in the right graph show DNA-DNA distance in the presence of the 13-19 a.a. $[GGAKRHRK]^{4+}$ fragments. In the absence of the tail fragments DNA-DNA distance is determined by the dimension of the simulation cell; presence of the histone tail fragments makes DNA-DNA distance 8-10 Å shorter.

same DNA and water content but in the absence of oligopeptides. DNA structure and dynamics, DNA-DNA interaction and its interplay with the histone tail fragments binding are described.[10,13] MD simulations allow capturing typical features of the histone tail-counterion-DNA structure, interaction and dynamics.

Binding of the tail fragments to the DNA is dynamic. The charged side chains of the lysines and arginines play a major role in mediating DNA-DNA attraction by forming bridges and coordinating to phosphate groups and electronegative sites in the minor groove. Correlation of the DNA-DNA distance with the presence and association of the histone tail between the DNA molecules is observed. Figure 3 compares distance between DNA molecules in the absence (dashed lines with maximum at 35-40 Å) and in the presence of the tails (all other data). In the systems without tails DNA molecules tend to distribute as far as possible from each other. In the systems with the histone tail fragments, the most populated DNA-DNA distances are in the range 24-34 Å (Fig. 3).

Detailed investigation of the histone tail-DNA interaction shows that tail bridging is performed through numerous contacts of Lys^+ and Arg^+ with the phosphate groups combined with penetration of the charged lysine side chains into the minor groove of the DNA. Experimental and computational work is in progress which investigates interplay between covalent modification of the histone tails and compaction of chromatin, NCP-NCP and DNA-DNA interaction.

References

1. K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, T. J. Richmond, *Crystal structure of the nucleosome core particle at 2.8 Å resolution*, Nature **389**, 251-260, 1997.
2. A. P. Wolffe, *Chromatin: Structure and function* 3rd Edition, pp. 447, Academic Press, San Diego, CA, 1998.
3. C. L. Woodcock, S. Dimitrov, *Higher-order structure of chromatin and chromosomes*, Curr.Opin.Genet.Dev. **11**, 130-135, 2001.
4. J. C. Hansen, *Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions*, Annu.Rev.Biophys.Biomol.Struct. **31**, 361-392, 2002.
5. P. J. Horn, C. L. Peterson, *Chromatin higher order folding: Wrapping up transcription*, Science **297**, 1824-1827, 2002.
6. B. D. Brower-Toland *et al.*, *Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA*, Proc.Natl.Acad.Sci.U.S.A. **99**, 1960-1965, 2002.
7. M. L. Bennink *et al.*, *Unfolding individual nucleosomes by streyching single chromatin fibers with optical tweezers*, Nature Struct.Biol. **8**, 606-610, 2001.
8. N. Korolev, A. P. Lyubartsev, A. Laaksonen, *Electrostatic background of chromatin fiber stretching*, J.Biomol.Struct.Dyn. **22**, 215-226, 2004.
9. N. Korolev, O. V. Vorontsova, L. Nordenskiöld, *Physicochemical analysis of electrostatic foundation for DNA-protein interactions in chromatin transformations*, Prog.Biophys.Molec.Biol. **in press**, 2007.
10. N. Korolev, A. P. Lyubartsev, L. Nordenskiöld, *Computer modeling demonstrates that electrostatic attraction of nucleosomal DNA is mediated by histone tails*, Biophys.J. **90**, 4305-4316, 2006.
11. S. Mangenot, E. Raspaud, C. Tribet, L. Belloni, F. Livolant, *Interactions between isolated nucleosome core particles. A tail bridging effect?*, Eur.Phys.J.E **7**, 221-231, 2002.
12. A. Bertin, A. Leforestier, D. Durand, F. Livolant, *Role of histone tails in the conformation and interaction of nucleosome core particles*, Biochemistry **43**, 4773-4780, 2004.
13. N. Korolev, L. Nordenskiöld, *H4 histone tail mediated DNA-DNA interaction and effects on DNA structure, flexibility and counterion binding. A molecular dynamics study*, Biopolymers **in press**, 2007.

Water Percolation Governs Polymorphic Transition and Conductivity of DNA

Ivan Brovchenko[1], **Aliaksei Krukau**[1], **Alla Oleinikova**[1], **and Alexey Mazur**[2]

[1] Physical Chemistry Department,
University of Dortmund, 44227 Dortmund, Germany
E-mail: {brov, alex, alla}@heineken.chemie.uni-dortmund.de

[2] CNRS UPR9080, Institut de Biologie Physico-Chimique,
13, rue Pierre et Marie Curie, Paris, 75005, France
E-mail: alexey@ibpc.fr

We report on the simulation studies of the percolation transition of water at the surface of the DNA double helix. At low hydrations, only small water clusters are attached to the DNA surface, whereas it is homogeneously covered by a spanning water network at high hydrations. Formation of a spanning water network occurs via a percolation transition at the hydration level close to the midpoint of polymorphic transition between A- and B-forms of DNA. This percolation transition results in sigmoid like acceleration of long-range ion transport in good agreement with experiment.

The presence of water at the surface of biomolecules is necessary for their conformational stability and functions. Dynamics and functions of biomolecules are often completely restored, when they are covered by about a monolayer of hydration water[1,2]. Formation of such monolayer occurs via quasi-2D percolation transition between two essentially different states of hydration water: ensemble of finite clusters and spanning hydrogen-bonded (HB) network of hydration water, which homogeneously envelopes a biomolecule[3,4]. The microscopic origin of the relationship between the percolation of hydration water and biological functions is unclear. One may expect that the appearance of a spanning network of hydration water strongly affects the dynamic properties of the hydrated biomolecule and the charge transport along the biological surfaces.

Water is crucial for the biological functions of the double helical DNA[5]. Experiments show that there is a strict relationship between the state of DNA and a hydration number Γ, measured as the number of water molecules per nucleotide. Biologically relevant B-DNA becomes unstable at hydrations $\Gamma < 20$ and midpoint of the transition between A- and B-forms of DNA is about $\Gamma = 15$.[5] In the same hydration range, conductivity DNA shows sigmoid like dependence on Γ.[6] To clarify the possible role of water percolation in the polymorphic transitions of DNA, we performed computer simulations of a rigid, dodecamer fragment of double helical DNA (CGCGAATTCGCG), fixed in space in canonical B-form, in the range of hydrations $12 \leq \Gamma \leq 30$.[7,8] The Cornell *et. al.* force field[9] was used for DNA and water was modeled by TIP3P potential.[10] The molecular dynamics (MD) simulations were carried out with the internal coordinate MD method adopted for DNA[11] with a time step of 0.01 ps. The standard duration of production runs was 10 ns with the coordinates saved every 1 ps. The clustering and percolation of water were analyzed by methods earlier developed in our group.[3,4,12–14]

The probability distribution $P(S_{max})$ of the size S_{max} of the largest water cluster on the B-DNA surface under different hydration levels is shown in Figure 1.A. The two peaks

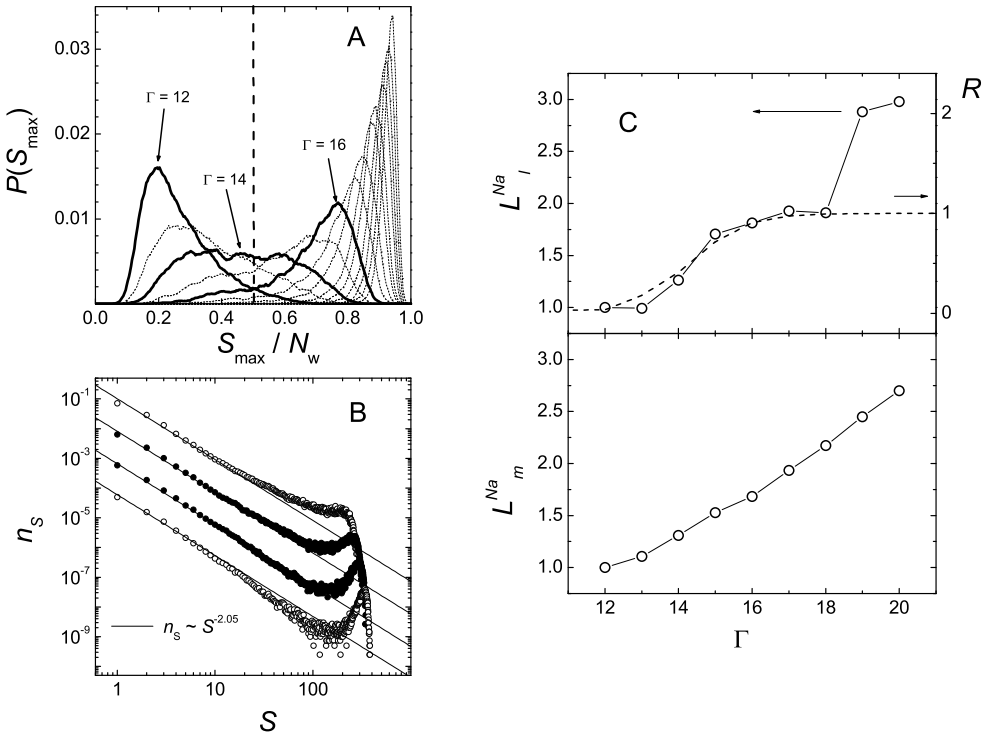


Figure 1. **A**: Probability distribution $P(S)_{max}$ of the size S_{max} of the largest water cluster on DNA surface with hydration levels Γ from 10 to 24. **B**: Distribution n_S of the size of water clusters at the surface of DNA molecule for representative hydration levels Γ from 14 (top) to 17 (bottom). The universal power law expected for 2D percolation transition is shown by solid lines. **C**: *Upper panel*: Relative mobility L_l^{Na} of ions, in time interval 200 ps $\leq$ t $\leq$ 500 ps (open circles, left axis), and percolation probability *R* of hydration water (dashed line, right axis) as function of the hydration level Γ. *Lower panel*: Dependence of the relative mobility L_m^{Na} of ions, in time interval 50 ps $\leq$ t $\leq$ 100 ps, on the hydration level Γ.

at small and large S_{max} corresponds to non-spanning and spanning HB networks, respectively. The most symmetric distribution $P(S_{max})$ is observed at a hydration level where spanning and non-spanning states of the largest cluster are equally populated, and this hydration corresponds to the midpoint of the percolation transition. The spanning probability $R(\Gamma)$ estimated from these distributions[7] is shown by dashed line in Figure 1.C (upper panel, right axes). The midpoint of the percolation transition, that is the position of the inflection point of $R(\Gamma)$ being fitted to sigmoid function, is close to $\Gamma \approx 14.3$. The true percolation threshold can be located by using the cluster size distribution n_S, which is a probability to find a cluster, containing *S* molecules. At the true percolation threshold, $n_S \sim S^{-2.05}$ in the widest range of *S*. HB water clusters at DNA surface follows this universal power law in the widest range of *S* at hydration levels $\Gamma \approx 15$ and 16 (closed symbols in Figure 1.B) and the true percolation threshold of water occurs at $\Gamma = 15.5 \pm 0.5$. At the threshold hydration, the minor groove of B-DNA is already filled with water, and the percolation transition results in formation of a spanning water network in the major

groove. The threshold hydration is strikingly close to the midpoint of the low hydration transition of DNA from the B-form to other forms.[5] This observation suggests that water percolation may be responsible for the low hydration polymorphism of DNA. So, DNA exists in biologically relevant B-form only when it is covered by a spanning water network.

To explore effect of water percolation on the mobility of Na^+ ions in the DNA hydration shell, we have analyzed time dependence of their mean square displacements $\left\langle \Delta r^2 \right\rangle (t)$. Due to confinement effect and due to heterogeneity of the DNA surface, such dependence is non-linear. We characterized ions diffusion by a dimensionless relative mobility $L(\Gamma)$ obtained as the slope of the linear fit of $\left\langle \Delta r^2 \right\rangle (t)$ in a given time interval normalized by the corresponding slope for $\Gamma = 12$. Mobility $L_m^{Na}(\Gamma)$ of Na^+ ions approximately linearly grows with hydration for short-time translations, $t \leq 100$ps (Figure 1.C, lower panel), but the relative mobility L_l^{Na}, evaluated in the time period 200 ps $\leq$ t $\leq$ 500 ps (Figure 1.C, upper panel), corresponding to the ion displacement up to 6.5 Å, behaves radically different. From hydration Γ = 12 up to Γ = 18, L_l^{Na} exhibits a sigmoid like dependence closely following the spanning probability $R(\Gamma)$. The overall increase of L_l^{Na} on passing the percolation threshold is about 100%. The step like increase of ion mobility at Γ = 18 may be attributed to the escape of ions from the DNA surface. Two step increase of the L_l^{Na} with hydration in our simulations remarkably agrees with the experimentally observed dependence of the radiation-induced conductivity of hydrated DNA fibers.[6] Ability of the spanning network of hydration water to provide ion transport along bio-surfaces may explain its crucial role in biological function.

References

1. J. A. Rupley, G. Careri, Adv. Protein Chem. **41**, 37–172, 1991.
2. I. D. Kuntz, W. Kauzmann, Adv. Protein Chem. **28**, 239–345, 1974.
3. A. Oleinikova, N. Smolin, I. Brovchenko, A. Geiger and R. Winter, J. Phys. Chem. B **109**, 1988-1998, 2005.
4. N. Smolin, A. Oleinikova, I. Brovchenko, A. Geiger and R. Winter, J. Phys. Chem. B **109**, 10995–11005, 2005.
5. W. Saenger, *Principles of Nucleic Acid Structure* (Springer-Verlag, Berlin, 1984).
6. J. M. Warman, M. P. de Haas and A. Rupprecht, Chem. Phys. Lett. **249**, 319–322, 1996.
7. I. Brovchenko, A. Krukau, A. Oleinikova, and A. K. Mazur, Phys. Rev. Lett. **97**, 137801, 2006.
8. I. Brovchenko, A. Krukau, A. Oleinikova, and A. K. Mazur, J. Phys. Chem. B **111**, 3258–3266, 2007.
9. W. D. Cornell et. al., J. Am. Chem. Soc. **117**, 5179–5197, 1995.
10. W. L. Jorgensen, J. Chandreskhar, J. D. Madura, R. W Impey and M. L. Klein, J. Chem. Phys **79**, 926–935, 1972.
11. A. K. Mazur, J. Am. Chem. Soc. **124**, 14707–14715, 2002.
12. I. Brovchenko, A. Krukau, N. Smolin, A. Oleinikova, A. Geiger and R. Winter, J. Chem. Phys. **123**, 224905, 2005.
13. A. Oleinikova, I. Brovchenko, N. Smolin, A. Krukau, A. Geiger and R. Winter, Phys. Rev. Lett. **95**, 247802, 2005.
14. A. Oleinikova, I. Brovchenko, Mol. Phys **104**, 3841-3855, 2006.

Molecular Dynamics in Excited States: Landau-Zener Model of Nitric Oxide Geminate Recombination to Nitrile Hydratase

Karina Kubiak and Wieslaw Nowak*

Institute of Physics, N. Copernicus University,
ul. Grudziadzka 5, 87-100 Torun, Poland
**E-mail: wiesiek@fizyka.umk.pl*

Molecular dynamics (MD) simulations on two coupled electronic surfaces are employed to investigate the geminate recombination of nitric oxide to photoactive biotechnological enzyme nitrile hydratase (NHase). NHase enzymatic activity is triggered by photodissociation of NO molecule[1,2]. The crossing between the ground and the excited state surfaces is treated using the Landau-Zener model[3,4]. The NO geminate recombination curve and recombination rate were calculated. Results suggest that the NO recombination is a picosecond time-scale process.

1 Introduction

The microbial enzyme nitrile hydratase (NHase, EC. 4.2.1.84, CAS registration no. 82391-37-5) is widely used in biotechnological production of amides[5,6] from nitriles. Out of two types of NHases (Fe and Co-type) only the Fe-type NHase is photosensitive. This enzyme loses activity in the dark conditions when the endogenous nitric oxide molecule (NO) blocks the active centre. This process is totally reversible and the enzyme recovers catalytic ability upon light irradiation[1,2]. The electronic mechanism of NO controlled photoactivity remains unknown, despite several model systems of NHase have been investigated[7,8] and theoretical models have been calculated[9–14]. In the earlier study of the photosensitivity phenomenon Nowak *et al.*, described structural changes upon NO ligand binding to the iron center ("inverted doming")[9]. The MD calculations of NHase gave an insight into the channel dynamics[15], however the kinetics of recombination is still mysterious.

In this paper the NO recombination kinetics rate is theoretically calculated for the protonated enzyme model. Similarly to Li *et al.*[16] classical Landau-Zener (LZ) MD model of recombination is employed.

2 Computational Methods

In the LZ method the system may evolve on two alternative, ground and excited state crossing hypersurfaces. At the crossing point, the LZ probability P is calculated (1) and the decision whether to switch the energy surface is made[3,4]

$$P = e^{-\frac{4\pi\varepsilon_{12}^2}{\hbar V|s_1 - s_2|}}, \tag{1}$$

where V is a velocity, ε_{12}^2 is energy difference between two energy states and $|s_1 - s_2|$ is a difference in the energy curves slopes.

All computations were performed for Fe-type NHase (2AHJ) in the MOIL package[16,17]. The parametrization of the NHase active site is described elsewhere[15]. Only crystallographic water molecules were present. Initially minimized system (the enzyme with NO molecule bounded via Morse potential, 3000 steps) was heated up to 300 K (50 ps) and equilibrated for 50 ps. From short 100 ps run 6 random structures were chosen as starts for production simulations of the NO recombination. For each start a 10 ps trajectory at 250, 275, 300, 325 and 350 K was generated. In each of 30 trajectories NO molecule was photodissociated (by excitation) at the very beginning. The SHAKEB protocol was used, data were collected in each 1 fs step. The temperature was held constant by velocity scaling, cutoffs for electrostatic and van der Waals interactions were 12 Å and 9 Å, respectively. The nonbonding interactions were recalculated in each step.

The excited state curve $U(r)$ was modeled by (2) and the binding potential $B(r)$ was approximated by the Morse function (3) (see Fig. 1):

$$U(r) = A_{rep}e^{-\beta r} - B_{rep}, \tag{2}$$

and

$$B(r) = D_{mor}e^{-2\alpha(r-r_{eq})} - 2e^{-\alpha(r-r_{eq})}. \tag{3}$$

Parameter r_{eq} is the equilibrium Fe–N(NO) bond length of 1.65 Å, parameters D_{mor}, α, A_{rep}, β and B_{rep} were 30 kcal/mol, 2 Å^{-1}, 80 kcal/mol, 1 Å^{-1} and 4 kcal/mol, respectively. Other parameters required for excited state MD were adopted from[16].

3 Results and Discussion

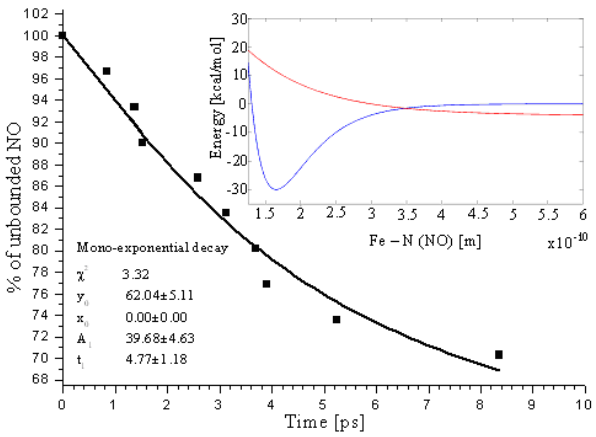


Figure 1. Kinetics of NO recombination to NHase active site. Points indicate a percent of recombined ligands. A mono-exponential decay $y = y_0 + A_1 \exp\left(-\frac{x-x_0}{t_1}\right)$ is fitted. **Inset**: Potential curves used in simulations of NO recombination.

For each of 10 ps non-equilibrium trajectories on excited state energy surface the Fe–N(NO) distance, the repulsion energy value (E_{rep}), the switching moment from repulsion

to Morse potential and a value of the Morse potential ($\mathrm{E_{morse}}$) were monitored. If the distance Fe–N(NO) was smaller than 2.5 Å or if the $\mathrm{E_{morse}}$ was smaller than -10 kcal/mol, we assumed that the recombination happened. A percent of recombinated ligands vs time gave data for estimation of recombination rate (see Fig. 1).

We assume that these data points are the best described by a mono-exponential decay since that assumption gives the smallest value of the χ^2 ($\chi^2 = 3.3$). Two and three exponential decays fitted to our MD data gave χ^2 of 4.7 and 8.7, respectively and times t_2 and t_3 were extremely long (data not shown). Fitted parameters (Fig. 1) indicate that the average lifetime of the free NO is 4.8 ps (the decay rate constant is 0.21 ps^{-1}). NO collisions with βArg56 or (less frequently) with αGln90 induced the recombination. Other collisions with the active site-solvent channel walls didn't result in the rebinding.

4 Conclusions

For the first time a geminate recombination of the NO molecule to Fe-type NHase active site was studied using the Landau-Zener approach and MD in excited state. We found that there is perhaps one energy barrier (4 kcal/mol) and the recombination rate is characterized by 4.8 ps lifetime of free NO. Residues βArg56 and αGln90 are critical for NO recombination. Better statistics is required for more quantitative estimates of NHase photophysics.

Acknowledgments

This research was supported by Polish Ministry of Education and Science, grant no. 2P04A 07229 and by UMK grant no. 377-F.

References

1. T. Noguchi, J. Honda, T. Nagamune, H. Sasabe, Y. Inoue, I. Endo, *Photosensitive nitrile hydratase intrinsically possesses nitric oxide bound to the non-heme iron center: evidence by Fourier transform infrared spectroscopy*, FEBS Letters **358**, 9 (1995).
2. M. Odaka, K. Fuji, M. Hoshino, T. Noguchi, M. Tsujimura, S. Nagashima, M. Yohda, T. Nagamune, Y. Inoue, I. Endo, *Activity regulation of photoreactive nitrile hydratase by nitric oxide*, J. Am. Chem. Soc., **119**, 3785 (1997).
3. L. Landau, Z. Phys. Sov., **2**, 46, (1932).
4. C. Zener, Proc. Roy. Soc. London, Sect. A (Math. Phys. Sci.), **137**, 696, (1932).
5. M. Kobayashi, T. Nagasawa, H. Yamada, *Enzymatic synthesis of acrylamide: a success story not yet over*, Tibtech, **10**, 402 (1991).
6. H. Yamada, S. Shimizu, M. Kobayashi, *Hydratases involved in nitrile conversion: screening, characterization and application*, Chemical Record **1**, 152 (2001).
7. H. L. Jackson, S. C. Shoner, D. Rittenberg, J. A. Cowen, S. Lovell, D. Barnhart, J. A. Kovacs, *Probing the influence of local coordination environment on the properties of Fe-type nitrile hydratase model complexes*, Inorg. Chem. **40**, 1646 (2001).
8. J. Shearer, H. L. Jackson, D. Schweitzer, D. K. Rittenberg, T. M. Leavy, W. Kaminsky, R. C. Scarrow, J. A. Kovacs, *The first example of a nitrile hydratase model complex that reversibly binds nitriles*, J. Am. Chem. Soc. **124**, 11417 (2002).

9. W. Nowak, Y. Ohtsuka, J. Hasegawa, H. Nakatsuji, *Density Functional Study on Geometry and Electronic Structure of Nitrile Hydratase Active Site Model*, Int. J. Quantum Chem. **90**, 1174 (2002).
10. A. J. Boone, Ch. H. Chang, S. N. Greene, T. Herz, N. G. J. Richards, *Modelling the spin-dependent properties of open-shell Fe(III)-containing systems: towards a computational description of nitrile hydratase*, Coord. Chem. Rev. **238–239**, 291 (2003).
11. T. C. Harrop, P. K. Mascharak, *Fe(III) and Co(III) Centers with Carboxamido Nitrogen and Modified Sulfur Coordination: Lessons Learned from Nitrile Hydratase*, Acc. Chem. Res. **34**, 253 (2004).
12. Ch. Chang, A. J. Boone, R. J. Bartlett, N. G. J. Richards, *Toward a Computational Description of Nitrile Hydratase: Studies of the Ground State Bounding and Spin-Dependent Energetics of Mononuclear, Non-Heme Fe(III) Complexes*, Inor. Chem. **43**, 458 (2004).
13. L. Peplowski, K. Kubiak, W. Nowak, *Insights into catalytic activity of industrial enzyme Co-nitrile hydratase. Docking studies of nitriles and amides*, J. Mol. Model. **13**, 725 (2007).
14. L. Peplowski, K. Kubiak, S. Zelek, W. Nowak, *A Comparative DFT Study of Substrates and Products of Industrial Enzyme Nitrile Hydratase* Int. J. Quant. Chem., (2007), *in print*.
15. K. Kubiak, W. Nowak, *Molecular dynamics of photoactive protein nitrile hydratase*, Biophys. J., *in preparation*.
16. H. Li, R. Elber, J. E. Straub, *Molecular dynamics simulation of No recombination to myoglobin*, J. Biol. Chem. **268**, 17908 (1993).
17. R. Elber, C. Simmerling, R. Goldstein, H. Li, G. Verkhiverer, C. Keasar, J. Zhang, A. Ulitsky, *MOIL: A program for simulations of macromolecules*, Comput. Phys. Commun. **91**, 159 (1995).

Exploring Conformational Space and Dynamics of RNA Hairpins by MD Simulations: Structure-Function Correlation of HIV-1 Genome Regulatory Elements

K. Kulinska and T. Kulinski

Institute of Bioorganic Chemistry,
Polish Academy of Sciences, 61-704 Poznan, Poland
E-mail: {*kasiak, tadkul*}*@ibch.poznan.pl*

To assess the reliability of MD derived structural models of the dynamic multifunctional RNA hairpin, a comparison between back calculated from MD trajectories and experimentally observed proton chemical shifts has been carried out. Our results confirmed coexistence of alternative conformations of the apical loop of HIV-1 TAR RNA hairpin in a solution. The comparison between the calculated and observed proton chemical shifts proved to be a good tool for validation and refinement of MD derived structures of dynamically inter-converting RNA conformational sub states.

1 Introduction

Insight into dynamical behavior of RNA molecules by experimental techniques is limited. Elucidating high-resolution RNA structure of dynamically inter-converting conformational sub states poses significant challenges to NMR spectroscopy and X-ray crystallography, which remain limited in applicability to static average structure determination of well folded conformations. Functionally active RNA conformations may not always be most populated in a solution. Transiently populated conformational sub states may be captured during protein recognition and stabilized by binding divalent ions or other biomolecular complex components. To elucidate structure-function relationships in RNA, it is necessary to go beyond the static structure and understand the dynamics of RNA in a solution. Molecular modeling and molecular dynamics simulation (MD) methods complement limitations of experimental techniques. However, the determination of RNA conformational space by simulations is related to limited time and conformation range, which may be probed by MD in comparison to real RNA dynamics. Roughness of the folding energy surface, kinetic traps and stable misfolds are the factors which require attention not only during in vitro RNA folding procedures, but also during computer simulations. Adequately, special consideration of infrequent events in structural transitions is needed for MD experiments due to relatively low statistics reachable in current status of software and hardware available for these calculations. In our exploring of the conformational space accessible to the dynamic RNA hairpin by molecular modeling and extensive MD simulations, we adopted several alternative approaches, which revealed the coexistence of two stable conformers. Then we investigated whether our MD derived structural models agree with the experimental NMR parameters.

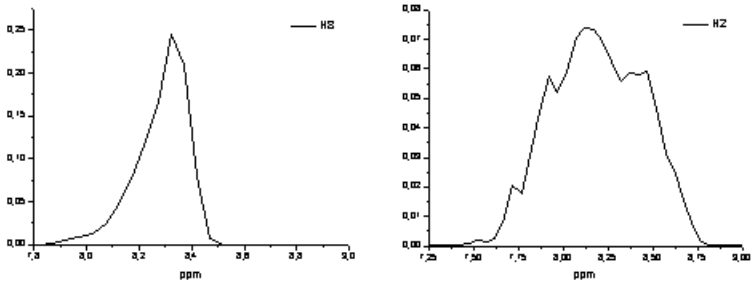


Figure 1. The distribution of CCS back-calculated from MD trajectories for H8 and H2 protons of adenosine A35. Experimentally measured values are 8.33 and 8.16, respectively.

2 Alternative Conformations of the Apical Loop of HIV-1 TAR RNA Hairpin

TAR RNA hairpin located in the extremely conserved domain of the 5' untranslated leader region (5' UTR) of the HIV-1 genome is involved in many steps of the virus replication cycle. Its functional diversity as viral regulatory element is determined by structural versatility. The structure of the TAR apical loop has been addressed by several experimental and computational methods. NMR studies revealed a compact loop structure with considerable conformational flexibility resulting in ensemble of NMR structural models[1]. Structural studies by biochemical methods (UV melting and thermodynamics, chemical structure probing, lead induced cleavage) of TAR RNA hairpins with a selectively mutated or 2-aminopurine substituted apical loops demonstrated that the apical loop of TAR is structured and we postulated that it is stabilized by the C30 - G34 cross-loop base pair[2]. Our recent molecular modeling and MD simulation studies revealed the coexistence of alternative conformations of the TAR apical loop: one stabilized by the C30 - G34 cross-loop base pair and the other stabilized by the interactions characteristic for U-turn motif[3].

3 Validation of MD Derived Structural Models by Comparing the Calculated and Experimentally Observed NMR Data

Initial models built by fitting TAR sequence to selected RNA motifs were adjusted by imposing constrains for the specific intra-loop interactions during short MD runs. All constraints were then removed and several full MD trajectories were collected running simulations for up to 8 ns each, depending on the convergence of the simulated structures, at different temperatures: 300K, 320K, 340K and 350K. The cluster analysis of trajectories yielded a range of conformational motifs characterized by different internal interactions and exhibiting diverse stability and convergence within the MD trajectories in a total of over 100 ns of simulation time. The reliability of the structural ensembles obtained during extensive MD simulations was assessed by NMR chemical shifts (CCS) calculations. CCS were back-calculated from the structures with the program NUCHEMICS, considering ring-current effects, magnetic-anisotropy terms, and ignoring charge contribution[4]. Each proton shift, calculated for structures within MD trajectories, contributed to the ensemble

of all calculated structures and was subsequently analyzed in ORIGIN in terms of statistical distribution. The structural analysis and interpretation of the TAR RNA 1H chemical shifts were carried out by comparing the calculated to experimental values[5]. Fig. 1 gives an example of the distribution of CCS back-calculated from MD trajectories for H8 and H2 protons of adenosine A35 base. Experimentally measured values are 8.33 and 8.16, respectively.

4 Conclusions

RNA hairpins exhibit remarkable versatility with regard to how conformational changes are utilized to modulate diverse cellular processes. RNA conformational switching is usually induced by cofactors such as proteins, small molecules, divalent ions or other RNAs. The study of the range of conformations accessible to RNA structural motifs is important for understanding the mechanisms of recognition processes and regulation by RNA. We observed good agreement between the back-calculated and experimental CCS for both MD derived structural models of the apical loop of TAR. Our results validate coexistence of both conformers of TAR RNA hairpin in a solution. Alternative structures are consistent with the range of previously obtained diverse experimental results. Both conformers can be functionally important: the conformer stabilized by cross-loop base pair can be responsible for TAR-proteins recognition, whereas the conformer stabilized by the interactions characteristic for U-turn motif can mediate RNA-RNA and RNA-DNA loop-loop interactions. The comparison between the calculated and observed CCS proved to be a good tool for validation and refinement of MD derived structures of dynamically inter-converting RNA conformational sub states.

Acknowledgments

The computations were performed with a grant of computer time provided by the Poznan Supercomputing and Networking Centre.

References

1. F. Aboul-ela, J. Karn, G. Varani, *The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein* , J. Mol. Biol. **253**, 313, (1995).
2. T. Kulinski et al., *The apical loop of the HIV-1 TAR RNA hairpin is stabilized by a cross-loop base pair* , J. Biol. Chem. **278**, 38892-38901, (2003).
3. T. Kulinski, K. Kulinska, J. Sarzynska, R.W. Adamiak, *Conformational dynamics of RNA hairpins - regulatory elements of the HIV-1 genome* , Biophysical Journal **88**, 402A, (2005).
4. S.S. Wijmenga, M. Kruithof, C.W. Hilbers, *Analysis of H-1 chemical shifts in DNA: Assessment of the reliability of H-1 chemical shift calculations for use in structure refinement* , Journal of Biomolecular Nmr **10**, 337-350, (1997).
5. K.S. Long, D.M. Crothers, *Characterization of the solution conformations of unbound and Tat peptide-bound forms of HIV-1 TAR RNA*, Biochemistry **38**, 10059-10069, (1999).

Anisotropic Internucleosome Interactions and Geometrical Constraints Favour the Two-Start Helical Structure of Chromatin

Giovanni La Penna[1], **Giorgio Cinacchi**[2], **and Angelo Perico**[3]

[1] Institute for chemistry of organo–metallic compounds
National research council, via Madonna del Piano 10, 50019 Sesto Fiorentino (FI), Italy
E-mail: glapenna@iccom.cnr.it

[2] Department of chemistry and industrial chemistry
University of Pisa, via Risorgimento 35, 56126 Pisa, Italy
E-mail: g.cinacchi@sns.it

[3] Institute for macromolecular studies
National research council, via De Marini 6, 16149 Genova, Italy
E-mail: perico@ge.ismac.cnr.it

The structures of chromatin at the high density characteristic of the silent phase are strongly influenced by the packing of nucleosome core particles (NCPs), the anisotropic attractive interactions between two of them and constraints, such as the DNA bending, imposed to the wrapped and linker DNA segments. In this work, coarse–grained models of chromatin are studied. For a pair of NCPs, a simple single–site anisotropic potential energy function is designed on the basis of the experimental data reported for the ordered phases of NCPs. This potential energy function is employed in random–walks of chromatin models where the NCP DNA wrapping is modulated in length, while the linker segments are modulated in both length and curvature. These models support the two–start helical organization for chromatin in the absence of linker histones. The geometry of two-start helical configurations is characterized by poorly bent linkers and by a moderate reduction of wrapped DNA in the NCP.

1 Introduction

The investigation of the supramolecular structure and dynamics of chromatin is of extreme importance to understand gene regulation and cell development. The genetic information in eukaryotic chromosomes is organized in chromatin, a chain–like supramolecular assembly formed by a long filament of DNA wrapped around globular octameric aggregates of eight histone proteins[1]. The repeat unit of this chain includes a nucleosome core particle (*i.e.* the complex of a protein histone octamer with DNA wrapped around it, NCP hereafter) and a DNA linker, which connects two NCPs. Other histone proteins in the linker region are present in the natural chromatin and absent in several model systems studied in vitro.

Two classes of organizations have been proposed for the compact state of chromatin, *i.e.* the so–called 30 nm fiber[2,3]: the one–start solenoidal helix model[4]; the two–start helix model[5]. Recent crystallographic results for the tetra–nucleosome molecule in the absence of linker histones[6], provide strong evidence for the two–start helix organization: the structure of the tetra–nucleosome is consistent with an idealized model where two left–handed twisted ribbons follows a straight fiber axis. These x-ray data are also consistent with the high propensity of NCPs of forming columns[7]. Theoretical and computational

models of the chromatin fiber have been proposed, but two–start helices with the geometry of the x-ray data were never identified among the proposed structures.

In this work, a single–site anisotropic Lennard–Jones interaction potential, calibrated on the experimental phase transition for NCPs, is used to perform random walks of chromatin models with a variable amounts of wrapped and linker DNA segments and with linker DNA with variable curvature.

2 Method

The geometrical construction of regular chains of many NCPs is summarized in the following. The pathway of DNA is built starting by wrapping n_w bp around the disc representing the histone octamer (radius of 4.5 nm and height of 6 nm). When the DNA finishes to wrap, a straight linker of n_{s1} bp follows. At a certain point, DNA starts to bend (n_b bp) around an origin **O** chosen at random within a cubic portion of space. Two curvatures have been tested, with no significant differences in results. The DNA has no kinks. A second straight linker segment of n_{s2} bp follows. Before the next NCP particle is generated, a further rotation of $\Delta\Phi$ modifies the orientation, otherwise imposed by the equilibrium twisting of DNA.

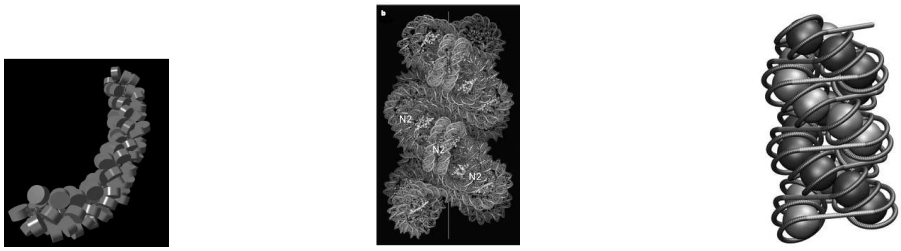
Monte Carlo trajectories have been collected using a temperature randomly chosen within 0 and 10000 K and using the NCP–NCP interactions only. Among the collected configurations, the linker–linker interactions have been computed by using a Lennard–Jones site for each DNA bp, with a σ parameter of 1 nm. Only those configurations with negative linker–linker potential energy and with low NCP–NCP interaction and high compactness have been selected and analyzed.

3 Results

The distribution of the supplemental bending angle (Θ) and of the torsional angle (Φ) of the selected chains has been analyzed. This distribution always displays a maximum around $\Theta \sim 5\pi/6$ and $\Phi \sim \pi/2$. Structures representative of this region display the two–start organization, with linker DNA almost perpendicular to the fiber axis. One of these structures is compared with the experimental x-ray data and with a previous simulation in Fig. 1.

The structures obtained by previous MC simulations using similar parameters (panel a) could display only a partial two–start organization, because of the low regularity. The representative structure obtained in this work (c) is very close to the construction based on x-ray data (b). The NCP planes are only slightly tilted away from the fiber axis compared to the tetra–nucleosome. The linker DNA segments are straight, but a local bending can be suggested by linker–NCP interactions that have been ignored so far. The wrapped length of the most representative configurations is about 130 bp, that is in close agreement with the value of 129 bp found in the tetra–nucleosome.

Therefore, this simple chromatin model accounts for most of the important features of the available structural information for chromatin in the absence of linker histones. The role of the interactions included in this model can be summarized: NCP–NCP interactions



(a) (b) (c)

Figure 1. Comparison between the most representative structure in this model (c), MC simulations (a) and the x-ray data (b): one snapshot of the simulation in Ref.[8] (a); the construction obtained from the tetra–nucleosome x-ray data[6] (b); one of the structures representative of the maximum population in the two–angles map (this work, c). In (c) the large spheres represent the histone octamers (dark gray the even numbers, light gray the odd numbers to emphasize the two–start organization). The small spheres represent DNA bp. The radii of these latter spheres is 1/2 of the used value for graphical purposes.

favour columns (like those experimentally observed); wrapped and linker DNA favour the coiling of columns; these columns can not be coiled enough to obtain a dense one–start organization (solenoid); on the other hand, two less coiled columns intercalated allow the two–start organization; crossed linkers allow a moderate DNA–NCP overlap; the unfolding of the two–start organization is not intricate.

Acknowledgments

This work has been done with the financial support of the FIRB 2001 project RBA01Y3SN of MIUR (Italy).

References

1. J. Widom, *Structure, dynamics and function of chromatin in vitro*, Ann. Rev. Biophys. Biomol. Struct., **27**, 285–327, 1998.
2. H. Schiessel, *The physics of chromatin*, J. Phys.-Cond. Matt., **15**, R699–R774, 2003.
3. P. J. J. Robinson and D. Rhodes, *Structure of the 30 nm chromatin fibre: A key role for the linker histone*, Curr. Op. Struct. Biol., **16**, 336–343, 2006.
4. J. T. Finch and A. K. Klug, *Solenoidal model for superstructure in chromatin*, Proc. Natl. Acad. Sci. USA, **73**, 1897–1901, 1976.
5. C. L. F. Woodcock, L.–L. Y. Frado, and J. B. Rattner, *The higher-order structure of chromatin: Evidence for a helical ribbon arrangment*, J. Cell. Biol., **99**, 42–52, 1984.
6. 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.
7. S. Mangenot, A. Leforestier, D. Durand, and F. Livolant, *Phase diagram of nucleosome core particles*, J. Mol. Biol., **333**, 907–916, 2003.
8. G. Wedemann and J. Langowski, *Computer simulation of the 30-nanometer chromatin fiber*, Biophys. J., **82**, 2847–2859, 2002.

Mechanism of Fibril Formation of Aβ_{16-22} Peptides

Mai Suan Li[1], **Phuong Hoang Nguyen**[2], **and D. Thirumalai**[3]

[1] Institute of Physics, Polish Academy of Sciences
Al. Lotnikow 32/46, 02-668 Warsaw, Poland
E-mail: masli@ifpan.edu.pl

[2] Institute of Physical and Theoretical Chemistry, J. W. Goethe University,
Marie-Curie-Str. 11, D-60439 Frankfurt, Germany
E-mail: phuong@theochem.uni-frankfurt.de

[3] Biophysics Program, Institute for Physical Science and Technology and
Department of Chemistry and Biochemistry,
University of Maryland, College Park, MD 20742
E-mail: thirum@glue.umd.edu

Using all-atom simulations with the GROMOS96 force field 43a1 in explicit water we have studied the kinetics of fibril formation of Aβ_{16-22} peptides. These short peptides are the hydrophobic core of longer amyloid peptides which are believed to play the key role in the Alzheimer's disease. It is found that the incorporation of a newly added peptide to the preformed template obeys the two-stage dock-lock mechanism. This is in the qualitative agreement with the recent experiments.

1 Introduction

Understanding the mechanism of oligomerization of proteins and peptides remains one of the most exciting problems in structural biology because a large body of evidence suggests that amyloid fibrils and associated oligomeric intermediates are related to a number of diseases, including Alzheimer's, Parkinson's, Huntington's, and prion diseases[1]. The recent experiments of different groups (see, e.g. Ref. 2 and references therein) suggest that the addition of soluble Aβ and Sup35 amyloid monomers to preformed fibril structures is a two-stage dock-lock process. First the added peptide docks into the template and at this stage its structure gets elongated. In the slow lock phase its orientation fluctuates a lot before reaching the fibril state. Motivated by these interesting experiments we have studied the process of incorporation of a Aβ_{16-22} peptide into the preformed template using the all-atom simulations by GROMACS software[3]. Namely, we have considered[4] the reaction $(\mathrm{A}\beta_{16-22})_{n-1} + \mathrm{A}\beta_{16-22} \rightleftharpoons (\mathrm{A}\beta_{16-22})_n$, where $n = 4, 5$ and 6. In this contribution we presents the results for the $n = 5$ case (pentamer) where a monomer adds to the preformed fibril state of 4 peptides. Our results support the two-stage picture in which the lock phase is much slower than the initial dock phase.

2 Method and Results

The procedure of generating the template of 4 preformed peptides (see Fig. 1a) may be found in Ref. 4. The conformation of monomeric Aβ_{16-22} was extracted from the structure of Aβ_{10-35} peptide available in the Protein Data Bank (ID: 1hz3). Then we have added it

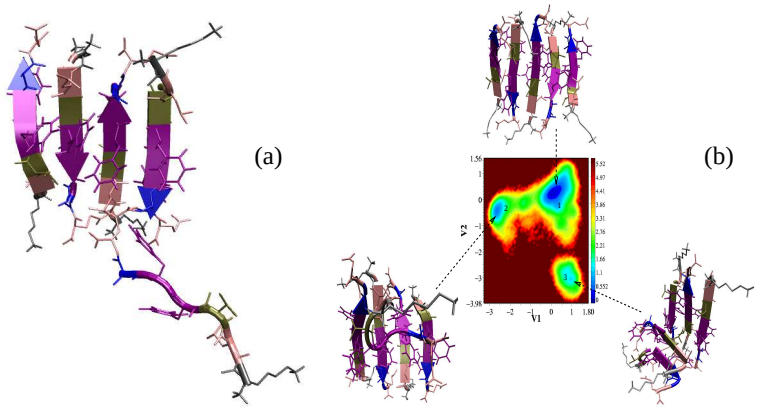


Figure 1. (a) A initial conformation of the pentamer, where one peptide is randomly added to the template of four ordered monomers. (b) The free energy landscape as a function of V_1 and V_2. The typical conformations of main local minima are shown.

randomly to the template (1a) and generated four trajectories of 329, 360, 540 and 737 ns using the GROMOS96 force field 43a1 in explicit water. A typical initial configuration is shown in Fig. 1a. The volume of the simulation box is 128 nm^3 which corresponds to the peptide concentration of 65 mM.

In order to monitor the fibril formation process we use the "liquid crystal" order parameter P_2. If $P \geq 0.9$ then the system is considered to be in the fibril-like state[4]. We used the dihedral principal component analysis[4] to compute the free energy landscapes (FEL) using the first two eigenvectors V_1 and V_2. To show that the added peptide incorporates into the preformed monomers by the dock-lock mechanism we monitor the time dependence of the β-content, $\beta(t)$, which has been computed using the definition, given in Ref. 4.

Fig. 1b shows the free energy as a function of V_1 and V_2. The fibril-like arrangement is found in the basin of the most prominent minimum 1 which has $\approx 67\%$ of population. Since three local minima are separated by low barriers of only a few k_BT the system is not stable under thermal fluctuations. The stability gets enhanced as the number of peptides increases[4].

The time dependence of the order parameter P_2 for the entire system and the preformed 4 peptides is shown in Fig. 2. Remarkably the template fluctuates a lot to accommodate the added monomer. This is probably because the number of monomers is lower than the size of the critical nucleus which remains largely unknown for the Aβ_{16-22} system but our recent analysis has shown that this size should be at least larger than six. If we define the time for adding a new monomer, t_{add} as the time needed to get $P_2 = 0.9$ for the entire system then $t_{add} \approx 110$ ns for the pentamer (the result is averaged over four trajectories).

From the time dependence of the content of the beta content, $\beta(t)$ (Fig. 3a), it follows that the nascent peptide docks into the the preformed subsystem first. Namely, initially $\beta(t)$ of the disordered monomer is lower than that of the preformed subsystem and they become compatible at the end of the dock phase. In order to get into the fibril state the added peptide spends a lot of time on its reorientation while the preformed template fluctuates to accommodate it. As evident from Fig. 3b the accumulated value $\overline{\beta}(t)$ of the nascent

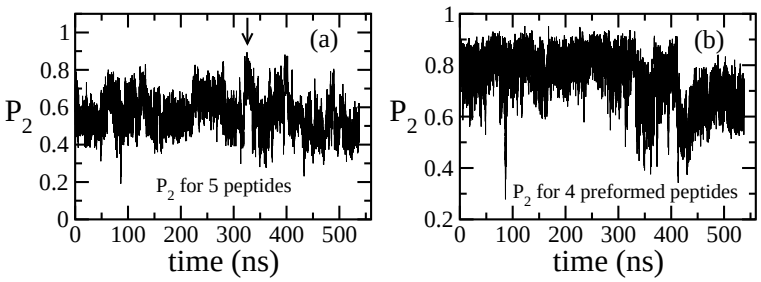


Figure 2. (a) Time dependence of P_2 for the whole system of 5 peptides. The arrow refers to $P-2 \approx 0.9$ which corresponds to the adding time. (b) The same as in (a) but for the preformed template.

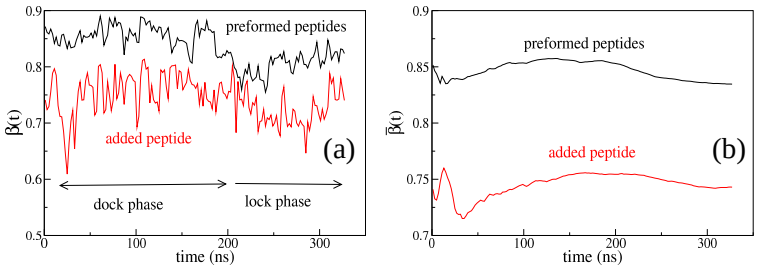


Figure 3. (a) The time dependence of the beta content of the preformed four peptides (black) and the added monomer (red). (b) The same as in (a) but for $\overline{\beta}(t) = \frac{1}{t}\int_0^t \beta(x)dx$.

monomer approaches the value of the template very slowly, the second lock phase should have much longer time scale compared to the dock stage.

In conclusion, we have shown that the fibril formation of short peptides obeys the two-stage dock-lock mechanism. This may be also valid for longer peptides and proteins.

Acknowledgments

This work was supported by the Polish Komitet Badan Naukowych grant No 1P03B01827 and National Institute of Health Grant R01 GM076688-05.

References

1. F. Chiti, and C. M. Dobson, *Protein misfolding, functional amyloid, and human disease*, Annual Rev. Biochemistry **75**, 333–366, 2006.
2. R. S. Collins, A. Douglass, R. D. Vale, and J. S. Weissman, *Mechanism of prion propagation: Amyloid growth occurs by monomer addition* PLOS Biol. **2**, 1582-1590, 2004.
3. H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *A message-passing parallel molecular dynamics implementation*, Comp. Phys. Comm. **91**, 43–56, 1995.

4. P. H. Nguyen, M. S. Li, G. Stock, J. E. Straub, and D. Thirumalai, *Monomer adds to preformed structured oligomers of Aβ-peptides by a two-stage dock-lock mechanism*, Proc. Natl. Acad. Sci. USA **104**, 111–116, 2007.

Multiple Beta-Sheet Molecular Dynamics of Two Abl-SH3 Domain Peptides

Inta Liepina[1], **Salvador Ventura**[2], **Cezary Czaplewski**[3], **and Adam Liwo**[3]

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

[2] Institut de Biotecnologia i de Biomedicina, Universitat Autonoma de Barcelona
E-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*

Six parallely placed ten stranded antiparallel β-sheets of DLSFMKGE (10x6xMK), and DLSFKKGE (10x6xKK) peptides, immersed in periodic water boxes, were subjected to molecular dynamics simulations (MD) for 90 ns and 40 ns respectively, to study the amyloid fibril formation tendencies of the two peptides. MD simulation showed that the 10x6xMK β-sheet sheet stack is stable, but 10x6xKK β-sheet stack is not. 10x6xMK β-sheet is stable because of hydrophobic interactions of metionine-phenilalanine and leucine of the neighbouring sheets. Met, Phe, Leu make a hydrophobic core for the stack of β-sheets. During MD run the Met, Phe, Leu of neighbouring sheets act as conformational switch moving β-sheets by two amino acid step towards each other.

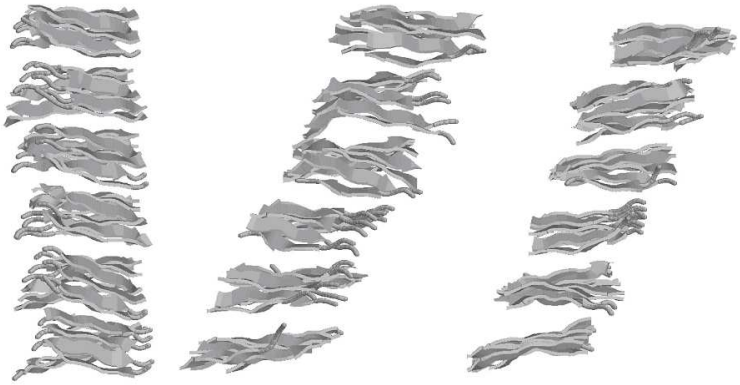
1 Introduction

Short protein sequence stretches drive the protein aggregation in amyloid fibrils[1–3]. Based on a homology search we have identified an aggregation-prone region in the Abl-SH3 domain of Drosophila with sequence DLSFMKGE (MK), and less amyloidogenic human homologous region with sequence DLSFKKGE (KK). The antiparallel flat β-sheets consisting of ten strands of MK and KK were constructed. We created two multi-sheet systems: (1) six parallely placed 10-strand β-sheets of MK (10x6xMK) (Fig.1a), (2) six parallely placed 10-strand β-sheets of KK (10x6xKK) (Fig.2a). Each of these β-sheet systems was surrounded by a 10 Å layer of water molecules over the solute and subjected to molecular dynamics (MD) simulations with the Amber 8.0 force field in the NPT (constant number of molecules, pressure, and temperature) scheme. The MD simulations were started from the temperature of 10 K and the temperature was gradually raised by the step of the ten degrees till 300 K, then the simulations were run at the constant temperature of 300 K. The 10x6xMK system was simulated by 90 ns of MD, the 10x6xKK system was simulated by 40 ns of MD.

2 Results

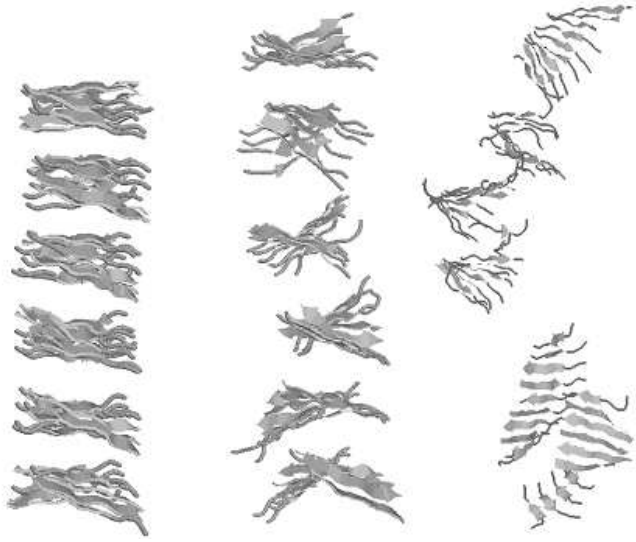
The 10x6xMK system maintains β-sheets, and the β-sheets remained together in a stack during all the time of MD simulation (Fig.1) and showed strong hydrogen bonding (Fig.3). During the simulation the β-sheets of the 10x6xMK system shifted towards each other by

two amino acids (Fig.1b, Fig.1c). The 10x6xKK system maintains β-sheets, but β-sheets do not remain together, they are hanging by a thread (Fig.2b), and moving away each from other (Fig.2c).



a) b) c)

Figure 1. Multisheet system of DLSFMKGE peptides (10x6xMK) keeps together all the time of MD simulation. The 10x6xMK system at a) 351 ps, b) 21705 ps, c) 74499 ps of MD.



a) b) c)

Figure 2. Multisheet system of DLSFKKGE peptides (10x6xKK) does not remain together during MD simulation. The 10x6xKK system at a) 382 ps, b)11408 ps, c) 36771 ps of MD.

3 Conclusions

The MD simulation of multisheet systems revealed that: 1.The 10x6xMK β-sheet stack is stable, but the 10x6xKK β-sheet stack is not. 2.The 10x6xMK β-sheet is stable because of hydrophobic interactions of metionine and phenylalanine side chains and the leucine side chain of the neighbouring sheets. Met, Phe, Leu make a hydrophobic core for the stack of β-sheets. 3.During the MD run, the Met, Phe, and Leu residues of neighbouring sheets act as a conformational switch moving the β-sheets by two amino acid step towards each other. 4.Replacement of Met by Lys destroys the hydrophobic core, which is the stability factor of the β-sheets stack. The 10x6xKK system maintains β-sheets, but loses interactions between β-sheets. 5.The calculations of six β-sheets confirm the conclusion drawn for single sheet systems: parallelly placed β-sheets stabilize each other.

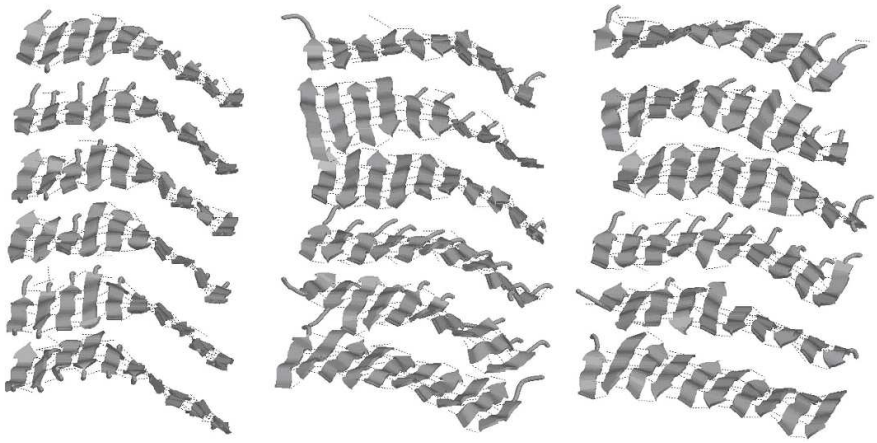


Figure 3. DLSFMKGE peptides (10x6xMK) show strong hydrogen bonding during the all MD run: a) 351 ps, b) 21705 ps, c) 74499 ps.

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, calculations are carried out at the Gdansk Academic Computer Center TASK.

References

1. 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.
2. 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.
3. I. Liepina, S. Ventura, C. Czaplewski, A. Liwo, *Molecular dynamics study of amyloid formation of two Abl-SH3 domain peptides*, J.Pept.Sci. **12**, 780-9, 2006.

Parallelization of ECEPP/3 in SMMP

Jan H. Meinke[1] **and Ulrich H. E. Hansmann**[1,2]

[1] John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {j.meinke, u.hansmann}@fz-juelich.de

[2] Department of Physics,
Michigan Technological University,
Houghton, MI 49931, USA

The power consumption of modern processors makes it difficult to increase their clock speed further. Even in the PC market CPU manufacturers now include multiple compute cores on a single chip to improve performance and keep up with Moore's law, a trend likely to continue. This is even more important in high performance computing, where cooling and electricity bills are becoming a large issue. The compactness and low power consumption of an IBM BlueGene, e.g, is only possible because of CPUs with moderate clock rates.

Protein folding is computationally hard. To take advantage of the increasing number of compute cores and reduce the time to solution, we need to parallelize our codes, especially the time intensive calculation of the energy function, and minimize the serial portions of the code. We are presenting the parallelization implemented in SMMP for the ECEPP/3 force and the resulting scaling behavior on various platforms including BlueGene/L. Combining the parallel energy function with parallel tempering, simulations scale up to thousands of processors.

1 Introduction

Moore predicted that the number of transistors on the cheapest chips doubles every 24 months.[1] This prediction has held for over 40 years. Combined with increasing clock speeds, this development let to ever faster CPUs, faster-running serial codes, and larger power consumption. One way to decrease the power consumption is to run the processor at a lower clockspeed. Using the same chip, doubling the clock speed results in an approximately 8 times higher power consumption. In the mid 1990s, an Intel Pentium I running at 133 MHz used about 11 W. Modern processors can use more than 100 W. With the introduction of multi-core processor to the consumer PC market in 2005, Intel moved away from providing processors that run serial code faster and forced many application developers to think about parallel programming. In high-performance computing, where the number of processors can go into the thousands, power consumption and cooling are important factors. With the development of BlueGene/L, IBM decided to use comparatively slow processors — with a correspondingly low power consumption — combined with fast networking, to build a very compact system that scales to more than a 100000 processors and a peak performance of more than 350 TFLOPS. This high performance comes at a price. Comparing the single processor performance of a BG/L with an AMD Opteron at 2.4 GHz, we find that our application runs more than 7 times faster on the Opteron than BG/L. To make up for this loss in performance and take advantage of our BlueGene/L JUBL and the increasing size of our smaller PC clusters, we parallelized the calculation of the energy in our protein simulation package SMMP.[2,3] In combination with parallel tempering, we now regularly run our simulation on 4096 processors. Here, we

present the application scaling on 4 different platforms and give some pointers for efficient parallelization on BlueGene/L.

2 Distributing the Work

For these measurements we used the ECEPP/3 force field.[4] The energy function that needs to be calculated is

$$E_{\mathrm{ECEPP}/3} = \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}} + \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + \sum_{l} U_l (1 \pm \cos(n_l \xi_l)) \; , \tag{1}$$

where i and j are indices of atoms, r_{ij} is the distance between atom i and atom j, and l is the index of a dihedral angle in the protein chain. All other variables (q_i, A_{ij}, B_{ij}, C_{ij}, and D_{ij}) are parameters of the force field.

In SMMP, every atom is associated with a dihedral angle. We used this relation to distribute the interactions as evenly across processors as possible without regard to spatial proximity.

3 Setup and Scaling

We ran our benchmark on 4 different platforms: JUMP, an IBM p690 cluster with 32 processors and 112 GB of shared memory per node; JUBL, an IBM BlueGene/L with 8 racks and a total of 16384 Power4 processor at 700 MHz; JULI a PC cluster using dual-core PowerPC 970MP processors at 2.5 GHz with an InfiniPath network and NICOLE, an Opteron based PC cluster with a clock speed of 2.4 GHz using Infiniband networking. Except for the setup of the communicators used for the energy calculation on BG/L, we used the same source code for all measurements. We performed 50 sweeps of a Monte Carlo simulation of the designed protein TOP7[5] starting from a stretched chain. Data was written to disk every 10 sweeps. On JUBL, we used multiple replicas in parallel with the indicated number of processors per replica to fill a half plane (512 processors).

Figure 1 shows walltime and scaling for the various machines. The execution time on a single processor ranges from about 18 min on JULI to about 2 h on JUBL. The lowest execution time ranges from 81 s on JUMP to 269 s on JUBL with 64 processors per replica. The maximum speedup is 25 on JUBL with 64 processors.

For JUMP, JULI, and NICOLE we used MPI's default processor assignment. On JUBL, however, this approach leads to a sub-optimal distribution of the processors. BG/L has a cubic geometry. By default, the rank of a processors increases first along x, than y, and finally z. This leads to a planar distribution of processors (cf. left picture in figure 2). Instead of the default, one should make communicators as cubic as possible (see right picture in figure 2) unless the problem geometry suggests a different approach.

4 Conclusions

For protein simulations, distributing the calculation of the interactions evenly across a number of processors can be done efficiently without worrying about the spatial distribution of

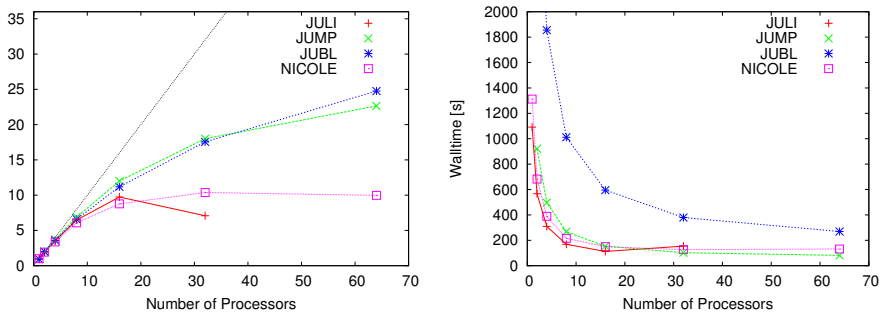


Figure 1. Strong scaling behavior. This figure shows the walltime and corresponding parallel scaling vs. number of processors on JULI, JUMP, and JUBL. For the benchmark, we performed 50 sweeps of a Monte Carlo simulation of Top7,[5] a 92 residue protein with 1477 atoms. Data and configurations were written to disk after every 10 sweeps. On JUBL, we used multiple replicas in parallel with the indicated number of processors per replica to fill a half plane.

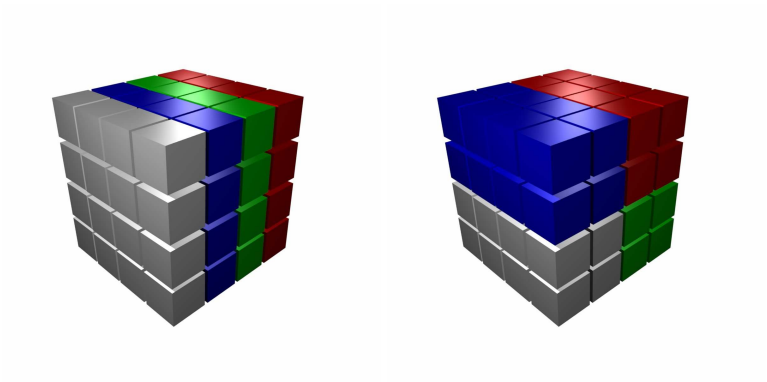


Figure 2. Two processor arrangements on BG/L. To combine parallel tempering with a parallel calculation of the energy for each replica, we define our own communicators for the energy calculation. On the left-hand side, we show the default arrangement for a consecutive number of processors on a $4 \times 4 \times 4$ partition with four replicas. Each replica is assigned to a plane of the partition. On the right, we show the most cubic arrangement possible. The cubic arrangement scales significantly better (up to $25\times$) than the planar one (up to $18\times$).

the atoms involved. The low cost per processor makes BlueGene/L an attractive platform for protein simulations. Using a cubic arrangement of 64 processors, we achieve a speedup of up to $25\times$ on BG/L. With the large number of processors available on JUBL, we can run simulations with 64 replicas at a quarter of the cost and at the same speed as on JUMP.

Acknowledgments

Development, computations and measurements were performed on computers of the John von Neumann Institute for Computing (NIC). U.H. received support from the National Science Foundation (USA) under grant number CHE-0313618.

References

1. Gordon E. Moore, *Cramming more components onto integrated circuits*, Electronics, **38**, no. 8, 114–117, 1965.
2. Frank Eisenmenger, Ulrich H. E. Hansmann, Shura Hayryan, and Chin-Kun Hu, *[SMMP] A modern package for simulation of proteins*, Comput. Phys. Commun., **138**, 192–212, aug 2001.
3. Frank. Eisenmenger, Ulrich. H. E. Hansmann, Shura Hayryan, and Chin-Kun Hu, *An enhanced version of SMMP—An Open-Source Software Package for Simulation of Proteins*, Comput. Phys. Commun., **174**, 422–429, 2006.
4. G. Nemethy, 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**, no. 15, 6472 – 6484, 1992.
5. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, *Design of a Novel Globular Protein Fold with Atomic-Level Accuracy*, Science, **302**, no. 5649, 1364 – 1368, 2003.

The Role of Metals in Misfolding and Aggregation Processes: X-ray Spectroscopy and Numerical Simulations

V. Minicozzi[1], S. Morante[1], G.C. Rossi[1], F. Stellato[1], and K. Jansen[2]

[1] University of Rome Tor Vergata and INFN, Via della Ricerca Scientifica 1, 00133 Rome, Italy
E-mail: {velia.minicozzi, silvia.morante, giancarlo.rossi, francesco.stellato}@roma2.infn.it

[2] NIC Desy-Zeuthen, Platanenallee 6, 15738 Zeuthen, Germany
E-mail: Karl.Jansen@desy.de

Recent X-ray Absorption Spectroscopy experiments designed to study the metal coordination mode in Cu-Aβ and Zn-Aβ complexes have been able to detect visible differences between Zn and Cu samples that are suggestive of dissimilar structural roles of the two ions. These findings indicate that metals can bind A$beta$-peptides in an intra- as well as in an inter-peptide coordination mode, hinting at a different aggregation propensity. First principle investigations of Cu–Aβ and Zn–Aβ complexes, based on *ab initio* molecular dynamics simulations of the Car–Parrinello type are set up to elucidate the molecular basis for this difference.

1 Introduction

Amyloidosis is a family of pathologies caused by the transition of certain endogenous proteins and peptides from their physiological soluble configuration to a pathological fibrillar state. The term describes a heterogeneous group of diseases (more than 20) characterized by extra-cellular deposition of fibril material[1,2]. Among them the Alzheimer disease (AD), a progressive and devastating neuro-degenerative pathology[3], characterized by misfolding of well identified peptides called Aβ-peptides. In AD, like in many other amyloidosis, an important, but not yet fully elucidated, role seems to be played by transition metals (mainly Cu^{+2} and Zn^{+2}) present in fairly large amount in neurological plaques. There exist in the literature conflicting statements about the ability of Cu^{+2} and Zn^{+2} in providing protection against or acting as promoters of plaques formation. The interest in elucidating the role of metals strongly increased after noticing that Cu^{+2} and Zn^{+2} chelators are capable of solubilizing Aβ-aggregates. It is clear enough that the nature of the metal binding mechanism has significant consequences on the protein folding/unfolding and aggregation processes[4]. However, the experimental techniques used until very recently were not able to unambiguously reconstruct the local atomic structure around the metal.

The availability of third generation synchrotron radiation sources has significantly enlarged the range of application of X-ray Absorption Spectroscopy (XAS) in investigating structural properties of biological systems. XAS can be profitably used to study the environment of metal ions complexed with proteins and peptides in physiological conditions, owing to its chemical selectivity and sensitivity to the local atomic structure around the absorber. Furthermore an accurate analysis of the extended X-ray absorption fine structure (EXAFS) region of the spectrum in terms of single plus multiple scattering contributions has been proved to allow a clear-cut identification of the amino acid (a.a.) residues primarily bound to the metal[5].

At the more fundamental level, describing the peculiar electronic properties of peptide-metal complexes in terms of quantum chemistry is necessary to really clarify the structural

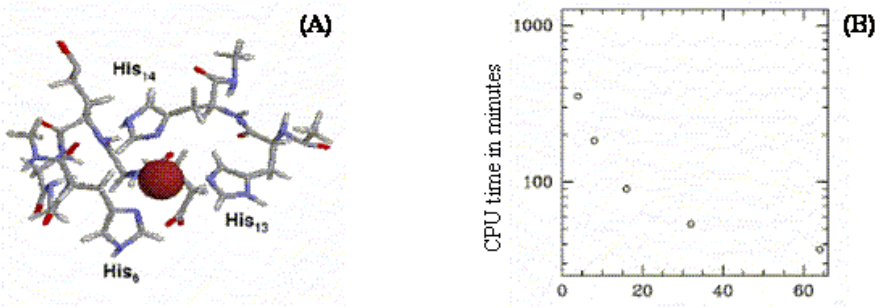


Figure 1. (A) The starting geometrical structure around Cu. (B) The scaling of the CPU running time of Espresso on the SGI Altix machine.

role of metal ions in misfolding and aggregation.

Density functional theory is the natural framework where a quantum mechanical description of the electronic structure of such compounds is possible. Special adjustments and refinements of the theory for application to bio-molecules are requested. The use of the plane-wave basis is particularly well suited for delocalized systems, like the organo-metallic molecules we are interested in, especially if coupled with the use of ultra-soft pseudo-potentials for the description of the electronic atom core. In this way the number of necessary plane waves can be kept at a tolerable level and Car–Parrinello (CP) *ab-initio* molecular dynamics (MD) can be efficiently carried out[6,7]. For reviews on the CPMD method see for instance[8,9].

2 XAS experiments and simulations

I) Extensive XAS experiments were performed at the DESY-Hamburg EMBL facility on one of the two most frequently naturally produced β-amyloids, the Aβ_{1-40} peptide[10]. They show evidence that according to whether the peptide is complexed with Cu^{+2} or Zn^{+2} a different metal binding site structure is formed. In fact, while the geometry around Cu is stably consistent with an intra-peptide binding with three metal-coordinated Histidine residues, the Zn coordination mode depends on specific solution conditions. In particular, different sample preparations lead to different geometries around the absorber compatible with either an intra- or an inter-peptide coordination mode. This result reinforces the hypothesis that assigns different physiological roles to the two metals, with Zn favoring peptide aggregation and, hence, possibly plaque formation. Further XAS measurements[11] on selected, shorter portions of the whole Aβ_{1-40}-amyloid, in complex with Cu^{+2} and Zn^{+2}, have allowed to identify the specific a.a.'s primarily bound to the metal. Analysis of spectral data suggests the following structural geometries.

i) Cu^{+2} is penta-coordinated with 3 Histidines, and 2 oxygen atoms.

ii) Zn^{+2} is penta-coordinated with 4 Histidines and 1 oxygen atom.

Since only three Histidine residues are present in each Aβ-peptide the peculiar structure of the Zn environment indicates that Zn acts as a bridge between two peptides in an inter-molecular coordination arrangement.

II) *Ab initio* CPMD simulations represent an invaluable tool to validate the detailed structural model suggested by experiments. We then constructed an atomistic model system for the Cu^{+2}–Aβ-peptide complex. In order not to have an impossibly large

system, of the first six a.a.'s the lateral chains of only the first (Aspartic acid), third (Glutamic acid) and sixth (Histidine) a.a., as well as of the 13^{th} and 14^{th} Histidine residue (His_{13} and His_{14}) are simulated in full atomic details. For the rest of the 1-6 a.a.'s only the backbone is retained. Furthermore none of the remaining a.a. is included. The C-terminal is capped with an $NHCH_3$ group, while the N-terminal of the 1-6 fragment is left open (with an ending NH group) to allow possible Cu binding with loss of the amide hydrogen. Structural information from XAS[10,11] and NMR[12] data were exploited in setting up the geometry around Cu which is schematically shown in Fig. 1 (panel A). The whole system, solvated with 125 water molecules, is contained in a box with a volume V=15x20x20 $Å^3$. In this way density is set equal to 1 g/cm^3. All in all (i.e. including water) the simulated system comprises 494 atoms and 1369 electrons. A similar system with Zn replacing Cu is under test.

III) Simulations are performed using the CPMD Espresso code[7]. In the Table we show the performances of the code, all taken in the 16 nodes configuration, on two PC-clusters (Fermi1 in Rome and BEN in Trento) and the SGI Altix 4700 machine (in Munich). Fig. 1

machine	**nodes**	$\mathbf{t}_{CPU}$ (s)
Fermi1	16	5880
BEN	16	4560
Altix	16	1080

(panel B) shows the scaling with the number of nodes of the CPU times on the Altix machine. The performances are for a not fully optimized code. The interesting outcome of this feasibility study is that with a moderately large number of nodes (say 64) on the SGI machine a fairly long CP trajectory (3-4 ps) of the large system described before can be collected in less that two months. Simulations of this length will allow to clarify the possible involvement of the N-terminal region of the Aβ-peptide in binding the metal, as suggested by XAS data[10,11].

References

1. M.B. Pepys, Philos. Trans. R. Soc. Lond. B Biol. Sci. 356, 203 (2001).
2. F. Chiti and C.M. Dobson, Ann. Rev. Biochem. 75, 333 (2006).
3. D.J. Selkoe, Physiol. Rev. 81, 741 (2001).
4. R.A. Cherny et al., Biometals 16, 91 (2003).
5. C. Meneghini and S. Morante, Biophys. J. 75, 1953 (1998); S. Morante, J. Synchrotron Rad. 8, 975 (2001); S. Morante et al., J. Biol. Chem. 279, 11753 (2004).
6. R. Car and M. Parrinello, Phys. Rev. Lett. 55, 2471 (1985).
7. P. Giannozzi, F. de Angelis and R. Car, J. Chem. Phys. 120, 5903 (2004).
8. M.C. Payne et al., Rev. Mod. Phys. 64 1045 (1992).
9. D. Mark and J. Hutter, in "Modern Methods and Algorithms of Quantum Chemistry", Ed. J. Grotendorst, NIC, Jülich, `http://www.fz-juelich.de/nic-series`.
10. F. Stellato et al., Eur. Biophys. J. 35(4), 340 (2006).
11. V. Minicozzi et al., in preparation.
12. J. Danielsson et al., FEBS J. 274, 46 (2007).

Folding and Aggregation of Proteins with Monte Carlo Simulations

Sandipan Mohanty[1] **and Ulrich H. E. Hansmann**[1,2]

[1] John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {s.mohanty,u.hansmann}@fz-juelich.de

[2] Department of Physics, Michigan Technological University, Houghton, Michigan, U.S.A.
E-mail: hansmann@mtu.edu

An implicit water all-atom model is used to study folding, aggregation of small proteins. Physically reasonable results obtained for a variety of applications indicate healthy global properties of the interaction potential.

1 Introduction

The prevailing picture of the process of protein folding is that proteins fold into their native 3D structures because those states have the minimum free energy among all conformations the protein chain can take, and hence are thermodynamically most probable. However, it has proven to be a considerably greater challenge to explicitly formulate an effective interaction potential, such that for every given protein, the minimum free energy structure calculated from the force field corresponds to the correct experimental structure. Different potentials give very different relative weights to the α-helix and β-strand regions of the protein conformation space. A potential that successfully folds α-helical peptides often has problems with β-sheet peptides, and *vice versa*. Also, proteins fold and unfold at physiologically relevant temperatures, and most potentials need further callibration in order to give more realistic temperature dependence of observable quantities.

In this article, folding[1,2] and aggregation[3] studies with one particular model for protein folding will be summarized. The protein molecules are represented in full atomistic detail, whereas the solvent molecules are left out to reduce the number of degrees of freedom of the system. The potential used was developed through repeated folding simulations of small peptides. It does not use any information about the known experimental structures of the peptides, and spontaneous folding from random initial conformations, and not the description of properties of the folded state through simulations of small perturbations around it, has been the chief objective.

2 Model and Methods

The model discussed here includes all atoms of the polypeptide chains, including all hydrogen atoms. 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.

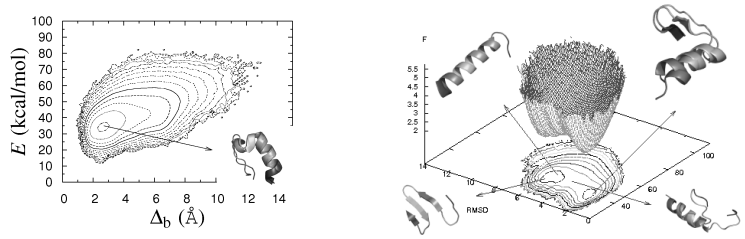


Figure 1. Free-energy estimated from the probability of occurence of different states in high statistics Monte Carlo simulations plotted as a function of energy and backbone RMSD Δ_b for, (a) For a helical protein 1RIJ where is one dominating minimum corresponding to the native state. Inset: optimal superimposition of simulation structure on the experimental structure. (b) For a protein with more complex native conformation, the free-energy surface may show many competing minima. The overlap of the global free-energy minimum with the PDB structure, as well as the 3D shape of some other minima are also shown

The force field consists of four terms, representing an $1/r^{12}$ excluded volume repulsion between every pair of atoms, a weak local electrostatics term along the protein backbone, an orientation dependent hydrogen bond term, and an effective hydrophobic attraction between non-polar side chains. For a detailed description of the model and the force field, the reader is referred to[1]. The rugged energy landscape of proteins was sampled using simulated tempering and parallel tempering Monte Carlo methods. Both single angle and a semi-local multiple-angle updates were employed on protein conformations. In aggregation studies, rigid body translations and rotations of molecules were used additionally. All simulations start with random initial conditions of the molecules. The simulations were done using the program package PROFASI[4], a C++ implementation of the model.

3 Results

We studied the folding behaviour of several small proteins of helical (Trp-cage, Fs, 1RIJ), β-sheet (GB1p, GB1m2, GB1m3, betanova, LLM, beta3s) and mixed secondary structure elements (BBA5). Unbiased simulations in the model, starting from random initial conformations are able to identify the native states of each of these proteins with one and the same choice of parameters for the energy function. The free-energy surfaces obtained in the model are found to be of different characters, depending on the eventual native fold of the protein[2]. Small proteins like Trp-cage and 1RIJ, with simple helical secondary structures have simple free-energy landscapes with one dominating minimum corresponding to the native state, as for example, in Fig. 1 (a). β–sheet proteins as well as proteins with mixed secondary structures show much more complex landscapes with several minima with significant free-energy barriers between them. One example of such a surface is shown in Fig. 1 (b). Estimates of the folding populations at experimental temperatures as well as the change of stability of peptides due to mutations agrees well with experiments.

We have also studied multiple chain systems of the Aβ_{16-22} peptide, a segment of the Alzheimer's β-amyloid peptide, that is experimentally known to form amyloid fibrils with an in-register anti-parallel cross-β organization of the strands. In simulated tempering

Figure 2. Oligomers of Aβ_{16-22} from 9 chain simulations. Many stable oligomeric forms are found, even a 9 chain barrel form (bottom right). In simulations, small oligomers show a weak preference towards ordered, anti-parallel β-strand organization. But such preference is seen to grow for the larger oligomeric species.

simulations of systems of 1, 3, 6 and 9 chains of Aβ_{16-22} peptides, the Aβ_{16-22} peptides self-assemble into β–sheet rich oligomers[3]. The isolated Aβ_{16-22} peptide is found to be unstructured, while multi-chain systems develop into a variety of different oligomeric species with a marked increase in β–sheet content, cf. Fig. 2. Of particular interest is the spontaneous formation of a 9 stranded β-barrel as one of the oligomeric species. A study of the population of parallel vs anti-parallel pairs of strands shows that larger oligomers tend to contain less defects.

4 Concluding Remarks

The simple and physically well motivated form of the interaction potential given in[1] appears to result in good global properties of the protein energy landscape. If simulations are interpreted carefully, keeping in mind the known weaknesses of the model, they could be used to extract meaningful physical predictions about the molecules in a wide variety of applications.

Acknowledgments

This work was partly supported by National Science Foundation (USA) grant (CHE-0313618) and the National Institute of Health (USA) grant No. GM62838.

References

1. A. Irbäck and S. Mohanty : Biophys. J. **88**, 1560 (2005).
2. S. Mohanty and U.H.E. Hansmann : Biophys. J. **92**, 3573 (2006).
3. G. Favrin, A. Irbäck, and S. Mohanty : Biophys. J. **87**, 3657 (2004).
4. A. Irbäck and S. Mohanty : J. Comp. Chem., **27**,1548 (2006), Internet address: `http://cbbp.thep.lu.se/activities/profasi/`.

Channel Transport and Molecular Motors without Brownian Ratchets

Walter Nadler[1] **and Wolfgang R. Bauer**[2]

[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

Brownian rotors are a large and important class of molecular motors in biological systems. Usually the function of these molecular machines is described as being a Brownian ratchet. However, Brownian ratchets need non-equilibrium fluctuations of an asymmetric potential for its function. We demonstrate that the Brownian ratchet paradigm does not apply to Brownian rotors. Rather, recent analytical results on facilitated transport through biological channels (W. R. Bauer & W. Nadler, PNAS 103, (2006) 11446) can be employed to understand also these purely concentration gradient-driven systems.

1 Introduction

Rotary molecular motors play a central role in biological cells where they transduce electrochemical or chemical energy into mechanical force. An archetype is the F0 portion of the F-ATP synthase which converts the electrochemical energy difference of protons or sodium cations across the inner membrane of mitochondria into a mechanical torque[1]. It consists of a rotary ring, which works as a proton or sodium carrier across the membrane[2].

Surprisingly, such rotary molecular motors are usually described as Brownian ratchets[1,3]. However, Brownian ratchets need non-equilibrium fluctuations of a force field to function[4,5]. But the only non-equilibrium aspect of biological molecular rotary motors is a concentration gradient across the membrane.

This situation is reminiscent of the problem of facilitated and asymmetric transport in biological channels[6,7]. There also non-equilibrium fluctuations of a force field were discussed to describe that phenomenon[8]. However we were able to show that a thorough analysis of the diffusive transport is sufficient[9]. We will sketch that theory and discuss how it can be applied to the situation of Brownian rotors. To compare the basic biological situations, see also Fig. 1.

2 Theory

The dynamics of the density of the molecules inside a channel, $\rho(x,t)$, is determined by the Smoluchowski equation[10],

$$\partial_t \rho(x,t) = D\partial_x \left[\partial_x - F(x)\right] \rho(x,t) \ , \tag{1}$$

where x is the channel coordinate and D is the diffusion coefficient. $F(x)$ is the force that describes 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

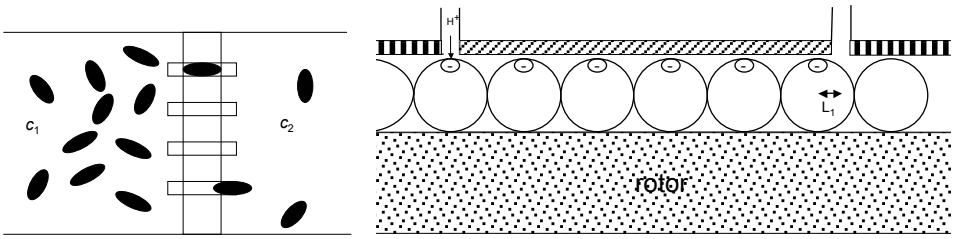


Figure 1. Basic biological situations: (*left*) A membrane separates two baths with molecular concentrations c_1 and c_2. The baths are connected by channels (hatched rectangles) allowing only access of a single molecule. (*right*) Brownian rotor driven by a H^+ gradient. The relevant component of the motor is a ring carrying identical hairpin-like protomers in which the amino acid Asparagine works as proton carrier. In order to emphasize the analogy to the channel situation, the rotor is drawn linearly, and periodic boundary conditions are assumed.

baths hold the molecular densities constant at $\rho(0,t) \equiv c_1$ and $\rho(L,t) \equiv c_2$, respectively, as seen e.g. in Fig. 1a.

Extending an old approach by Hardt[11] we have shown recently[12] that in very general situations described by a Smoluchowski equation (1) the flow J of non-interacting particles across some region is given by a macroscopic version of Fick's law[13]

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

where τ is the mean first passage time[10] to cross the region and n is a measure of the steady state particle number in that region.

In reality transported molecules are interacting, in particular they cannot pass each other. In order to allow for that fact the above approach has to be extended. We now interpret $\rho(x,t)$ as the probability density that a channel contains a particle at x. The empty channel has to be added as an additional state, the probability of which we will denote as ρ_0. This additional empty channel state leads to a cyclic state model, see Fig. 2, since a molecule can enter the empty channel from either side. The corresponding transition rates between the empty channel state and the states in which a molecule is attached to either end $\rho_1 = \rho(0,t)$, $\rho_2 = \rho(L,t)$ are proportional to the concentration of molecules of the adjacent baths which we denote again as c_1 and c_2. In the steady state the equation system

$$J_{0\to 1} = c_1\, k_+^{(1)}\, \rho_0 - k_-^{(1)}\, \rho_1\ , \tag{3}$$

$$J_{1\to 2} = \frac{n}{\tau}\,(\rho_1 - \rho_2)\,, \tag{4}$$

$$J_{2\to 0} = k_-^{(2)}\, \rho_2 - c_2\, k_+^{(2)}\, \rho_0\ , \tag{5}$$

holds, where Eq. (4) derives from Eq. (2). $k_\pm^{(i)}$, $i = 1, 2$, are the reaction rate constants describing attachment of molecules to and dissociation from either end of the channel. In the steady state all flows of Eqs. (3-5) are identical. In addition, we have to consider that the total probability is conserved. Including such a normalization condition[9] we obtain an expression for the flow similar to Eq. (2) where n is replaced by a nonlinear function of n

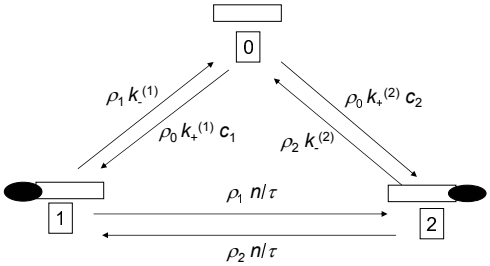


Figure 2. Cyclic state model of molecular transport through the channel that allows for interacting particles. Three states of the channel are depicted: 0 refers to the empty channel, 1 and 2 are channel states with a single molecule attached either of at one or the other end. The respective unidirectional flows between the states are shown above the corresponding arrows.

(and other parameters),

$$J = \frac{f\,(n)}{\tau}\ (c_1 - c_2)\ . \tag{6}$$

This function shows a nonlinear dependence on the concentrations c_1 and c_2 that exhibits analogies to a Michaelis-Menten-type[14] behavior:

$$f(n) = \frac{n}{1 + n\,(c_1 + c_2) + \Delta n\,(c_1 - c_2) + \frac{n}{\tau}\left[\frac{1}{k_-^{(1)}} + \frac{1}{k_-^{(2)}} + 2n\left(\frac{c_2}{k_-^{(1)}} + \frac{c_1}{k_-^{(2)}}\right)\right]}\ . \tag{7}$$

Consequences from this expression on transport facilitation, optimality of parameters, and asymmetric transport are analyzed in detail in Ref. 9.

3 Discussion

In Brownian rotors the transport of H^+ bound to the protomers, see Fig. 1b, is coupled directly to the movement of the rotor. Therefore, applying the above theory for determining the H^+-current also gives expressions for the rotation. Note, however, that in the above considerations no potential difference between the baths was assumed. But Brownian rotors under workload are best described by an additional macroscopic potential gradient[3]. Our model can be readily adapted to such a situation when $\Phi(0) = \Phi_1 \neq \Phi(L) = \Phi_2$. In particular, flow vanishes if the chemical potentials of the baths $\mu_i = \Phi_i + \ln(c_i)$ are equal. Using potential corrected specific particle numbers $\tilde{n}$ and first passage times $\tilde{\tau}$[9], Fick's diffusion law in Eq. (2) can be generalized to

$$J = \frac{\tilde{n}}{\tilde{\tau}}\,(e^{\mu_1} - e^{\mu_2}) \quad . \tag{8}$$

Also all other results Ref. 9, e.g. Eq. (7), can be generalized to this situation by replacing concentrations c_i by the activities e^{μ_i} and introducing the corresponding potential corrected parameters. In this way a straightforward analysis of Brownian rotors is readily possible[15].

Acknowledgments

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

References

1. W. Junge (2004) *Photosynth. Res.* **80** 197.
2. See `http://www.biologie.uni-osnabrueck.de/biophysik/junge/` for an illustrative animation.
3. H. Wang (2005) *J. Phys.: Condens. Matter* **17**, S3997.
4. Frey, E. & Kroy, K. (2005) *Ann. Phys. (Leipzig)* **14**, 20.
5. P. Hänggi, F. Marchesoni, & F. Nori (2005) *Ann. Phys. (Leipzig)* **14**, 51.
6. Schwarz, G., Danelon, C., & Winterhalter, M. (2003) *Biophys. J.* **84**, 2990.
7. Alcaraz, A., Nestorovich, E.M., Aguilella-Arzo, M., Aguilella V. M., & Bezrukov, S.M. (2004) *Biophys. J.* **87**, 943.
8. Kostin, I. & Schulten, K. (2004) *Phys. Rev. Lett.* **93**, 238102.
9. Bauer, W.R. & Nadler W. (2006) *Proc. Natl. Acad. Sci. USA* **103**, 11446.
10. Gardiner, C.W. (1985) *Handbook of Stochastic Methods*, 3rd ed. (Springer, Berlin).
11. Hardt, S. (1981) *Bull. Math. Biol.* **43**, 89.
12. Bauer, W.R. & Nadler W. (2005) *J. Chem. Phys.* **122**, 244904.
13. D A. Fick (1855) *Ann. Phys.* **94**, 59.
14. L. Michaelis & M. L. Menten (1913) *Biochem. Z.* **49**, 333.
15. Bauer, W.R. & Nadler W. (2007) manuscript in preparation.

Aggregate Size and Shape Distributions in Amyloid-β Peptide Solutions

Luitgard Nagel-Steger[1], **Borries Demeler**[2], **and Dieter Willbold**[1]

[1] Institute for Neurosciences and Biophysics, INB-2, Research Centre Jülich, D-52425 Jülich, and Institut für Physikalische Biologie & BMFZ, Heinrich-Heine-Universität Düsseldorf, D-40225 Düsseldorf
E-mail: nagelst@biophys.uni-duesseldorf.de

[2] Dept. of Biochemistry, University of Texas Health Science Center 7703 Floyd Curl Drive, US- San Antonio, Tx, USA
E-mail: demeler@biochem.uthscsa.edu

A peptide with 42 amino acid residues (Aβ (1–42)) plays a key role in the pathogenesis of the Alzheimer's disease. Aggregation of this peptide leads to the formation of assemblies of different sizes and shapes, among which the amyloid fibril is the most prominent member. We used sedimentation velocity centrifugation to analyze Aβ (1–42) peptide solutions at different time points during aggregation regarding their hydrodynamic properties. New data evaluation software allowed us to determine not only sedimentation coefficients, but also at the same time the related frictional ratios of species present in a multicomponent system. Shortly after sample preparation, a sharp peak dominated the measured s-value distribution. Two-dimensional spectrum analysis assigned this species an s-value of $28 \cdot 10^{-13}$ s, a molecular weight of $1.23 \cdot 10^6$ and a frictional ratio of 1.44. Under the assumption of a rigid rod model this structural information is in good agreement with a fibril of 50 nm length build out of two neighboring β-sheets with about 110 monomers each, arranged in a parallel in-register fashion. Incubation for 1, 2 and 5 days causes further aggregation, which results in decreasing frictional ratios, implying a growth to more spherical-like particles.

1 Introduction

A peptide with 42 amino acid residues (Aβ (1–42)) plays a key role in the pathogenesis of the Alzheimer's disease. It is highly prone to self aggregation, leading to the formation of fibrils which are deposited in amyloid plaques in the brain of affected individuals[1]. In previous years increasing evidence arose that probably smaller oligomeric assemblies[2,3] play a more decisive role as neurotoxic agents than the mature fibril. Information about size and shape of Aβ peptide assemblies formed during aggregation is therefore of high relevance.

Analytical ultracentrifugation is a method for retrieving structural information about macromolecules by direct observation of their hydrodynamic properties in a centrifugal field. Advanced data analysis permits the determination of s-value, molecular weight and shape distributions for multicomponent systems.

Our objective is to evaluate the applicability of analytical ultracentrifugation in combination with the newly developed data evaluation software for generating size and shape distributions of solutions of the Aβ peptide. The method shall be used to monitor the effects of peptide concentration, solvent and added cosolutes upon aggregation.

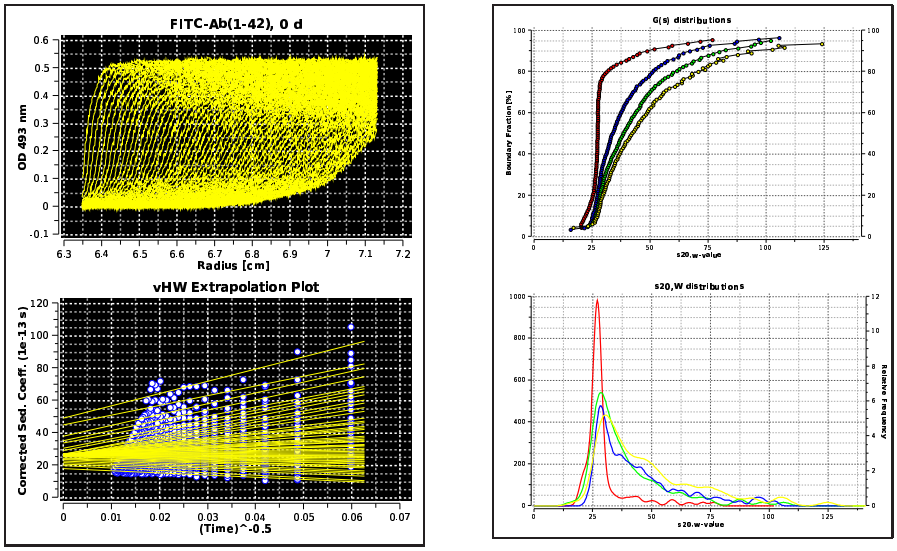


Figure 1. Van-Holde-Weischet Analysis. Left: Raw data (top) and corresponding vHW-analysis plot (bottom) for Aβ aggregation mixture 2 h after sample preparation. Sample volume was 300 μl. Right: Upper graph: G(s) distribution of sample after 0 d (red), 1 d (blue), 2 d (green), and 5 d (yellow) incubation at 20 ℃. Lower graph: s-value distribution of sample after 0 d (red), 1 d (green), 2 d (blue), and 5 d (yellow) incubation at 20 ℃.

2 Methods

The aggregation mixture contained 70 μM unlabeled and 14 μM FITC-labeled Aβ (1–42) in 10 mM sodium phosphate buffer, pH 7.4 with 6 % final DMSO concentration. Sedimentation velocity experiments were performed with an X-LA analytical ultracentrifuge (Beckman-Coulter), equipped with absorption optics. Samples were measured in standard double-sector aluminum cells at 20,000 rpm, 20 ℃. Radial step size was set to 0.001 cm.

Raw data from the analytical ultracentrifuge were processed and evaluated using the UltraScan software package[4] running on a 44 node AMD Opteron cluster under Linux. After determination of suitable start parameters by a van-Holde-Weischet analysis[5] the 2-dimensional spectrum analysis[6] (2DSA), implemented in the software package, was started. 2DSA solves the inverse problem of fitting sedimentation velocity data to a linear combination of finite element solutions of the Lamm equation. Each term of the linear combination reflects a solute in the 2-dimensional space over s and f/f_0. Finally Monte-Carlo simulations were used to identify statistically significant solutes.

3 Results

Data analysis of sedimentation velocity runs performed at 20,000 rpm, 20 ℃ at 4 time points during aggregation of Aβ (1–42) according to van-Holde-Weischet as depicted in Fig. 1 showed an increase of the mean s-value over the incubation time of 57 %. Aggregation of the amyloid-β peptide starts with a narrow distribution with an s_{avg}-value of 30 S. The distribution is dominated by a species of 28 S, a molecular weight of $1.22 \cdot 10^6$

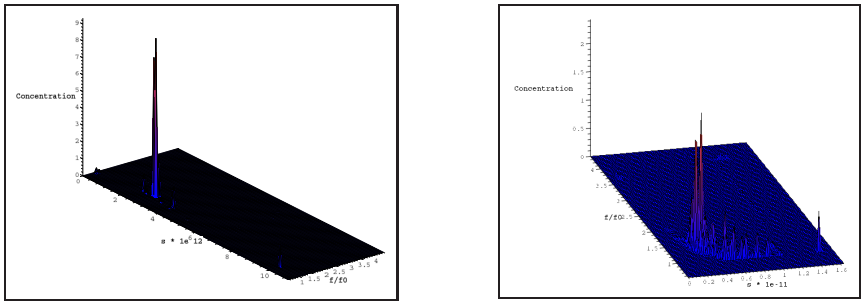


Figure 2. 3D-Plots of s-value *vs* f/f_0 distributions. Left: Aβ(1–42) after 2 h incubation at 20 °C. Right: Aβ(1–42) after 5 d incubation at 20 °C.

and a frictional ratio of 1.44, which comprises about 40 % of the sedimentation boundary (Fig. 2, left). This aggregate species is reproducibly detected in repeated experiments. It can be modeled by a rod-like particle with an axial ratio of 8.6. Under the assumption of a rigid rod model and molecule dimensions for β-strands this structural information is in very good agreement with a fibril of 50 nm length build out of two neighboring β-sheets with about 110 monomers each, arranged in a parallel in-register fashion, as proposed by[7].

Upon incubation, the measured distribution is asymmetrically broadening to higher s-values with a still distinguishable maximum at about 28 S. During the 5 d incubation of the sample, the f/f_0-value of detected solutes decreases from 1.6 to around 1 with s-values increasing from 28 S to 150 S. Particles are becoming more spherical during the incubation time within the observed growth range (Fig. 2, right). The method is a tool to study the action of drugs on the aggregation of Aβ (1–42) peptide.

References

1. J.A. Hardy and G.A. Higgins, *Alzheimer's disease: the amyloid cascade hypothesis*, Science, **256**, 184–185, 1992.
2. W.L. Klein, *Fibrils, proteofibrils and Aβ-derived diffusible ligands: How Aβ causes neuron dysfunction and death in Alzheimer's disease*, Humana Press, Totowa, NJ, pp. 1–49, 2001.
3. D.M. Walsh, J.V. Klyubin, I. Fadeeva, W.K. Cullen, R. Anwyl, M.J. Wolfe, M.S. Rowan, and D.J. Selkoe, *Natural secreted oligomers of amyloid β protein potently inhibit hippocampal longterm potentiation in vivo*, Nature, **416**, 535–539, 2002.
4. B. Demeler, *UltraScan*, `http:www.ultrascan.uthscsa.edu`, version 9.0.
5. B. Demeler, and K.E. van Holde, *Sedimentation velocity analysis of highly heterogeneous systems*, Anal. Biochem., **335**, 279–288, 2004.
6. E.H. Brookes, R.V. Boppana, and B. Demeler, *Computing Large Sparse Multivariate Optimization Problems with an Application in Biophysics*, in: SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, vol. 6, p. 81, 2006.
7. A.T. Petkova, W.-M. Yau, and R. Tycko, *Experimental Constraints on Quaternary Structure in Alzheimer's β-Amyloid Fibrils*, Biochemistry, **45**, 498–512, 2006.

PSO@Autodock : A Novel Bio-Algorithm-Based Fast Flexible Docking Tool for Virtual Screening

Vigneshwaran Namasivayam and Robert Günther

Institute of Biochemistry, University of Leipzig,
Brüderstrasse 34, 04103 Leipzig, Germany
E-mail: {vignesh, robguent}@uni-leipzig.de

Virtual Screening (VS) has become an essential element in the drug discovery. Based on the 3D structure of the receptor, molecular docking studies can be applied to screen large libraries of compounds for candidate molecules specific for the particular target. Molecular docking simulation methods describe protein-ligand interactions at atomic level within a force field-based approach. However, due to their high computational costs, molecular docking simulation methods have rarely been applied to VS. We present PSO@Autodock, a novel molecular docking molecular tool for virtual screening. It is based on Particle Swarm Optimization (PSO) and has been implemented in AutoDock3 (AD3), a widely used docking program. In comparison to standard AD3, PSOAutodock requires less than 20% of computing time for predicting the correct protein-ligand complex. Thus, simulating the docking of 2000 compounds with full flexible treatment of the ligands employing PSO@Autodock can now be finished over night on a standard personal computer.

1 Introduction

Virtual Screening (VS) techniques applying computational molecular docking methods have proven to be a viable alternative to experimental High Throughput Screening (HTS). Given the 3-D structure of the protein target, molecular docking methods allow screening large libraries of compounds and provide detailed information on the protein-ligand interactions.

Currently, most VS strategies employ geometric matching methods that dock an ensemble of ligands and rank them based on empirical scoring functions. On contrary, simulation methods apply force fields to describe the molecular docking process in more detail. However, this is computational demanding. Simulation methods rely on the assumption that the protein-inhibitor complex with the lowest calculated binding energy (ΔE_{Bind}) is the closest to the native one. Thus, predicting the protein-inhibitor complex can be considered as multidimensional optimization problem. Various optimization algorithms have been proposed to solve this flexible docking problem. AutoDock3[1] (AD3) is one of the widely used docking programs of this type and utilizes a Lamarckian Genetic Algorithm (LGA) solve this optimization problem. It allows flexible docking of the ligand while treating the receptor as rigid. In AD3 the binding energy of the ligand is predicted based on Eq. 1

$$\Delta E_{Bind} = \Delta E_{vdW} + \Delta E_{elec} + \Delta E_{hbond} + \Delta E_{desolv} + \Delta E_{tors} \quad (1)$$

2 PSO@Autodock

PSO@Autodock, our fast flexible docking tool for virtual screening is based on Particle Swarm Optimization[2] (PSO) and has been implemented in AD3. It employs *var*CPSO-ls

(***v***elocity ***a***daptive and ***r***egenerative **C**onstriction **P**article **S**warm **O**ptimization with **l**ocal **s**earch) algorithm. PSO is a form of swarm intelligence and has been developed to describe the social behavior of flocking birds. If one member of the swarm (particle) sees a desirable path to go (i.e. the global minimum of ΔE_{Bind}) the rest of the particles will follow quickly even if they are on the opposite side to the particle in the multidimensional hyperspace. In *var*CPSO-ls, each particle is initialized at random position in the real valued search space. The position of the particle *i* is represented by the vector x_i = (x_{i1}, x_{i2}, ..., x_{iN}) where *N* is the dimension of the problem. The velocity of each particle is given by the vector v_i = (v_{i1}, v_{i2}, ..., v_{iN}). Furthermore, each particle has two kinds of memory that influence the movement in the search space. The cognitive memory p_i = (p_{i1}, p_{i2}, ..., p_{iN}) stores the best previous position visited by each individual particle. The social memory p_g = (p_{g1} , p_{g2}, ..., p_{gN}) contains the position of the best point in search space visited by all particles in the swarm. In each swarm move, the particle velocity is updated according to Eq. 2

$$v_i(t+1) = \chi \cdot \{v_i(t) + rand(0,1) \cdot c_1 \cdot (p_i - x_i(t)) + rand(0,1) \cdot c_2 \cdot (p_g - x_i(t))\}, \quad (2)$$

with the constriction factor $\chi = \frac{2}{2-\sqrt{\phi^2-4}}$. ϕ is computed with cognitive and social parameters c_1 and c_2, according to $\phi = c_1 + c_2$, $\phi > 4.0$.
After the velocity has been calculated, the positions of the particles are updated according to Eq. 3

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (3)$$

In *var*CPSO-ls, the constriction factor χ controls the magnitude of the particle velocity leading to a fast convergence. *var*CPSO-ls is characterized by three important features, which help to reduce the computational cost while maintaining the diversity of the swarm and preventing from premature convergence in local minima. (i) *velocity adaptive*: The velocity of the particle is updated only if the current fitness of the particle is lower than its fitness at previous iteration. (ii) *regenerative*: If a particle moves out of the search space, the particle is removed from the swarm and regenerated with a new random position. (iii) *local search*: At each swarm move, the best particle of the swarm S_{best} undergoes local search according to Solis and Wets.

3 Performance of PSO@Autodock

We accessed the performance of *var*CPSO-ls implemented in PSO@Autodock employing a suite of seven different protein-ligand complexes and compared it to the default LGA of AD3 (Table 1). The optimization process was performed with 1.5 million evaluations for LGA and 50,000 evaluations for PSO@Autodock. Fig.(1) indicates that *var*CPSO-ls of PSO@Autodock clearly outperforms the default LGA of AD3. In order to compare the performance on VS we selected the complex of HIV protease 1 with XK263 (1hvr.pdb) for screening the NCI diversity data set[3] (1990 compounds with unique scaffolds). Since PSO@Autodock uses the same objective function (Eq. 1), the correlation coefficient between the predicted ΔE_{Bind} of LGA and that of *var*CPSO-ls is 0.94.
The performance speed of PSO@Autodock is comparable to docking programs tailored for VS, like Glide (version4.0, Schrödinger Inc, San Diego, CA 92122) or GOLD (version3.0,

	LGA		*var*CPSO-ls		Top Hits in VS (*var*CPSO-ls)	
	ΔE_{Bind}	RMSD	ΔE_{Bind}	RMSD	NCI Number	ΔE_{Bind}
PDB ID	kcal/mol	Å	kcal/mol	Å		kcal/mol
3ptb	-8.15	0.32	-8.15	0.29	NCI371876	-11.34
2cpp	-7.40	1.76	-7.32	1.09	NCI101825	-10.57
2mcp	-5.45	1.09	-5.02	0.95	NCI12181	-10.99
1stp	-10.93	0.46	-10.85	0.39	NCI3354	-11.96
1hvr	-21.31	0.69	-21.34	0.29	NCI343227	-14.62
4hmg	-8.39	0.78	-8.30	0.80	NCI179187	-10.18
4dfr	-13.03	1.10	-12.64	0.96	NCI371884	-13.24

Table 1. Binding energies and rmsd values predicted by LGA and *var*CPSO-ls. Shown are also the top hit of the NCI diversity data set[3] screening employing PSO@Autodock.

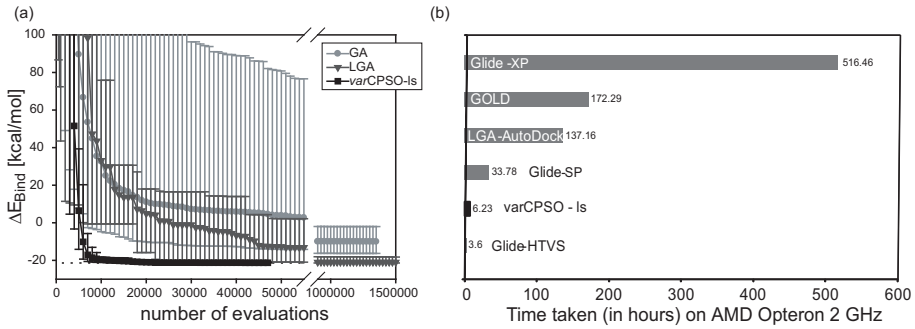


Figure 1. (a) Performance of *var*CPSO-ls and genetic algorithms(GA and LGA) on 1hvr.pdb. (b) Computing times for screening the NCI diversity data set on 1hvr.pdb.

CCDC, Cambridge UK). Thus, PSO@Autodock employing the novel *var*CPSO-ls algorithm is well suited for virtual screening of large libraries within an accurate force field approach.

Acknowledgments

We are indebted to SFB 610 for financial support.

References

1. G.M. Morris, D.S. Goodsell, R.S. Halliday, W.E. Hart, R. Belew, A.J. Olson, *Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function*, J. Comp. Chem. **19**, 1639, 1998.
2. J. Kennedy, and R.C. Eberhardt, *Particle Swarm Optimization*, Proc. IEEE Intl. Conf. Neural Networks, Piscataway, NJ , 1942, 1995.
3. `http://dtp.nci.nih.gov/`

Efficient Parallel Tempering with Multiple Gaussian Modified Ensembles

M. Magiera[1], T. Neuhaus[2], and U. H. E. Hansmann[2,3]

[1] Department of Theoretical Physics, University Duisburg-Essen, 47048 Duisburg, Germany

[2] John von Neumann Institute for Computing, FZ Jülich, D-52425 Jülich, Germany
E-mail: {t.neuhaus, u.hansmann}@fz-juelich.de

[3] Department of Physics, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931-1295, USA

We present some aspects of multiple Gaussian modified ensemble simulations (MGME) for the efficient Monte Carlo studies of temperature driven first order phase transitions.

It is a matter of fact, that most of nature's phase transitions are of discontinuous i.e., first order nature. In first order phase transitions one finds mixed phases, which for *finite sized* systems have finite and possibly small probability to be found in experiments as well as in computer simulations. Also frequent experimentation with nano-devices, as well as with soft matter like proteins, highlights the need for understanding the thermodynamics of finite sized systems[1]. These systems often are made from mixed phases, inhomogeneous in nature, one example being the aggregation of protein look alike molecules in a box, or the simple nucleation of droplets.

In this short note we turn to one of the simplest $2d$ spin models with a mixed phase, namely the $2d$ Potts model at a number of spin states $Q = 20$ and $Q = 256$, which both display strong first order behaviour. The $2d$ Potts model is a perfect testing ground for algorithmic developments as due to Baxter's and others work many observables at the first order phase transition are exactly known. MGME Monte Carlo simulations[2] are designed to calculate the density of states $g(E)$, that is the number of configurations at energy values E, for all values of the energy E.

The idea of MGME simulations consists in writing the partition function as a product

$$Z = \prod_{i=1}^{N_{rep}} Z_i \tag{1}$$

of $i = 1, ..., N_{rep}$ partition functions Z_i in such a way, that the normalised probability distribution functions $PDF_i(E)$ in the energy E of the replicas fulfil

$$0.63 \approx \sum_E \min[PDF_i(E), PDF_{i+1}(E)]. \tag{2}$$

Of course this overlap criterium is taken in-between all neigbouring factors i and $i + 1$ of the product form Z, which then ensures a broad sampling of all energy values on the interval $-2V, ..., 0$, where the Potts energy is $E = -\sum_{nn} \delta_{q_i,q_j}$ and V is the volume. Quite naturally also, one will distribute the various partition functions Z_i on the nodes of a parallel computer, and, utilising local spin updates, as well as parallel tempering swaps[3] in-between neighbouring partition functions Z_i and Z_{+1}, one will record the

multi-histogram $\mathrm{MHIST}(E)$, for the probability of energy E occurring in Z. The multi-histogram $\mathrm{MHIST}(E)$ is then used to determine a stochastic estimate of $g(E)$. Modifications using Bennett's method[4] will be included into a forthcoming publication. The peculiar choice of the constant 0.63 in eq.(2) yields swap acceptance rates $P_{acc} \approx 0.5$ for unimodular, that is almost Gaussian PDF's.

In detail we factorise Z into $Z = Z_{\mathrm{lowT}} \times Z_{\mathrm{MGME}} \times Z_{\mathrm{highT}}$, where the first and third factors take care of broad histogram sampling in low and high temperature phases. Z_{MGME} covers the mixed phase at β_T for energy values $e_o V \leq E \leq e_d V$. For $Q = 20$ we have the energy density values $e_d = -0.62652917$ and $e_o = -1.82068443$, while $\beta_T = \ln[1 + \sqrt{Q}]$ generally. All of these three partition functions are itself products of Z_i functions and with $\beta = 1/T$ we write

$$Z_{\mathrm{lowT}} = \prod_{i}^{N_1} Z_{\mathrm{lowT,i}} \quad ; \quad Z_{\mathrm{lowT,i}} = \sum_{conf.} e^{-\beta_i E}\, \Theta(\frac{e_o + e_d}{2} - e) \tag{3}$$

$$Z_{\mathrm{MGME}} = \prod_{i}^{N_2} Z_{\mathrm{MGME,i}} \quad ; \quad Z_{\mathrm{MGME,i}} = \sum_{conf.} e^{-\beta_T E - [(E - E_0^i)/\Delta E_0]^2} \tag{4}$$

$$Z_{\mathrm{highT}} = \prod_{i}^{N_3} Z_{\mathrm{highT,i}} \quad ; \quad Z_{\mathrm{highT,i}} = \sum_{conf.} e^{-\beta_i E}\, \Theta(e - \frac{e_o + e_d}{2}) \tag{5}$$

with $N_{rep} = N_1 + N_2 + N_3$ and Θ denotes the Heaviside step function.

One important measure of computational efficiency in broad histogram sampling methods is the ergodicity time scale τ_{erg}[6,7], sometimes also called tunnelling autocorrelation time, for the completion - in the mean - of one single "tunnelling event" from the low to

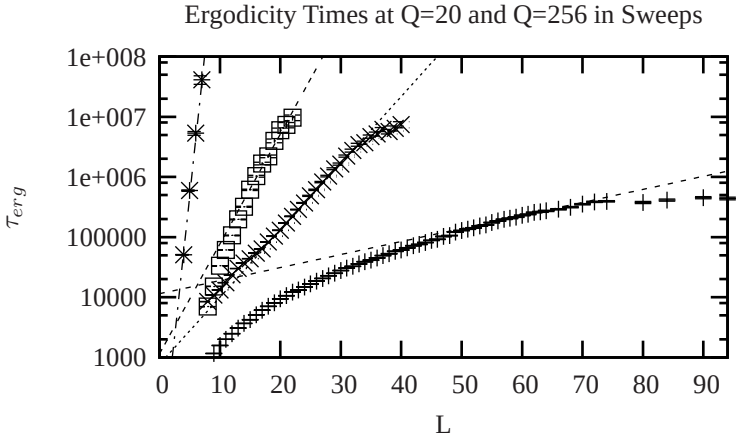


Figure 1. Parallel tempering ergodicity times τ_{erg} for $Q = 256$ (stars) and $Q = 20$ (squares) Potts models *without* MGME improvement. With MGME simulation we obtain the two data sets with small slopes for $Q = 256$ (large crosses) and $Q = 20$ (somewhat smaller crosses). All straight lines have slopes that can be calculated from the free energy of a classical lens shaped droplet at the droplet/strip transition on lattices with p.b.c.[5].

high - or high to low - temperature regions in phase space. We display our final τ_{erg} results from high statistics simulations, at about 1 Giga-Sweeps for each data point - and in a logarithmic scale, in the $Q = 20$ and $Q = 256$ Potts models in Fig.1. The figure contains four data sets. Two of these, the ones with the largest slopes in the figure, correspond to parallel tempering simulations at $Q = 256$ and $Q = 20$ *without* MGME improvement. They exhibit an asymptotic, that is large system size L, supercritical slowing down with τ_{erg} diverging as

$$\tau_{erg} \propto e^{+\,1.1346\;2\;L\;\sigma}. \tag{6}$$

The quantity σ hereby denotes the known planar interface tension in-between ordered and disordered phases at β_T in the Potts model. It is clear from these data, why standard parallel tempering at first order phase transitions does not yield useful information. However *with* MGME improvement, see the two data sets with small slopes in the figure, one finds residual supercritical slowing with τ_{erg} diverging as

$$\tau_{erg} \propto e^{+\,(1.1346-1)\;2\;L\;\sigma}, \tag{7}$$

and thus:

- MGME simulations for first order phase transition studies have a similar level of computational efficiency as Wang Landau[8] and multicanonical ensemble simulations[9].

In conclusion, we think that MGME simulations are useful for the study of mixed phases of finite sized thermodynamics systems.

Acknowledgments

The calculations were done on computers of the John von Neumann Institute for Computing, Forschungszentrum Jülich, Jülich, Germany, and on some privately owned (TN) workstations. This work was supported, in part, by a research grant (CHE-0313618) from the National Science Foundation (USA).

References

1. D.H.E. Gross, *Microcanonical Thermodynamics*, (World Scientific, Singapore, 2001).
2. T. Neuhaus and J.S. Hager, Phys. Rev. E **74** (2006) 036702.
3. K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65** (1996) 1604.
4. C.H. Bennett, J. of Comp. Phys., **22** (1976) 245.
5. T. Neuhaus and J.S. Hager, J. Stat. Phys. **113** (2003) 47.
6. W. Nadler and U. Hansmann, Phys. Rev. E **75** (2007) 026109.
7. S. Trebst, M. Troyer and U. Hansmann, J. Chem. Phys. **124** (2006) 174903.
8. F. Wang and D.P. Landau, Phys. Rev. Lett. **86** (2001) 2050; F. Wang and D.P. Landau, Phys. Rev. E **64** (2001) 056101.
9. B.A. Berg and T. Neuhaus, Phys. Lett. B **267** (1991) 249; B.A. Berg and T. Neuhaus, Phys. Rev. Lett. **68** (1992) 9.

A Network-Based Approach to Biomolecular Dynamics

Frank Noé[1]**, Jeremy C. Smith**[2]**, and Christof Schütte**[1]

[1] DFG Research Center Matheon,
FU Berlin, Arnimallee 6, 14195 Berlin, Germany
E-mail: {noe, schuette}@math.fu-berlin.de

[2] Interdisciplinary Center for Scientific Computing,
University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
E-mail: biocomputing@iwr.uni-heidelberg.de

Molecular dynamics simulation generates large quantities of data that must be interpreted using physically meaningful analysis. A common approach is to describe the system dynamics in terms of transitions between coarse partitions of conformational space. In contrast to previous work that partitions the space according to geometric proximity, we examine here clustering based on kinetics, merging configurational microstates together so as to identify long-lived, *i.e.* dynamically metastable, states. The method is applied to microsecond molecular dynamics simulations of Ala_{12}. The system clearly exhibit metastability, with some kinetically distinct metastable states being geometrically very similar.

1 Introduction

Given the large amount of molecular dynamics (MD) simulation data available for macromolecules, models are needed to analyze the data that capture the essential dynamical features and can be physically interpreted. Here we use transition networks which encode the transitions of the system between a set of discrete states. In a transition network, a vertex corresponds to a discrete conformational state, and an edge corresponds to a transition between two states. Each edge may be weighted by the associated transition rate or probability.

A central question when generating discrete state models is how best to partition state space into discrete states. It is often desirable that the partition reflects the dynamical behavior of the system. In particular, biomolecular function often depends on the ability to undergo transitions between long-lived, or "metastable" states. These inter-state transitions are rare events in which a molecular system stays for long periods of time within one state, before rapidly switching to another.

Geometric clustering approaches do generally not fulfill these claims. Therefore, here we use the Perron Cluster Cluster Analysis (PCCA)[2,3], a dynamical data-based kinetic clustering method to partition the state space. The method is applied to a 4-microsecond molecular dynamics simulation of a 12-mer polyalanine. Using different numbers of metastable states, a hierarchical picture for the kinetics is provided.

2 Kinetic Clustering Method

The state space of the system is first partitioned into fine discrete sets, called "micro-states" (typically on the order of 10^3 to 10^4). This may be done by characterizing the torsion

rotamer state or by using a geometric clustering method with a large number of clusters on the Cartesian coordinates of the molecule with rotation and translation removed. See Ref.[4] for details.

A metastable state is a set of micro-states in which the system stays for a long time before leaving it. Following this idea, a straightforward definition for a partition into metastable states would be the following: Given a discrete trajectory $[i(0), i(\tau), i(2\tau), ...]$ in space of micro-states $\{1, ..., m\}$, find a partition of micro-states into C metastable sets, such that the number of transitions within sets is maximized. This formulation of the problem is known in graph theory as a C-min-cut problem but is, unfortunately, NP hard. A similar partition, however, can be obtained very efficiently with the Perron cluster cluster analysis (PCCA) method which was first introduced in[1]. Its basic idea is the following: The Markov time evolution of the probability distribution of states, $\mathbf{p}(t)$, is given by $\mathbf{p}(t+\tau) = \mathbf{p}(t)\mathbf{T}(\tau)$. This equation can also be expressed in terms of an expansion in left eigenvectors of $\mathbf{T}$, $\mathbf{q}_i$, and corresponding eigenvalues, λ_i, and for longer time increments, $n\tau$:

$$\mathbf{p}(t + n\tau) = \sum_{i=1..m} c_i \lambda_i^n \mathbf{q}_i. \quad (1)$$

For systems fulfilling detailed balance there is only a single eigenvalue $\lambda_1 = 1$, while $\lambda_i < 1$ for all $i > 1$. Thus, for $n \to \infty$ (after infinity time), all terms but the first one vanish and Eq. (1) converges to: $\mathbf{p}(\infty) = c_1\mathbf{q}_1$, where c_1 is simply a normalization factor. Thus, the left eigenvector with the largest eigenvalue, $\lambda_1 = 1$, specifies the stationary distribution (and has therefore only nonnegative entries). All other eigenvectors, $\mathbf{q}_2...\mathbf{q}_m$, are associated with decaying processes, their speed of decay being determined by the magnitudes of the associated eigenvalues. As apparent from Eq. (1), eigenvectors associated with eigenvalues close to 1 correspond to processes which decay very slowly and thus are, by definition, related to transitions between metastable states.

The improved PCCA method[2,3] used here identifies C representative micro-states with maximum pairwise distances in the coordinates of the first C eigenvectors. Each microstate can then be expressed as a convex combination of these representative micro-states and thus be assigned a degree of similarity with each representative micro-state according to the convex coordinates. Therefore, clustering is made by assigning each microstate to the cluster containing the representant with which it has the strongest similarity.

3 Results

A complex network of metastable states is identified. Use of only a few metastable states (*e.g.* 2, 3, ...) leads to the distinction between the "main state" (the basin with the globally-minimal energy and several "kinetic traps" which are thermally accessible and have half-life times similar to, or greater than, that of the main state. By increasing the number of metastable states the main state is further decomposed into conformational subsets. An interesting feature of the networks is that some states which are structurally very similar are not kinetically contiguous, but rather interchange *via* structurally very different intermediates. This clearly demonstrates the usefulness of the kinetic clustering method in biomolecular simulation studies.

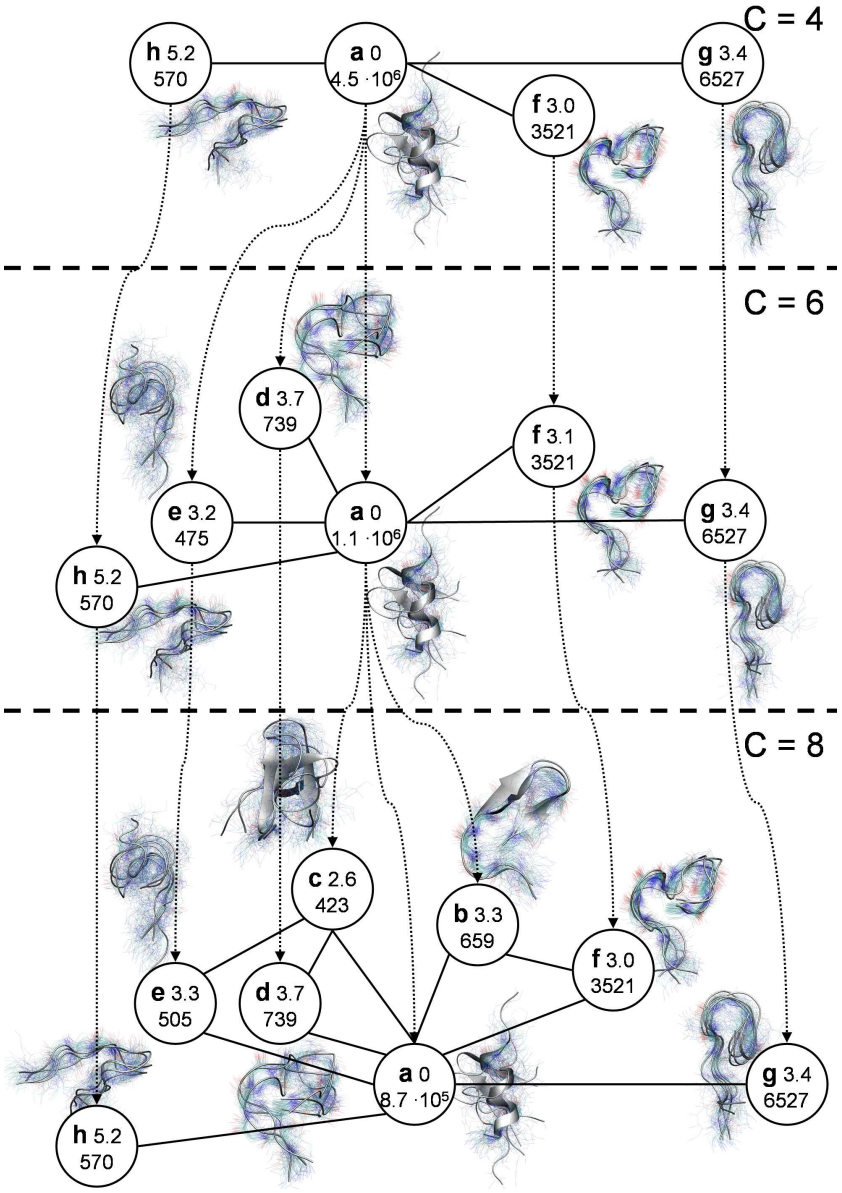


Figure 1. Hierarchical Transition Network analysis for the Ala_{12} peptide for 4, 6 and 8 metastable sets. Each bullet and the structure next to it corresponds to one metastable set. The bullets contain the state name (a letter), the free energy in kcal/mol (upper number) and the mean lifetime in picoseconds (lower number). Each structure is shown by a few representative tubes and an overlay of 100 examples randomly drawn from the ensemble of structures of each state, shown as line representations of the backbone. A pair of states is connected if at least one transition between these states was observed in the trajectory. The hierarchical relationship between the three networks is indicated by the dotted arrows. Each arrow starts at the metastable state in the higher-order network which contains the majority of micro-states in the state the arrow points to. For example, the micro-states of state a in the $C = 4$ network are split into three sub-states, a, d and e, in the $C = 6$ network.

References

1. C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid monte carlo," *J. Comp. Phys.*, vol. 151, pp. 146–168, 1999.
2. M. Weber, "Improved perron cluster analysis," *ZIB Report*, vol. 03-04, 2003.
3. P. Deuflhard and M. Weber, "Robust perron cluster analysis in conformation dynamics," *ZIB Report*, vol. 03-09, 2003.
4. F. Noé, I. Horenko, C. Schütte, J.C. Smith, "Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States" *J. Chem. Phys.*, vol. 126, p. 155102, 2007.

Steered Molecular Dynamics as a Virtual Atomic Force Microscope

Wieslaw Nowak*, Sebastian Wasilewski, and Lukasz Peplowski

Theoretical Molecular Biophysics Group,
Institute of Physics,
N. Copernicus University,
ul. Grudziadzka 5, 87-100 Torun, Poland
**E-mail: wiesiek@fizyka.umk.pl*

Two examples of non-standard applications of computer modeling of protein dynamics are discussed. External harmonic force attached to selected atoms closely simulates a force exerted by a tip in atomic force microscope in single molecule stretching experiments. Preliminary results of forced unfolding of oncogenic protein gankyrin and forced ligand undocking from photoactive industrial enzyme nitrile hydratase are presented.

1 Introduction

Computer modeling based on simple laws of classical mechanics are successfully used in interpretation of single molecule atomic force microscopy experiments[1–4]. The steered molecular dynamics (SMD)[5] or its variants are more and more popular. However, the methodology of doing effective computer experiments is still far from being well established. In this paper preliminary results of applications of the steered molecular dynamics method (SMD) to stretching single proteins are presented. Gankyrin (GAN) is recently discovered oncogenic protein which is highly expressed in many tumors[6]. Its modular structure is close to ankyrins studied earlier[5], but the molecule has only 7 structurally similar subunits. A small size makes GAN a perfect model system for testing SMD computer models. No experimental AFM data or MD simulations have been published so far.

In nitrile hydratase (NHase), biotechnological enzyme used for production of acrylamide, there are two subunits separted by wide channel leading to the unusual active site containing Fe or Co metal ions[8]. The details how the substrate reaches deeply buried catalytic center nor how the product leaves this cavity are not known. The rational modeling new industrial enzymes, for example, having a better stereoselectivity, requires knowledge on steric determinants of NHase activity. We propose that forced unfolding of ligands, both substrates and products, using virtual AFM is promising strategy for further studies of enzymatic mechanisms.

2 Methods

For gankyrin (GAN) simulations the 1TR4 pdb structure was used, for Co-Nhase a 1IRE structure was adopted. NAMD code[7] with CHARM27 force field running on a local linux cluster was used in SMD simulations. To perform Co-Nhase modeling an extensive set of new parameters for non-corrine active site has been developed[8]. Two limiting models of GAN simulations were tested: (a) GAN in a large TIP3 water box and periodic boundary

conditions, 1.45 ns simulation time, $v = 0.1A/ps$, spring constant $K = 7kcal/mol/Ang^2$ (b) initial GAN structure surrounded by a droplet of water, 145 ns simulation time, $v = 0.001A/ps$, $K = 14kcal/mol/Ang^2$. A Langevin coupling to the thermal bath at 300 K was assumed. External forces were attached to CA atom of LEU G8, while GLY G214 CA atom was held fixed. The pulling speed v seems to be the most critical parameter affecting SMD simulations. We estimated this parameter that the dimension of GAN approximately doubled after a period of time planned for the experiment. Data were analyzed using computer graphics and the VMD code[9].

3 Results and Discussion

3.1 Forced Unfolding of Gankyrin

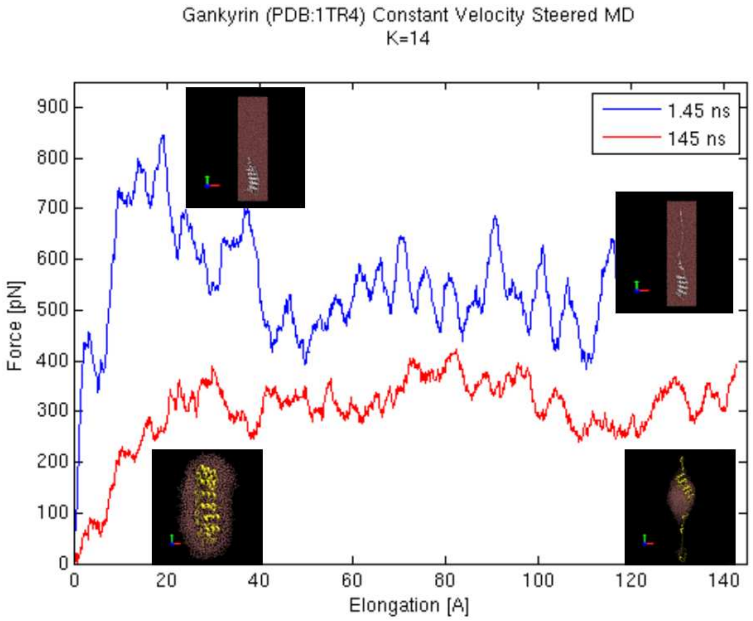


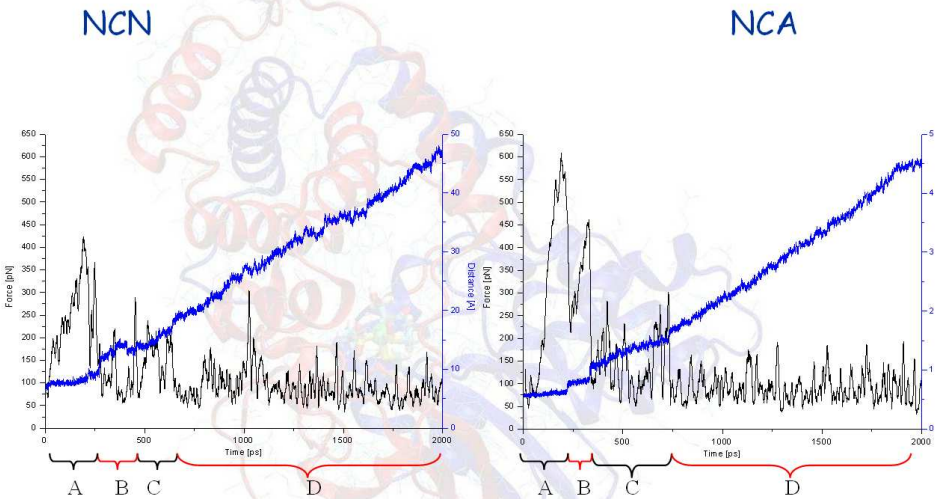
Figure 1. Dependence of force-elogation spectra on SMD pulling velocity in gankiryn.

AFM experiments performed for 24 unit ankyrin[10] have shown that for the initial unfolding of this nanospring a force of the 250 pN is sufficient. Consecutive subunits unfold under even smaller stress of 20-50 pN. Results of our two GAN simulations are shown in Fig. 1. One can immediately see that in 1.45 ns simulation (a) the maximum observed force of 800 pN is far too high. When small pulling speed is applied, in 145 ns simulation a factor of two lower forces are observed. They are still slightly higher than in the related experiment, but this seems to be acceptable. The problem with long simulations is a proper treatment of water, insets in Fig. 1 show that the in long simulation a strong lag in water position

is observed. The water shell becomes quite thin in certain regions of the highly stretched GAN, and this may also result in artifacts lowering unfolding forces. Our experiments show that some algorithms weakly bounding water with the protein undergoing constant elongation (or other large conformational change) have to be developed for effective use of SMD approach. PBC with huge water boxes seem to be too expensive and unpractical.

3.2 Forced Undocking of Nitrile Hydrates Ligands

It is interesting to check what minimal forces are expected if we pull out the well known ligand from the protein interior. This limiting value is perhaps affected not only of the pulling speed, but also by the nature of ligand-protein interactions and an assumed trajectory of the extracted ligand. In Fig. 2 force-travel distance curves are presented for undocking of a substrate (nicotinonitrile, NCN) and a product (nicotinoamide, NCA) from the best docked position within the NHase interior[8]. A rather slow pulling out is a different process for both molecules, despite their similar sizes. We attribute these differences to H-bonds of NCA with appropriate residues of the NHase channel. Critical positions in space determined in these two computer experiment should be good starting points for systematic mutagenesis studies of Co-NHase.



phase	time	V	f_{max}
A	0-280	76	423
B	290-340	263	237
C	365-410	-4	98
D	450-2000	211	188(302)

phase	time	V	f_{max}
A	0-230	27	605
B	230-330	50	459
C	330-730	124	301
D	730-2000	236	192

Figure 2. Dependence of SMD force on simulation time and distance traveled by substrate (NCN) and product (NCA) in Co-NHase. Four phases of motion A-D are recognized.

4 Conclusions

The model protein GAN unfolding process studied by the SMD method on two time-scales (different by a factor of 100) shows quantitatively distinct features in force-extention curves. Forced undocking of NHase ligand gives useful hints for determining the best opportunities for tailored enzyme construction. New methodology of keeping droplet of a water close to the protein is required before a virtual AFM may be used as a common tool for studying mechanical properties of single biomolecules.

Acknowledgments

This research was supported by Polish Ministry of Education and Science, grant no. 2P04A 07229.

References

1. W. Nowak, P. Marszalek, *Molecular Dynamics Simulations of Single Molecule Atomic Force Microscope Experiments, Current Trends in Computational Chemistry*, 47-83, 2005.
2. G. Lee, W. Nowak, J. Jaroniec, Q. Zhang and P.E. Marszalek, J. Am. Chem. Soc. **126**, 6218-6219, 2004.
3. G. Lee, W. Nowak, J. Jaroniec, Q. Zhang and P.E. Marszalek, Biophys. J. **87(3)**, 1456-65, 2004.
4. Z. Lu, W. Nowak, G. Lee, P. Marszalek and W. Yang, J. Am. Chem. Soc. **126**, 9033-41, 2004.
5. M. Sotomayor, D.P. Corey and K. Schulten, Structure **13**, 669-682, 2005.
6. S. Dawson, H. Higashitsuji, A.J. Wilkinson, J. Fujita and R.J. Mayer, Trends Cell Biol. **16**, 229-233, 2006.
7. J.C. Phillips et al., J. Comp. Chem. **26**, 1781, 2005.
8. L. Peplowski, K. Kubiak and W. Nowak, J. Mol. Modeling **13**, 725, 2007.
9. W. Humphrey et al., J. Molec. Graphics **14**, 33, 1996.
10. G. Lee, K. Abdi, Y. Jiang, P. Michaely, V. Bennett, P.E. Marszalek, Nature **440**, 246-9, 2006.

The Zinc-Finger Motif of *T.thermophilus* Ribosomal Protein S14 and the Functionality of *E.coli* Ribosomes

Georgios Papadopoulos[1], Maria A. Xaplanteri[2], Fotini Leontiadou[3], Theodora Choli-Papadopoulou[3], and Dimitrios L. Kalpaxis[2]

[1] Department of Biochemistry and Biotechnology, University of Thessaly, Ploutonos 26, 41221 Larissa, Greece
E-mail: papg@chem.auth.gr

[2] Laboratory of Biochemistry, School of Medicine, University of Patras, 26500 Patras, Greece

[3] Laboratory of Biochemistry, School of Chemistry, Aristotle University of Thessaloniki, 54006 Thessaloniki, Greece

Protein S14 (TthS14) of the 30S ribosomal subunit from *Thermus thermophilus* contains a CXXC-X12-CXXC motif that coordinates a zinc ion. Here we report the results of experiments and molecular dynamics simulations (MDS) on the structural and functional importance of the Zn-finger and the commonly conserved cysteine 24 residue at the first position of the motif. We replaced C24 with serine and incorporated the mutants in *Escherichia coli* ribosomes. The modified ribosomes showed: a) a capability in binding tRNA at the P-and A-sites equal to that obtained with ribosomes incorporating wild-type TthS14 (20% lower compared to native *E.coli* ribosomes), b) reduced capability of the 30S subunit for association with 50S subunits after replacement of the native *E.coli* S14 by wild-type, and particularly by mutant TthS14, c) a peptidyl transferase activity in the chimeric ribosomes bearing mutant TthS14 which is unexpectedly much lower than that in ribosomes incorporating wild-type TthS14. Since the catalytic center of the ribosome is located within the 50S subunit, it seems that the perturbing effect of the S14 mutation on the catalytic center propagates by adjacent inter-subunit bridges or the P-site tRNA. This hypothesis was verified by MDS, which revealed subtle as well as large structural differences between the *E.coli* 30S-subunit head with wild type TthS14 and that with C24S mutant TthS14.

1 Introduction

Proteins in all organisms are synthesized by ribosomes. Bacterial ribosomes are assembled from two subunits (30S and 50S) with more than 50 different proteins in complex with large RNA molecules. The head of the 30S subunit contains the ribosomal protein S14 in direct contact with proteins S3 and S10 and 16S rRNA.[1] Biochemical and crystallographic studies[2,3] indicated that S14 in *T.thermophilus* ribosomes binds a zinc ion coordinated by four cysteine residues (24, 27, 40, 43) in a CXXC-X12-CXXC motif. Zinc-fingers have been implicated in nucleic acid recognition and binding.[4] There are, however, other bacteria, for example *E.coli*, which lack a Zn-finger motif but demonstrate rRNA binding.

Clearly, a more detailed study is needed on the significance of the Zn-motif in relation to other structural features. Here we investigated the functional importance of the S14 Zn-finger motif and its role in the assembly of the 30S subunit, in subunit association, tRNA binding and PTase activity. We performed *in vivo* experiments using *E.coli* ribosomes incorporating wild-type S14 from *T.thermophilus* (wtTthS14). We also employed targeted mutagenesis of C24 in the ThS14 Zn-finger motif to probe the role of this element in cell

growth and ribosome structure and function. Finally, we used Molecular Dynamics Simulations (MDS) to analyze the experimental results and unravel the underlying processes.

2 Materials and Methods

The biochemical methodology of this work was described elsewhere.[6] Here we discuss the MDS details. The TthS14 structure was taken from PDB:1FJG and docked in the *E.coli* 30S-subunit head (PDB:2AVY). After solvation in a sphere with 49475 water molecules and neutralization with 343 Na ions, the structure was relaxed and equilibrated for 500 ps at 300 K (NVT). We used NAMD2[5] and the CHARMM27 force field. Only atoms of the head of the 30S subunit (A929 to U1390 segment of 16S rRNA, proteins S3, S7, S9, S10, S13, TthS14 and S19, Mg and Zn ions) have been included in the simulation. A Cys patch with deprotonated -SH groups was applied, adopting a non-bonded model for Zn-ligand interactions. The same procedure was followed for the replacement of EcoS14 by TthS14-C24S. At the end of each equilibration, an average over the final 50 ps was taken to obtain a more representative structure, which was then used for further analysis. The $\left|\vec{r}_{gc}^{\,mut} - \vec{r}_{gc}^{\,wt}\right|$ distances of the geometrical centers for all residues were calculated for the final structures.

3 Results and Discussion

Table 1 summarizes the results of TthS14 incorporation in *E.coli* ribosomes and 70S reconstruction experiments. The results suggest an S14 role in the association of ribosomal subunits. Disturbances in the S14 structure apparently induce changes in other components of the 30S subunit head, for example residues of the S13 protein and 16S rRNA, which are directly implicated in subunit association.[7] Our MDS results showed significant shifts (Fig. 1) for several 16S rRNA groups in inter-subunit bridges. We found that the Zn ion moves towards two oxygen residues (O2 and O2') of U1202, and away from C24. This behavior can be expected only for TthS14-C24S. In the wtTthS14 case, it is rather striking and it may be attributed to the non-bonded model for the Zn^{2+}-ligand interactions, adopted in this study. For both species, the Zn ion stays close to C27, C40, and C43.

Assembled protein	Degree of incorporation (%)	70S ribosomes (%)
EcoS14	100.0	81.6 ± 5.1
wtTthS14	88.5	75.0 ± 3.2
TthS14-C24S	57.4	66.0 ± 4.5

Table 1. Degree of incorporation of TthS14 and TthS14-C24S proteins into *E.coli* ribosomes and ability of the resulting ribosomes to form 70S complexes.

The tRNA affinity was not affected by the C24S mutation, as confirmed by the MDS. All 16S rRNA residues implicated previously[1,8] in tRNA binding to the P-site, except m^2G966, relaxed to almost identical geometries. Regarding PTase activity our experiments show an eight times smaller value for ribosomes with S14 C24S. Provided that there are not direct contacts of S14 to the catalytic center, it is tempting to speculate, that the observed

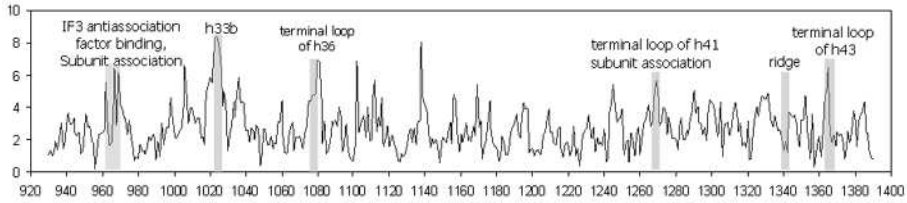


Figure 1. Displacements of 16S rRNA groups after mutation of TthS14 C24 into S24. Regions implicated in subunit association and tRNA binding are in grey.

by MDS conformational changes initiated by the C24S mutation are transmitted through adjacently located intersubunit bridges (bridge B1a or B1b) or through the P-site bound tRNA and cause a less reactive reorientation in the catalytic center.

Acknowledgments

This work was supported by Grant 99ED-605 from the GSRT, Greece, and the European Social Fund. We thank C. M. McCallum for providing the Cys-patch.

References

1. A. Korostelev, S. Trakhanov, M. Laurberg and H. F. Noller, *Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements*, Cell **126**, 1065–1077, 2006.
2. D. E. Brodersen, W. M. Clemons Jr, A. P. Carter, B. T. Wimberly and V. Ramakrishnan, *Crystal structure of the 30S ribosomal subunit from Thermus thermophilus: Structure of the proteins and their interactions with 16S rRNA*, J. Mol. Biol. **316**, 725–768, 2002.
3. P. Tsiboli and T. Choli, *Studies on S14 protein from Thermus thermophilus possessing zinc finger-like motifs*, Biol. Chem. Hoppe Seyler **376**, 127–130, 1995.
4. W. D. Kohn, C. T. Mant and R. S. Hodges, *Alpha-helical protein assembly motifs*, J. Biol. Chem. **272**, 2583–2586, 1997.
5. L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, *NAMD2: greater scalability for parallel molecular dynamics*. J. Comput. Phys. **151**, 283–312, 1999.
6. M. A. Xaplanteri, G. Papadopoulos, F. Leontiadou, T. Choli-Papadopoulou and D. L. Kalpaxis, *The Contribution of the Zinc-Finger Motif to the Function of Thermus thermophilus Ribosomal Protein S14*, J. Mol. Biol., in press, 2007.
7. A. R. Cukras and R. Green, *Multiple effects of S13 in modulating the strength of inter-subunit interactions in the ribosome during translation*, J. Mol. Biol. **349**, 47–59, 2005.
8. V. Berk, W. Zhang, R. D. Pai and J. H. D. Cate, *Structural basis for mRNA and tRNA positioning on the ribosome*, Proc. Natl. Acad. Sci. USA **103**, 15830–15834, 2006.

The Locally Enhanced Sampling Study of Large Ligands Diffusion inside Enzyme. Acrylonitrile and Acrylamide Journey in Nitrile Hydratase

Lukasz Peplowski, Karina Kubiak, and Wieslaw Nowak*

Institute of Physics, N. Copernicus University,
ul. Grudziadzka 5, 87-100 Torun, Poland
**E-mail: wiesiek@fizyka.umk.pl*

Nitrile Hydratese (NHase) is a metaloenzyme with non-standard active site containing noncorin Co^{3+}. In industry it is used for a large scale conversion of toxic nitriles into useful amides. In this research NHase from *Pseudonocardia Thermophila* JCM 3095 (1IRE) is investigated. In order to understand its excellent catalytic activity the possible transport routes of substrates and products have to be determined[1]. Transient states are not easily elucidated using experimental techniques, so computer modeling of molecular dynamics (MD) helps a lot. The main goals of finding cavities inside the protein and entry/exit pathways for a substrate (acrylonitrile) and the product (acrylamide) have been achieved. In our opinion a very convenient tool for that type of study is MD with the Locally Enhanced Sampling (LES) Hamiltonian[2]. The LES method, by multiplying non-interacting ligand copies, allows for better probing of the conformational space than the standard MD method, despite known problems with the energy equipartition[3].

1 Introduction

Nitrile hydratase (NHase) is microbial metaloenzyme containing non-corrin Co^{3+} or non-heme Fe^{3+} metal ion in the nonstandard active site. NHases are widely used in biotechnology for catalysis of the conversion of toxic nitriles to amides[4]. Products of this enzyme are used in pharmacy (nicotinamide is known as vitamin B3), as soil conditioners, components of diapers, paints and in paper industry[5].

Several crystallographic articles reported high sequence and structure similarity between NHases[4,6]. All native structures solved with high resolution confirmed that the protein is composed of two subunits - α(23 kDa) and β(26 kDa). It have nonstandard active cite composed of four residues: αCys109, αCys112, αSer113, αCys114. The αCys112 and αCys114 residues were found to be post-translationally oxidized to cysteine-sulfinic acid CysSO2H (CSD) and cysteine-sulfenic acid CysSOH (CEA), respectively. Metal ion is coordinated by three sulfur atoms from αCys109, CEA and CSD, and the two amide nitrogen atoms from αSer113 and CEA[4,6].

Although there are 14 structures of NHase known none of them contains a substrate or a product of the catalysis. Only two theoretical reports are known about substrate/product interaction inside protein matrix[7,8] and only one paper presents a comparison of the whole series of substrates and corresponding products[1]. Substrate and products entry/exit paths are not known either. In the x-ray structure of NHase a wide channel is visible[9]. Proper understanding of NHase activity requires new data on structural determinants of large ligands transport within this complex molecular system.

In this paper, for the first time, we describe newtonian molecular dynamics of substrate (acrylonitrile, **ACN**) and product (acryloamide, **ACA**) inside nitrile hydratase protein on 50 ns timescale.

2 Methods

Molecular dynamics simulations on a complex of *Pseudonocardia Thermophila* JCM 3095 Co-NHase (pdb code 1IRE)[6] and a docked **ACA** and **ACN** ligands[8] were performed using NAMD 2.6[10] with Locally Enhanced Sampling (LES) method[2]. Thus, starting structures have ligands located in the active site pocket. CHARMM27 force field was applied and three 6 ns long simulations with LES factor 10 (**ACN** Les10) or 10 and 15 (**ACA** Les10, **ACA** Les15). Periodic boundary conditions and a TIP3 water box with at least 7Å distance from protein atoms to the border were employed. The cutoffs for electrostatics and van der Waals interactions have been set on 12Å. The main simulations were preceded by 100 ps of water equilibration at 300K with a frozen protein, 1000 steps of minimization and 50 ps of heating from 0K to 300K. During the 6 ns production phase the Langevin dynamics protocol with temperature held at 300K has been used.

The idea of the LES method seems to be quite simple[2]. The molecular system is divided into two subsystems a big one for example a protein with surrounding water and a small one in our case **ACN** (7 atoms) or **ACA** (10 atoms). The small subsystem may be cloned giving several non-interacting copies of ligands. These copies are subject to nonbonding potential from the protein. The big subsystem feels an average potential from a swarm of ligands. Since ligands don't interact they may occupy the same place in the space. Due to its large number, a better sampling of the conformational space than the standard one-copy MD is achieved. Thus our 10 copies LES 6ns simulation is equivalent to normal 60 ns MD. In visualization and analysis the VMD 1.8.6 package[11] and home made TCL scripts have been used.

3 Results and Discussion

In both systems studied only one major ligand diffusion route along the NHase channel was observed. However, the detailed paths for the nitrile and the amide are different. Close to the exit **ACN** uses for its motion the upper part of channel, while **ACA** occupies mainly the lower part (see Fig. 1, the lower part is closer lying to the metal ion). Interactions of substrates and ligands with the NHase interior are different. **ACN** came out to the solvent after 1.5 ns but **ACA** remained buried in both 6 ns Les10 and Les15 simulations. Differences in ligand-protein interactions are also seen in statistics of collisions (atom-atom distance closer than 2.5Å) occurring during simulations. Data are presented in Fig. 2. **ACN** very often interacts with αLeu88, βPhe41 and quite often with βPhe118. It seems that these three residues stabilize nitrile on the NHase surface in the neighborhood of the channel entrance.

In **ACA** simulations αGln89, βLeu48, βPhe51 and αTrp52 most often collide with catalytic product. These residues compose the lower part of entry to the channel. αGln89 is particularly important because this residue stabilize a network of hydrogen bonds near active site[12]. Our calculation also shows that this residue very often forms H-bonds with the **ACA** amide group. In our opinion this is the reason why **ACA** did not leave the channel during the simulation despite the presence of the 7Å wide passage.

The space exploited by LES copies inside the NHase may be divided into two subcavities. The deeper one corresponds to the active site pocket and ligands stay there for about 1 ns. In the so called *rez-de-chaussée* more shallow cavity **ACN** remains only for 0.5 ns,

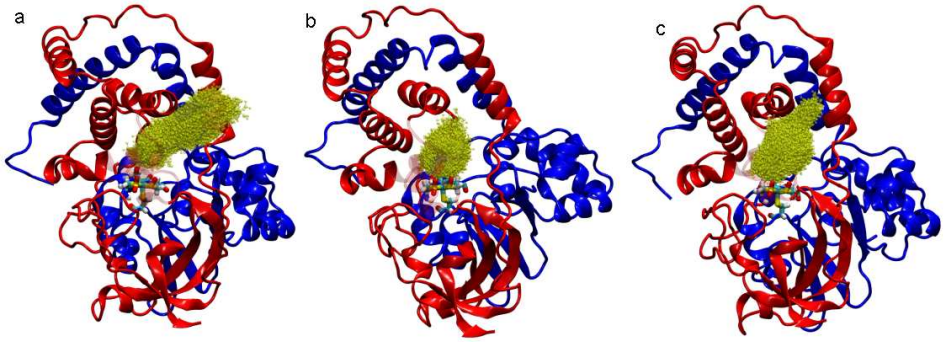


Figure 1. Diffusion paths of aliphatic ligands: (a) substrate **ACN** Les10, (b) product **ACA** Les10, (c) **ACA** Les15.

then it goes on the proteins surface and stays there by 4.5 ns interacting with αLeu88 and βPhe41. In both cavities this small aliphatic substrate has a very large conformational freedom and 180 deg. rotations are sometimes observed. It is worth to note that in **ACN** simulations we observe only weak interactions with CSD and CEA but no collisions with the other parts of the active are noticed. This may suggest that the substrate interacts with the active center via water molecules.

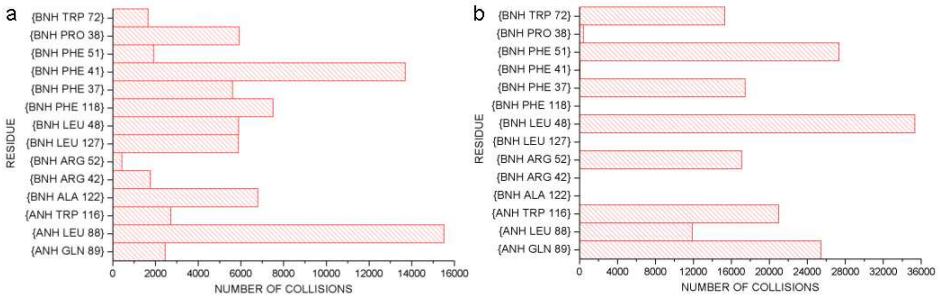


Figure 2. Statistics of the collisions for the **ACN** Les10 (a) and **ACA** Les10 (b).

In **ACA** Les10 simulation the product stays only in the deeper subcavity. Since more copies in the LES method lead to lowering of energy barriers, in Les15 trajectories **ACA** visits also *rez-de-chaussée* subcavity, but instantly goes back to deeper one (compare Fig. 1b and 1c).

4 Conclusions

LES simulations of the NHase from *Pseudonocardia Thermophila* JCM 3095 and its natural ligand reveals differences in the substrate and the product interactions with enzyme. Our calculations show that the **ACN** substrate prefers the upper part of the entrance and is

strongly stabilized by surface residues αLeu88, βPhe41 and βPhe118. The amide product stays at the other region of the entrance. Strong H-bond interactiona of **ACA** with αGln89 are clearly observed. Mutational studies show as well that this residue is critical for enzymatic activity of the NHase[12]. The LES method is a useful computational tool in qualitative studies of enzyme function.

Acknowledgments

This research was supported by Polish Ministry of Education and Science, grant no. 2P04A 07229 and grant supported by Marszalek Wojewodztwa Kujawsko-Pomorskiego Krok w przyszlosc stypendia dla doktorantow. We thank Slawomir Orlowski, Msc for writing the TCL script for collision counting.

References

1. L. Peplowski, K. Kubiak, S. Zelek, W. Nowak, *A comparative DFT study of substrates and products of industrial enzyme Nitrile Hydratase*, Int. J. Quantum Chem., DOI:10.1002/qua.21357, 2007.
2. R. Elber, M. Karplus, *Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin*, J. Am. Chem. Soc. **112**, 9161, 1990.
3. S. Orlowski, W. Nowak, *Oxygen Diffusion in Minihemoglobin from Cerebratus Lacteus a Locally Enhanced Sampling Study*, Theoret. Chem. Acc. **117**, 253, 2007.
4. S. Ngashima, *et al.*, *Novel non-heme iron center of nitrile hydratase with a claw setting of oxygen atoms.* Nat. Struc. Biol. **5**, 347, 1998.
5. H. Yamada, S. Shimizu, M. Kobayashi, *Hydratases involved in nitrile conversion: screening, characterization and application* Chem. Rec. **1**, 152, 2001.
6. A. Miyanaga, S. Fushinobu, K. Ito, T. Wakagi, *Crystal structure of cobalt-containing nitrile hydratase* Biochem Biophys Res Commun. **288**, 1169, 2001.
7. L. V. Desai, M. Zimmer, *Substrate selectivity and conformational space available to bromoxynil and acrylonitrile in iron nitrile hydratase* Dalton Trans. **6**, 872, 2004.
8. L. Peplowski K. Kubiak, W. Nowak, *Insights into Catalytic Activity of Industrial Enzyme Co-Nitrile Hydratase. Docking Studies of Nitriles and Amides* J. Mol. Model. **13**, 725, 2007.
9. T. Arakawa, *et al.*, *Structure of thiocyanate hydrolase: a new nitrile hydratase family protein with a novel five-coordinate cobalt(III) center.* J. Mol. Biol. **366**, 1497, 2007.
10. J. C. Phillips, *et al.*, *Scalable molecular dynamics with NAMD* J. Comp. Chem. **26**, 1781, 2005.
11. W. Humphrey, A. Dalke, K. Schulten, *VMD - Visual Molecular Dynamics* J. Molec. Graphics. **14**, 33, 1996.
12. H. Takarada, *et al.*, *Mutational study on alphaGln90 of Fe-type nitrile hydratase from Rhodococcus sp. N771* Biosci. Biotechnol. Biochem. **70**, 881, 2006.

Receptor Specific Forcefield: Improving Classical Forcefields with Quantum Mechanical Calculations

Horacio Emilio Perez Sanchez[1], **Bernhard Fischer**[1], **Holger Merlitz**[2], **Kaori Fukuzawa**[3], **and Wolfgang Wenzel**[1]

[1] Forschungszentrum Karlsruhe,
Institut für Nanotechnologie, Forschungszentrum Karlsruhe, Germany
E-mail: horacio.sanchez@int.fzk.de

[2] Department of Physics, Xiamen University, China

[3] CREST, Japan Science and Technology Agency, Saitama, Japan

We report results for the in-silico screening of a database of 10000 flexible compounds against various crystal structures of the thymidine kinase receptor complexed with 10 known substrates. The ligands were docked using FlexScreen, a recently developed docking tool based on the stochastic tunneling method. We used a first-principle based scoring function. For rigid receptor conformations we find large deviations in the rank of the known inhibitors depending on the choice of receptor conformation. These data demonstrate that the failure to dock originates from the neglect of receptor degrees of freedom and is not attributable to deficiencies in the scoring function or the docking algorithm. We then performed a screen in which critical receptor sidechains were permitted to change their conformation and found improved scores for those inhibitors that did not dock well in any of the previous screens. So, the consideration of receptor sidechain flexibility in FlexScreen improves the quality of the screening approach. We also demonstrate how the inclusion of QM-calculations of receptor-ligand complexes with the Fragment Molecular Orbital Method (FMO)[1], can be used to improve a classical forcefield. In comparing this QM-forcefield for protein and ligand with a standard ab-initio forcefield (ESFF) we can demonstrate a performance gain.

1 Methods

Docking Method: Stochastic optimization with STUN[2]: Non-linear transformation to the potential energy surface using

$$E_{\mathrm{STUN}}(x) = \ln\left(x + \sqrt{x^2+1}\right) \ , \tag{1}$$

with $x = \gamma\,(E - E_0)$, $\gamma = 0.05$ Mol/kJ and E_0 is the lowest energy encountered during the simulation.

Scoring Function:

$$S = \sum_{Protein} \sum_{Lig.,fl.SC.} \left(\frac{R_{ij}}{r_{ij}^{12}} - \frac{A_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{r_{ij}}\right) + \sum_{h-bonds} \cos\,\Theta_{ij} \left(\frac{\tilde{R}_{ij}}{r_{ij}^{12}} - \frac{\tilde{A}_{ij}}{r_{ij}^{10}}\right) \tag{2}$$

Partial charges q_i are usually evaluated with InsightII and ESFF forcefield, Lennard-Jones parameters R_{ij}, A_{ij} from OPLSAA or from AutoDock and Hydrogen bond parameters $\tilde{R}_{ij}$, $\tilde{A}_{ij}$ from AutoDock.

2 Results

We investigate the accuracy of the predicted ligand-receptor conformation for 83 complexes of the high resolution ASTEX/CCDC dataset for which crystal structures with an experimental accuracy of better than 2 Å are available. For each receptor-ligand complex we perform 10 independent simulations. The resulting conformations are ordered by energy according to the scoring function. The median RMS deviation between the predicted and the experimental structure is 0.83 Å, only ten ligands fail to reach the binding mode within the experimental resolution.

2.1 Astex Data Set Results

With these results and the docking results[3] of three other programs (Glide, Gold and FlexX) we compare: 1) RMSD as a sign of the docking accuracy, and 2) The docking reliability as the percentage of having a RMS better than 2.0 Å.

	FlexScreen	Glide	Gold	FlexX:
FlexScreen wins/total		26/56	18/25	44/56
Results < 2.0 Å in %	80	71	76	57

FlexScreen performed (almost) equally good or better in accuracy and reliability in comparison with all other automated docking methods for which data is available.

Between the remaining difficulties for our approach we find that: 1) Steric clashes in the experimental X-ray structure between ligand and the receptor, 2) Water molecules which have a direct contact to the ligand are sometimes necessary to find the experimental binding mode, 3) Deficiency of the scoring function for solvation energies, 4) Ligands binding to metal complexes: Metal complexes have their specific group geometry and should be considered in *FlexScreen*.

2.2 Docking Study to Human Estrogen α

The ERα (pdb code 1ERE) was previously characterized at the HF/STO-3G level using the fragment molecular orbital (FMO) method[1].
As QM calculations are computationally very expensive, even with the FMO-technique, the binding energies for the ligands were calculated with respect to the most important fifty amino acid residues of the receptor and is therefore also used for our docking runs.

In this study we investigated the influence on QM-based based parameters for the ligands and the receptor on the binding energy accuracy.
We distinguished three cases; 1) QM partial charges for protein and ligands, 2) ESFF partial charges for protein and ligands, 3) QM partial charges for protein and ESFF partial charges for the ligands.

With only one receptor structure *FlexScreen* could well reproduce the binding modes of the quantum mechanical calculations. This is possible, because *FlexScreen* supports side-chain flexibility: the side-chains can accommodate to different ligands.

Comparison with FMO Binding Energies: With a correlation coefficient $R = 0.94$ the correlation is highest the more parameters are from the qm calculations.

Comparison with Experimental Binding Energies: We also compare the calculated binding energies of *FlexScreen* with experimental relative binding affinities (relative to the binding affinity of 17-β-Estradiol (RBA)). Also in comparison to the experimental RBA the correlation is highest the more parameters are from the qm characterization of the protein and the ligands. Case 1 and Case 3 have a higher correlation coefficient than Case 2, for which solely the ESFF-forcefield is used. In overall we get the following correlation coefficients:

	Case 1	Case 2	Case3
QM Energies	0.94	0.71	0.79
RBA	0.68	0.37	0.52

3 Discussion

A mixed setup as in Case 3 is especially interesting for high-throughput screening, because a significant part of the improvement is retained, when only the receptor is treated with QM-based parameters, while the ligands are parameterized with a purely classical model. The protein preparation may take days of calculations, but for each ligand the calculation time is reduced to a minimum. As an additional improvement a setup as in Case 3 seems also to improve the docking accuracy for the binding mode.

Acknowledgments

We thank the Fond der Chemischen Industrie, the BMBF, the Deutsche Forschungsgemeinschaft (grant WE) and the Kurt Eberhard Bode Stiftung for financial support.

References

1. K. Fukuzawa, K. Kitaura, M. Uebayasi, K. Nakata, T. Kaminuma, and T. Nakano, *Ab initio Quantum Mechanical Study of the Binding Energies of Human Estrogen Receptor alfa with Its Ligands: An Application of Fragment Molecular Orbital Method*, J. Comp. Phys. **26**, 1–10, 2005.
2. W. Wenzel and K. Hamacher, *Stochastic Tunneling Approach for Global Minimization of Complex Potential Energy Landscapes*, Phys. Rev. Lett. **82**, 3003–3007, 1999.
3. R. A. Friesner, and P. S. Shenkin, *Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy*, J. Med. Chem. **47**, 1739–1749, 2004.

Folding of Two Helical Peptide with Free Energy Methods and Molecular Dynamics

Aina Quintilla and Wolfgang Wenzel

Forschungszentrum Karlsruhe, Institute für Nanotechnologie,
Postfach 3640, 76021 Karlsruhe, Germany
E-mail: wenzel@int.fzk.de

Studying the folding dynamics of a protein using Molecular Dynamics might require an extra criterium when selecting the studied protein. We present a study of the free energy landscape a two-helix protein using the free-energy protein forcefield PFF01.The free energy landscape of this protein is very simple, suggesting it as candidate for all-atom molecular dynamics simulations. In independent simulations we find the formation of the correct secondary structure and several folding events into the tertiary structure.

1 Introduction

The investigation of protein folding mechanism is one of the most important problems of biophysical chemistry Many β-hairpin systems have been investigated both experimentally and theoretically of small two-helix peptides that are known to fold experimentally into well-defined tertiary structure. Since two-helix proteins constitute a minimal model, in which to investigate the interplay of hydrophobic collapse, secondary structure formation and the formation of native contacts, the identification of such systems may be helpful to elucidate the protein folding mechanism.

In this work we investigate the folding of 1WQE, which exhibits a parallel two-helix bundle. It folds reproducibly with free-energy forcefield into a stable tertiary structure, with very simple free-energy funnel. We demonstrate through molecular dynamics simulation that the lack of competing metastable conformations makes these protein an ideal candidate for folding studies to elucidate the interplay of secondary and tertiary structure formation.

2 Method

An all-atom (except apolar CH_n groups) free-energy protein forcefield[1,2] (PFF01) parametrizes the internal free energy of the protein (excluding backbone entropy).

The elimination of high energy barriers in the free energy surface is the basis of the basin hopping technique[3], also known as Monte Carlo with minimization.

Starting from the same unfolded conformation as above, we performed all-atom implicit water molecular dynamics simulation using AMBER99 forcefield. We generated five trajectories with 50 ns total simulation time each, three at 300 K and two at 325 K, after independent minimization and equilibration.

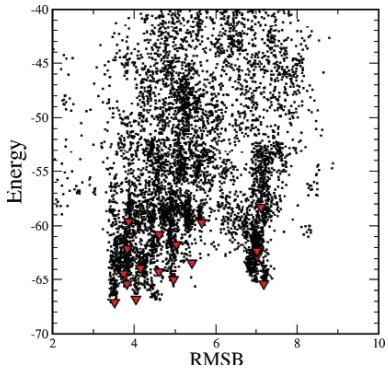


Figure 1. Plot of the energy vs the RMSB in all accepted conformations in the free-energy simulations, the triangles show the best conformations of the 20 simulations. There are only two structural clusters in the free-energy landscape with characteristic RMSB deviations of around 3 and 6 Å to the experimental model.

3 Results

Figure 1 shows energy versus RMSB for all accepted configurations at the end of basin hopping cycles (from all simulations). The triangles indicate the final configurations of the individual simulations. We clearly see two broad funnels of conformations, which terminate into low-energy structures with 3.4 Å and about 7.0 Å RMSB deviation to the native conformation, respectively. The configuration corresponding to the non-native funnel is inconsistent with the formation of the correct number of native disulfide bridges of this peptide.

There is only one, very broad folding funnel consistent with the native disulfide bridge topology. For this reason, the protein studied here may be ideal example to follow the kinetics of protein folding with molecular dynamics or replica exchange methods.

The internal free-energy estimate does not contain backbone entropy; stabilization of one particular conformation with respect to all others does not mean that this conformation is stable with respect to the unfolded ensemble. To settle this question kinetic or thermodynamic simulations must be performed.

We have therefore performed all-atom implicit water molecular dynamics simulations for this protein as described in the methods section. The results for the deviation of the actual conformation from the native structure and the two helices are shown in Figure 2A. The simulations equilibrate quickly into a rapidly fluctuating ensemble with an average overall rmsd deviation between 5 and 8 Å. When we analyze the rmsd deviation of the helical segments however (Helix1: 1-11, Helix 2: 15-21), we find that the entire simulation is dominated with conformations that are within 1-2 Å of the respective fragment of the protein.

We have also analyzed the helix propensity as a function of time for each amino acid as a function of time, as measured by DSSP. Figure 2B demonstrates a very strong helical content for both segments, but the propensity of helix formation may be forcefield dependent. We also analyze the sulfur-sulfur distance between CYS8-CYS18 and CYS4-22 as a function of time (lower panel in Figure 2A). These distances fluctuate strongly, but on

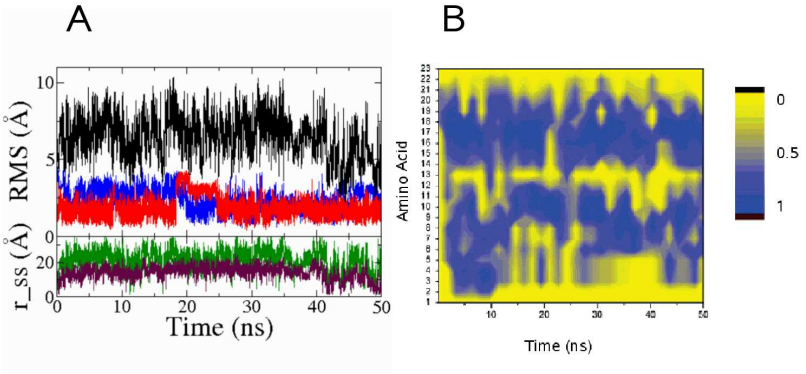


Figure 2. **A:** Analysis of the molecular dynamics trajectories as a function of simulation time. The top panel shows the rmsd of the actual conformation to the native conformation (black) and for the helical fragments only (red: helix 1-11, blue: helix 15-21). The lower panel always shows the deviation of the sulfur-sulfur distance for a potential disulfide bridge (at 2 A distance) for the amino acids forming the first (green, CYS8-CYS18) and the second disulfide bridge (brown, CYS4-CYS22). **B:** Time average over a 100-ps moving window of the helix propensity of each amino acid in the molecular dynamics simulations. Blue: maximal propensity, yellow: no helical contentIn the native conformation the first helix spans amino acids 1-11, and the second helix spans amino acids 14-21,respectively.

occasion, however, some of the sulfur atoms approach each other to within 3-4 Å, i.e., close enough for a disulfide bridge to form. On isolated instances, folding events occur in which both pairs of sulfur atoms approach one another, while both helices are preformed. In those occurrences (which last several ps), the simulations attain all-atom RMSDs to native smaller than 3.5 Å. The intrahelix rmsd vary between 2.1 and 2.5 Å for helix 1 and between 0.8 and 1.0 Å for helix 2 in this time frame.

4 Concluding Remarks

According to our MD simulations, secondary structure formation precedes hydrophobic collapse. The next step would be to substitue the cysteine residues by hydrophobic residues leading to hydrophobic collapse of the already formed helical ensemble into a well-defined tertiary structure without need of disulfide bridges for the peptide to be stable and help to guide the design of stable hydrophobic cores for such small proteins, which would have implications for important challenges in protein design, e.g., for zinc-finger design.

References

1. T. Herges, W. Wenzel, *Reproducible in-silico folding of a landscape in a transferable all-atom forcefield* Phys. ReV. Lett. **94**, 018101, 2005.
2. T. Herges, W. Wenzel, *An All-Atom Force Field for Tertiary Structure Prediction of Helical Proteins* Biophys. J. **87**, 3100–3109, 2004.
3. A. Nayeem, J. Vila, H. A. Sheraga *Acomparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structrues of poloypeptides: [met]-enkephalin* J. Comp. Chem. **12**, 594-605, 1991.

Multiscale Simulation of Protein Cluster Dynamics – the Encounter Complex

Jakob Schluttig[1], **Denitsa Alamanova**[2], **Volkhard Helms**[2], **and Ulrich Schwarz**[1]

[1] Interdisciplinary Center for Scientific Computing, Im Neuenheimer Feld 267 (BIOQUANT), Ruprecht-Karls-University of Heidelberg, 69120 Heidelberg, Germany
E-mail: jakob.schluttig@bioquant.uni-heidelberg.de

[2] Center for Bioinformatics, Saarland University, 66041 Saarbrücken, Germany
E-mail: volkhard.helms@bioinformatik.uni-saarland.de

Most proteins in the cell are active in complexes with two to several hundreds of components. As only very small assemblies can be studied in an all-atom framework, coarse-grained approaches are required to model the association and dissociation dynamics of larger protein assemblies. We approach this problem by using Langevin equations, which allow us to simulate the cluster dynamics in a very efficient way. In order to make contact to specific systems of interest, we use the concept of an encounter complex and extract the relevant association and dissociation rates from biomolecular simulations.

Today proteomics provides us with almost complete lists of proteins in different biological systems of interest. In order to further advance our quantitative understanding of these biological systems, we now have to address the interactions of their proteins in space and time. Indeed most proteins are biologically active in complexes and thus it is crucial to understand their association and dissociation dynamics.

Although very complex if investigated in detail, conceptually the dynamics of protein association can be viewed as a sequence of different steps (see Fig. 1). Initially the binding partners undergo pure diffusion. After reaching a certain proximity, the proteins are steered towards each other, usually by electrostatic forces. This leads to the formation of the encounter complex, which can be viewed as a local minimum on an effective free energy landscape. This minimum results often from the electrostatic and hydrophobic attraction on the one hand and from the final barrier of desolvation on the other hand. This view immediately suggests two important measures to speed up the simulation of protein association: first the diffusion step can be described by a simple Langevin simulation of a particle with the respective shape. Second, overcoming the final energy barrier can be described by an effective rate of association which in principle can be extracted from an effective free energy landscape with the help of transition state theory. In order to scale the system to larger complexes, one then has to simulate many cycles of association and dissociation using appropriate Langevin equations.

Although the notion of an encounter complex has been successfully used before to describe association reactions in chemistry and physics, for specific biological systems it has to be validated by biomolecular simulation. Moreover, biomolecular simulations are required to provide detailed values for the association and dissociation rates for specific systems of interest. In the following we review some of the work regarding the encounter complex, with an eye towards the question how it can be used to provide a bridge between Langevin and biomolecular simulations.

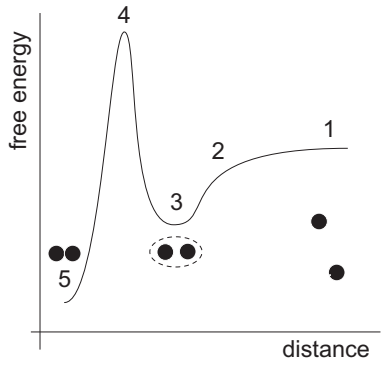


Figure 1. Schematic effective free energy landscape of protein association. 1) free diffusion, 2) electrostatic steering region, 3) encounter complex, 4) barrier due to desolvation and other effects, 5) final complex.

Analytical considerations with respect to the encounter complex have a long tradition. Early works attempted to find mathematical descriptions for the encounter step, i.e. for the transport part of the reaction. One prominent result is the rate of encounter between a colloid of finite size and an ensemble of small colloids derived by Smoluchowski (Z. Phys. Chem. 1916). Later, Debye (Trans. Electrochem. Soc. 1942) calculated reaction rates in ionic solutions. Eigen did the first step towards the consideration of the encounter complex in a biological context, particularly enzyme physics (Quant. Stat. Mech. 1974). He discussed the classic work by Debye and the two limits of pure electrostatics (Langevin 1903, Onsager 1924) and pure geometry (Smoluchowski). Berg and Purcell introduced the standard model for this field: ligands diffuse to a sphere coated with receptor patches (in this case disc like) and are immediately captured upon encounter. Interestingly, for typical values from cell receptor applications, already a very low surface coverage ($\sim 10^{-3}$) leads to nearly optimal outcome. Later, Zwanzig discussed cooperative effects between the receptor patches (PNAS 1990) and derived a correction to the Berg and Purcell result, which perfectly matched simulations by Northrup (JPC 1988).

Bell first connected considerations about the formation of the encounter and the final complex with the additional possibility of dissociation in a model for cell adhesion clusters (Science 1978). Describing both by using stochastic on- and off-rates, he derived basic expressions for the rate of the total reaction, which were discussed both in the diffusion-limited and the reaction-limited case. DeLisi and Wiegel discussed the Berg-Purcell model with a finite reaction rate as introduced by Bell and including electrostatic interactions (PNAS 1981). They claim that, although the particular kinetics of association and dissociation can be affected, the equilibrium properties remain the same. Shoup and Szabo used the concept of a radiation boundary condition to model the formation of the final complex from the encounter state in a mathematically rigorous way (BPJ 1982). Their treatment implied electrostatic interactions and was not restricted to the diffusion-limited case. For the latter, however, they were able to reproduce the results of Berg and Purcell. In the following years, Goldstein and Thompson worked out more details.

In many cases, experimental rates were found to be larger than predicted by the theoretical work. As a consequence, more specific properties like the particular geometry of

the receptor patches were considered in the modelling work. Shoup and Szabo discussed the influence of rotational diffusion and orientation constraints and found that these effects can strongly shorten the mean first passage times for a reaction between a ligand and a receptor patch like in the Berg-Purcell model (BPJ 1981). By using computer simulations, Northrup (PNAS 1992) showed that the rate enhancement seems to be caused by an entrapment of the encounter complex by surrounding water, which makes it possible for the two reacting molecules to test a large number of alignments without leaving the encounter state. Barzykin and Sushin claimed that disk like patches used in the Berg-Purcell model lead to substantially lower reaction rates than the use of hemispherical patches (BPJ 2001). In another paper by the same authors, they suggest that anisotropic diffusion can enhance the reaction rate (BPJ 2001). Recently Korn and Schwarz[1] used the purely geometrical interpretation of the encounter complex to study the efficiency of cell adhesion in hydrodynamic flow, where convection competes with diffusion. Erdmann and Schwarz[2] used the concept of a position-dependent rebinding rate to study the role of cell-substrate distance in cell adhesion.

Schlosshauer and Baker extended the work of Shoup and derived the binding rate for two spherical molecules which both can only bind with hemispherical reactive patches at a finite reaction rate (JPC B 2002). In a recent study, Alsallaq and Zhou (BPJ 2007) again extended this 'hemisphere' model by introducing a 'crater' model consisting of a spherical protein with a crater to which another spherical protein fits snugly. There the formation of a stereospecific complex was disfavored by the loss of translational and rotational freedom, and small translations and rotations between the protein subunits destroyed the interactions, leading to a sharp transition between the bound and the unbound state. The energy lanscape was described as funnel-like, with the deep well of the bound state surrounded by a broad shallow basin.

Miyashita et al. (PNAS 2004) investigated the effect of electrostatic interactions on the binding reaction between cytochrome c2 and a bacterial reaction center. The mechanism involved an encounter complex stabilized by electrostatic interactions, followed by a transition state similar to those found by Zhou, leading to the bound complex active in electron transfer. The study involved determination of a set of transition state structures by fitting experimental kinetic data over a wide range of protein-protein configurations. The transition state ensemble, obtained from structures having the highest correlation coefficients in comparison with the experimental data, had the cytochrome displaced by about 10 Å from its position in the x-ray crystal structure. The observed similarity between the structures of the encounter state, transition state, and bound complex accounted for the rapid rate of association responsible for fast diffusion-controlled electron transfer.

Subsequently, Spaar and Helms[3] used Brownian Dynamics simulations in order to study the association of barnase and barstar. The individual positions and orientations of one protein relative to the other were interpreted as a probability distribution allowing the calculation of the entropy landscape. The free energy landscape was obtained by summing the electrostatic, desolvation, and entropy contributions. A characteristic minimum at about 10 Å distance between the two binding patches denoted the position of the encounter complex.

In summary, biomolecular simulations have identified two systems (cytochrome c : reaction center and barnase : barstar) in which the electrostatic attraction in combination with the desolvation barrier leads to a well-pronounced funnel-shaped free energy surface with a

final transition state barrier between encounter and final complexes. Other protein-protein pairs may have less pronounced features of this kind. This poses the challenge to biomolecular simulations to identify which effective model may apply to particular protein:protein pairs. Yet if a system of this kind has been identified, then methods from stochastic dynamics (like transition state theory and Langevin simulations) can be employed to scale up the system. Only if biomolecular and stochastic simulations are combined, we will be able to model the association and dissociation kinetics of large macromolecular complexes.

Acknowledgments

This project is financed by the Volkswagen Foundation (I/80 469 to V.H. and I/80 470 to U.S.).

References

1. C. Korn and U. S. Schwarz., Phys. Rev. Lett. **97**, 138103-1–4, 2006.
2. T. Erdmann and U. S. Schwarz, Eur. Phys. J. E **22**, 123–137, 2007.
3. A. Spaar and V. Helms, J. Chem. Theory Comput. **1**, 723–736, 2005.

Aggregation of Fragments of the Islet Amyloid Polypeptide as a Phase Transition: A Cluster Analysis

Gurpreet Singh, Ivan Brovchenko, Alla Oleinikova, and Roland Winter

Department of Chemistry, Biophysical Chemistry
University of Dortmund, D-44227 Dortmund, Germany
E-mail: roland.winter@uni-dortmund.de

Aqueous solutions of amyloidogenic polypeptides undergo a phase separation into a water-rich and peptide-rich (fibrillar) phase already at very small peptide concentrations. Studies of the fibrillar phase (stable phase or its metastable intermediates) can be performed by simulations of solution with peptide concentrations deeply inside the two-phase region. In such states, clustering (aggregation) of molecules is extremely sensitive to the system size. We performed MD simulations of 12 amyloidogenic fragments of IAPP (residues 15-19) in liquid water, starting from different random configurations. Analysis of peptide clustering and aggregation evidences features typical for a phase separation. In some simulation runs, the formation of a stable aggregate is hampered by the small system size. We propose to use a clustering analysis to select the configurations relevant for macroscopic systems.

Formation of amyloid fibrils may be the cause of various diseases. Understanding of the driving forces and molecular mechanisms of the fibril formation should elucidate the possibilities to inhibit formation of toxic amyloid fibrils. Aggregation of polypeptides is a cooperative process, which occurs when their concentration in water exceeds some critical value.[1] The critical concentration depends on the characteristics of the considered polypeptide, temperature, ionic strength, pH, etc. The kinetics of aggregation speeds up with increasing peptide concentration and can be facilitated by adding seeds of the peptide-rich phase. All these features are typical for a first-order phase transition in binary mixtures. In the two-phase region the system is separated into a water-rich phase with the critical peptide concentration mentioned above and a peptide-rich phase, which appears as well-ordered solid-like fibrillar structure.

Currently, simulation studies of the fibrillar phase (as stable phase or its metastable intermediates) can be performed by simulations of an aqueous solution with constant peptide concentrations deeply inside the two-phase region. In these states, clustering of molecules may be drastically distorted by the finite size of the simulation system,[2] however, and may even hamper formation of stable peptide aggregates. To explore the importance of this effect in simulation studies of peptide aggregation, we have simulated an aqueous solution of amyloidogenic fragments of IAPP (residues 15-19) at T = 330 K and P = 1 atm pressure. The concentration of peptides in the system was about 120 mg/ml (12 peptides and about 3070 to 3087 water molecules), i.e, deeply inside the expected immiscibility region. Simulations were carried out with the GROMACS-3.2.1 software package and TIP3P water model. Several starting configurations were generated by random insertion of the peptides in a cubic box of 125 nm^3 such that each peptide is at least 7 Å away from its nearest neighbor. The duration of production runs taken were from 50 to 150 ns.

Two peptides were considered as belonging to the same cluster, when the number of the contacts between the atoms of the side chain hydrophobic residues (hydrophobic contacts) exceeds 20. Such contact exists, when a distance between two heavy atoms involved

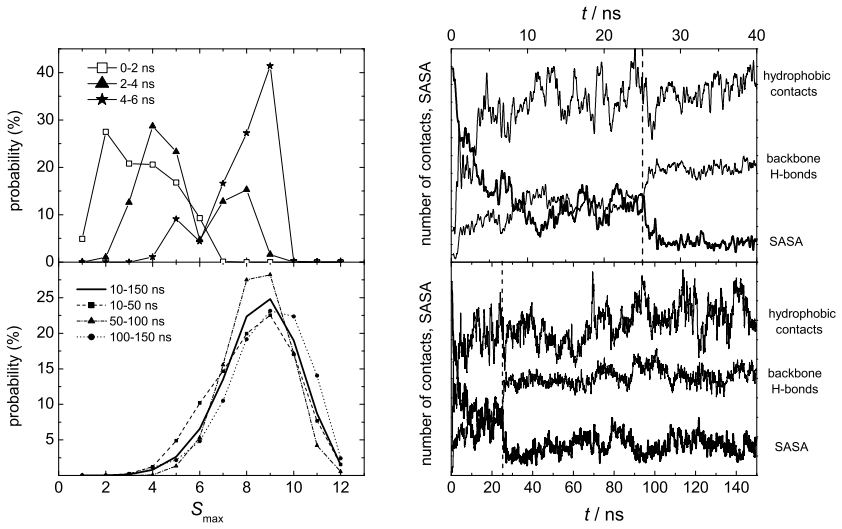


Figure 1. Left panel: probability distributions of the size S_{max} of the largest peptide cluster, averaged over various time intervals of the simulation run. Right panel: Time dependencies of the number of hydrophobic contacts, number of backbone H-bonds and SASA. The equilibration period (up to about 25 ns, indicated by the vertical line) is shown in an enlarged time scale in the upper right panel.

does not exceed the sum of their VdW radii plus 2.8 Å. Additionally, we have analyzed H-bonded peptide clusters, with at least 2 inter-peptide backbone H-bonds being a connectivity criterion. Clustering of the peptides was characterized by calculation of the distribution of sizes *S* of the clusters, measured as a number of peptides in a cluster. This distribution gives, in particular, the average cluster size and allows analysis of the largest peptide cluster. The secondary structure to each residue is assigned using SEGNO.[3] Each peptide is assigned to a particular secondary structure using the criteria that at least three consecutive residues should have the same secondary structure, and no other consecutive secondary structure is present. Two peptides are considered to be forming β-sheets if both peptides are assigned as β-strands and if they are connected by at least two backbone H-bonds.

Upon equilibration, the number of inter-peptide hydrophobic contacts achieve saturation already at $t < 10$ ns (see right panel in Fig. 1). During this time interval, the distribution of the size S_{max} of the largest cluster evolves, but in the time interval 10 ns $< t <$ 150 ns, it remains almost intact (see left panel in Fig. 1). The number of inter-peptide H-bonds and, accordingly, the solvent accessible surface area (SASA) require a longer simulation time to achieve their equilibrium values (about 25 ns). In the simulation run considered, the peptides form a rather stable aggregate, which contains on average 8.6, i.e, the majority, of the totally 12 peptides. This aggregate should be considered as an analogue to the stable peptide-rich fibrillar phase, which appears due to the phase separation in the macroscopic aqueous solution of peptides. Therefore, simulation studies of such peptide aggregates may give useful information concerning the formation of ordered fibrils in experimental situations.

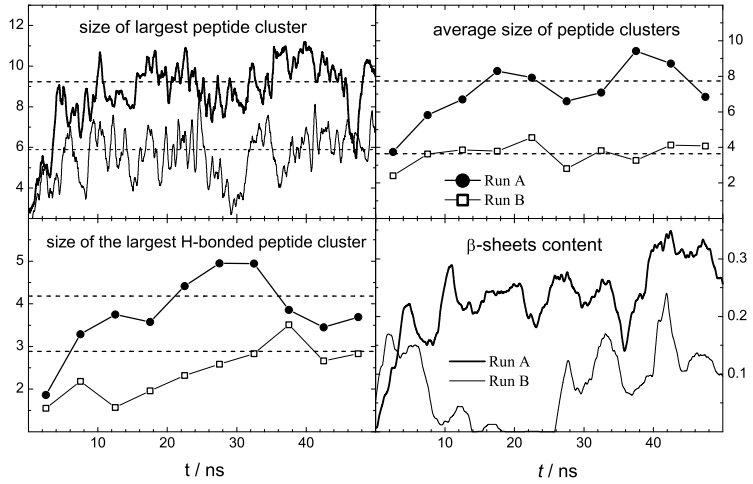


Figure 2. Time dependencies of the size of the largest cluster, of the size of the largest H-bonded cluster, of the average cluster size and of the β-sheet content in two simulation runs: A and B.

However, the large peptide aggregates do not necessarily appear in all simulation runs performed for the same system starting from different initial configurations. Moreover, in some simulation runs, a large peptide aggregate may dissolve after a long simulation time. In Fig. 2, we show two simulation runs, which yield quite different clustering of peptides after an initial equilibration period of about 15 ns. In Run A, there is a large peptide cluster which contains on an average the majority of all peptides. In Run B, the largest cluster contains on an average, less then half of all peptides, and they should be considered as dissolved in water. Such behavior is typical for simulation studies of a small system, whose density (concentration) is kept deeply inside the two-phase region.[2] Namely, the peptide-rich phase may be in a condensed state or in a dissolved state, but both these states are *stable* in simulations and should substitute each other in the course of a very long simulation run. Obviously, the structural characteristics of these two states are quite different. In particular, β-sheet content is essentially higher in the condensed state (see Fig. 2). Clearly, only the condensed state is relevant to the ordered peptide-rich phase seen in experiments, whereas the dissolved state of peptides is relevant for extremely small systems only. Our results show a strong effect of a small size of the simulation system on the aggregation behavior of peptides. We propose to use a full clustering analysis to select the proper condensed state of peptides, which is relevant for large (macroscopic) systems and, therefore, for comparison with experiments.

Acknowledgments

Financial support from the International Max-Planck Research School in Chemical Biology and from the Zentrum für Angewandte Chemische Genomik is gratefully acknowledged.

References

1. J. D. Harper and P. T. Lansbury, Ann. Rev. Biochem., **66**, 385-407(1997).
2. L. G. MacDowell, P. Virnau, M. Muller, and K. Binder, J. Chem. Phys., **120**, 5293-5308(2004).
3. M. V. Cubellis, F. Cailliez, and S. C. Lovell, BMC Bioinformatics, **6**, S8(2005).

A Load Balanced Force-Domain Decomposition Algorithm for Parallel Molecular Dynamics Simulations

Godehard Sutmann[1] **and Florian Janoschek**[2]

[1] 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

[2] Stuttgart University
Institute for Computational Physics, Pfaffenwaldring 27
D - 70569 Stuttgart, Germany

1 Introduction

Classical molecular dynamics simulations are often considered as the *method par excellence* to be ported to parallel computers, promising a good scaling behavior. On the one hand parallel algorithms exist which enable good scaling[1]. On the other hand the complexity of the problem at hand, often scales like $\mathcal{O}(N)$, enabling a linear increase of problem size with memory. However, this point of view applies only to a limited class of problems which can be tackled by molecular dynamics. E.g. in the case of homogeneous periodic systems, where particles interact via short range interactions, the most efficient algorithm is a domain decomposition scheme, guaranteeing local communication between processors and therefore allowing good parallel scaling. In combination with linked-cell lists, the problem scales like $\mathcal{O}(N)$ both in computational complexity and memory demand, so that an ideal behavior in both strong and weak scaling might be expected.

On the other hand, this ideal behavior breaks down if different problem classes are considered, e.g. the case of long range interactions, where not only local communications between processors are required. Another class of counter examples is the case of inhomogenous systems, which occur e.g. in open systems, where the particle density is considerably larger in the center of the system than in the diffuse halo or e.g. in systems consisting of different thermodynamic phases as is the case for the coexistence of liquid/gas or solid/gas phases. In this case, domain decomposition algorithms often fail. Due to a more or less regular geometric decompoition of space, processors are responsible for different numbers of particles, often introducing a strong load-imbalance, which leads to inefficient CPU usage on some processors and therefore to a bad parallel scaling.

For these classes, other parallel decomposition schemes are often applied, like atom- or force-decomposition schemes[1]. While the former one distributes equally particles onto processors, the latter one decomposes the interaction matrix among processors. It is the latter case which is considered in the present paper. This method also allows a geometric approach which consists of equally partitioning of the force-matrix $\mathsf{F} \in \mathbb{R}^{3N\times 3N}$ onto processors. Taking into account Newton's law of action and counteraction either the upper triangular part of F may be decomposed into equal areas[2], or the whole matrix is decomposed[3,4], while assigning only a subset of interactions onto each PE, in order to fulfil the skew symmetry of the interactions.

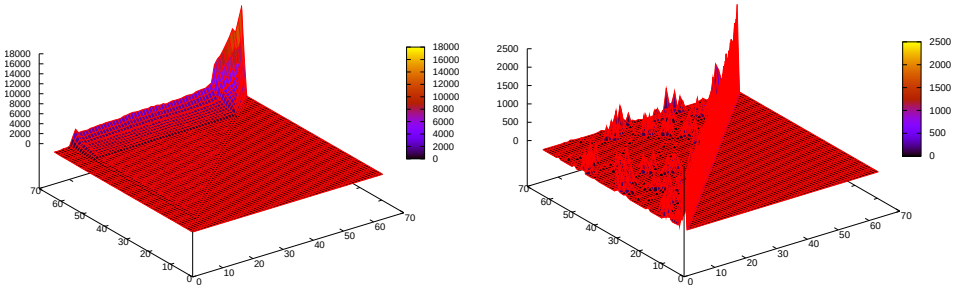


Figure 1. Communication patterns between processors for random distribution of particles (left) and sorted particles according to a space-filling Hilbert curve (right). Note the different magnitudes of data volume. Shown is the case for $P = 64$. Left part of each figure shows the upper triangular part of the force matrix. Each row corresponds to data stored on individual PEs.

In general there are two points in force decomposition methods, where communication is required:

1. after propagating the particle position in the integration step, the position of particles must be transferred to remote processors in order to calculate mutual interactions.

2. after calculating interactions, the partial forces, calculated on different processors acting on a tagged particle must be collected onto the host processor of that particle, in order to propagate its positions and momenta.

2 Method

2.1 Communication

Traditionally, coordinates and forces are transferred by all-to-all communication steps. Positions are usually distributed by an `mpi_allgatherv` command while forces are summed up simply by an `mpi_allreduce` procedure. This is certainly the most simple way to proceed. Since most MPI implementations internally make use of a tree-wise communication protocol, the global communication will scale like $\mathcal{O}(\log_2(P))$, if P is the number of processors. However, for a big system and a large number of processors, there will be a lot of redundant data transferred to processors. I.e. global communication operations do not take into account whether transferred data are really needed on remote processors. Therefore two alternative methods are considered here.

The first approach still gathers all position coordinates from remote PEs, in order to calculate interactions with local particles. However, the forces are selected according to whether they have been calculated or not. This avoids sending a lot of redundant information across the network, although the communication protocoll gets a little bit more involved. In the case of the upper triangular force matrix decomposition, forces must be sent from local PEs to the ones with larger rank. In this case $P - i_p - 1$ communication steps have to be performed, if i_p is the rank of the local PE. In this case two `mpi_gatherv` operations have to be performed: (i) to transfer the non-zero partial forces, (ii) to transfer the particle indices, onto which forces act.

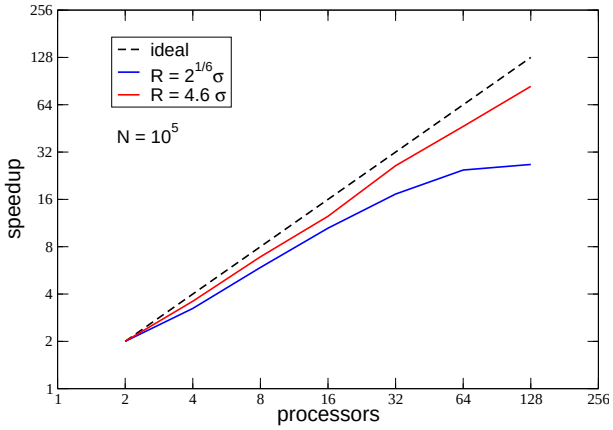


Figure 2. Parallel speedup for the force-domain decomposition method for the case of $N = 10^5$ particles and two different cutoff radii for interparticle interactions.

The second approach is more involved. The basic principle is to combine a domain decomposition with a force decomposition scheme. Domain decomposition is achieved by sorting the particles according to their positions along a space filling Hilbert curve. This ensures that most particles which are local in space are also local in memory. For short range interactions this implies that most interaction partners are stored already on the same PE. According to this organisation, the force matrix becomes dominant around the diagonal and off-diagonal areas are sparse. The next step is to calculate interaction lists, which are distributed onto all PEs. Indices of interaction partners are stored in Verlet lists, which are valid for a number of time steps. Therefore, these lists only have to be created and distributed from time to time (e.g. every 20 time steps[5]). According to these lists it is known which particles from remote PEs are needed on local PEs. The same is true for calculated partial forces. Since the amount of data is very small with respect to global communications of positions and forces, the communication overhead of this method is strongly reduced, enabling a very much better parallel scaling. Since randomisation of particle positions occur on a diffusive time scale, the space filling curve has to be updated only one or two orders of magnitude less than the interaction lists, thus introducing only a small overhead.

2.2 Load-Balancing

Another contribution of the present paper is to combine the proposed methods with an efficient load-balancing strategy. This is developed for the force-stripped row method, which partitiones equal areas in the interaction matrix to different processors. For inhomogenous systems, this approach becomes quite inefficient, since the number of interactions within such areas may vary considerably. Therefore, two different approaches are discussed and implemented here. The first consists in distributing the number of interactions equally onto the pocessors. This may give good results for the case, where particles are located randomly in memory. However, if a subset is random and another subset is ordered (as it might occur for the case of a solid/liquid system) in memory, cache effects will destroy the

equal load on PEs. Therefore, a second approach consists in distributing the work, proportional to the time which is spent on every PE. This strategy turns out to be most efficient, leading to a very fast distribution of equal load between the processors.

3 Results

The above methods are tested for a simple system consisting of Lennard-Jones particles. Molecular dynamics simulations were carried out for systems with $N = 10^5$ particles. Fig. 1 shows the amount of data which has to be communicated between processors for the cases of traditional force-decomposition (FD) methods and the force-domain-decomposition (FDD) technique, which uses the space filling Hilbert curve. In FD interacting particles are randomly distributed in memory (at least for long simulations particles get uncorrelated). Therefore, coordinates have to be sent from a processor p_i to all other processors $p_j < p_i$, while forces have to be redistributet in the opposite direction from p_j to all other processors $p_i > p_j$. Sorting reduces significantly the amount of data to be sent. First of all the size of the data buffers become smaller, second the matrix becomes sparse, i.e. a given processor p_i does not send coordinates to all other processors $p_j < p_i$, but only to those where these coordinates are really needed to compute interactions. Since sorting concentrates interaction partners close to the diagonal, only small amounts of data have to be sent to remote processors. Calculations were carried out for the FDD method, considering a system of $N = 10^5$ particles with different size of the interaction radius R_c. For decreasing R_c, the amount of data to be sent to remote PEs gets smaller. On the other hand also the work performed on a single PE decreases, because of a smaller number of interaction evaluations. Therefore, for small R_c communication becomes a bottleneck for a large number of processors, because of the accumulation of latency. Therefore, the speedup curve saturates for 128 processors. In the case of a larger cutoff ($R_c = 4.6\,\sigma$) the computation dominates communication and the program scales up to 128 processors. Note, that for a global communication, like in traditional FD, communication would have a significant larger contribution and the program would not scale as well as in the present case.

References

1. S. Plimpton. Fast parallel algorithms for short range molecular dynamics *J. Comp. Phys.*, 117:1, 1995.
2. R. Murty and D. Okunbor. Efficient parallel algorithms for molecular dynamics simulations. *Parall. Comp.*, 25:217–230, 1999.
3. V. E. Taylor, R. L. Stevens, and K. E. Arnold. Parallel molecular dynamics: Communication requirements for parallel machines. In *Proc. of the fifth Symposium on the Frontiers of Massively Parallel Computation*, pages 156–163, 1994.
4. S. Plimpton and B. Hendrickson. A new parallel method for molecular dynamics simulation of macromolecules. *J. Comp. Chem.*, 17:326, 1996.
5. G. Sutmann and V. Stegailov. Optimization of neighbor list techniques for molecular dynamics simulations. *J. Mol. Liq.*, 125:197–203, 2006.

Forcefield Validation with the Rosetta Protein Decoy Set

Abhinav Verma and Wolfgang Wenzel

Institute for Nanotechnology,
Research Center Karlsruhe, 76344 Karlsruhe, Germany
E-mail: {*verma, wenzel*}*@int.fzk.de*

We have recently extended our helical protein forcefield PFF01 to a more generalized protein forcefield PFF02 in our efforts towards a universal free energy forcefield for all atom protein folding and prediction. Here we selectivity of various proteins PFF02 with a Rosetta decoy set consisting of 32 proteins. The results conclude good selectivity of PFF02 for structure prediction with an average z-score of -3.46 and an average root mean square deviation of 2.14 Å.

1 Introduction

All atom protein folding and structure prediction have been one of the central problems in biophysical chemistry. Transferable potentials are needed to address these questions for a wide range of proteins[1]. We have recently extended our helical protein forcefield PFF01[2] to a more generalized protein forcefield PFF02[3] following the thermodynamic hypothesis[4], that most proteins are in thermodynamic equilibrium with the environment. With PFF02, we could demonstrate folding of small hairpin polypeptides into their native-like conformations[5,6].

The accuracy and predictivity of free energy protein forcefields can be investigated using decoy sets[7], a method that works even for proteins that are too large or too complex to be folded from random initial conformations. For the selectivity of PFF02, we study a decoy set generated using Rosetta[8] consisting of 32 proteins.

2 Method

A decoy set is a large library of protein conformations generated to approximately span all relevant low energy regions of the free energy landscape. To measure the predictivity and selectivity of a forcefield, the conformations in the library (decoy set) must be ranked according to their energy. If near native conformations emerge lowest in the free-energy function, the force field differentiates between native and near-native conformations. In the limit of completeness of the decoy set, which is rarely reached in practice, this test alone is sufficient to show that the force field stabilizes the native conformation of the protein against all competing metastable conformations and corresponds to the global optimum of the free-energy force field.

For decoy sets generated with unbiased methods, the computation of the Z-score (the difference between energies of near-native decoys to the mean energy of the decoy set in units of its standard deviation) gives a quantitative measure of the selectivity of the force field. The Z-score is defined as

$$Z = \frac{E_{\text{ref}} - \langle E \rangle}{\sigma} \tag{1}$$

where E_{ref} is the reference energy, *i.e.*, the energy of the native conformation, $\langle E \rangle$ is the average energy of the decoy set and σ is the standard deviation of the decoy set. The Z-score simply measures the mean energy distance from the native state of protein in terms of the standard deviations of the decoy set. The lower the Z-score, the better is the discrimination between native and non-native conformations in the decoy set. The histograms showing the distribution of decoys over energy range are shown in Figure 1(inset).

3 Results

For this study we investigated, which of the proteins of the the large all atom Rosetta decoy sets[9] could be stabilized by PFF02. The proteins in this decoy set range between 32-85 amino acids in size and span all secondary structural classes.

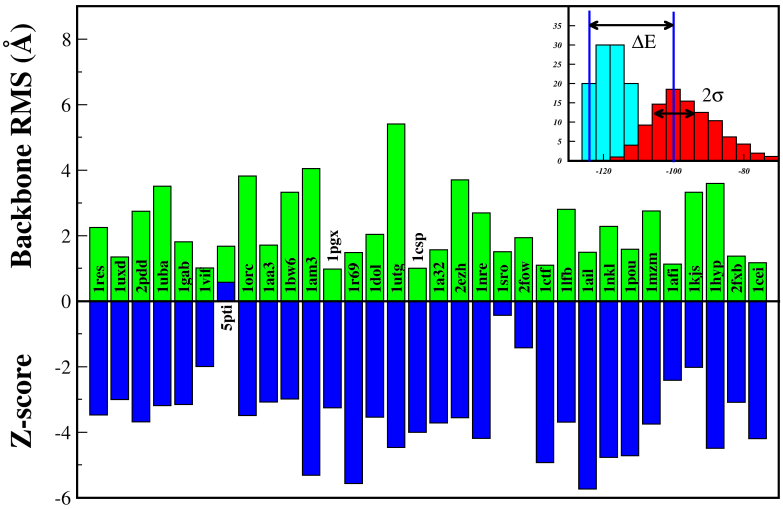


Figure 1. RMSD of the lowest energy conformation (Green) and Z-scores (Blue) of proteins in the Rosetta decoy set. Inset shows a sample distribution of decoys for a protein. The bars in cyan represent the distribution of near-native decoys generated from native structure and red bars represent all the decoys from the decoy set.

For the calculation of Z-scores we generated near-native conformations for 32 proteins of the latest Rosetta decoy library. We excluded only proteins that are stabilized by transition metal clusters or other ligands as such interactions are yet to be implemented in the present force field. The resulting near-native conformations deviate 1-4 Å from the experimental conformation, except for 1am3 and 1utg, where deviations of 4.05 and 5.4 Å

respectively are observed (top panel of Figure 1, Table 1 for all data). Since both of these proteins are dimeric, this difference arises because the molecules are relaxed here in isolation. The average deviation between experiment and near-native conformation in the force field for the set of 32 proteins was 2.14 Å, the figure also indicates that there is little correlation between the size of the protein and the accuracy with which the local minimum of the force field agrees with the experimental conformation.

In order to arrive at a meaningful comparison of the energies we relaxed the approximately 2000 decoys for each of the proteins in the decoy library in PFF02. This procedure maps each decoy to a local minimum of the force field of similar structure, the average change in RMSD between the starting and relaxed conformation was less than 0.02 Å. This means that the decoys are not changed in the relaxation process.

PDB ID	Z-Score	RMSD (Å)	PDB ID	Z-Score	RMSD (Å)
1a32	-3.72	1.57	1nre	-4.19	2.69
1aa3	-3.08	1.71	1orc	-3.49	3.82
1afi	-2.41	1.13	1pgx	-3.26	0.98
1ail	-5.73	1.49	1pou	-4.72	1.58
1am3	-5.32	4.05	1r69	-5.57	1.48
1bw6	-2.98	3.32	1res	-3.47	2.25
1cei	-4.19	1.17	1sro	-0.43	1.51
1csp	-4.01	1.00	1uba	-3.19	3.96
1ctf	-4.93	1.10	1utg	-4.47	5.41
1dol	-3.54	2.04	1uxd	-3.00	1.35
1gab	-3.16	1.81	1vif	-2.00	1.01
1hyp	-4.49	3.59	2ezh	-3.56	3.70
1kjs	-2.02	3.32	2fow	-1.43	1.94
1lfb	-3.69	2.80	2fxb	-3.09	1.37
1mzm	-3.75	2.75	2pdd	-3.69	2.74
1nkl	-4.77	2.28	5pti	0.58	1.68

Table 1. Zscores and RMSD(lowest energy) for the 32 proteins of Rosetta decoy set in PFF02.

The Z-scores for 29 out of the 32 proteins in the decoy set are less than -2.0 (top panel of Figure 1). This indicates a good selectivity of the force field for these proteins. The average the score of -3.46 is lower than that of any previously reported alternate scoring function for the same decoy set. The average Z-score for the same set of proteins in PFF01 was -3.06[10]. This indicates the improvement of the force field for this set of proteins which spans all kinds of secondary structural elements, with the only exception of 5PTI. Since the Rosetta decoy sets were specifically generated to span a wide range of near-native and non-native conformations for each protein. These data indicate that PFF02 stabilizes near-native conformations of a large family of small and medium-size proteins of all secondary structure classes as its global optimum.

4 Concluding Remarks

In this work, a 32 protein Rosetta decoy set was used to test the selectivity and predictivity of a recently modified protein forcefield PFF02. The results indicate that PFF02 has good selectivity with an average z score of -3.46. Also the average RMSD for the lowest energy conformation for all these proteins is only 2.14 Å showing good predictability for a wide range of proteins. PFF02 thus emerges as a positive step towards a universal and transferable forcefield for all atom protein folding and tertiary structure prediction.

Acknowledgments

We thank 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 and at the Barcelona Supercomputer Center.

References

1. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–6, 2001.
2. T. Herges and W. Wenzel. An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.*, 87:3100–9, 2004.
3. A. Verma and W. Wenzel. Towards a universal free-energy approach for all-atom protein folding and structure prediction. Submitted, 2007.
4. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181: 223–30, 1973.
5. A. Verma and W. Wenzel. Predictive and reproducible de-novo all-atom folding of a β-hairpin loop in an improved free energy forcefield. *J. Phys. Cond. Matt*, in press, 2007.
6. W. Wenzel. Predictive folding of a β-hairpin in an all-atom free-energy model. *Europhys. Lett.*, 76:156–162, 2006.
7. B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Molec. Biol.*, 258:367, 1996.
8. R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. M. E. Strauss, and D. Baker. Rosetta in CASP4: progress in ab-initio protein structure prediction. *Proteins*, 45:119–126, 2001.
9. J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 53:76–87, 2003.
10. A. Verma and W. Wenzel. Protein structure prediction by all-atom free-energy refinement. *BMC Structural Biology*, 7:12, 2007.

Knots in Macromolecular Systems: Concepts and Challenges

Peter Virnau

Institut für Physik, Johannes Gutenberg–Universität Mainz
Staudinger Weg 7, 55099 Mainz, Germany
E-mail: virnau@uni-mainz.de

The following survey of computational and experimental research activities on knots in synthetic and biological macromolecules outlines previous accomplishments and current challenges in the field.

1 Introduction

Knots are one of mankind's oldest and most practical devices. Appreciated by fisherman, seafarers, surgeons and rescue professionals around the world, knots have long captured the imagination of natural scientists, too. In 1867, Lord Kelvin suggested that atoms may consist of knots formed by the ether,[1] which inspired physicists like Maxwell and Tait to lay the foundations of knot theory.

In mathematics, knots are only well–defined in closed curves and usually categorized according to the minimal number of crossings in a projection onto a plane. The search for an algorithm which can distinguish between all knots is ongoing and still one of the grand challenges in mathematical knot theory. From a practical point of view, however, several algorithms, like the Alexander polynomial or the HOMFLY polynomial, are able to distinguish between simple knots and suffice for most purposes. In this spirit, open chains can also be analyzed by simply connecting end points in a defined way.

After descending into mathematical obscurity for some time, interest in the subject was revived in the 1960s, when Frisch, Wassermann[2] and Delbrück conjectured that all sufficiently long polymers have to contain knots - experimental length scales were, however, not known at this time. From a theoretical point of view, this statement is intriguing because knots are not included in the standard theory for self-avoiding polymers, but affect equilibrium properties and dynamics.

2 Computer Simulations of Knots in Polymers

Simulations are particularly well-suited to address these questions because the statistics of knots can be unveiled by analyzing independent polymer configurations generated by the computer. Starting with the ground-breaking work of Vologodskii et al.,[3] a variety of coarse-grained polymer models were tested.

With little or no excluded volume (random walk) and no inherent stiffness, a single three-dimensional polymer displays a large number of very small and localized knots. Each monomer can be positioned anywhere around its predecessor which promotes entanglements on the local scale. If excluded volume is considered, however, the chain becomes unknotted.[4] A single bead–spring polymer with 1000 monomers in good solvent

conditions, e.g, only contains knots in about 1% of all configurations.[5] The fraction of the polymer which is occupied by the knot in one of these rare events is considerably larger than in random walks, but still small with respect to the total chain–length. If a polymer collapses into a globular state and has enough time equilibrate, or if it is confined into a capsid, knots become frequent again[5] and spread out all over the globule.[5] In such a dense phase, the same bead–spring model contains knots in 80% of all configurations - most of which are already quite complex. In this case, the free energy of the polymer is dominated by energetic contributions which do not distinguish between unentangled and entangled states which are otherwise favored by conformational entropy. In the swollen phase, on the other hand, the polymer would like to have access to as many states as possible, and the free energy is dominated by entropic contributions which disfavor entanglements.

Analogies between microscopically knotted polymers and macroscopically knotted strings and robes are far reaching. Not only do single polymers and robes become knotted in crowded conditions, they also share similar material properties. A knotted robe, e.g., withstands only a fraction of the traction which it can withstand if the robe remains unknotted - a well-known fact among mountaineers and anglers. Knotted polymers are also far less stable under tension and tend to break at the entrance to the knot.[6]

3 Knots in DNA, Proteins and Synthetic Polymers

Knowledge gained from computer simulations is particularly relevant for biological systems whose fundamental constituents are long biopolymers. Knots were first discovered in bacterial DNA in 1976.[7] Since then, many knots were also created artificially, e.g., by the action of topoisomerase I on circular DNA. In good solvent conditions, DNA behaves like a model polymer and contains almost no knots (0.5-4% knots in a 10000 base pair strand depending on salt concentration.)[8] From the simulation of knotted homopolymers, one might expect that DNA also becomes highly knotted in confinement. This condition is, however, problematic for DNA in crowded environments like the cell, and indeed, nature has developed successful strategies to circumvent this effect. Human DNA, e.g., wraps around histone proteins. In ds-DNA viruses, the rather rigid DNA forms a spool when it is fed into the capsid (with one end remaining attached to the loading channel.)

Topology is even more relevant in proteins because the three-dimensional structure of a protein directly determines its functionality. Once again, knots are rare[9–12] although the reason is not well understood. It has been hypothesized that knotted structures are difficult to fold[11] and would essentially preserve their unknotted state after the initial collapse.[12] However, experiments suggest that certain knotted proteins may fold and unfold reversibly under a change of solvent conditions.[13] Secondary structure and inherent stiffness of the protein backbone may also simply shift the length scale at which knots occur in comparison to globular polymers.[12] Overall, the problem remains essentially unsolved. Interestingly, knots in proteins are usually preserved in structural homologs throughout evolution which indicates that they are relevant for the functionality of the protein and have existed since the beginning of life. In our recent comprehensive analysis of the Protein Data Bank,[9] we were also able to identify one counter-example and demonstrate how the presence of a knot may alter the enzymatic activity of the protein.

The synthesis of man–made molecules with knots has also been an important topic in the chemical literature.[14] These so-called “knotanes” consist of small knotted molecules obtained by supramolecular-template techniques.

4 Challenges

In the near future, I would expect that closed knotted nanoparticles based on knotted polymer globules will be synthesized.[15] These particles should be highly knotted, in the order of 10 nm in size and offer a reactive surface which may turn them into interesting candidates for the delivery of drugs. Similarly, it should become feasible to genetically alter proteins such that they become knotted. From a theoretical point of view, it would certainly be interesting to quantify why there are so few knots in proteins and how knotted proteins actually fold. 140 years after Kelvin, knotty problems persist.

References

1. W. T. Thompson, *On vortex atoms*, Philos. Mag. 34, 15-24 (1867).
2. H. L. Frisch and E. J. Wassermann, *Chemical Topology*, J.Am.Chem.Soc. 83, 3789-3795 (1961).
3. M.D. Frank-Kamenetskii, A.V. Lukashin, and A.V. Vologodskii, *Statistical mechanics and topology of polymer chains*, Nature **258**, 398-402 (1975).
4. K. Koniaris and M. Muthukumar, *Knottedness in ring polymers*, Phys. Rev. Lett. **66**, 2211-2214 (1991).
5. P. Virnau, Y. Kantor, and M. Kardar, *Knots in globule and coil phases of a model polyethylene*, J. Am. Chem. Soc. **127**, 15102-15106 (2005).
6. A. M. Saitta, P. D. Soper, E. Wasserman, and M. L. Klein, *Influence of a knot on the strength of a polymer strand*, Nature 399, 46-48 (1999).
7. L. F. Liu, R. E. Depew, and J. C. Wang, *Knotted single-stranded DNA Rings - novel topological isomer of circular single-stranded DNA formed by treatment of Escherichia Coli omega protein*, J.Mol.Biol. 106, 439-452 (1976).
8. V. V. Rybenkov, N. R. Cozzarelli, and A. V. Vologodskii, *Probability of DNA knotting and the effective diameter of the DNA double helix*, Proc. Natl. Acad. Sci. U.S.A. 90, 5307-5311 (1993).
9. P. Virnau, L. Mirny, and M. Kardar, *Intricate knots in proteins: Function and evolution*, PLOS Comp. Biol., **2**, e122 (2006).
10. M. L. Mansfield, *Are there knots in proteins?* Nat. Struct. Mol. Bio. 1, 213-214 (1994).
11. W. R. Taylor, *A deeply knotted protein and how it may fold*, Nature 406, 916-919 (2000).
12. R. C. Lua and A. Y. Grosberg, *Statistics of knots, geometry of conformations, and evolution of proteins*, PLOS Computational Biology 2, e45 (2006).
13. S. E. Jackson and A. L. Mallam, *Folding studies on a knotted protein*, J. Mol. Bio. 346, 1409-1421 (2005).
14. O. Lukin and F. Vögtle, *Knotting and threading of molecules: Chemistry and chirality of molecular knots and their assemblies*, Angew. Chem. Int. Ed. 44, 1456-1477 (2005).
15. private communication with Prof. Vögtle.

Steered Classical and Quantum Path-Integral Molecular Dynamics Simulations of Strongly Coupled Protons Motions in Porphycene

Ł. Walewski[1,2], P. Bała[3], and B. Lesyng[1]

[1] Department of Biophysics, Faculty of Physics, University of Warsaw,
Żwirki i Wigury 93, 02-089 Warszawa, Poland
E-mail: lesyng@icm.edu.pl

[2] Interdisciplinary Centre for Mathematical and Computational Modelling,
Pawińskiego 5a, 02-106 Warszawa, Poland
E-mail: ljw@icm.edu.pl

[3] Department of Mathematics and Computer Science, N. Copernicus University,
Chopina 12/18, 87-100 Toruń, Poland
E-mail: bala@mat.uni.torun.pl

MD simulations of many body quantum-classical systems is of big importance for the description of functioning of many (bio)molecular and nanosystems. Porphycene containing two strongly coupled quantum protons in a classical molecular scaffold is a model system which is used by us to develop novel methodological QCMD approaches. The system is also of practical importance for molecular nanotechnologies. Energy profiles for the proton motions in porhycene have already been studied by us using an SCC-DFTB approach[1], however, classical MD simulations were not sufficiently precise to determine effective barriers for the proton transfer. Since the quantum dynamical motions of the protons have considerable impact on the motions of the classical atoms and vice versa, in the current approach the whole system is simulated using a steered Path Integral Molecular Dynamics (PIMD) method with an on the fly Car-Parrinello DFT propagation scheme for all nuclei, and with the adiabatic QD of the electrons[2]. Such approach together with thermodynamic sampling of the conformational space provides realistic energy profiles along the reaction pathways for the protons motions. The method requires massively parallel computer systems to carry out simulations for real systems.

1 Introduction

In previous studies proton transfer reaction pathways in porphycene were determined[3]. The pathways *trans* → *ts-A* → *cis-A* and *trans* → *ss-A* → *trans* are low energy pathways. They contain barriers of 5 kcal/mol and 7.5 kcal/mol, respectively. The former participates in the stepwise mechanism of the double proton transfer reaction that requires two consecutive transitions of this type. The latter describes the concerted reaction mechanism. They are both accessible for the system in 300 K. The high energy pathways – *trans* → *ts-B* → *cis-B* and *trans* → *ss-B* → *cis-B* contain barriers of 43 and 64 kcal/mol, respectively, and are statistically irrelevant at 300 K. The aim of the computations carried out at NIC was to get mean-field energy profiles for selected proton-transfer pathways and to compare results of the standard Car-Parrinello Molecular Dynamics (CPMD) method with the Path Integral Molecular Dynamics (PIMD) approach. This comparison can reveal the influence of nuclear quantum effects, such as tunneling, on the reaction energy barriers.

Table 1. Relative energies between stationary points on the porphycene potential energy surface (PES) computed with different DFT functionals. Plane wave expansion cutoff (E_{cut}) is in Ry, energy in kcal/mol. Calculations with 120 and 70 Ry cutoffs where carried out using the Trouiller-Martins normconserving pseudopotentials, while those with 40 and 25 Ry cutoffs with the Vanderbilt ultrasoft ones. Structure names follow ref.[3], scheme 2 and 3.

Functional	E_{cut}	Structure						
		trans	*cis-A*	*ts-A*	*ss-A*	*cis-B*	*ts-B*	*ss-B*
B3LYP/TZ2P[a]	–	0.0	2.4	4.9	7.6	28.8	43.0	64.4
PZB	70	0.0	2.5	4.9	7.5	26.9	40.9	29.5
	120	0.0	2.1	3.9	5.9	28.5	40.7	30.7
BLYP	40	0.0	2.0	3.7	5.4	28.2	40.3	30.6
	25	0.0	2.4	3.5	5.1	27.9	39.9	30.4
	120	0.0	1.5	2.4	3.5	30.3	39.5	32.2
PBE	40	0.0	1.6	2.3	3.1	30.3	39.3	32.2
	25	0.0	1.5	1.9	2.5	30.2	39.0	32.2
	120	0.0	1.6	2.5	3.6	30.4	39.7	32.3
BP	40	0.0	1.6	2.4	3.3	30.2	39.6	32.1
	25	0.0	1.8	2.2	2.3	29.7	39.0	31.6

[a] Values cited from ref.[3] for comparison.

2 Selection of the Exchange-Correlation Potential

Locations of known stationary points on the potential energy surface (PES) of porphycene were optimized using a number of Density Functional Theory (DFT) approximations. Since Perdew, Zunger, Becke[4,5] gradient corrected exchange-correlation functional (PZB) was already successfully applied for the proton transfer reaction in malonaldehyde by other group[6] it was treated as the most trustworthy one. Other widely used functionals (BLYP, PBE and BP) were also tested. All results were compared with reference values of Kozlowski[3] *et al.* Plane wave expansion of valence orbitals and atomic pseudopotentials for the core electrons were used with different cutoff values. For results see Table 1.

In most cases lowering the expansion cutoff results in lowering the relative energies of the transition states and local minima of the energy landscape. The energy values with higher cutoff are closer to the reference values. The PZB functional with 70 Ry cutoff gives nearly the same results as B3LYP with TZ2P basis set in the low energy region, i.e. for the "A" pathways. In high energy region (the "B" pathways) the differences of 2 kcal/mol occur for *cis-B* and *ts-B* states, which is acceptable, however for *ss-B* a large difference of 35 kcal/mol appears. Although the discrepancy is large, the "B" pathway is not accessible for the system at room temperature, and we accept the PZB results with 70 Ry. The BLYP values with 120 Ry cutoff are also acceptable since the deviation from the reference energy is only 1.5 kcal/mol. It should be noted that reduction of the cutoff value causes significant gain in terms of the computing time. With this in mind one may consider the BLYP functional with 25 Ry cutoff as a promising choice for faster, although less accurate computations. The PBE and BP functionals significantly underestimate the energy barrier values for the "A" pathways (more than 50%), and are quite accurate in high energy region, which is however less interesting for our studies.

The PZB functional appeared to be the most accurate in describing the characteristic

points on the potential energy surface in comparison to previous *ab initio* studies. Therefore it was chosen for further calculations.

3 Free Energy Profiles

The calculation of the free energy profiles was carried out *via* thermodynamic integration in the "blue moon ensemble"[7]. In this approach the reaction pathway is divided into steps covering relevant values of the reaction coordinate. Then for each reaction step a constrained molecular dynamics run is carried out with the reaction coordinate being fixed at certain value by molecular constraints. Free energy derivatives are averaged over each run and integrated to give a free energy profile. This procedure was applied twice: with all atomic nuclei treated classically with CPMD, and with all atomic nuclei treated quantum-dynamically with PIMD. In the latter case the nuclei were quantized using 16 imaginary time-slices (see Figure 1 for visualization). The NVT constrained dynamics simulations were carried out in a 17 Å cubic box.

Two low energy reaction pathways in porphycene were studied. Calculations for *trans* → *ts-A* → *cis-A* and *trans* → *ss-A* → *trans* pathways were carried out using the CPMD and PIMD approaches. The former pathway was divided into 17 and 9 reaction steps, applying CPMD and PIMD, respectively. For the latter case, 10 reaction steps were used for only one half of the pathway, because the *trans* → *ss-A* transition is symmetrical to the *ss-A* → *trans* one. The total of number of 4000 steps of the constrained dynamics simulation was carried out for each reaction step, in each case. The time-step was 3 a.u. Fictitious electron mass was set to 350 a.u. for CPMD, and 300 a.u. for PIMD. Temperature of the Nose-Hoover thermostat was set to 300 K.

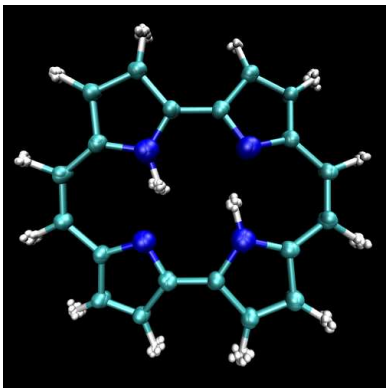
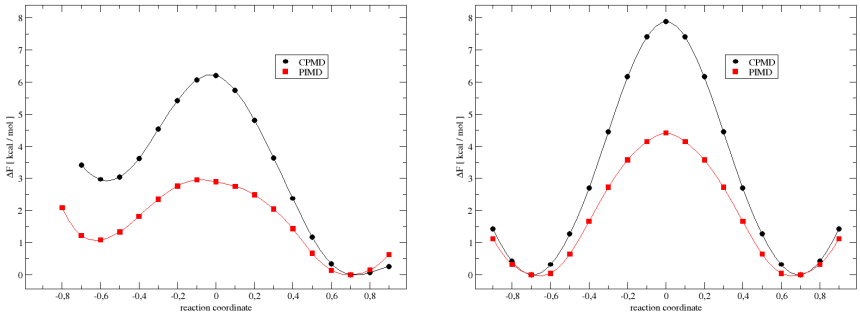


Figure 1. Path integral configuration of *trans* porphycene. Each nucleus is represented by cyclic polymer of 16 *beads*.

Figure 2 shows free energy profiles for the two pathways in porphycene computed from the averaged constraint forces. Averaging was carried out over the converged parts of the trajectories in order to avoid nonequilibrium regions.

The static calculations of the barrier height for the double proton transfer in porphycene on the *trans* → *ts-A* → *cis-A* pathway is 4.9 kcal/mol (see Table 1). Inclusion of classical thermal vibrations at 300 K increases the barrier by 1.4 kcal/mol to 6.3 kcal/mol. This is rather unexpected since the thermal vibrations should rather reduce the barrier. This may be the evidence of larger error of the method originating from poor equilibration of the system on one hand and small data set for the statistical averaging on the other. Taking this into account one would assume the error bar to be around 1 kcal/mol. Taking into account the quantum nature of the nuclei within the PIMD model reduces the barrier height to 1.8 kcal/mol (Fig. 2(a)). This can be treated as an expected result.

Interestingly the *cis-A* configuration turns out to be much more stable in the PIMD model than in the CPMD one. The free energy difference is about 3 kcal/mol, which is



(a) *trans* → *ts-A* → *cis-A* (b) *trans* → *ss-A* → *trans*

Figure 2. CPMD and PIMD free energy profiles for two reaction pathways in porphycene.

a significant effect even assumed a 1 kcal/mol error. This results in equal barrier heights for the transitions in both directions: the *trans* → *cis-A* barrier is approximately equal to the *cis-A* → *trans* barrier which yields 1.8 kcal/mol. This makes the *trans* and *cis-A* states practically indistinguishable, which is not the case for the CPMD model where the 'forward' barrier is 6.3 kcal/mol while the 'reverse' one yields 3.0 kcal/mol.

The *trans* → *ss-A* → *trans* pathway leads through the classical barrier of 7.5 kcal/mol. Applying the CPMD integration the barrier height is slightly higher and equals 7.9 kcal/mol which is consistent with the tendency observed previously. The PIMD computations show the low barrier of 4.4 kcal/mol.

4 Conclusions

The obtained so far results show that inclusion of nuclear quantum effects reduces the free energy barrier for the proton transfer reaction in porphycene. The *trans* → *ts-A* → *cis-A* pathway showed a 4.5 kcal/mol reduction of the barrier height. Qualitatively, it is expected, however, exact values of the free energy barriers require further refinement, carrying out longer simulations. Some difficulties in averaging the free energy derivatives in the process of thermodynamic integration result mostly from not fully equilibrated constrained dynamics trajectories.

The goal of comparing the free energy profiles of the two low-energy proton-transfer reactions in porphycene, i.e. concerted vs stepwise one between two symmetrical *trans* isomers, could not be completed yet, since we were not able to carry out the PIMD calculations of the *trans* → *ss-A* → *trans* pathway based on the NIC granted resources. Computations are being carried out.

The presented results may also suggest that a more precise description of the double proton transfer mechanism would require even a more sophisticated approach. This could account for a two-particle quantum state, which is the two-proton wave function defined on a spatial grid. The dynamics of such system should be governed by the time-dependent Schrödinger equation. The Hamiltonian should then contain terms describing the interactions between the two-particle, proton wave function and the classical ionic environment forming the molecular skeleton.

Acknowledgments

Computations were carried out at NIC Jülich (grant EPL00) and ICM Warsaw University. Authors express their gratitude for granting the access to the computing resources.

References

1. Ł. Walewski, P. Bała, M. Elstner, Th. Frauenheim, and B. Lesyng, *Fast QM/MM Method and Its Application to Molecular Systems*, Chem. Phys. Lett., **397**, 451–458, 2004.
2. The CPMD Consortium, http://www.cpmd.org/, *CPMD Consortium page*.
3. P. M. Kozlowski, M. Z. Zgierski, and J. Baker, *The inner-hydrogen migration and ground-state structure of porphycene*, J. Chem. Phys., **109**, 5905–5913, 1998.
4. J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B, **23**, 5048–5079, 1981.
5. A. D. Becke, *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. A, **38**, 3098–3100, 1988.
6. M. E. Tuckerman and D. Marx, *Heavy-Atom Skeleton Quantization and Proton Tunneling in "Intermediate-Barrier" Hydrogen Bonds*, Phys. Rev. Lett., **86**, 4946–4949, 2001.
7. E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, *Constrained reaction coordinate dynamics for the simulation of rare events*, Chem. Phys. Lett., **156**, 472–477, 1989.

Side-Chain Ordering in Homopolymers

Yanjie Wei[1], **Walter Nadler**[1], **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, D-52425 Jülich, Germany
E-mail: u.hansmann@fz-juelich.de

In order to study the relation between backbone and side-chain ordering in proteins, we have performed multicanonical simulations of five amino acid homopolymers. Glu10, Gln10, Asp10, Asn10, and Lys10 were selected to cover a wide variety of possible interactions between the side chains. All polymers undergo helix-coil transitions. We found that peptides with long side chains that are capable of hydrogen bonding, i.e. Glu10 and Gln10, exhibit a second transition at lower temperatures connected with side-chain ordering. This occurs in gas phase as well as in solvent. However, in polymers with short side chains capable of hydrogen bonding, i.e. Asp10 and Asn10, side-chain ordering takes place over a wide temperature range and exhibits no phase transition-like character. Again, these results are qualitatively independent of the environment. Side chain ordering in Lys10, whose side groups are long and polar, also takes place over a wide temperature range and exhibits no phase transition-like character in both environments.

1 Introduction

In the last two decades, energy landscape and folding funnel paradigms[1] have led to an emerging understanding of the protein folding process. However, these concepts describe only the general characteristics of folding. Many details still remain poorly understood. One aspect is the role of side-chain ordering in the folding process. Since amino acids are distinguished by the side chains, the study of side-chain ordering can help us understand the protein folding problem. Using the multicanonical Monte Carlo method[2,3], we have studied the side-chain ordering of five homopolymers Glu_{10}, Gln_{10}, Asp_{10}, Asn_{10}, and Lys_{10}[4,5]. It is found that, besides the helix-coil transition, a side-chain ordering transition is taking place for particular polypeptides only. The de-coupling of side-chain and backbone ordering transitions is independent of the environment.

2 Methods

We used the ECEPP/3 force field[6] as implemented in the program package SMMP[7]. Here the intramolecular interactions are approximated by a sum consisting of electrostatic energy, a Lennard-Jones term, a hydrogen-bonding term and a torsional energy term. The protein-solvent interactions are approximated by a solvent accessible surface term[8].

The above defined energy function leads to a landscape that is characterized by a multitude of minima separated by high barriers. In order to enhance sampling we rely on the multicanonical approach[2,3]. Here, configurations are weighted with a term $w_{MU}(E)$ determined iteratively such that the resulting histogram obeys

$$P_{MU}(E) \propto n(E) w_{MU}(E) \approx const\ , \tag{1}$$

where $n(E)$ is the spectral density of the system.

3 Results and Discussions

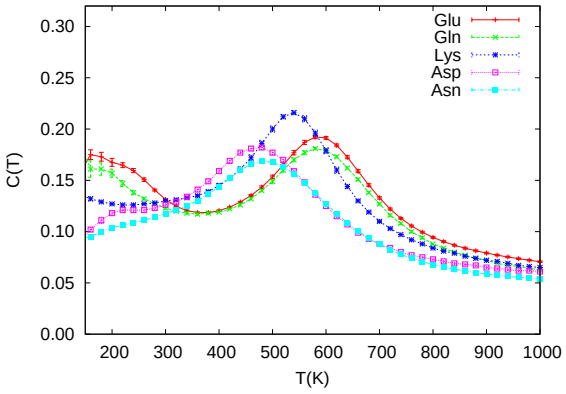


Figure 1. Specific heat $C(T)$ as function of temperature T for the five homopolymers in gas phase. This figure is taken from Ref. 5.

We first studied the molecules in gas phase. Fig. 1 displays the specific heat per molecule for the five polymers as a function of temperature. In each case one observes a peak at a temperature T_1 in the range $450\ K$ to $600\ K$. The individual peak temperatures T_1 and other properties are listed in Table 1. These peaks correspond to the helix-coil phase transitions which separate a high temperature region where the backbone has no ordering from a region where temperatures are low enough to allow the formation of an α-helix[4,5].

The transition temperatures are largest for Glu_{10} and Gln_{10}, and lowest for Asn_{10} and Asp_{10}, with the individual values being very close to each other in each block. The ordering of the transition temperature can be understood from the particular energetic and entropic contributions of the various side chains[5].

Interestingly, a second peak in the specific heat is observed for Glu_{10} and Gln_{10}. The peak temperatures T_2 are around $170\ K$ (see table 1). This second peak is related to the side-chain ordering in the form of hydrogen bonding between side groups[4,5].

Asn_{10} and Asp_{10}, whose side chains are only one CH_2-group shorter than those of Glu_{10} and Gln_{10}, show no second peak in the specific heat. The reason for this remarkably different behavior is the apparent existence of a threshold length for the side chains. If these chains are not long enough, the number of degrees of freedom is simply too small to allow for the fluctuations observed for Glu_{10} and Gln_{10}. Consequently, the ordering of the side chains in Asn_{10} and Asp_{10} does not have a transition-like character but is spread out over a wide range of temperatures. Side-chain ordering behavior for Lys_{10} is similar to Asn_{10} and Asp_{10}. Since its side groups can participate only in weak hydrogen bonds, they act neutrally at the helix-coil transition and finally order themselves by wrapping around the helical cylinder[5].

The structural features in the low-temperature phase found here correspond to early results from theoretical investigations on the ground state structure of such homopolymers in vacuum, performed by H. Scheraga's group, see e.g. Ref. 9.

Table 1. Properties of the helix-coil transitions observed for the five homopolymers: Transition temperatures T_1 and T_2, half width ΔT, and the specific heat per molecule at the transition temperatures, $C(T_1)$ and $C(T_2)$. The number in parenthesis is the uncertainty in the last digit. The data are taken from Ref. 5.

	vacuum			solvent		
	T_1[K]	ΔT[a]K]	$C(T_1)$ [b]	T_1[K]	ΔT[a]K]	$C(T_1)$ [b]
Glu_{10}	587(14)	161	0.193(2)	477(7)	88	0.296(3)
Gln_{10}	584(14)	163	0.181(1)	484(8)	97	0.268(3)
Lys_{10}	538(8)	151	0.216(2)	447(10)	98	0.266(3)
Asn_{10}	485(19)	193	0.169(2)	424(9)	105	0.249(4)
Asp_{10}	471(19)	170	0.182(2)	415(6)	82	0.300(3)
	T_2[K]	ΔT[c]K]	$C(T_2)$ [b]	T_2[K]	ΔT[c]K]	$C(T_2)$ [b]
Glu_{10}	166(16)	203	0.176(4)	111(10)	166	0.152(2)
Gln_{10}	181(23)	187	0.161(6)	120(17)	146	0.133(3)

[a]Determined at $C = [C(T_1) + C(T_{min})]/2$, with T_{min} from either $C'(T_{min}) = 0$ or $C''(T_{min}) = 0$.
[b]Unit is Kcal/mol.
[c]Determined at $C = [C(T_2) + C(T_{min})]/2$, with T_{min} from $C'(T_{min}) = 0$.

So far, we have focused on the molecules in gas phase. However, in nature proteins are usually solvated, and their function often depends on the solvent environment. For this reason we have extended our investigation to that of solvated homopolymers.

As in the case of gas phase simulations we observed for all five molecules a helix-coil transition indicated by peaks in the specific heat, see Fig. 2, their properties also being listed in Table 1. As can be seen clearly, the width of these peaks is much narrower, and their height is larger than in gas phase, indicating a sharper transition.

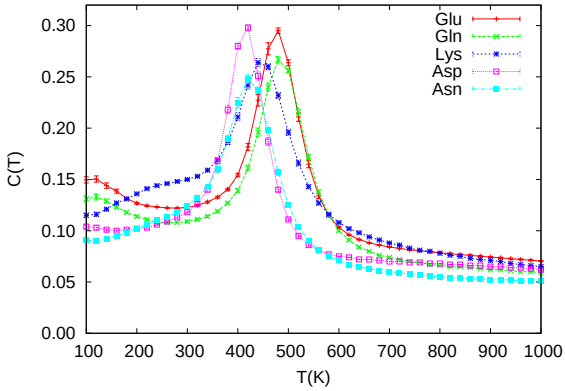


Figure 2. Specific heat $C(T)$ as function of temperature T for the five solvated homopolymers. The figure is taken from Ref. 5.

All transition temperatures are shifted to lower values. The reason is the competition between the formation of backbone hydrogen bonds that stabilize the α-helix, and interac-

tion between the peptide and the solvent in the coil phase. While in vacuum the transition to the coil phase is driven solely by entropy, here also the peptide-solvent interaction favors the coil phase. These effects collaborate so that the transition takes place at a lower temperature and becomes sharper. However, the relative ordering of the transition temperatures among the polymers remains practically the same as in the gas phase.

Interestingly, again a second peak in the specific heat is observed for Glu_{10} and Gln_{10}, the exact properties also being listed in Table 1. However, here the reason for such a behavior is not the formation of hydrogen bonds among the side chains, because the total number of hydrogen bonds is limited to those that stabilize the helix backbone. A closer inspection shows that this second peak in the specific heat of Glu_{10} is mainly due to fluctuations in the solvent energy[4]. For Gln_{10}, it is the interplay of fluctuations in solvent energy with the fluctuations in ECEPP/3 energy that causes the second peak[5].

No such second peaks in the specific heat can be observed for Asn_{10} and Asp_{10}. Their fluctuations in solvent energy are cancelled by the strong anti-correlations between solvent energy fluctuations and ECEPP/3 energy fluctuations for these two peptides. For similar reasons Lys_{10} also does not exhibit particular features with respect to side-chain ordering. As before in gas phase, its side-chain ordering does not have a transition-like character but is spread out over a wide range of temperatures.

Detailed reports of this work are available in Refs. 4 and 5.

Acknowledgments

We thank H. Scheraga for motivating this research. Support by a research grant (CHE-0313618) of the National Science Foundation (USA) is acknowledged.

References

1. J. N. Onuchic, Z. Luhey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
2. B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).
3. U. H. E. Hansmann and Y. Okamoto, J. Comp. Chem. **14**, 1333 (1993).
4. Y. Wei, W. Nadler, and U. H. E. Hansmann, J. Chem. Phys. **125**, 164902 (2006).
5. Y. Wei, W. Nadler, and U. H. E. Hansmann, J. Phys. Chem. B **111**, 4244 (2007).
6. M. J. Sippl, G. Némethy, and H. A. Scheraga, J. Phys. Chem. **88**, 6231 (1984), and references therein.
7. F. Eisenmenger, U. H.E. Hansmann, Sh. Hayryan, and C.-K. Hu, Comp. Phys. Comm. **174**, 422 (2006).
8. T. Ooi, M. Obatake; G. Nemethy, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 3086 (1987).
9. J. F. Yan, G. Vanderkooi and H. A. Scheraga, J. Chem. Phys. **49**, 2713 (1968).

Dihedral Angle Patterns in Coil Regions of Protein Structures

Olav Zimmermann[1] **and Ulrich H. E. Hansmann**[1,2]

[1] John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {*olav.zimmermann, u.hansmann*}*@fz-juelich.de*

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

We report first results on a Decision Tree classifier for subclassifying β-turn types from sequence. It is based on previous work where we developed a software DHPRED that predicts dihedral angle regions for each residue in a polypeptide chain. For 5 out of 8 β-turn (sub)classes we obtain good prediction results with Matthew's correlation coefficients between 0.3 and 0.6, thus providing additional geometric constraints in those critical regions which determine the fold topology of a protein. Three β-turn classes however can not be classified with sufficient accuracy. We discuss possible sources of misclassifications and outline further research directions.

1 Introduction

Many structure prediction strategies involve the use of dihedral angle constraints from secondary structure prediction. From the usual 3-class predictions useful dihedral bounds can however only be derived for α-helices and β-sheets. We recently developed a program (DHPRED) to predict dihedral angle ranges for all residues, including those in coil regions[1]. The most abundant structural motifs within the coil regions of proteins are β-turns which are defined as tetrapeptide motifs with a maximum C-α distance of 7 $\mathring{A}$. They have been originally described by Venkatachalam[2] and Lewis[3] and classified by the dihedral angle regions of their central two residues. β-turns are of prime importance for the tertiary structure of proteins as most of them characterize positions where the peptide chain reverses its direction. Several algorithms have been described to predict the location of β-turns from sequence, however to our knowledge there is no prediction program available which can distinguish between the different types of β-turns. In this study we use the dihedral angle regions predicted by DHPRED to develop a classification algorithm for individual β-turn classes.

2 Methods

As reference for the training of the classification algorithm we use a nonredundant set of protein structures from the Protein Data Bank (PDB) with pairwise sequence identities $\leq 25\%$ and x-ray resolutions ≤ 2.0 $\mathring{A}$. The assignment of the different turn types described by Lewis is performed by the PROMOTIF software[4]. The Decision Tree algorithm[5] as implemented in[6] is employed for multi-class classification. For this initial study we use

only minimal information. For the central two residues we encode (a) the amino acid type as either Glycine, Proline or all other, (b) the 3-class secondary structure prediction from PSIPRED[7] and (c) the dihedral angle regions predicted by DHPRED. We use the definitions in Table 1 for prediction outcome types:

Prediction	Observation	
	+1	-1
+1	TP (True Positive)	FP (False Positive)
-1	FN (False Negative)	TN (True Negative)

Table 1. Definition of prediction outcome types.

and employ as performance measures the accuracy $Acc = (TP + TN)/(TP + TN + FP + FN)$, specificity $Spec = TN/(TN + FP)$, sensitivity $Sens = TP/(TP + FN)$ and Matthew's correlation coefficient MCC $= \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

3 Results and Discussion

For the ca. 1,000 proteins of the nonredundant PDB-set, PROMOTIF assigns about 9,000 β-turns. We devide the turns randomly into two sets of same size and use one set for training of the Decision Tree classifier and the other for testing. Table 2 shows the confusion matrix of the test set and the distribution of the different β-turn classes.

		predicted								
	β-turn class	I	I'	II	II'	IV	VIa	VIb	VIII	sum
observed	I	1099	65	17	5	449	2	1	34	1672
	II	72	305	26	0	77	0	0	3	483
	I'	21	55	89	1	26	0	0	1	193
	II'	26	1	8	12	74	1	0	1	123
	IV	744	104	35	16	697	4	16	50	1666
	VIa	4	0	0	0	5	10	15	0	34
	VIb	0	0	0	0	5	2	34	0	41
	VIII	107	3	2	2	195	0	0	43	352
	sum	2073	533	177	36	1528	19	66	132	4564

Table 2. Confusion matrix for β-turn classifier.

Table 3 shows the performance of the Decision Tree classifier using different measures. While the Matthew's correlation coefficients for the classes show reasonable (I, VIa) or good (I', II, VIb) classification performance, the classes II', IV and VIII can not be identified correctly. The confusion matrix shows that the pseudo-class IV, which is just defined as all β-turns which do not belong to any other class, is responsible for most of the misclassifications.

β-turn class	Acc	Spec	Sens	MCC
I	65.7%	0.66	0.66	0.31
I'	63.1%	0.94	0.63	0.55
II	46.1%	0.98	0.46	0.46
II'	9.8%	0.99	0.10	0.17
IV	41.8%	0.71	0.42	0.13
VIa	29.4%	1.00	0.29	0.39
VIb	82.9%	0.99	0.83	0.65
VIII	12.2%	0.98	0.12	0.16

Table 3. Performance measures for β-turn classifier: accuracy (Acc.), specificity (Spec.), sensitivity (Sens.), and Mathew's correlation coefficient (MCC).

Our initial results show the general feasibility to predict β-turn types from the amino acid sequence. This study does not address the separation of β-turns from other secondary structure elements but such algorithms are already available[8]. In the future we will direct our research to assemble a comprehensive library for the prediction of local structure features. For the β-turn classification we will e.g. use a derivate of the DHPRED software which is targeted to different dihedral regions as used in Lewis' definition of β-turns and employ Support Vector Machine algorithms for the classification.

Acknowledgments

This work is supported in part by an NIH research grant (GM62838).

References

1. O. Zimmermann and U. H. E. Hansmann, *Support Vector Machines for prediction of dihedral angle regions*, Bioinformatics **22**, 3009–3015, 2006.
2. C. M. Venkatachalam, *Stereochemical criteria for polypeptides and proteins V. Conformation of a system of three linked peptide units.*, Biopolymers **6**, 1425–1436, 1968.
3. P. N. Lewis, F. A. Monany, and H. A. Scheraga, *Chain reversals in proteins*, Biochim. Biophys. Acta **303**, 211–229, 1973.
4. E. G. Hutchinson and J. M. Thornton, *PROMOTIF - A program to identify structural motifs in proteins*, Protein Science **5**, 212–220, 1996.
5. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, (Prentice Hall, Englewood Cliffs, NJ, 2003).
6. P. Norvig, Python implementation of the Decision Tree algorithm decribed in[5], `http://aima.cs.berkeley.edu/python/`.
7. D. T. Jones, *Protein secondary structure prediction based on position-specific scoring matrices*, J Mol Biol **292**, 195–202, 1999.
8. Q. Zhang, S. Yoon and W. J. Welsh, *Improved method for predicting β-turn using support vector machine*, Bioinformatics **21**, 2370–2374, 2005.

Workshop Program

Wednesday 02.05.2007

07:40	Pick up at Hotels in Jülich
08:00-09:00	Registration
09:00-09:20	Welcome by Prof. Dr. A. Bachem (CEO, FZ Jülich)
09:20-09:30	Introductory Remarks (Prof. U. H. E. Hansmann, NIC-CBB)

Morning Session (Chair: Ulrich H. E. Hansmann)

09:30-10:15	Harold Scheraga (Cornell University, Ithaca, USA): Evolution of Protein Simulation
10:15-11:00	Robert B. Russell (EMBL, Heidelberg, Germany): Pushing Details into the Interactome
11:00-11:30	Coffee break
11:30-12:15	Koichi Takahashi (tMSI, Berkeley, USA): The E-Cell Project and Challenges in Computational Systems Biology
12:30-14:00	Lunch

Afternoon Session (Chair: Jan H. Meinke)

14:00-14:45	Adrian Roitberg (University of Florida, Gainesville, USA): Mixed Quantum Mechanics/Classical Mechanics Methods for the Study of Enzymatic Reactions
14:45-15:30	Marek Cieplak (Polish Academy of Sciences, Warsaw, Poland): Stretching to Understand Proteins
15:30-16:00	Coffee break
16:00-16:30	Bernd Berg (FSU, Tallahassee, USA): Residual Entropy of Ice I from Multicanonical Simulations
16:30-17:00	Abhinav Verma (FZK, Karlsruhe, Germany): All-Atom Protein Folding and Structure Prediction in a Transferable Universal Free-Energy Force-Field
17:00-17:30	Walter Nadler (MTU, Houghton, USA): Channel Transport and Molecular Motors without Brownian Ratchets
17:30-18:00	Inta Liepina (Latvian Institute of Organic Synthesis, Riga, Latvia): Multiple Beta-Sheet Molecular Dynamics of Two Abl-SH3 Domain Peptides
18:00	Welcome Reception and Poster Session
20:00	Open Discussion: Supercomputing and the Physics of Cells
20:00/21:30	Bus to Hotels in Jülich

Thursday 03.05.2007

08:10	Pickup at Hotels in Jülich

Morning Session (Chair: Olav Zimmermann)

09:00-09:45	Jeffrey Skolnick (Georgia Institute of Technology, Atlanta, USA): Prediction of Protein Structure, Function and Druggability on a Proteomic Scale
09:45-10:30	Christodoulos Floudas (Princeton University, Princeton, USA): Advances In De Novo Protein Design
10:30-11:00	Coffee break
11:00-11:30	Mai Suan Li (Polish Academy of Sciences, Warsaw, Poland): Mechanism of Fibril Formation of Short Peptides
11:30-12:00	Velia Minicozzi (University of Rome "Tor Vergata", Rome, Italy): The Role of Metals in Misfolding and Aggregation Processes X-ray Spectroscopy and Numerical Simulations
12:00-12:30	Wolfhard Janke (University of Leipzig, Leipzig, Germany): Microcanonical Analyses of Peptide Aggregation Processes
12:30-14:00	Lunch

Afternoon Session (Chair: Tatjana Eitrich)

14:00-14:45	Rebecca Wade (EML, Heidelberg, Germany): Bridging from Molecular Simulation to Biochemical Networks
14:45-15:30	Adam Liwo (University of Gdansk, Gdansk, Poland): Mesoscopic Dynamics with the UNRES Force Field – a Tool for Studying the Kinetics and Thermodynamics of Protein Folding
15:30-16:00	Coffee break
16:00-16:30	Peter Virnau (Johannes Gutenberg-Universität, Mainz, Germany): Knots in Proteins, DNA and Polymers: Statistics, Function and Evolution
16:30-17:00	Siegfried Höfinger (MTU, Houghton, USA): Comparing Semi-Empirical versus Classic Charge Assignments in BioMolecules and their Effect on Electrostatic Potentials
17:00-17:30	Klaas Hellingwerf (University of Amsterdam, Amsterdam, NL): Photosensory Proteins as Vehicles to Bridge Computational Biophysics and Systems Biology to Initiate the Field of Synthetic Biology
17:30-18:00	Léo Degrève (Universidade de São Paulo, Ribeirao Preto, Brazil): A Molecular Dynamics Study of the Basic Fibroblast Growth Factor – Fibroblast Growth Factor Receptor Complex
18:15	Bus to Castle Obbendorf (Hambach) for Dinner
18:30	Dinner in Celebration of 20 Years of NIC (sponsored by IBM) Commemorative Speech by Prof. Dr. Dr. Thomas Lippert, Director of NIC and ZAM
22:00	Bus to Hotels in Jülich

Friday 04.05.2007

8:10	Pickup at Hotels in Jülich

Morning Session (Chair: Sandipan Mohanty)

09:00-09:45	Dave Thirumalai (University of Maryland, College Park, USA): Measuring Energy Landscape Parameters of Biomolecules and their Complexes
09:45-10:30	Bogdan Lesyng (Warsaw University, Warsaw, Poland): Causality and Correlation Analyses of Molecular Dynamics Simulation Data
10:30-11:00	Coffee break
11:00-11:30	Nikolay Korolev (Nanyang Technological University, Singapore): DNA Packaging and Electrostatic Interactions
11:30-12:00	Georgios Papadopoulos (University of Thessaly, Larisa, Hellas): The Zinc-Finger Motif of T.thermophilus Ribosomal Protein S14 and the Functionality of E.coli Ribosome
12:00-12:30	Giovanni La Penna (National Research Council, Florence, Italy): Anisotropic Internucleosome Interactions and Geometrical Constraints Favour the Two-Start Helical Structure of Chromatin
12:30-13:30	Lunch

Afternoon Session (Chair: Ulrich H. E. Hansmann)

13:30-14:15	José Onuchic (UCSD, La Jolla, USA): Mechanisms of Protein Assembly and Folding: Lessons from Minimalist Models
14:15-15:00	Jörg Langowski (Deutsches Krebsforschungszentrum, Heidelberg, Germany): Chromatin Dynamics in Silicio
15:00-15:05	Concluding Remarks (Prof. Ulrich H. E. Hansmann, NIC-CBB)
afterwards	Bus to Cologne Social Event: Guided Tour Cologne Cathedral and Treasure Chamber, Visit of a Traditional Brewhouse
21:30	Bus to Jülich

List of Participants

- Anand, Priya; Panjab University; Chandigarh; India
- Andrews, Maximilian; University of Dortmund; Dortmund; Germany
- Arashiro, Everaldo; FZJ, NIC; Jülich; Germany
- Behringer, Hans; Bielefeld University; Bielefeld; Germany
- Berg, Bernd; FSU; Tallahassee; USA
- Bischler, Ulrike; VolkswagenStiftung; Hannover; Germany
- Bisson, Melanie; Heinrich-Heine-Universität; Düsseldorf; Germany
- Buchen, Elisa; Heinrich-Heine-Universität; Düsseldorf; Germany
- Burri, Raghunadha Reddy; University of Dortmund; Dortmund; Germany
- Cannavacciuolo, Luigi; FZJ, IFF; Jülich; Germany
- Cao, Ye; University of Bonn; Bonn; Germany
- Cieplak, Marek; Polish Academy of Sciences; Warsaw; Poland
- Degrève, Léo; Universidade de São Paulo; Ribeirão Preto; Brazil
- Di Fiore, Stefano; RWTH; Aachen; Germany
- Eitrich, Tatjana; FZJ, NIC; Jülich; Germany
- Ensikat, Hans Jürgen; University of Bonn; Bonn; Germany
- Ferro, Noel; University of Bonn; Bonn; Germany
- Fink, Florian; University of Regensburg; Regensburg; Germany
- Fischer, Bernhard; FZK; Karlsruhe; Germany
- Floudas, Christodoulos; Princeton University; Princeton; USA
- Fredel, Fabian; FZJ, NIC; Jülich; Germany
- Gorecki, Adam; University of Warsaw; Warsaw; Poland
- Gowtham, S.; MTU; Houghton; USA
- Greulich, Philip; Universität zu Köln; Köln; Germany
- Griewel, Axel; Universität Hamburg; Hamburg; Germany

- Gronwald, Wolfram; University of Regensburg; Regensburg; Germany
- Grudinin, Sergei; FZJ, INB-2; Jülich; Germany
- Hallay-Suszek, Marta; Warsaw University; Warsaw; Poland
- Hammen, Volker; UFZ; Halle (Saale); Germany
- Hansmann, Ulrich; FZJ/NIC and MTU; Jülich; Germany
- Hayat, Sikander; University of Saarland; Saarbrücken; Germany
- Hayryan, Shura; Academia Sinica; Taipei; Taiwan (ROC)
- Held, Martin; FU-Berlin; Berlin; Germany
- Hellingwerf, Klaas; University of Amsterdam; Amsterdam; NL
- Heuer, Andreas; University of Münster; Münster; Germany
- Höfinger, Siegfried; MTU; Houghton; USA
- Hoffmann, Bernd; FZJ, IBN-4; Jülich; Germany
- Hoffmann, Kurt; RWTH; Aachen; Germany
- Hsu, Hsiao-Ping; Johannes Gutenberg-Universität; Mainz; Germany
- Jain, Pooja; University of Nottingham; Nottingham; United Kingdom
- Janke, Wolfhard; University of Leipzig; Leipzig; Germany
- Johnson, Gary; U. S. Department of Energy; Washington, DC; USA
- Kar, Parimal; MTU; Houghton; USA
- Klenin, Konstantin; FZK; Karlsruhe; Germany
- Koch, Kerstin; University of Bonn; Bonn; Germany
- Kokh, Daria; Wuppertal University; Wuppertal; Germany
- Kondov, Ivan; FZK; Karlsruhe; Germany
- Korolev, Nikolay; Nanyang Technological University; Singapore; Singapore
- Krukau, Aliaksei; University of Dortmund; Dortmund; Germany
- Kubiak, Karina; Theoretical Molecular Biophysics Group; Torun; Poland
- Kulinski, Tadeusz; Polish Academy of Sciences; Poznan; Poland
- La Penna, Giovanni; National Research Council; Florence; Italy
- Lammert, Heiko; Universität Münster; Münster; Germany

- Langowski, Jörg; Deutsches Krebsforschungszentrum; Heidelberg; Germany
- Lesyng, Bogdan; Warsaw University; Warsaw; Poland
- Li, Mai Suan; Polish Academy of Sciences; Warsaw; Poland
- Liepina, Inta; Latvian Institute of Organic Synthesis; Riga; Latvia
- Lindgren, Matteus; Umeå University; Umeå; Sweden
- Liwo, Adam; University of Gdansk; Gdansk; Poland
- Meerbach, Eike; FU-Berlin; Berlin; Germany
- Meinke, Jan; FZJ, NIC; Jülich; Germany
- Minicozzi, Velia; University of Rome ”Tor Vergata”; Rome; Italy
- Mohanty, Sandipan; FZJ, NIC; Jülich; Germany
- Möhl, Christoph; FZJ, IBN-4; Jülich; Germany
- Nadler, Walter; MTU; Houghton; USA
- Nagel-Steger, Luitgard; Heinrich-Heine-Universität; Düsseldorf; Germany
- Namasivayam, Vigneshwaran; University of Leipzig; Leipzig; Germany
- Neuhaus, Thomas; FZJ, NIC; Jülich; Germany
- Noe, Frank; FU Berlin; Berlin; Germany
- Nowak, Wieslaw; Theoretical Molecular Biophysics Group; Torun; Poland
- Nöh, Katharina; FZJ, IBT-2; Jülich; Germany
- Ołdziej, Stanisław; University of Gdansk; Gdansk; Poland
- Onuchic, José; UCSD; La Jolla; USA
- Pachov, Georgi; EML; Heidelberg; Germany
- Papadopoulos, Georgios; University of Thessaly; Larisa; Hellas
- Pentas, Stefanos; Aristotle University of Thessaloniki; Thessaloniki; Hellas
- Peplowski, Lukasz; Theoretical Molecular Biophysics Group; Torun; Poland
- Perez Sanchez, Horacio Emilio; FZK; Karlsruhe; Germany
- Quintilla, Aina; FZK; Karlsruhe; Germany
- Reich, Lothar; MPI for Colloids and Surfaces; Potsdam; Germany
- Roitberg, Adrian; University of Florida; Gainesville; USA

- Russell, Robert B.; EMBL; Heidelberg; Germany
- Saufi, Ahmad; Heinrich-Heine-Universität; Düsseldorf; Germany
- Scharein, Benjamin; Heinrich-Heine-Universität; Düsseldorf; Germany
- Scheiba, Rafael; Heinrich-Heine-Universität; Düsseldorf; Germany
- Scheraga, Harold; Cornell University; Ithaca; USA
- Scherer, Nicole; Heinrich-Heine-Universität; Düsseldorf; Germany
- Schluttig, Jakob; Uni Heidelberg; Heidelberg; Germany
- Singh, Gurpreet; University of Dortmund; Dortmund; Germany
- Singh, Sandesh; University of Bonn; Bonn; Germany
- Skolnick, Jeffrey; Georgia Institute of Technology; Atlanta; USA
- Sulkowska, Joanna I.; Polish Academy of Sciences; Warsaw; Poland
- Sutmann, Godehard; FZJ, ZAM; Jülich; Germany
- Szymczak, Piotr; Warsaw University; Warsaw; Poland
- Takahashi, Koichi; tMSI; Berkeley; USA
- Thirumalai, Dave; University of Maryland; College Park; USA
- Verma, Abhinav; FZK; Karlsruhe; Germany
- Virnau, Peter; Johannes Gutenberg-Universität; Mainz; Germany
- Vollmar, Melanie; Heinrich-Heine-Universität; Düsseldorf; Germany
- Wade, Rebecca; EML; Heidelberg; Germany
- Walewski, Łukasz; University of Warsaw; Warsaw; Poland
- Wei, Yanjie; MTU; Houghton; USA
- Westlund, Per-Olof; Umeå University; Umeå; Sweden
- Xiao, Xiaolin; FZJ, NIC; Jülich; Germany
- Younesi, Erfan; University of Bonn; Bonn; Germany
- Zimmermann, Olav; FZJ, NIC; Jülich; Germany

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 Lövenich
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

GALA Grünenthal Applied Life Science Analysis
Achim Kless and Johannes Grotendorst (Editors)
NIC Series Volume 30
ISBN 3-00-017349-8, November 2006, 204 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

Parallel Computing: Current & Future Issues of High-End Computing
Proceedings of the International Conference ParCo 2005
G.R. Joubert, W.E. Nagel, F.J. Peters,
O. Plata, P. Tirado, E. Zapata (Editors)
NIC Series Volume 33
ISBN 3-00-017352-8, October 2006, 930 pages

From Computational Biophysics to Systems Biology 2006 Proceedings
U.E.H. Hansmann, J. Meinke, S. Mohanty, O. Zimmermann (Editors)
NIC Series Volume 34
ISBN-10 3-9810843-0-6, ISBN-13 978-3-9810843-0-6,
September 2006, 224 pages

Dreistufig parallele Software zur Parameteroptimierung von Support-Vektor-Maschinen mit kostensensitiven Gütemaßen
Tatjana Eitrich
NIC Series Volume 35
ISBN 978-3-9810843-1-3, March 2007, 262 pages

All volumes are available online at

http:// www.fz-juelich.de/nic-series/.