



Computational Biology and Biophysics at NIC

Ulrich H.E. Hansmann

published in

NIC Symposium 2006,
G. Münster, D. Wolf, M. Kremer (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. 32, ISBN 3-00-017351-X, pp. 13-20, 2006.

© 2006 by John von Neumann Institute for Computing
Permission to make digital or hard copies of portions of this work for personal or classroom use is granted provided that the copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise requires prior specific permission by the publisher mentioned above.

<http://www.fz-juelich.de/nic-series/volume32>

Computational Biology and Biophysics at NIC

Ulrich H. E. Hansmann

John von Neumann Institute for Computing (NIC)
Research Centre Jülich, 52425 Jülich, Germany
E-mail: u.hansmann@fz-juelich.de
Dept. of Physics, Michigan Technological University
Houghton, MI 49931, USA

We summarize shortly the research program of the newly installed research group "Computational Biology and Biophysics" at NIC. This group exists since July 2005 and investigates the Biophysics and Biochemistry of biological macromolecules by means of high performance computing.

1 Introduction

As of July 1st 2005 the "Complex Systems" group at NIC has been replaced by the new research group "Computational Biology and Biophysics" anticipating that this area of research will in the next years have more and more a need for high performance computing. This is because a new challenge has emerged after the successful deciphering of whole genomes: for most sequences we do not know the function of the corresponding proteins, the workhorses in a cell that are responsible for transporting molecules, catalyzing biochemical reactions, or fighting infections.

Proteins are only functional if they assume specific shapes. Despite decades of research it is still an open question how these structures emerge from a protein's chemical composition (the sequence of amino acids as specified in the genome). An answer to this question could lead to a deeper understanding of various diseases that are caused by the miss-folding of proteins, and enable the design of novel drugs with customized properties.

Computer experiments offer one way to gain such knowledge but are extremely difficult for realistic protein models¹: all-atom models of proteins lead to a rough energy landscape with a huge number of local minima separated by high energy barriers. Consequently, sampling of low-energy conformations becomes a hard computational task, and physical quantities cannot be calculated accurately from simple low-temperature molecular dynamics or Monte Carlo simulations. The difficulties become even more pronounced if the structure of a protein depends on its interaction with other bio-molecules. Overcoming these obstacles may be one of the defining challenges in high performance computing for the next few years and will require the use of massive parallel computers such as JUMP and the new BlueGene computer JUBL in Jülich.

Research in the new group is concerned with the development and test of algorithms for these machines that allow atomistic simulations of stable domains in proteins (usually of order 50-200 residues), i.e. for overcoming the protein-folding problem². Protein-protein interactions are the topic of another line of research. Especially interesting are the conditions under which proteins mis-fold and aggregate. As abnormally folded and aggregated proteins are related to the outbreak of various diseases, such simulations may

provide insight into the mechanism of their pathogenesis. Protein-ligand binding and protein interaction networks belong to the same research direction and provide an interface for collaborations with bioinformatics groups.

Related to the above described research is the development and publication of new software for simulations of protein. These programs will be included in future updates of SMMP³, the freely available program package that was developed by my group.

2 Algorithms for Protein Simulations

The key-idea behind the novel techniques employed by us is to replace the canonical weights, that suppress the crossing of an energy barrier of height ΔE by a factor $\propto \exp(-\Delta E/k_B T)$ (k_B is the Boltzmann constant and T the temperature of the system), with such weights that allow the system to escape out of local minima. Often the weights are chosen in such a way that a Monte Carlo or molecular dynamics simulation will lead to a uniform distribution of a pre-chosen physical quantity. For instance, in multicanonical sampling⁴ the weight $w(E)$ leads to a distribution

$$P(E) \propto n(E)w(E) = \text{const}, \quad (1)$$

with $n(E)$ the density of states. A free random walk in the energy space is performed that allows the simulation to escape from any local minimum. From this simulation one can calculate the thermodynamic average of any physical quantity A by re-weighting⁵:

$$\langle \mathcal{A} \rangle_T = \frac{\int dx \mathcal{A}(x) w^{-1}(E(x)) e^{-E(x)/k_B T}}{\int dx w^{-1}(E(x)) e^{-E(x)/k_B T}}. \quad (2)$$

Here, x labels the configurations. The weights $w(E)$ are not *a priori* known and estimators have to be determined by an iterative procedure described in Refs. 4,6. The first application of this technique to protein studies can be found in Ref. 7.

The computational effort increases in multicanonical simulations with the number of residues as $\approx N^4$. While this is a much better numerical performance than in canonical simulations where one would expect a supercritical slowing down (i.e. the computer time would grow as $\propto e^{aN}$ with a an unknown constant), this scaling limits the size of systems that can be studied. In general, the computational effort in generalized-ensemble algorithms scales as $\propto X^2$ where X is the variable in which one wants a flat distribution. This is because these algorithms generate an unbiased $1D$ random walk in the ensemble coordinate. In the multicanonical algorithm the coordinate is the potential energy $X = E$. Since $E \propto N^2$ the scaling relation for multicanonical simulations is recovered. Hence, a better scaling of the computer time can be obtained by choosing a more appropriate ensemble coordinate than the energy. We have demonstrated this recently for the 36 residue villin headpiece sub-domain HP-36⁸ using the helicity as system coordinate. We are now extending this approach to proteins that have not only helices as secondary structure elements. This requires to explore possible parameters for generalized-ensembles other than the energy or the helicity. Examples are the simple scoring function of Chang et al.⁹ or the so-called hydrophobic ratio of Silverman¹⁰.

All generalized-ensemble techniques are designed to explore low energy configurations but avoiding at the same time entrapment in local minima. In *energy landscape paving* (ELP), a new optimization method that proved to be very promising in protein studies¹¹,

this is achieved by performing low-temperature Monte Carlo simulations with a modified energy expression that steers the search away from regions already explored:

$$w(\tilde{E}) = e^{-\tilde{E}/k_B T} \quad \text{with} \quad \tilde{E} = E + f(H(q, t)) . \quad (3)$$

Here, T is a (low) temperature, \tilde{E} serves as a replacement of the energy E , and $f(H(q, t))$ is a function of the histogram $H(q, t)$ in a pre-chosen ‘‘order parameter’’ q . Within ELP the weight of a local minimum state decreases with the time the system stays in that minimum till it is no longer favored. The system will then explore higher energies till it falls into a new local minimum. Obviously, for $f(H(q, t)) = f(H(q))$ the method reduces to the various generalized-ensemble methods² (for instance for $f(H(q, t)) = \ln H(E)$ to multicanonical sampling).

Another way of enhancing the sampling of low-energy protein configurations that is especially interesting for parallel computing is parallel tempering (also known as replica exchange)¹², a technique that was first introduced to protein folding in Ref. 13. In its most common form, one considers an artificial system built up of N *non-interacting* copies of the molecule, each at a different temperature T_i . In addition to standard Monte Carlo or molecular dynamics moves that affect only one copy, parallel tempering introduces a new *global* update¹²: the exchange of conformations between two copies i and $j = i + 1$ with probability

$$w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) . \quad (4)$$

This exchange of conformations leads to a faster convergence of the Markov chain than is observed in regular canonical simulations. Note that parallel tempering does not require Boltzmann weights. The method can be combined easily with other generalized-ensemble techniques as was demonstrated first in Ref. 13.

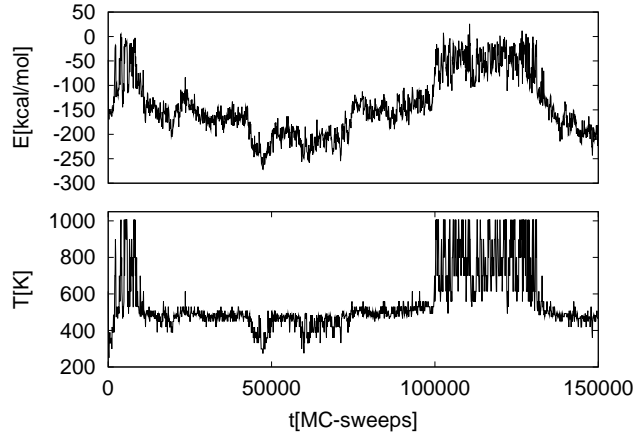


Figure 1. Time series of energy and temperature for a Parallel Tempering simulation of the protein HP-36

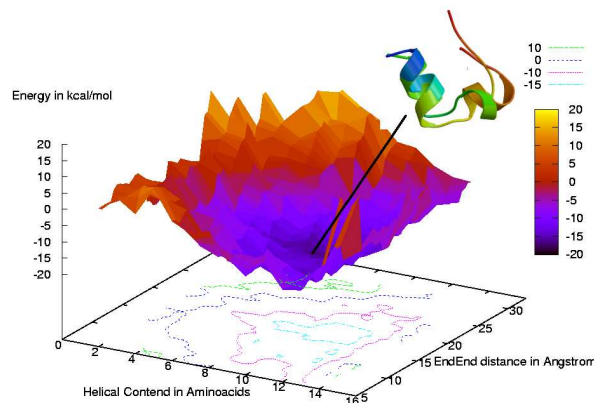


Figure 2. Energy landscape of the 20-residue trp-cage protein

I show as an example in Fig. 1 the time series of temperature and energy of one arbitrary chosen replica as obtained in a parallel tempering simulation of the 36-residue protein HP-36 (the figure is taken from Ref. 14). Note how the resulting random walk in temperature leads to one in energy that enables escapes out of local minima. In this way sampling of low-energy structures will be enhanced. A simple implementation of this and other modern protein simulation techniques can be found in the free program package SMMP (Simple Molecular Mechanics for Proteins)³ which is available from www.phy.mtu.edu/biophys/smmp.htm. The present version allows only simulation of isolated molecules but we are now re-writing the program to allow simulation of more than one (interacting) macromolecules. The package is written in FORTRAN but we are currently working on a C++ version. Test, modification and/or optimization of SMMP for Grid-computing are also planned.

3 Physics of Folding

Current applications focus on probing the mechanism of folding in small proteins and the conditions under which proteins mis-fold and aggregate. It is now widely believed that the energy landscape of proteins (in contrast to random heteropolymers) resembles a partially rough funnel with a free energy gradient toward the native structure (for a review, see, for instance, Ref. 15). Folding occurs by a multi-pathway kinetics and the particulars of the funnel landscape determine the transitions between the different thermodynamic states^{16,17}. Fig. 2 shows as an example a two-dimensional projection of the folding funnel of the 20-residue trp-cage protein as determined in a computer simulation. Configurations found at the bottom of the funnel resemble closely the experimentally determined structure (the figure is taken from Ref. 18).

In the above described case we have used that the protein is built up out of only α -helices. This allows in a simple way the definition of an "order parameter" for the fold-

ing process. The situation is different for $\alpha\beta$ -proteins such as Fsd-Ey, the LysM-domain and Chymotrypsin Inhibitor 2. These molecules have both α -helices and β -sheets as secondary structure elements and are therefore of higher complexity. While they allow a more general study of small proteins, the problem is that there is no obvious reaction coordinate describing folding. However, such coordinate can be extracted *a posteriori* from generalized-ensemble simulations using the fact that these techniques allow one to sample whole ensembles of low-energy structures and to construct the corresponding energy landscape.

Analyzing the data from simulations of the 28-residue protein Fsd-Ey and the 48-residue LysM domain with clustering techniques, our group attempts to sample the ensemble of local minima of both proteins. For each pair is probed whether there is a path between them that does not require crossing a free energy barrier of pre-set height. In this way, one obtains a connectivity network for the protein energy landscape. While it is interesting in itself to study the topology of these networks, the main emphasis is on identifying the “optimal” path(s) that lead from high energy configurations down to the native state. Using dimension reduction techniques we try to identify the true degrees of freedom in the protein motion along the optimal path in the connectivity network. While protein motion is in general non-linear, we start the investigation with principal component analysis (PCA) albeit this is a globally linear method and leads to a higher dimensional than necessary sub-space. We hope that the combination of our sophisticated sampling techniques with PCA will help identifying the true degrees of freedom and reaction coordinates for describing the folding process. We use these techniques to test whether the energy landscape of Fsd-Ey, the LysM-domain, Chymotrypsin Inhibitor 2, and apo calbindin D9K can be described with the funnel concept, how the tertiary structure formation is related to collapse and secondary structure formation, whether there are nucleation sites, and whether entropic or energetic factors guide the path(s) toward the native structure. The relative stability of secondary structure elements is another question that we want to probe.

The above mentioned tools is also used by us to research the effect of various solvent representations on protein simulations. We use the data from simulations of Fsd-Ey and later the LysM-domain to study the energy landscape of these proteins as a function of the solvent representation. Especially interesting is how the distribution of low-energy states depends on the solvent model and how it differs from the gas phase model. In this way, we will study systematically the accuracy of the model, and explore potential avenues for their betterment. Separating the effects of intramolecular and hydration interactions, such research allows one also to study to which extent folding is determined by intrinsic properties of the protein.

4 Mis-folding and Aggregation

Particularly interesting and important are situations where proteins fold incorrectly as abnormal protein folding and aggregation appears to be involved as a general mechanism in a number of diseases such as Alzheimer’s, Huntington’s or spongiform encephalopathies (prion-mediated)¹⁹. The most common of these diseases is Alzheimer’s. Associated with its neuropathology are amyloid deposits, composed mainly of the β -amyloid peptide (β A). It is found in body fluids in a soluble form that has partial α -helical structure. In Alzheimer’s disease, β A undergoes a conformational change toward a β -sheet structure



Figure 3. Low-energy configurations of the peptide EKYLRT

in which it is insoluble and assembles in fibrils 60-90 Å in diameter. Fibrillar amyloids form lesions 10-200 μm in diameter known as senile plaques. These plaques are surrounded by degenerating and swollen nerve terminals, and found in extra-cellular space of the brain. The neurotoxicity of the βA-peptide is related to the degree of β-aggregation. A similar situation is observed in a family of inherited neurodegenerative diseases that includes Huntington disease²⁰. These polyglutamine (polyQ) disorders are characterized by long (> 35) glutamine repeats in the affected proteins forming protein aggregates that show a fibrillar morphology similar to that observed in Alzheimer²¹. Hence, the analysis of the structural changes in polyQ molecules or the βA-peptide, and their subsequent aggregation, could contribute to developing understanding of the biogenesis of the corresponding neurological disorders¹⁹. A possible mechanism for the growth of the toxic fibrils may be that the incorrectly folded protein induces mis-folding in close-by molecules. For instance, the peptide EKAYLRT likes to form a β-strand when in the vicinity of an other β-strand (Fig3b), while further away (or isolated) it tends to form an α-helix (Fig. 3a). The figure is taken from Ref. 22.

We start our research with investigating the mechanism of β-sheet versus α-helix formation in polyQ peptides. Chains of increasing length are simulated in order to compare our results with the observed pathogenic threshold of ~ 35 – 40 glutamines. We expect to find as local minima the soluble α-helix form and the insoluble β-sheet structure, but other structures may also exist at room temperature as local minima in the free energy landscape. The relative weight of the different structures as a function of chain length are determined and the separating free energy barriers measured. This will allow us estimating the life times of these conformers and to identify possible pathways between these local free-energy minima. Principal component analysis will be used to identify the true degrees of freedom describing the motion along these pathways.

The autocatalytic properties of βA or polyQ fibrils let us expect that surface effects play an important role in the formation of β-sheets and the aggregation of the β-sheet

form. Hence, we simulate the molecule in the presence of hydrophobic (which will model previously aggregated molecules) or hydrophilic surfaces. We are especially interested in observing how the free-energy landscape of the peptide is modified through the presence of the surface, and how this change in the energy landscape depends on the characteristics (especially its hydrophobicity) of the surface. We expect that such a detailed investigation of the free energy landscape of βA and its change with environment will lead to a better understanding of the mechanism of β -formation in this peptide. Simulations and analysis will then be repeated for the 42-residue β -A peptide, and the mechanism of mis-folding and aggregation compared for both molecules.

5 Closing Remarks

I have outlined a group of research projects in the newly found research group "Computational Biology and Biophysics" that uses high performance computing to study proteins and their interactions. Their center piece is the continuing development of novel algorithms (the "generalized-ensemble" approach) toward the final goal of structure prediction of stable domains in proteins (usually of order 50-200 residues). One challenge in the next years will be to extend these lines of research to larger and medically relevant proteins. Other research will focus on the interaction of proteins with different biological molecules (flexible docking) in order to understand how biomolecules interact and regulate each other in a cell. Application of the current research may also include the use of proteins for assembling nanostructures.

Acknowledgments

The presented work is in part also supported through research grants of the National Science Foundation (CHE-0313618) and the National Institutes of Health (GM62838), both USA.

References

1. U.H.E. Hansmann, *Comp. Sci. Eng.* **5** (2003) 64.
2. *Protein folding in silico - The Quest for Better Algorithms*, in: A. Hartmann and H. Rieger (eds), *New Optimization Algorithms in Physics*, VCH-Wiley (2004).
3. F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *Comp. Phys. Comm.* **138** (2001) 192.
4. B. Berg and T. Neuhaus, *Phys. Lett.* **B267** (1991) 249; B. Berg and T. Neuhaus, *Phys. Rev. Lett.* **68** (1992) 9.
5. A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.* **61** 2635 (1988); **63** (1989) 1658(E), and references given in the erratum.
6. U.H.E. Hansmann and Y. Okamoto, *Physica A* **212** (1994) 415.
7. U.H.E. Hansmann and Y. Okamoto, *J. Comp. Chem* **14** (1993) 1333.
8. C.J. McKnight, D.S. Doehring, P.T. Matsudaria and P.S. Kim, *J. Mol. Biol.* **260** (1996) 126.

9. I. Chang, M. Cieplak, R.I. Dima, A. Maritan and J.R. Banavar, *Proc. Nat. Acad. Sci. USA* **98** (2001) 14350.
10. B.D. Silverman, *Proc. Nat. Acad. Sci. USA* **98** (2001) 4996.
11. U.H.E. Hansmann and L.T. Wille, *Phys. Rev. Lett.*, **88** (2002) 068105.
12. K. Hukushima and K. Nemoto, *J. Phys. Soc. (Japan)*, **65** (1996) 1604; G.J. Geyer, *Stat. Sci.* **7** (1992) 437.
13. U.H.E. Hansmann, *Chem. Phys. Lett.* **281** (1997) 140.
14. C.-Y. Lin, C.-K. Hu and U.H.E. Hansmann, *Proteins: Structure, Function and Genetics*, **52** (2003) 436.
15. K.A. Dill and H.S. Chan, *Nature Structural Biology* **4** (1997) 10.
16. J.D. Bryngelson and P.G. Wolynes, *Proc. Natl. Acad. Sci. (USA)* **84** (1987) 7524.
17. J.N. Onuchic, Z. Luhey-Schulten and P.G. Wolynes, *Annu. Rev. Phys. Chem.* **48** (1997) 545.
18. A. Schug, W. Wenzel and U.H.E. Hansmann, *J. Chem. Phys.*, **122** (2005) 194711.
19. J-C. Rochet and P.T. Lansbury, *Curr Op Struc Biol* **10** (2000) 60.
20. L. Masino and A. Pastore, *Brain Res. Bull.* **56** (2001) 183.
21. P.A. Temussi, L. Masino, and A. Pastore, *EMBO J.* **22** (2003) 355.
22. Y. Peng and U.H.E. Hansmann, *Phys. Rev. E*, **68** (2003) 041911.