John von Neumann Institute for Computing

Protein Simulations on Massively Parallel Computers

J. H. Meinke, S. Mohanty, W. Nadler, Th. Neuhaus,
O. Zimmermann, U. H. E. Hansmann

published in

NIC Symposium 2008,
G. Münster, D. Wolf, M. Kremer (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **39**, ISBN 978-3-9810843-5-1, pp. 9-16, 2008.

http://www.fz-juelich.de/nic-series/volume39

Protein Simulations on Massively Parallel Computers

Jan H. Meinke[1], Sandipan Mohanty[1], Walter Nadler[1,2], Thomas Neuhaus[1], Olav Zimmermann[1], and Ulrich H.E. Hansmann[1,2]

[1] John v. Neumann Institute for Computing (NIC), Research Centre Jülich, 52425 Jülich, Germany
E-mail: u.hansmann@fz-juelich.de

[2] Dept.of Physics, Michigan Technological University, Houghton, MI 49931, USA

We summarize shortly the research done in the "Computational Biology and Biophysics" group at NIC. This research group was founded in July 2005 and spearheads the use of modern supercomputers for research into the Biophysics and Biochemistry of biological macromolecules.

1 Introduction

The need for high performance computing in Biochemistry and Systems Biology is well-recognized. For most sequences in the now deciphered genomes, the structures and functions of the corresponding proteins are not known. Even less is known about their interaction and regulation. Reliable tools that allow one to study these phenomena in computer experiments would open the way for understanding the molecular foundations of the workings of whole cells. This is because proteins are the workhorses in a cell, transporting molecules, catalyzing biochemical reactions, or fighting infections. Hence, improved computational tools could lead to a deeper understanding of various diseases that are caused by the mis-folding of proteins, and enable the design of novel drugs with customized properties.

Computer simulations of even small proteins have remained a challenging computational task. This is because the complex form of the forces within and between molecules leads to a rough energy landscape with a huge number of local minima acting as traps. As a rule, computational cost to accurately calculate physical quantities in simple room-temperature Molecular Dynamics or Monte Carlo simulations increases exponentially with the number of residues. Overcoming these obstacles requires both new sampling techniques and the use of modern massive parallel computers (such as the new BlueGene/P computer JUGENE in Jülich) that soon will approach the Petaflop range.

Research in the group is concerned with the development of algorithms, and their implementation into software optimized for modern supercomputer architectures, for overcoming these numerical difficulties in protein studies. Our aim is to develop protocols that allow atomistic simulations of stable domains in proteins (usually of order 50-200 residues). Successful method and software development is at its best when it is guided by the needs of specific applications. For this reason, we have studied extensively the folding mechanism in a number of small proteins. Finally, we are interested in the combination of our techniques with knowledge-based approaches. Such combinations are the backbone of the structure prediction algorithms that we have tested successfully at the CASP7 (Critical Assessment of protein Structure Prediction) competition in summer 2006. Our newly developed techniques are included in the recently published update of SMMP[1,2], the freely available program package co-developed by this group.

2 Algorithms for Protein Simulations

The key-idea behind our novel techniques is to replace canonical simulations, where crossing of an energy barrier of height ΔE is suppressed by a factor $\propto \exp(-\Delta E/k_B T)$ (k_B is the Boltzmann constant and T the temperature of the system), with schemes that both ensure sampling of low-energy configurations *and* avoid trapping in local minima. For instance, in multicanonical sampling[3] the weight $w(E)$ in a Monte Carlo or molecular dynamics simulation is set so that the distribution of energies $P(E)$ is given by:

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

where $n(E)$ is the spectral density. In this way, a free random walk in the energy space is performed that allows the simulation to escape from any local minimum. The thermodynamic average of a physical quantity A can now be calculated by re-weighting:[4]

$$< \mathcal{A} >_T \;\; = \;\; \frac{\int dx\; \mathcal{A}(x)\; w^{-1}(E(x))\; e^{-E(x)/k_B T}}{\int dx\; w^{-1}(E(x))\; e^{-E(x)/k_B T}} \; . \tag{2}$$

Here, x stands for configurations. Note that the weights $w(E)$ are not *a priori* known, and estimators have to be determined by an iterative procedure described in Refs.[3,5].

In parallel tempering (also known as replica exchange method)[6], first introduced to protein science in Ref.[7], standard Monte Carlo or molecular dynamics moves are performed in parallel at different values of a control parameter, most often the temperature. At certain times the current conformations of replicas at neighbouring temperatures T_i and $T_{j=i+1}$ are exchanged with probability

$$w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))) \; , \tag{3}$$

with $\beta = 1/k_b T$. For a given replica the swap moves induce a random walk from low temperatures, where barriers lead to long relaxation times, to high temperatures, where equilibration is rapid, *and back*. This results in a faster convergence at low temperatures.

However, these sophisticated novel techniques are hampered still by slow relaxation due to barriers and bottlenecks. Their successful application to protein simulations requires that the weight functions or temperature distributions are chosen optimally. In order to unify the analysis of these techniques, we derived equivalent one-dimensional stochastic processes from the underlying Master equations and analyzed the conditions under which these representations are valid descriptions of the random walk in order parameter or control parameter space[8] . They allow a unified discussion of the stationary distribution on each space, as well as of the stationary flow across it. We demonstrated that optimizing the flow is equivalent to minimizing the first passage time for crossing the space, and discussed the consequences of our results for optimizing simulations, particularly under conditions of broken ergodicity. Based on these results we were able to determine an analytical expression for the optimal number of replicas in PT simulation[9]. In addition, we have investigated the theoretical basis for combining molecular dynamics simulation with parallel tempering and developed optimized replica exchange move sets for that[10].

Another example of our algorithmic research is concerned with temperature driven first order phase transition where traditional parallel tempering implementations fail. We have developed an efficient new parallel tempering algorithm[11], that eliminates the supercritical slowing down associated with the nucleation barrier at the transition and which is capable

of determining the density of states function. Our algorithm is much simpler than multi-canonical ensemble simulations, which on the input side need an a-priori unknown weight function - or - Wang Landau simulations, which require a tedious parameter fine tuning. In addition, multiple Gaussian modified ensemble simulations[11] are perfectly suited for parallel computer architectures. Future applications of the method will include studies of condensates with ensembles of chain molecules.

3 Protein Simulation Programs for Parallel Computers

The above algorithms are implemented in SMMP[1], our program package for simulation of protein. The code is free and open source. The latest version[2] features a Python interface and allows simulations of more than one protein as necessary for studies of aggregation or ligand binding. The implementation of additional force fields are other newly added features. SMMP is available from either the program library of *Computer Physics Communications* or from *www.phy.mtu.edu/biophys/smmp.htm*

Emphasis was put on the parallelization of SMMP: we now regularly run our simulation on 4096 processors. 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. We ran our benchmark on 4 different platforms: JUMP, an IBM p690 cluster with 32 Power4+ processors at 1.7 GHz and 112 GB of shared memory per frame and a total of 1312 processors; 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 (92 residues, 1477 atoms) 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, then y, and finally z. This leads to a planar distribution of processors. Instead of the default, one should make communicators as cubic as possible unless the problem geometry suggests a different approach. 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.

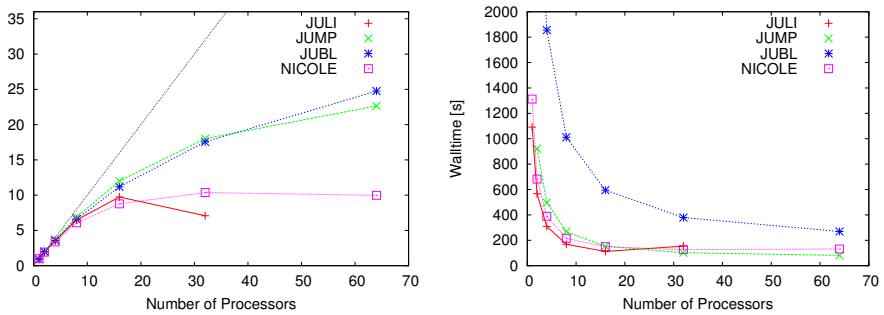

Figure 1. Strong scaling behaviour. This figure shows the walltime and corresponding parallel scaling vs. number of processors on JULI, JUMP, and JUBL. On JUBL, we used multiple replicas in parallel with the indicated number of processors per replica to fill a half plane.

4 Side Chain Ordering in Polymers and Proteins

Work of the group is characterized by the combination of algorithm development with application of these novel techniques to research the physics of proteins and their interaction. One area of interest is the distinct transitions in the folding process, their thermal order and relation. An important example is the role of side-chain ordering.

In recent studies on homopolymers[12, 13], we found for certain amino acids a de-coupling of backbone and side-chain ordering. The characteristics of the side chain ordering process did not depend on the details of the environment, i.e. whether the molecules were in gas phase or solvent, but solely on the side groups. It exhibited a phase transition-like character, marked by an accompanying peak in the specific heat. In a related investigation, we were able to establish the role of charged end groups in stabilizing and de-stabilizing secondary structures in gas phase[14]. These results are important for comparison with molecular beam experiments. Later, we have extended those investigations to proteins, starting with the villin headpiece subdomain HP-36[15]. This molecule has raised considerable interest in computational biology as it is one of the smallest proteins (596 atoms) with well-defined secondary and tertiary structure but at the same time still accessible to simulations. Our results indicate a thermal hierarchy of ordering events with side-chain ordering appearing at temperatures below the helix-coil transition, i.e. secondary structure formation, but above the final folding transition to the native state. We believe that the observed thermal hierarchy of folding reflects an underlying temporal sequence of these ordering processes in actual protein folding dynamics. We conjecture that side-chain ordering facilitates the search for the correct backbone topology. In contrast to homopolypeptides we do not find collective effects leading to a separate transition. The heterogeneity of the sequence seems to destroy the phase transition-like character of side-chain ordering.

5 Folding Mechanisms in Small Proteins

The above investigations were concerned with helical polypeptides. In an α-helix, hydrogen bonds are formed between residues i and $i + 4$ along the sequence. Because of

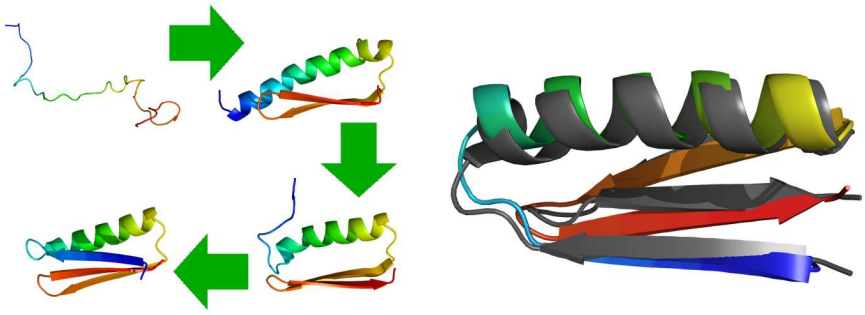

Figure 2. Left: The non-trivial unexpected pathway for the folding of the molecule CFr revealed in the simulations. Right: The free-energy mimimum structure seen in all-atom simulations (coloured) superimposed on the experimentally measured structure (gray).

this short sequence separation, these residues are constrained so that the helical hydrogen bonds form along with chain compactification. Small helical proteins indeed show simple funnel like folding free-energy landscapes[16]. The helix hydrogen bonds form in no particular order, although the two ends of a helix show greater tendency to dissolve and reform. β-hairpins are also "local" structures in that the hydrogen bonded residues are close in sequence. For the 3-stranded β-sheet beta3s, we find the folding of the two β-hairpins more cooperative than the folding of helices. Folding proceeded in a zipper mechanism from the turns towards the ends of the hairpins[16]. Once formed, the β-sheets show a greater resilience towards unfolding.

The simplest example of a protein with both helix and β-sheet elements is a structure with a helix and a β-hairpin. We have examined two such systems and found rather distinct mechanisms. The 23 residue BBA5 molecule has a small β-hairpin where the turn region is stabilized by a synthetic amino acid D-proline. This hairpin and the helix of BBA5 form on their own, and only later make hydrophobic contacts[16]. On the other hand, in a simulation of the protein FSDEY with a similar helix hairpin structure, we found that the hydrophobic residues of the helix line up on one side, to provide a template around which the hairpin forms[17]. The hairpin of FSDEY never forms independently of the helix in our simulations, nor did we find any clear evidence of a zipper like mechanism.

The formation of structures with non-local β-sheet contacts are highly non-trivial. We have done extensive folding simulations of one such molecule, CFr (PDB id: 2GJH), the C-terminal fragment of the designed 93 residue protein Top7 (PDB id: 1QYS). Along the sequence from N- to C- terminal, the secondary structure profile of the molecule CFr is : strand – helix – strand – strand (see Fig.2). The two strands at the C-terminal make a β-hairpin. The strands at the N- and C-termini are also adjacent in the 3 stranded β-sheet. None of the simple folding mechanisms discussed above could give rise to this arrangement. Instead, our simulations revealed an unanticipated mechanism (see Fig. 2) for the formation of this structure[18]. The N-terminal β-strand first folded into a non-native extension of the native helix. The β-hairpin at the C-terminus forms independently. When the helix and the C-terminal hairpin make the correct tertiary contacts, the non-native part of the helix unfolds to release the N-terminal residues. These subsequently make β-sheet

contacts with the hairpin and complete the native structure. By "caching" the N-terminal β-strand as a non-native extension of a helix, the molecule protects them from premature contacts with other regions with strong β-strand propensities, which would lead to misfolding, or very slow folding. The caching of the N-terminal strand, accelerates folding of CFr by avoiding many misfolded states. We speculate that this mechanism is employed in molecules where adjacent strands in a β-sheet have large sequence separation. The same mechanism could protect a nascent N-terminal β-strand which is synthesized early, from intermolecular interactions leading to aggregation, until the rest of the molecule is synthesized and properly folded. Simulations of FSDEY were done with SMMP[2], while we used PROFASI[19] for the other proteins mentioned in this section.

6 Thick Polymers

The results described above raise the question to what extent the properties of proteins can also be explained as polymer properties (i.e. are not sequence specific). To investigate this, we simulated "thick" polymers in a simple $3d$ homo-polymer chain model with thickness regularized by the global radius of curvature, and studied the interplay of "hard" geometric constraints and attractive Lennard Jones interactions. In an earlier work the long chain limit of thick polymers was studied[20] and was found to be consistent with field theoretic expectations. As a function of the global radius of curvature we now find a rich "landscape" of low temperature i.e., ground state configurations, which we classify. These include simple hypercubic crystals, ideal helices as introduced by Maritan[21], short sheet like structures, twisted circles and helical superstructures. We also observe a phase region where open ended polymers close into ring polymers with zero knot number and with short range interactions as being characteristic for a liquid. While these studies are of principal nature, some similarities with real proteins are seen. In particular we observe secondary structure formation as well as mixed phases of few conformational simplices for long chains. However, a direct mapping onto protein configurations is problematic and there is clear need for the inclusion of additional either protein specific - or more fundamental interactions into the simplified model. A possible candidate is a thick polymer model with attractive Lennard Jones interactions as well as dipole-dipole interactions on the chain.

7 Constraints for Structure Prediction

Our algorithms are limited to simulations of small proteins of order ≈ 50 residues. As the average size of proteins is around 250 residues, it is necessary to constrain the conformational search space for the purpose of structure prediction. Such constraints are generally obtained from known structures of similar sequences. As long range distance constraints (e.g. from fold recognition or correlated mutation analysis) lead to frustrated conformations, local constraints e.g. dihedral constraints from secondary structure predictions are preferred. State of the art algorithms like PSIPRED[22] achieve highly accurate secondary structure predictions, but do not allow mapping of these classes to dihedral angles. To overcome these problems we have developed DHPRED to predict for each residue its dihedral angle region[23]. The algorithm has a three layer structure and is based on classification by Support-Vector-Machines (SVM)[24], a supervised machine learning algorithm. The performance of DHPRED is comparable to PSIPRED but provides a direct mapping to dihedral

angles. In addition the dihedral angle regions for many coil residues can be identified thus providing local dihedral constraints for the entire protein chain.

Patterns of dihedral angles i.e. correlations between the dihedral state of neighbouring residues can also be observed in coil regions. Among these patterns β-turns[25] are the most abundant class. They are frequently encountered at topologically interesting positions where the chain changes its direction and as such are important prediction targets. We have developed a classifier to resolve the different classes of β-turns as they indicate different local topologies[26]. While certain classes can be easily distinguished, several other classes (I, IV, VIII) show considerable overlap. As these classes also share certain geometrical properties it is an open question whether this is due to an insufficient number of samples available or to the fact that the class boundaries are artificial.

The rapid growth of the protein databank PDB makes it increasingly likely that a structurally related protein to a sequence has already been entered even if the sequence similarity is too low to allow for direct identification. For these cases all successful structure prediction strategies use fold recognition, where structural features are derived from the target sequence and compared to a structure database. In the simplest case secondary structure alignment is used. As ordinary secondary structure has only three different elements there is a high likelihood that two proteins with the same secondary structure have vastly different topology. A considerable reduction of the possible topologies can be achieved by predicting which β-strands are part of an antiparallel β-sheet and which form a parallel one. Our program BETTY[27] implements an SVM-approach similar to the one described for DHPRED but without the iterative second layer. It classifies 88% of all β-residues correctly. Combined with PSIPRED 79.3% can be correctly classified into parallel-β, antiparallel-β, α-helix, and coil. We are planning to use this enhanced secondary structure alphabet as part of a fold recognition algorithm.

In order to test our techniques for structure prediction we have participated in the CASP 2006 competition where we have submitted structures for all but three of the 100 valid targets. As expected our approach is not competitive in cases where homology to resolved proteins allows application of knowledge-based techniques. However, we are within the top five groups in the free modelling section when ranked based on visual inspection (see `http://predictioncenter.org/casp7/meeting/presentations/Presentations_assessors/CASP7_FM_Clarke.pdf`). We are currently evaluating our protocols in order to determine how we can optimize the use of our techniques for structure prediction, and plan to set up a server that automates the predictions for the 2008 round of the CASP competition.

8 Closing Remarks

We have outlined research done by the "Computational Biology and Biophysics" group since its inception in July 2005. Using high performance computing and developing novel algorithms (the "generalized-ensemble" approach) the group has made substantial progress on the way toward the goal of structure prediction of stable domains in proteins (usually of order 50-200 residues). In future, the group will extend these lines of research to larger and medically relevant proteins.

Acknowledgments

Part of the presented work was supported also through research grants of the National Science Foundation (CHE-0313618) and the National Institutes of Health (GM62838).

References

1. F. Eisenmenger, U.H.E. Hansmann, Sh. Hayryan, C.-K. Hu, *Comp. Phys. Comm.* **138** (2001) 192; *Comp. Phys. Comm.* **174** (2006) 422.
2. J. H. Meinke, S. Mohanty, F. Eisenmenger and U. H. E. Hansmann, *Comp. Phys. Comm.*, (2007), doi: 10.1016/j.cpc.2007.11.004.
3. B. Berg and T. Neuhaus, *Phys. Lett.* **B267** (1991) 249; B. Berg and T. Neuhaus, *Phys. Rev. Lett.* **68** (1992) 9.
4. A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.* **61** 2635 (1988); **63** (1989) 1658(E), and references given in the erratum.
5. U.H.E. Hansmann and Y. Okamoto, *Physica A* **212** (1994) 415.
6. K. Hukushima and K. Nemoto, *J. Phys. Soc. (Japan)*, **65** (1996) 1604; G.J. Geyer, *Stat. Sci.* **7** (1992) 437.
7. U.H.E. Hansmann, *Chem. Phys. Lett.* **281** (1997) 140.
8. W. Nadler and U.H.E. Hansmann, Phys. Rev. E, **75** (2007) 026109.
9. W. Nadler and U.H.E. Hansmann, Phys. Rev. E, **76**, (2007) 065701(R).
10. W. Nadler and U.H.E. Hansmann, Phys.Rev. E **76** (2007) 057102.
11. T. Neuhaus, M. Magiera and U. Hansmann, Phys. Rev. E **76**, (2007) 045701(R).
12. Y. Wei, W. Nadler and U.H.E. Hansmann, J. Chem. Phys. **125** (2006) 164902.
13. Y. Wei, W. Nadler and U.H.E. Hansmann, J. Phys. Chem. B **111** (2007) 4244.
14. Y. Wei, W. Nadler and U.H.E. Hansmann, J. Chem. Phys., **126** (2007) 204307.
15. Y.Wei, W. Nadler and U.H.E. Hansmann, Backbone and Sidechain Ordering in a small Protein, J. Chem. Phys., in press.
16. S. Mohanty and U.H.E. Hansmann, *Biophysical Journal* **82** (2006) 3573.
17. S. Mohanty and U.H.E. Hansmann, *J. Chem. Phys.* **127** (2007) 035102.
18. S. Mohanty, J. Meinke, O. Zimmermann and U.H.E. Hansmann, Caching of Chameleon Segments: a New Folding Mechanism, submitted for publication.
19. A. Irbäck and S. Mohanty *J. Comp. Chem.* **27**, (2006) 1548.
20. T. Neuhaus, O. Zimmermann and U. Hansmann, Phys. Rev. E **75**, (2007) 051803.
21. A. Maritan, C. Micheletti, A. Trouato and J. Banavar, *Nature* **406**, (2000) 287.
22. D.T. Jones, J. Mol. Biol. **292** (1999) 195.
23. O. Zimmermann and U.H.E. Hansmann, Bioinformatics **22** (2006) 3009.
24. B. Schökopf and A.J. Smola, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, 2002.
25. P.N. Lewis, F.A. Monany, and H.A. Scheraga, Biochem. Biophys. Acta **303** (1973) 211.
26. O. Zimmermann and U.H.E. Hansmann, Dihedral Angle Patterns in Coil Regions of Protein Structures, in: Proceedings of CBSB07, U.H.E. Hansmann, J.H. Meinke, S. Mohanty, O. Zimmermann (Eds.), John v. Neumann Inst. for Computing, NIC-Series, vol. 36, Jülich, Germany, 2007, pp. 301–304.
27. O. Zimmermann, L. Wang and U.H.E. Hansmann, In Silico Biol. **7** (2007) 0037.