All atom protein folding with massively parallel computers

ABHINAV VERMA,
Forschungszentrum Karlsruhe
Institute for Scientific Computing
PO Box 3640, 76021 Karlsruhe, Germany

JUNG S. OH, KYU H. LEE
Supercomputational Materials Laboratory
Korean Institute of Science
Seoul, Korea

ALEXANDER SCHUG, KONSTANTIN KLENIN, WOLFGANG WENZEL
Forschungszentrum Karlsruhe
Institute for Scientific Computing
PO Box 3640, 76021 Karlsruhe, Germany

http://www.fzk.de/biostruct

Abstract: - Protein folding and structure prediction at the all atom-level remain important grand-computational challenges. We review the development of algorithms and forcefields for predictive all atom folding simulations of proteins with up to sixty amino acids using an evolutionary stochastic optimization technique. We have implemented a master-client model which scales near perfectly from 64 to 4096 nodes. Using a PC cluster we fold the sixty-amino acid bacterial ribosomal protein L20 to near-native experimental conformations. Starting from a completely extended conformation with 2048 nodes of the IBM BlueGene we predictively fold the fourty amino acid HIV accessory protein in less then 24 hours.

Key-Words: - Protein Folding, evolutionary algorithm, free-energy forcefield, HIV accessory protein

1 Introduction

Protein folding and structure prediction have been among the important grand computational challenges for more than a decade. In addition to obvious applications in biology the life sciences and medicine success for protein simulation strategies also impact the materials and an increasingly the nano-sciences. Among these challenges it is important to develop methods that are capable of folding proteins and their complexes from completely unbiased extended conformations to the biologically relevant native structure. This problem is difficult to sort of by the presently most accurate simulation techniques, which follow the evolution of the protein in its environment in a three-time. Since the microscopic simulation step in such molecular-mechanics methods is off the order of femtoseconds, while the folding or asociation process takes place on the order of milliseconds, such simulations remain limited in the system size by the large computational effort required[1]. It has been a great hope for almost a decade that emerging massively parallel computers the architectures, which are available now at the teraflop scale, and which will reach the teraflop skated in the foreseeable future, we be able to contribute to the solution of these problems. Unfortunately kinetic methods face enormous difficulties in the exploitation of the full computational power of these architectures, because they impose a sequence of steps onto the simulation process, which must be completed one after the other. The parallelization of the energy and force evaluation of a single time-slice of the simulation requires a very high communication bandwidth when distributed across thousands of nodes. This approach alone is therefore unlikely to fully utilize many thousand processors of emerging petaflop-architectures, let alone grid-applications with hundreds of thousands of processors.

In a fundamentally different approach we have developed models[2] and algorithms[3] which permit reproducible and predictive folding of small proteins from random initial conformations using free-energy forcefields. According to Anfinsen's thermodynamic hypothesis many proteins are in thermodynamic equilibrium with their environment under physiological conditions. Their unique three-dimensional native conformation then corresponds to

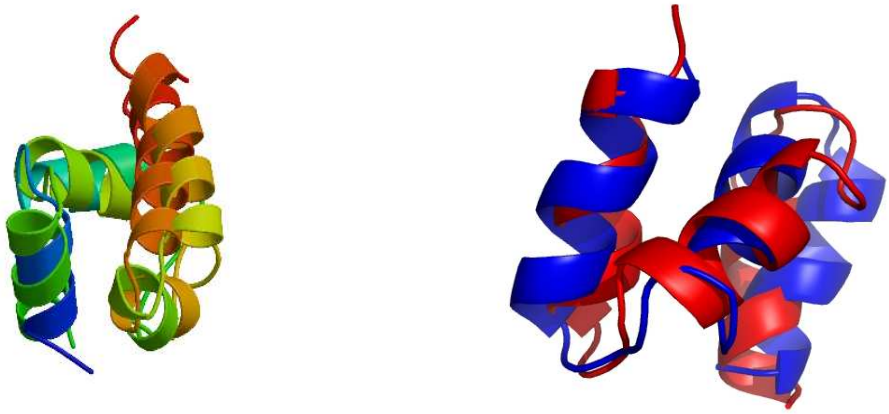

Figure 1: Overlay of the folded and the experimental conformation of the bacterial ribosomal protein L20 (left) and the HIV accesory protein (right)

the global optimum of a suitably free-energy model. The free-energy model captures the internal energy of a given backbone conformation with the associated solvent and side-chain entropy via an implicit solvent model. Comparing just individual backbone conformations these models assess the relative stability of conformations (structure prediction). In combination with thermodynamic simulation methods (Monte-Carlo or parallel tempering)[4], this approach generates continuous folding trajectories to the native ensemble.

Stochastic optimization methods[5], rather than kinetic simulation, can be used to search the protein free energy landscape in a fictitious dynamical process. Such methods explore the protein free-energy landscape orders of magnitude faster than kinetic simulations by accelerating the traversal of transition states the directed construction of downhill moves on the free-energy surface, the exploitation of memory effects or a combination of such methods. Obviously this approach can be generalized to use not just one, but several concurrent dynamical processes to speed the simulation further, but few scalable simulation schemes are presently available.

The development of algorithms that can concurrently employ thousands of such dynamical processes to work in concert to speed the folding simulation remains a challenge, but holds the prospect to make predictive all-atom folding simulations in a matter of days a reality.

The development of such methods is no trivial task for a simple reason: if the total computational effort (number of function evaluations N) is conserved, while the number of nodes (n_p) is increased, each process explores a smaller and smaller region of the conformational space. If the search problem is exponentially complex, as protein folding is believed to be [24], such local search methods revert to an enumerative search, which must fail. It is only the 'dynamical memory' generated in thermodynamic methods such as simulated annealing [20], that permit the approximate solution of the search problem in polynomial time. Thus, massively parallel search strategies can only succeed if the processes exchange information.

Here we review applications of a recently developed an evolutionary algorithm, which generalized the basin hopping or Monte-Carlo with minimization[6], method to many concurrent simulations. Using this approach we could fold the sixty amino acid bacterial ribosomal protein to its native ensemble[7, 8].

2 Methods

2.1. Forcefield

We have parameterized an all-atom free-energy forcefield for proteins (PFF01), which is based on the fundamental biophysical interactions that govern the folding process. We have also developed, or specifically adapted, efficient stochastic optimization methods[9] (stochastic tunneling, basin hopping, parallel tempering, evolutionary algorithms) to simulate the protein folding process. Forcefield and simulation methods are implemented in the POEM (Protein Optimization with free-Energy Methods) program package. We could demonstrate that the free-energy approach is several orders of magnitude faster than the direct simulation of the folding pathway, but nevertheless permits the full characterization of the free-energy surface that characterizes the folding process according to the prevailing funnel-paradigm for protein folding.

2.2. Optimization Method

Most stochastic optimization methods map is such of the complex potential energy landscape of the

problem onto a fictitious dynamical process that is guided by its inherent dynamics toward the low energy region, and ultimately the global optimum, of the landscape. In many
prior simulations the basin hopping technique proved to be a reliable workhorse for many complex optimization problems[10, 11], including protein folding[12], but employs only one dynamical process. This method simplifies the original landscape by replacing the energy of each conformation with the energy of a nearby local minimum. This replacement eliminates high energy barriers in the stochastic search that are responsible for the freezing problem in simulated annealing. In order to navigate the complex protein landscape we use a simulated annealing (SA) process for the minimization step[13]. Within each SA simulation, new configurations are accepted according to the Metropolis criterion, while the temperature is decreased geometrically from its starting to the final value. The starting temperature and cycle length determine how far the annealing step can deviate from its starting conformation. The final temperature must be small compared to typical energy differences between competing metastable conformations, to ensure convergence to a local minimum.
We have generalized this method to a population of P interdependent dynamical processes operating on a population of N conformations. The whole population is guided towards the optimum of the free energy surface with a simple evolutionary strategy in which members of the population are drawn and then subjected to a basin hopping cycle. At the end of each cycle the resulting conformation either replaces a member of the active population or is discarded. This algorithm was implemented on a distributed master-client model in which idle clients request a task from the master. Conformations are drawn with equal probability from the active population. The acceptance criterion for newly generated conformations must balance the diversity of the population against the enrichment low-energy decoys. We accept only new conformations which are different by at least 4 Å RMSB (root mean square backbone deviation) from all active members. If we find one or more members of the population within this distance, the new conformation replaces the all existing conformations if its energy is lower than the best, otherwise it is discarded. If the new conformation differs by at least the threshold from all other conformation it replaces the worst conformation of the population if it is better in total (free) energy. If a merge operation has reduced the size of the population, the energy criterion for acceptance is waived until the original number of conformations is restored.

3 Results

The simulation for the bacterial ribosomal protein L20 was performed in three stages: In the first stage we generate a large set of unfolded conformations, which was pruned to 266 conformations by energy and diversity. In stage two we 50 annealing cycles per replica, after which the population was pruned to the best N=50 decoys (by energy). We then continued the simulation for another 5500 annealing cycles. At the end of the simulations, the respective lowest energy conformations had converged to 4.3 Å RMSB with respect to the native conformation. Six of the ten lowest structures had independently converged to near-native conformations of the protein. The first non-native decoy appears in position two, with an energy deviation of only 1.8 kcal/mol (in our model) and a significant RMSB deviation.
The good agreement between the folded and the experimental structure is evident from Figure (1) (left panel), which shows the overlay of the native and the folded conformations. The good alignment of the helices illustrates the importance of hydrophobic contacts to correctly fold this protein. Figure (2) demonstrates the convergence of both the energy and the average RMSB deviation as the function of the number of total iterations (basin hopping cycles). Both simulations had an acceptance ratio approximately 30 %.

We have also folded the 40 amino acid HIV accessory protein(sequence: QEKEAIERLK ALGFEESLVI QAYFACEKNE NLAANFLLSQ, pdb-id: 1F4I)[43]. For timing purposes we have performed simulations using 64, 128, 256, 512, 1024, 2048 and 4096 processors on an IBM BlueGene in virtual processor mode. We find that the simulation scales perfectly with the number of processors, inducing less than 5% loss of efficiency when comparing P=64 with P=4096 processor simulations. The control loop is implemented employing a synchronous simulation protocol, where tasks are distributed to all processors of the machine. As the simulations finish, their conformations are transferred to the master, which decides whether to accept (average probability: 57%) the conformation into the active population or disregard the conformation. Then a new conformation is immediately given to the idle processor. Because the processors are processed sequentially some processors wait for the master before they get a new conformation. Fluctuations in the client execution times induce a waiting time before the next iteration can start. For the realistic simulation times chosen in these runs, the average waiting time is less than 10% of the execution time and nearly independent of the number of processors

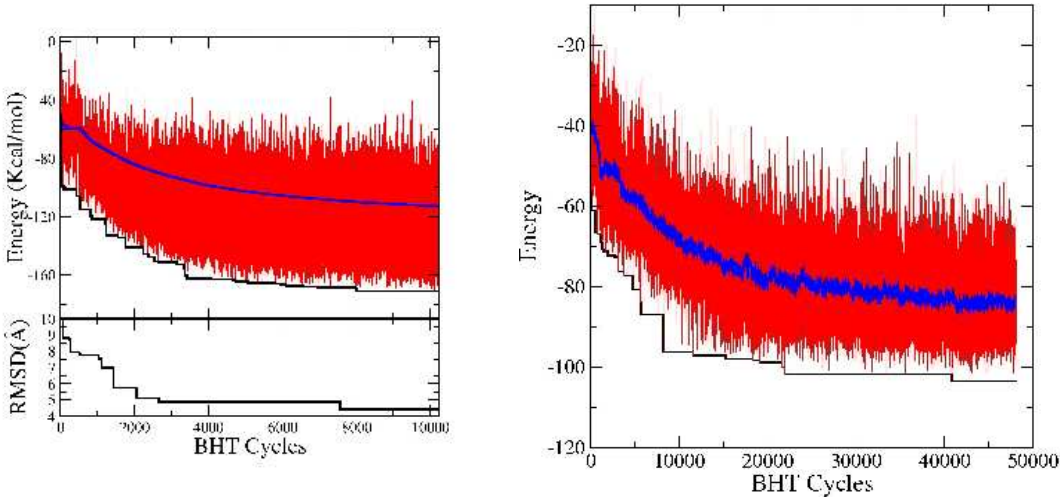

Figure 2: Instantaneous energy (red), mean energy (blue) and best energy (black) for the simulations of the bacterial ribosomal protein L20 (left) and the HIV accessory protein (right). For the bacterial ribosomal protein L20 the lower panel indicates the convergence of the RMS deviation of the lowest energy conformation.

used.

We next performed a simulation using 2048 processors starting from a single completely stretched "stick" conformation. The seed conformation had an average RMSB deviation of 21.5Å to the experimental conformation. We then performed 20 cycles of the evolutionary algorithm described above. Figure 1 shows the overlay of the folded and the experimental conformation. The starting conformation has no secondary structure and no resemblance of the native conformation. In the final conformation, the secondary structure elements agree and align well with the experimental conformation. Figure 2 shows that the best energy converges quickly to a near-optimal value with the total number of basin hopping cycles. The average energy trails the best energy with a finite energy difference. This difference will remain indefinitely by construction, because the algorithm is designed to balance diversity and energy convergence. The acceptance threshold of 4 Å RMS for the new population enforces that only one near-native conformation is accepted in the population, the average energy will therefore always be higher than the best energy.

4 Discussion

Using a scalable evolutionary algorithm we have demonstrated the all-atom folding two proteins: Using 50 processors of loosely connected PC cluster we succeeded to fold the 60 amino acid bacterial ribosomal protein to near-native conformations. time using 2048 processors of an IBM BlueGene we also folded the 40 amino acid HIV accessory protein from a completely extended conformation to within 4 A of the native conformation in about 24 hours turnaround. The results of this study provide impressive evidence that all-atom protein structure prediction with free-energy forcefields is becoming a reality. The key to convergence of the method lies in the exploitation of the specific characteristics of the free energy landscape of naturally occurring proteins. Following the current funnel paradigm[44,45] the protein explores an overall downhill process on the energy landscape, where the conformational entropy of the unfolded ensemble is traded for enthalpic gain of the protein and free energy gain of the solvent[46,7]. Using one- or low-dimensional indicators the complex folding process appears for many small proteins as a two-state transition between the unfolded and the folded ensemble with no apparent intermediates. This transition has been rationalized in terms of the funnel paradigm, where the protein averages over average frictional forces[14] on its downhill path on the free-energy landscape. In this context one cycle of the evolutionary algorithm attempts to improve many times each of the conformations of the active population. Because of the high dimensionality of the search problem ($D = 160$ free dihedral angles for 1F4I) most of these attempts fail, but those which succeed are efficiently distributed for further improvement by the evolutionary method.

The search for methods and models for de novo folding of small and medium size proteins from the

completely extended conformation at atomic resolution has been a "holy grail" and grand computational challenge for decades. The development of multi-teraflop architectures, such as the IBM BlueGene used in this study, has been motivated in part by the large computational requirements of such studies. The demonstration of predictive folding of a 40 amino acid protein with less than 24 hours turnaround time, is thus an important step towards the long time goal to elucidate protein structure formation and function with atomistic resolution. The free-energy approach employed here can complement Hamiltonian based simulation methods, such as molecular dynamics or replica exchange methods, to understand how proteins fold and interact. The mapping of the "folding problem" onto an optimization problem permits the use of methods that speed the exploration of the free-energy surface. The results reviewed above demonstrate that it is possible to parallelize the search process by splitting the simulation into a large number of independent conformations, rather than by parallelizing the energy evaluation

The present study thus demonstrates a computing paradigm for protein folding that may be able to exploit the petaflop computational architectures that are presently being developed. The availability of such computational resources in combination with free-energy folding methods can make it possible to investigate and understand a wide range of biological problems related to protein folding, misfolding and protein-protein interactions.

Acknowledgments

This work is supported by grants from the German national science foundation (DFG WE1863/10-2), the Secretrary of State for Science and Research through the Helmholtz-Society and the Kurt Eberhard Bode foundation. We acknowledge the use of facilities at the IBM Capacity on Demand Center in Rochester and KIST Supercomputional Materials Lab in Seoul. We are grateful for technical assistance from G. S. Costigan and C. S. Sosa from the IBM Capacity on Demand Center for technical assistance.

1. Duan, Y. and P.A. Kollman, *Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution.* Science, 1998. **282**: p. 740-744.
2. Herges, T. and W. Wenzel, *An All-Atom Force Field for Tertiary Structure Prediction of Helical Proteins.* Biophys. J., 2004. **87**(5): p. 3100-3109.
3. Schug, A., et al. *Stochastic Optimization Methods for Protein Folding*. in *Recent Advances in the Theory of Chemical and Physical Systems*. 2006: Springer.
4. Sugita, Y. and Y. Okamoto, *Ab initio replica-exchange Monte Carlo method for cluster studies.* Chem. Phys. Lett, 1999. **314**: p. 141-151.
5. Wenzel, W. and K. Hamacher, *Stochastic Tunneling Approach for Global Optimization of Complex Potential Energy Landscapes.* Phys. Rev. Lett., 1999. **82**: p. 3003-3007.
6. Nayeem, A., J. Vila, and H.A. Scheraga, *A Comparative Study of the Simulated-Annealing and Monte Carlo-with-Minimization Approaches to the Minimum-Energy Structures of Polypeptides: [Met]-Enkephalin.* J. Comp. Chem., 1991. **12**(5): p. 594-605.
7. Schug, A. and W. Wenzel, *An evolutionary Strategy for All-Atom folding of the sixty amino acid bacterial ribosomal proein L20.* {Biophys. Journal, 2006. **90**: p. 4273-4280.
8. Schug, A. and W. Wenzel, *Predictive in-silico all-atom folding of a four helix protein with a free-energy model.* J. Am. Chem. Soc., 2004. **126**: p. 16736-16737.
9. Schug, A., et al. *Stochastic Optimization Methods for Protein Folding*. 2005.
10. Carr, J.M. and D.J. Wales, *Global optimization and folding pathways of selected -helical proteins.* J. Chem. Phys., 2005. **123**: p. 234901.
11. Mortenson, P.N. and D.J. Wales, *Energy landscapes, global optimization and dynamics of poly-alanine Ac(ala)$_8$NHMe.* J. Chem. Phys., 2004. **114**: p. 6443-6454.
12. Verma, A., et al., *Basin Hopping Simulations for All-Atom Protein Folding.* J. Chem. Phys., 2006. **124**: p. 44515.
13. Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi, *Optimization by Simulated Annealing.* Science, 1983. **220**: p. 671-680.
14. Dill, K.A. and H.S. Chan, *From Levinthal to Pathways to Funnels: The "New View" of Protein Folding Kinetics.* Nature Structural Biology, 1997. **4**: p. 10-19.