Massively Parallel All Atom Protein Folding in a Single Day

Abhinav Verma, Srinivasa M. Gopal, Alexander Schug,
Jung S. Oh, Konstantin V. Klenin, Kyu H. Lee,
Wolfgang Wenzel

published in

Parallel Computing: Architectures, Algorithms and Applications ,
C. Bischof, M. Bücker, P. Gibbon, G.R. Joubert, T. Lippert, B. Mohr,
F. Peters (Eds.),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **38**, ISBN 978-3-9810843-4-4, pp. 527-534, 2007.
Reprinted in: Advances in Parallel Computing, Volume **15**,
ISSN 0927-5452, ISBN 978-1-58603-796-3 (IOS Press), 2008.

Massively Parallel All Atom Protein Folding in a Single Day

Abhinav Verma[1], Srinivasa M. Gopal[2], Alexander Schug[2], Jung S. Oh[3], Konstantin V. Klenin[2], Kyu H. Lee[3], and Wolfgang Wenzel[2]

[1] Institute for Scientific Computing
Research Centre Karlsruhe, D-76344, Karlsruhe, Germany
E-mail: verma@int.fzk.de

[2] Institute for Nanotechnology
Research Centre Karlsruhe, D-76344, Karlsruhe, Germany
E-mail: {gopal, klenin, schug, wenzel}@int.fzk.de

[3] Supercomputational Materials Lab
Korean Institute for Science and Technology, Seoul, Korea
E-mail: {soo5, khlee}@kist.re.kr

The search for efficient methods for all-atom protein folding remains an important grand-computational challenge. We review predictive all atom folding simulations of proteins with up to sixty amino acids using an evolutionary stochastic optimization technique. Implementing a master-client model on an IBM BlueGene, the algorithm 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, we predictively fold the forty amino acid HIV accessory protein in less then 24 hours on 2046 nodes of the IBM BlueGene supercomputer.

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 medicinal success, protein simulation strategies also impact the materials and an increasingly the nano-sciences.

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 achieve by the presently most accurate simulation techniques, which follow the time evolution of the protein in its environment. Since the microscopic simulation step in such molecular-mechanics methods is of the order of femtoseconds, while the folding or association process takes place on the order of milliseconds, such simulations remain limited in the system size due to the large computational efforts[1,2]. Direct simulation studies were successful however for a number of small proteins, while unfolding simulations starting from the native conformation have given insight into several aspects of protein thermodynamics[3] and folding[4].

It has been a great hope for almost a decade that emerging massively parallel computers architectures, which are available now at the teraflop scale, and which will reach the petaflop scale in the foreseeable future, will 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.

In a fundamentally different approach we have developed models[5] and algorithms[6,7] which permit reproducible and predictive folding of small proteins from random initial conformations using free-energy forcefields. According to Anfinsen's thermodynamic hypothesis[8] many proteins are in thermodynamic equilibrium with their environment under physiological conditions. Their unique three-dimensional native conformation then corresponds to the global optimum of a suitable 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[9] (structure prediction). In combination with thermodynamic simulation methods (Monte-Carlo or parallel tempering)[10], this approach generates continuous folding trajectories to the native ensemble.

Stochastic optimization methods[13], rather than kinetic simulations, 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[14], due to the directed construction of downhill moves on the free-energy surface, or 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. In a recent investigation, we found that parallel tempering scales only to about 32 replicas[10,11]. 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[12].

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[15], 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[13], 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 evolutionary algorithm, which generalized the basin hopping or Monte-Carlo with minimization method[7,21] to many concurrent simulations. Using this approach we could fold the sixty amino acid bacterial ribosomal protein to its native ensemble[17]. For the largest simulations we used a 4096 processor IBM BlueGene computer to investigate the folding of the 40 amino acid HIV accessory protein. We find that the algorithm scales from 64 to 4096 nodes with less than 10% loss of computational efficiency. Using 2048 processors we succeed to fold the protein from completely extended to near-native conformations in less than a single day.

2 Methods

2.1 Forcefield

We have parameterized an all-atom free-energy forcefield for proteins (PFF01)[5], which is based on the fundamental biophysical interactions that govern the folding process. This forcefield represents each atom (with the exception of apolar CH_n groups) of the protein individually and models their physical interactions, such as electrostatic interactions, Pauli exclusion, on the bonds attraction and hydrogen bonding. The important interactions with the solvent are approximated by an implicit solvent model. All of these interactions have been optimized to represent the free-energy of a protein microstate corresponding to a particular backbone conformation. For many proteins we could show that the native conformations correspond to the global optimum of this forcefield. We have also developed, or specifically adapted, efficient stochastic optimization methods[14,11,6] (stochastic tunnelling, 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.

With this approach we were able to predictively and reproducibly fold more than a dozen proteins, among them the trp-cage protein (23 amino acids)[18], the villin headpiece (36 amino acids)[19], the HIV accessory protein (40 amino acids)[9], protein A (40 amino acids) as well as several toxic peptides and beta-hairpin proteins (14-20 amino acids) in simulations starting from random initial conformations. The largest protein folded de-novo to date is the 60 amino-acid bacterial ribosomal protein L20[17]. 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[9,19].

2.2 Optimization Method

Most stochastic optimization methods map 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[7], including protein folding[9], but employs only one dynamical process. This method[21] 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[7]. Within each SA[13] 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.

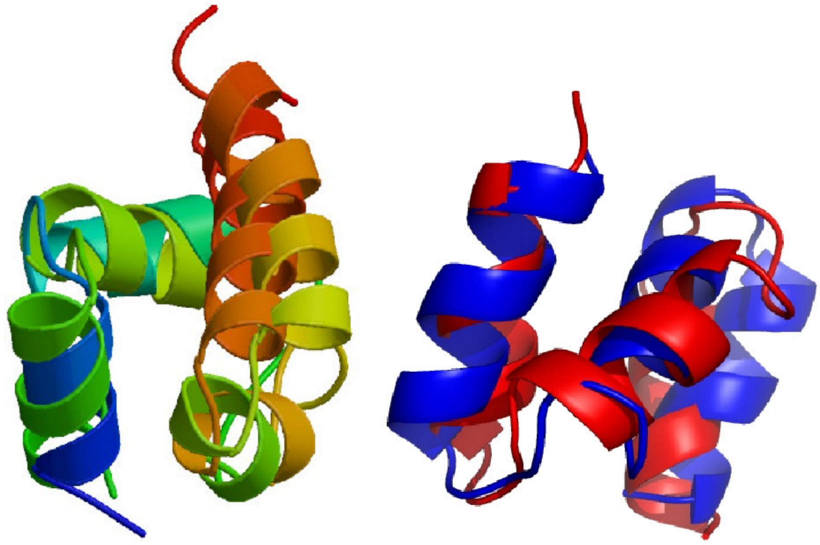

Figure 1. Overlay of the folded and the experimental conformation of the bacterial ribosomal protein L20 (left) and the HIV accessory protein (right)

We have generalized this method to a population of P interdependent dynamical processes operating on a population of N conformations[17]. 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.

This approach is easy to parallelize in a master client approach, where idle clients request tasks from the master. Since the new task is drawn at random from the population, new tasks can be distributed at any time. When the number of processors is much large than the size of the population, many processors attempt to improve the same conformation. However, because the dimensionality of the search space is so large (over 100 independent variables in the problems described here), most of these simulations fail to find good solutions. As a result, the algorithm remains effective even in this limit.

3 Simulations

3.1 Folding the Bacterial Ribosomal Protein L20

We have folded the 60 amino acid bacterial ribosomal protein(pdb-id:1GYZ). This simulation 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 perform 50 annealing cycles per replica on these conformation, 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 Fig. 1, 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 %.

3.2 Folding the HIV Accessory Protein

We have also folded the 40 amino acid HIV accessory protein(pdb-id: 1F4I). 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 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

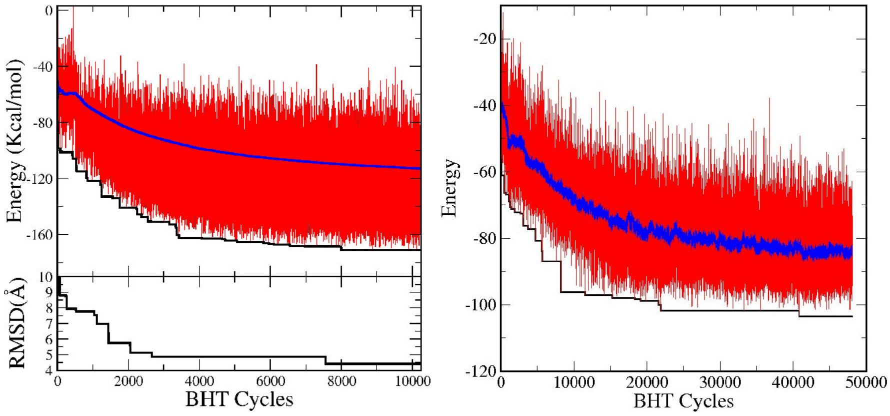

Figure 2. Instantaneous energy (red), mean energy (blue) and best energy (black) for the simulations of the bacterial ribosomal protein L20 (left) and the HIV accessory protein (right). For the bacterial ribosomal protein L20 the lower panel indicates the convergence of the RMS deviation of the lowest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

to balance diversity and energy convergence. The acceptance threshold of 4 Å RMSB 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 of two proteins: Using 50 processors of a loosely connected PC cluster we succeeded to fold the 60 amino acid bacterial ribosomal protein to near-native conformations, 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 Å 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. According to the current funnel paradigm[22] 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[3]. 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 on its downhill path on the free-energy landscape. In this context, 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[23]. 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 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. This more coarse-grained parallelization permits the use of a much larger number of weakly-linked processors. The algorithm scales very well, because each work-unit that is distributed to a client is very large. We do not parallelize over a single, but instead over thousands of energy evaluations. For this reason the algorithm can tolerate very large latency times and even collapse of part of the network. Since conformations are drawn randomly according to some probability distribution from the current population, client requests need never wait for the next task. In practice, however, redistribution of the same conformation to many processors may lead to duplication of work. Already for the study performed here (N=2048, N_{pop}=64), 32 processors attempt to improve each conformation on average. However, the success rate for improvement drops rapidly with the dimension of the search space. For this reason, the algorithm will perform well on even larger networks as the sizes of the proteins under investigation increase.

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, mis-folding and protein-protein interactions.

Acknowledgements

This work is supported by grants from the German national science foundation (DFG WE1863/10-2), the Secretary 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.

References

1. X. Daura, B. Juan, D. Seebach, W.F. van Gunsteren and A. E. Mark, *Reversible peptide folding in solution by molecular dynamics simulation*, J. Mol. Biol., **280**, 925, (1998).
2. C. D. Snow, H. Nguyen, V. S. Pande and M. Gruebele, *Absolute comparison of simulated and experimental protein folding dynamics*, Nature, **420**, 102–106, (2002).
3. T. Lazaridis, and M. Karplus, *"New view" of protein folding reconciled with the oldthrough multiple unfolding simulations*, Science, **278**, 1928–1931, (1997).

4. A. E. Garcia and N. Onuchic, *Folding a protein in a computer: An atomic description of the folding/unfolding of protein A*, Proc. Nat. Acad. Sci. (USA), **100**, 13898–13903, (2003).
5. T. Herges and W. Wenzel. *An all-atom force field for tertiary structure prediction of helical proteins*, Biophys. J., **87**, 3100–3109, (2004).
6. A. Schug, A. Verma, T. Herges, K. H. Lee and W. Wenzel, *Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein*, ChemPhysChem, **6**, 2640–2646, (2005).
7. A. Verma, A. Schug, K. H. Lee and W. Wenzel, *Basin hopping simulations for all-atom protein folding* , J. Chem. Phys., **124**, 044515, (2006).
8. C. B. Anfinsen, *Principles that govern the folding of protein chains*, Science, **181**, 223–230, (1973).
9. T. Herges and W. Wenzel, *Reproducible in-silico folding of a three-helix protein and characterization of its free energy landscape in a transferable all-atom forcefield*, Phys. Rev. Lett., **94**, 018101, (2005).
10. A. Schug, T. Herges and W. Wenzel, *All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method*, Proteins, **57**, 792–798, (2004).
11. A. Schug, T. Herges, A. Verma and W. Wenzel, *Investigation of the parallel tempering method for protein folding*, J. Physics: Cond. Mat., **17**, 1641–1650, (2005).
12. A. Garcia and J. Onuchic, *Folding a protein on a computer: hope or reality*, Structure, **13**, 497–498, (2005).
13. S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, *Optimization by simulated annealing*, Science, **220**, 671–680, (1983).
14. W. Wenzel and K. Hamacher, *Stochastic tunneling approach for global optimization of complex potential energy landscapes*, Phys. Rev. Lett., **82**, 3003–3007, (1999).
15. C. Levinthal, *Are there pathways for protein folding ?*, J. Chim. Phys., **65**, 44–45, (1968).
16. Z. Li and H. A. Scheraga, *Monte carlo minimization approach to the multiple minima problem in protein folding*, Proc. Nat. Acad. Sci. U.S.A., **84**, 6611–6615, (1987).
17. A. Schug and W. Wenzel, *An evolutionary strategy for all-atom folding of the sixty amino acid bacterial ribosomal proein L20*, Biophys. J., **90**, 4273–4280, (2006).
18. A. Schug, T. Herges and W. Wenzel, *Reproducible protein folding with the stochastic tunneling method*, Phys. Rev. Letters, **91**, 158102, (2003).
19. T. Herges and W. Wenzel, *Free energy landscape of the villin headpiece in an all-atom forcefield*, Structure, **13**, 661, (2005).
20. T. Herges, H. Merlitz and W. Wenzel, *Stochastic optimization methods for biomolecular structure prediction*, J. Ass. Lab. Autom., **7**, 98–104, (2002).
21. A. Nayeem, 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., **12(5)**, 594–605, (1991).
22. K. A. Dill and H. S. Chan, *From levinthal to pathways to funnels: The "new view" of protein folding kinetics*, Nature Structural Biology, **4**, 10–19, (1997).
23. Y. Duan and P. A. Kollman *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution*, Science, **282**, 740–744, (1998).