All-Atom *De Novo* Protein Folding with a Scalable Evolutionary Algorithm

ABHINAV VERMA,[1] SRINIVASA M. GOPAL,[2] JUNG S. OH,[3] KYU H. LEE,[3] WOLFGANG WENZEL[2]
[1]*Institute for Scientific Computing, Forschungszentrum Karlsruhe, Karlsruhe, Germany*
[2]*Institute for Nanotechnology, Forschungszentrum Karlsruhe, Karlsruhe, Germany*
[3]*Supercomputational Materials Lab, Korean Institute for Science and Technology, Seoul, Korea*

Abstract: The search for efficient and predictive methods to describe the protein folding process at the all atom level remains an important grand computational challenge. 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. Here we report the predictive all atom folding of the forty amino acid HIV accessory protein using an evo lutionary stochastic optimization technique. We implemented the optimization method as a master client model on an IBM BlueGene, where the algorithm scales near perfectly from 64 to 4096 processors in virtual processor mode. Starting from a completely extended conformation, we optimize a population of 64 conformations of the protein in our all atom free energy model PFF01. Using 2048 processors the algorithm predictively folds the protein to a near native conformation with an RMS deviation of 3.43 Å in <24 h.

Key words: *De Novo* protein folding; evolutionary algorithm; all atom folding

Introduction

Protein folding and structure prediction at the all atom level remain important computational challenges with many applica tions in the life and nano sciences. One important milestone in this direction is the development of methods that are capable to predictively fold proteins and peptides from unbiased unstruc tured conformations to the native ensemble. Direct simulation studies have demonstrated the folding of several small peptides and mini proteins from completely extended conformations, but remain limited in the system size by the large computational effort required.[1 6] Unfolding simulations, starting from the native conformation have given insight into protein thermody namics[7] and the transition state ensemble in a first principles approach even for larger proteins,[8,9] but typically fail to return to the native ensemble once the transition state has been crossed.

One great hope towards reproducible all atom folding is the development of algorithms that can exploit emerging massively parallel computational architectures that will deliver petaflop computational performance by the end of this decade. The straightforward parallelization of the energy and force evaluation in a single computational step is one possible approach to dis tribute the computational load, but remains limited to a rela tively small number of processors (presently roughly 16 128 nodes with a high speed interconnect, depending on program and architecture). This approach alone is therefore unlikely to efficiently exploit the many thousand processors of emerging petaflop architectures, let alone grid applications with hundreds of thousands of processors. An alternative approach has been to distribute short simulations on many independent nodes and to extrapolate the results[2] or to reconstruct the folding dynamics from coarse grained dynamics of populations of such simula tions.[10]

In yet another alternative approach, we have developed models[11] and algorithms,[12,13] which permit reproducible and pre dictive folding of small proteins (up to sixty amino acids) from random initial conformations using free energy forcefields. We exploit Anfinsen's thermodynamic hypothesis[14] that many pro teins are in thermodynamic equilibrium with their environment under physiological conditions. The unique three dimensional native conformation of the protein can then be predicted as the global optimum of a suitably free energy model. The free energy

Correspondence to: W. Wenzel; e mail: wenzel@int.fzk.de

Contract/grant sponsor: German national science foundation; Contract/ grant number: DFG WE1863/10 2

Contract/grant sponsor: Secretary of State for Science and Research through the Helmholtz Society and the Kurt Eberhard Bode foundation.

model captures the internal energy of a given backbone conformation and solvent and side chain entropy via an implicit solvent model. Comparing just individual backbone conformations these models assess the relative stability of conformations[15] (structure prediction). In combination with thermodynamic simulation methods (Monte Carlo or parallel tempering),[16 19] this approach generates continuous folding trajectories to the native ensemble.

To assess the relative stability of given backbone ensembles, stochastic optimization methods[20] can be used to search the protein free energy landscape in a fictitious dynamical process. Such methods thus offer the potential to explore the protein free energy landscape orders of magnitude faster than kinetic simulations by accelerating the traversal of transition states[21] the directed construction of downhill moves on the free energy surface,[22] the exploitation of memory effects or a combination of such methods.[23] 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.[17,18] 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.[19]

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 (SA),[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.

We have recently developed an evolutionary algorithm, which generalized the basin hopping or Monte Carlo with minimization,[13,15,23,25 29] method to many concurrent simulations. Using this approach we could fold the sixty amino acid bacterial ribosomal protein to its native ensemble.[30,31] This simulation exploited 50 concurrent dynamical processes and ran for about 6 months on a PC cluster. Here 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 $<10\%$ 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.

Methods

We have parameterized an all atom free energy forcefield for proteins (PFF01),[11] which is based on the fundamental biophysical interactions that govern the folding process. We could that near native conformations of several proteins correspond to the global optimum of this forcefield. We have also developed, or specifically adapted, efficient stochastic optimization methods[12,18,21] (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.

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),[32] the villin headpiece (36 amino acids),[33] the HIV accessory protein (40 amino acids),[15] protein A (40 amino acids) as well as several toxic peptides and β hairpin proteins (14 20 amino acids)[13] in simulations starting from random initial conformations. With 60 amino acids the four helix bacterial ribosomal protein L20[30] is the largest protein folded *de novo* to date. 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.[15,33]

Forcefield

The all atom (with the exception of apolar CH_n groups) free energy forcefield PFF01[11] parameterizes the internal free energy of a protein macrostate in a minimal thermodynamic approach.[11,15,34] The forcefield parameterizes the internal free energy of the protein (excluding backbone entropy) and contains the following nonbonded interactions:

$$V(\{\vec{r}_i\}) = \sum_{ij} V_{ij}\left[\left(\frac{R_{ij}}{r_{ij}}\right)^{12} \quad 2\left(\frac{R_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{ij}\frac{q_i q_j}{\varepsilon_{g(i)g(j)} r_{ij}} + \sum_i \sigma_i A_i + \sum_{\text{hbonds}} V_{hb}. \quad (1)$$

Here r_{ij} denotes the distance between atoms i and j and $g(i)$ the type of the amino acid i.

The Lennard Jones parameters (V_{ij}, R_{ij} for potential depths and equilibrium distance) depend on the type of the atom pair and were adjusted to satisfy constraints derived from a set of 138 proteins of the PDB database.[34 36] The nontrivial electrostatic interactions in proteins are represented via group specific dielectric constants ($\varepsilon_{g(i)g(j)}$ depending on the amino acids to which the atoms i and j belong). The partial charges q_i and the dielectric constants were derived in a potential of mean force approach.[37] Interactions with the solvent were first fit in a minimal solvent accessible surface model[38] parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly X Gly family of peptides.[39] A_i corresponds to the area of atom i that is in contact with a fictitious solvent. Hydrogen bonds are described via dipole dipole interactions included in the electrostatic terms and an additional short range term for backbone backbone hydrogen bonding (CO to NH) which

depends on the OH distance, the angle between N, H, and O along the bond and the angle between the CO and NH axis.[11]

In the folding process under physiological conditions, the degrees of freedom of a peptide are confined to rotations about single bonds. In our simulation we therefore consider only moves around the sidechain and backbone dihedral angles, which are attempted with thirty and seventy percent probability respectively. The moves for the sidechain angles are drawn from an equidistributed interval with a maximal change of 5°. Half of the backbone moves are generated in the same fashion, and the remainder is generated from a move library that was designed to reflect the natural amino acid dependent bias towards the formation of α helices or β sheets. The probability distribution of the move library was fitted to experimental probabilities observed in the PDB database.[40] While driving the simulation towards the formation of secondary structure, the move library introduces no bias towards helical or sheet structures beyond that encountered in nature.

Optimization Strategy

Protein simulations that do not conserve the overall energy of the system at least on average are complicated by the comparatively close packing of the atoms in the collapsed ensemble, which the protein encounters en route from the unfolded to the folded ensemble. The high density of the collapsed conformations means that many proposed moves of the dynamical scheme of the simulation have very high potential energy. Suitable optimization methods must therefore be able speed the simulation by avoiding high energy transition states, adapt large scale moves or accept unphysical intermediates. The basin hopping technique has proved to be a reliable workhorse for many complex optimization problems,[13,26,29] including protein folding,[15,23,27,28,41] but employs only one dynamical process.

This method[42] employs a relatively straightforward approach to eliminate high energy transition states of the PES (Fig. 1). The original potential energy surface is simplified 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 SA. In many applications the additional effort for the minimization step is more than compensated by the improved efficiency of the stochastic search. The basin hopping technique and derivatives[35] has been used previously to study the potential energy surface of model proteins[29] and polyalanines using all atom models.[27,28] We have used a SA process for the minimization step,[13] because analytical gradients for the SASA implicit solvent model of our forcefield are computationally very difficult to obtain. Within each SA[20] 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 with typical energy differences between competing metastable conformations, to ensure convergence to a local minimum.

We have generalize this method to a population of size N, which is iteratively improved by P concurrent dynamical processes.[30,31] 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. Similar strategies, employing a conformation

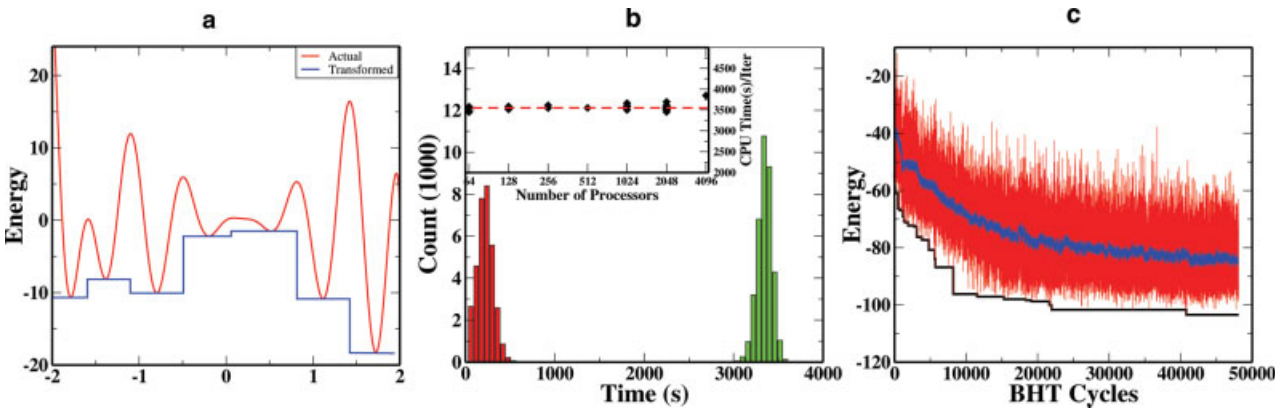

Figure 1. (a) Schematic illustration of the bain hopping technique: each point of the original potential energy surface (PES) is mapped to its nearest local minimum. The stochastic search of the effective PES is much faster than on the original one, because intervening transition states are reduced or eliminated. (b) Histogram of the distribution of client execution time (blue) and client idle time (red) in seconds for 20 iterations of the EA on 2048 processors. Inset: Wall clock time per iteration for the evolutionary algorithm as a function of the number of processors; a constant time dependence indicates perfect scaling. The red line indicates the average of all iterations for N 2048 processors as a guide to the eye. (c) Time series of the best energy (black), the average energy (blue) and the instantaneous energy (red) as a function of iterations for N 2048. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

stack, have previously been explored in simulations of the 23 amino acid BBA5 protein.[13,35]

This algorithm was implemented on a distributed master client model in which idle clients request a task from the master. The master maintains a list of open tasks comprising the active conformations of the population. The client then performs a SA simulation of specified length ($N = 40{,}000$ steps) on the conformation. The SA runs used a geometric cooling schedule reducing the temperature from 1200 to 2 K.

Conformations are drawn randomly according to some probability distribution from the active population. The acceptance criterion for newly generated conformations must balance the diversity of the population against the enrichment low energy decoys. Since one can in principle account for the number of times a given conformation was found (not employed here), there is no need to store duplicates. We therefore accept only new conformations which are different by at least 4 Å RMSB (root mean square backbone deviation) from all members of the active population. 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.

Results

Scalability

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. Here we report data for a population size $P = 64$ for simulations of the 40 amino acid HIV accessory protein (sequence: QEKEAIERLK ALG FEESLVI QAYFACEKNE NLAANFLLSQ, pdb id: 1F4I).[43] As demonstrated in Figure 1(a) the algorithm scales well up to 4096 processors.

The control loop is implemented employing a synchronous simulation protocol, where tasks are distributed to all processors of the machine, each drawing a member of the presently active conformation with equal probability. Each processor then performs a basin hopping simulation in which the present conformation is optimized independently of all others. For each step of the process the energy evaluation is optimized to compute only those energy terms in the model that have changed from the previous conformation, clashing conformations are rejected outright. For this reason the simulation time varies slightly from basin hopping simulation to basin hopping simulation even though the number of simulation steps is identical for each processor ($N = 40{,}000$). 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 (Fig. 1b) induce a waiting time before the next iteration can start. This waiting time is largest in the first few iterations, because a processor in subsequent iterations have slight starting offsets along the time axis, which increase the likelihood that the results are returned in the same sequence of processors that they were issued. In this scenario there would be no waiting time even in a synchronous processing mode.

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. An asynchronous implementation of the master loop would probably reduce these fluctuations further. We have also performed simulations on the smaller trp cage protein (20 amino acids) and the 56 amino acid protein G with qualitatively similar results (data not shown).

Folding Simulation

For the folding simulation the population was initially seeded with a single completely stretched "stick" conformation. The seed conformation had an average RMSB deviation of 21.5 Å to the experimental conformation. We then performed 20 cycles of the evolutionary algorithm described above.

Figure 1(c) shows the convergence of the energy as a function of the total number of basin hopping cycles. We find 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. The red line shows the instantaneous energy of the conformations that are returned from the client to the master after one BHT cycle. As expected they continue to fluctuate around the average energy of the population. Figures 2b and 2c shows the overlay of the folded and the experimental conformation. The starting conformation (Fig. 2a) 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. Aside from the native ensemble, only one different cluster of conformations with low energy occurs frequently in the simulations. One representative conformation for this family is shown in Figure 2(c), which has nearly the same secondary structure content, but the first helix is not aligned properly with the helix 2(ESLVIQAYF) and helix 3(NLAANFLLS). Helix 2 and helix 3 agree to 2.8 Å with the native conformation in this structure.

The success of the algorithm to fold the protein can be rationalized by analyzing the flow of information through the decision tree (See Chart 1). We have annotated the arrows of the tree to show the fraction of total new conformations flowing through the various branches. About 30% of the returning conformations are similar to at least one of the active conformations

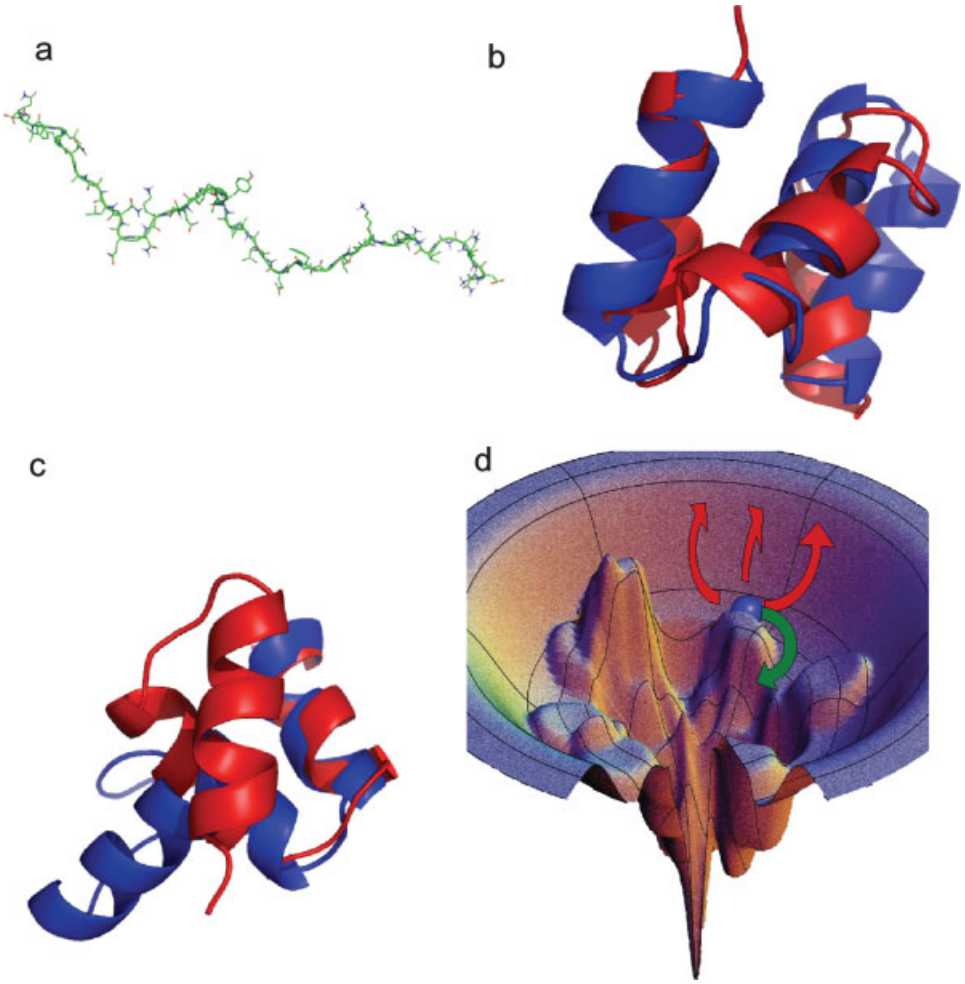

Figure 2. (a) Starting conformation for the evolutionary algorithm, (b) and (c) overlay of the experimental (red) and the folded structure (blue) for the N 2048 run. Both conformations agree in their secondary structure content, the backbone RMS deviation is 3.43 Å, originating from a misalignment of the first helix (shown in front), (d) Schematic illustration of the different routes of the evolutionary algorithm on a two dimensional model protein folding funnel. Most simulations explore the vicinity of the starting conformation, but with increasing dimension of the search space, many go astray (red), only a few find new conformations (green), that are refined in later iterations. This inherent limitation of the local search process, here the simulated annealing run, makes it possible to employ algorithms that start many simulations from the same conformation without wasting computational resources. The funnel landscape was taken from K. Dill's homepage (www.dillgroup.ucsf.edu).

and all of these are accepted into the active population (refinement). This implies that the SA step is highly successful to improve existing conformations. We find that 10% of the simulations lead to the replacement of more than one conformation (merge operation) in the decision tree, which indicates a narrowing of the folding funnel as the simulation proceeds. The protein is not just folded one, but many simulations converge to the same intermediate structure. The merge operation is therefore useful to avoid replication of the information.

From the remaining 72% conformations, 10% conformations (the same as the fraction of merge operations) are added to the population because it has shrunk. The algorithm thus succeeds to continuously reseed itself; this generates a high likelihood that the simulation does not get stuck in an uninteresting metastable area of the folding landscape. Nineteen percent of the new conformations are dissimilar to all other conformations of the population, but nevertheless better than the worst conformations. These new structural templates are then the candidates for further local refinement in the steps discussed above. About 43% of the basin hopping cycles go astray, which is commensurate with earlier basin hopping investigations. We note that the balance of refinement and new structural templates generate

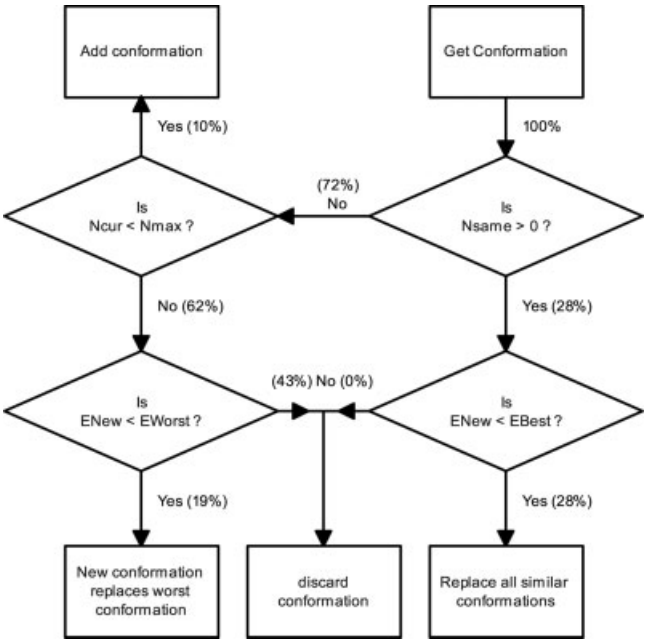

Chart 1. Schematic illustration of the decision tree for the evolutionary algorithm employed in this investigation: new conformations enter the decision tree with energy ENew, the number of conformations in the population with an RMSD $<$ CutOff RMSD is designated as Nsame. Nmax/Ncur are the maximal/current number of conformations in the population. The highest energy of all conformations in the population is designated by EWorst. The arrows in the tree are annotated by the total probabilities of the conformation flow in the folding simulation described in the text.

a dynamic population that slides as whole towards the global optimum of the free energy funnel.

Discussion

Using a scalable evolutionary algorithm we have demonstrated the all atom folding of the 40 amino acid HIV accessory protein from a completely extended conformation to within 4 Å of the native conformation in about 24 h turnaround time using 2048 processors of an IBM BlueGene. The evolutionary algorithm evolves not one, but an active population containing many conformations concurrently. Considering the limiting cases, it is a priori unclear, how such a strategy can succeed to efficiently fold the protein. For small population size (P) many processors (N) construct short trajectories emanating from the same conformation ($P \ll N$). If the energy gain for each such step is small compared with the total folding energy, many cycles will be required to complete the simulation even if many processors are available. A large fraction of the computational resources would be wasted in such a scenario. In the opposite limit ($N \ll P$) most conformations sample high energy regions of the free energy surface that are unrelated to the native conformation. Improvement of such conformations is irrelevant to the folding process. The latter limit is therefore unattractive for large scale distributed computational architectures, where N is large.

The key to convergence lies therefore in the exploitation of the specific characteristics of the protein 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.[7,46] 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[47–49] on its downhill path on the free energy landscape. In this context on cycle of the evolutionary algorithm in the $P \ll N$ limit attempts to improve many times each of the conformations of the active population.

Because of the effective friction and local frustration on the free energy landscape most of these simulations explore the vicinity of their respective starting points. Because of the actual high dimensionality (D) of the search space ($D = 160$ free dihedral angles for 1F4I) most of them terminate higher in free energy than their starting conformation. For a two dimensional free energy surface this is illustrated schematically in Figure 1(d). These conformations are rejected by the energy criterion. Most of the remaining simulations that improve upon the starting conformation stay within the distance acceptance threshold of the evolutionary algorithm and replace their starting conformation in the active population. The distance acceptance threshold thus ensures that the population is not overpopulated by nearly identical conformations of the same region in conformational space. In the rare event, that the simulation improves the energy and generates a genuinely new conformation, the energetically worst conformation of the active population is replaced. This conformation is the starting point for further local refinement in subsequent iterations.

This analysis reveals the mechanism for the effectiveness of the evolutionary algorithm. The move generator, in this case the SA run in the individual step, generates an "acceptable" new conformation with a probability $p(D)$ that falls rapidly with the dimension of the search space and the quality of the present population. As long as $p(D) < P/N$ each cycle of the evolutionary algorithm will improve each member of the active population at most once on average. As long as no genuinely better move generator exists (higher $p(D)$), all computation effort is, on average, efficiently directed towards folding the protein. Only when N becomes so large that the above relation no longer holds, several attempts per cycle will improve the same member of the active conformation, even though only one of these improvements can be kept according to the acceptance rules, leading to duplication and hence waste of computational resources. This is good news for the scalability of the evolutionary algorithm for larger proteins: Because $p(D)$ drops rapidly with the size of the protein, the number of processors that can be effectively employed for folding can be further increased using thousands, possibly hundreds of thousands of processors concurrently.

Summary

The search for methods and models for de novo folding of small and medium size proteins from the completely extended confor mation at atomic resolution has been a "holy grail" and grand computational challenge for decades.[50] 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 <24 h turnaround time is thus an important step towards the long time goal to elucidate protein structure formation and function with atomistic resolu tion. The free energy approach employed here, which presently lacks a detailed elucidation of the folding kinetics, can comple ment Hamiltonian based simulation methods, such as molecular dynamics or replica exchange methods, to understand how pro teins 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 present study demonstrates, equally importantly, it is possible to paral lelize the search process by splitting the simulation into a large number of independent conformations, rather than by paralleliz ing the energy evaluation. This more coarse grained paralleliza tion permits the use of a much larger number of weakly linked processors. 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 pos sible to investigate and understand a wide range of biological problems related to protein folding, misfolding and protein pro tein interactions.

Acknowledgments

We acknowledge the use of facilities at the IBM Capacity on Demand Center in Rochester and KIST Supercomputional Mate rials 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. Daura, X.; Juan, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. J Mol Biol 280, 925, 1998.
2. Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. Nature 2002, 420, 102.
3. Simmerling, C.; Strockbine, B.; Roitberg, A. J Am Chem Soc 2002, 124, 11258.
4. Pitera, J.; Swope, W. Proc Natl Acad Sci USA 2003, 100, 7587.
5. Yang, W. Y.; Pitera, J. W.; Swope, W. C.; Gruebele, M. J Mol Biol 2004, 336, 241.
6. Chen, J.; Im, W.; Brooks, C. L., III. J Am Chem Soc 2006, 128, 3728.
7. Lazaridis, T.; Karplus, M. Science 1997, 278, 1928.
8. Garcia, A. E.; Onuchic, N. Proc Natl Acad Sci USA 2003, 100, 13898.
9. Mayor, U.; Guydosh, N. R.; Johnson, C. M.; Grossmann, J. G.; Sato, S.; Jas, G. S.; Freund, S. M. V.; Alonso, D. O. V.; Daggett, V.; Fersht, A. R. Nature 2003, 421, 863.
10. Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. Proc Natl Acad Sci USA 2005, 102, 6801.
11. Herges, T.; Wenzel, W. Biophys J 2004, 87, 3100.
12. Schug, A.; Verma, A.; Herges, T.; Lee, K. H.; Wenzel, W. Chem Phys Chem 2005, 6, 2640.
13. Verma, A.; Schug, A.; Lee, K. H.; Wenzel, W. J Chem Phys 2006, 124, 044515.
14. Anfinsen, C. B. Science 1973, 181, 223.
15. Herges, T.; Wenzel, W. Phys Rev Lett 2005, 94, 018101.
16. Hansmann, U. H. E. J Chem Phys 2004, 120, 417.
17. Schug, A.; Herges, T.; Wenzel, W. Proteins 2004, 57, 792.
18. Schug, A.; Herges, T.; Verma, A.; Wenzel, W. J Phys Condens Mat ter 2005, 17, 1641.
19. Garcia, A.; Onuchic, J. Structure 2005, 13, 497.
20. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Science 1983, 220, 671.
21. Wenzel, W.; Hamacher, K. Phys Rev Lett 1999, 82, 3003.
22. Derreumaux, P. J Chem Phys 1997, 106, 5260.
23. Abagyan, R. A.; Totrov, M. J Comp Phys 1999, 151, 402.
24. Levinthal, C. J Chim Phys 1968, 65, 44.
25. Li, Z.; Scheraga, H. A. Proc Natl Acad Sci USA 1987, 84, 6611.
26. Leitner, D. M.; Chakravarty, C.; Hinde, R. J.; Wales, D. J. Phys Rev E 1997, 56, 363.
27. Mortenson, P. N.; Wales, D. J. J Chem Phys 2004, 114, 6443.
28. Mortenson, P. N.; Evans, D. A.; Wales, D. J. J Chem Phys 2002, 117, 1363.
29. Wales, D. J.; Dewbury, P. E. J. J Chem Phys 2004, 121, 10284.
30. Schug, A.; Wenzel, W. J Am Chem Soc 2004, 126, 16736.
31. Schug, A.; Wenzel, W. Biophys J 2006, 90, 4273.
32. Schug, A.; Herges, T.; Wenzel, W. Phys Rev Lett 2003, 91, 158102.
33. Herges, T.; Wenzel, W. Structure 2005, 13, 661.
34. Herges, T.; Merlitz, H.; Wenzel, W. J Ass Lab Autom 2002, 7, 98.
35. Abagyan, R. A.; Totrov, M. J Mol Biol 1994, 235, 983.
36. Herges, T.; Schug, A.; Wenzel, W. Int J Quant Chem 2004, 99, 854.
37. Avbelj, F.; Moult, J. Biochemistry 1995, 34, 755.
38. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.
39. Sharp, K. A.; Nicholls, A.; Friedman, R.; Honig, B. Biochemistry 1991, 30, 9686.
40. Pedersen, J. T.; Moult, J. J Molec Biol 1997, 269, 240.
41. Carr, J. M.; Wales, D. J. J Chem Phys 2005, 123, 234901.
42. Nayeem, A.; Vila, J.; Scheraga, H. A. J Comp Chem 1991, 12, 594.
43. Withers Ward, E. S.; Mueller, T. D.; Chen, I. S.; Feigon, J. Bio chemistry 2000, 39, 14103.
44. Onuchic, J. N.; Luthey Schulten, Z.; Wolynes, P. G. Annu Rev Phys Chem 1997, 48, 545.
45. Dill, K. A.; Chan, H. S. Nat Struct Biol 1997, 4, 10.
46. Becker, O. M.; Karplus, M. J Chem Phys 1997, 106, 1495.
47. Plaxco, K. W.; Baker, D. Proc Natl Acad Sci USA 1998, 95, 13591.
48. Zagrovic, B.; Pande, V. J Comput Chem 2003, 24, 1432.
49. Qiu, L.; Hagen, S. J. Chem Phys 2005, 312, 327.
50. Duan, Y.; Kollman, P. A. Science 1998, 282, 740.