Modern Methods for Theoretical Physical Chemistry of Biopolymers
Edited by E.B. Starikov, J.P. Lewis and S. Tanaka

CHAPTER 17

All-atom protein folding with stochastic optimization methods

A. Schug[1], A. Verma[2], K. H. Lee[3] and W. Wenzel[1]

*[1]Forschungszentrum Karlsruhe, Institute for Nanotechnology,
P.O. Box 3640, D-76021 Karlsruhe, Germany
[2]Forschungszentrum Karlsruhe, Institute for Scientific Computing,
P.O. Box 3640, D-76021 Karlsruhe, Germany
[3]Korean Institute for Science and Technology, Supercomputing Materials Laboratory,
P.O.Box 131, Cheongryang, Seoul 130-650, Korea*

Abstract

We recently developed an all-atom free energy forcefield (PFF01) for protein structure prediction with stochastic optimization methods. We demonstrated that PFF01 correctly predicts the native conformation of several proteins as the global optimum of the free energy surface. Here we review recent folding studies, which permitted the reproducible all-atom folding of the 20 amino-acid trp-cage protein, the 40 amino-acid three-helix HIV accessory protein and the 60 amino-acid bacterial ribosomal protein L20 with a variety of stochastic optimization methods. These results demonstrate that all-atom protein folding can be achieved with present day computational resources for proteins of moderate size.

17.1 INTRODUCTION

Ab initio protein tertiary structure prediction (PSP) and the elucidation of the mechanism of the folding process are among the most important outstanding problems of biophysical chemistry [1,2]. The many complementary proposals for PSP span a wide range of representations of the protein conformation, ranging from coarse-grained models to atomic resolution. The choice of representation often correlates with the methodology employed in structure prediction, ranging from empirical potentials for coarse-grained models [3,4] to complex atom-based potentials that directly approximate the physical interactions in the system. The latter offer insights into the mechanism of protein structure formation and promise better transferability, but their use incurs large computational costs that has confined all-atom protein structure prediction to all but the smallest peptides [5,6].

It has been one of the central paradigms of protein folding that proteins in their native conformation are in thermodynamic equilibrium with their environment [7]. Exploiting this characteristic the structure of the protein can be predicted by locating the global minimum of its free energy surface without recourse to the folding dynamics, a process which is potentially much more efficient than the direct simulation of the folding process. PSP based on global optimization of the free energy may offer a viable alternative approach, provided that suitable parameterization of the free energy of the protein in its environment exists and that the global optimum of this free energy surface can be found with sufficient accuracy [8].

We have recently demonstrated a feasible strategy for all-atom protein structure prediction [9–11] in a minimal thermodynamic approach. We developed an all-atom free-energy force field for proteins (PFF01), which is primarily based on physical interactions with important empirical, though sequence independent, corrections [11]. We already demonstrated the reproducible and predictive folding of four proteins, the 20 amino-acid trp-cage protein (1L2Y) [9,12], the structurally conserved headpiece of the 40 amino-acid HIV accessory protein (1F4I) [10,13] and the 60 amino-acid bacterial ribosomal protein L20 [14]. In addition we showed that PFF01 stabilizes the native conformations of other proteins, e.g., the 52 amino-acid protein A [5,15], and the engrailed homeodomain (1ENH) from *Drosophilia melangaster* [16].

17.1.1 Force field

We have recently developed an all-atom (with the exception of apolar CH_n groups) free energy protein force field (PFF01) that models the low-energy conformations of proteins with minimal computational demand [10,11,17]. In the folding process at physiological conditions the degrees-of-freedom of a peptide are confined to rotations about single bonds. The force field is parameterized with the following non-bonded interactions:

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

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 atom pair and were adjusted to satisfy constraints derived from a set of 138 proteins of the PDB database [17–19]. The non-trivial electrostatic interactions in proteins are represented via group-specific dielectric constants ($\varepsilon_{g(i),\ g(j)}$ depending on the amino-acid to which atom i belongs). The partial charges q_i and the dielectric constants were derived in a potential-of-mean-force approach [20]. Interactions with the solvent were first fit in a minimal solvent accessible surface model [21] parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides [22]. A_i corresponds to the area of atom i that is in contact with a ficticious 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].

17.1.2 Optimization methods

The low-energy free energy landscape of proteins is extremely rugged due to the comparatively close packing of the atoms in the native structure. Suitable optimization methods must therefore be able to speed up the simulation by avoiding high-energy transition states, adapt large-scale moves or accept unphysical intermediates. Here we report on four different optimization methods, the stochastic tunneling method [23], the basin hopping technique [24,25], the parallel tempering method [26,27] and a recently employed evolutionary technique. The stochastic tunneling method and the basin hopping approach are inherently sequential algorithms, which evolve a single configuration according to a given stochastic process. In contrast, parallel tempering and evolutionary techniques are inherently parallel optimization strategies that are well suited to presently available multiprocessor architectures with low bandwidth connections. Since all-atom protein structure prediction remains a computationally challenging problem it is important to search for suitable optimization methods that are capable of exploiting such architectures, i.e., a high degree of parallelism with very little and optimally asynchronous communication is desirable.

17.1.2.1 Stochastic tunneling method

The stochastic tunneling technique (STUN) [23] was proposed as a generic global optimization method for complex rugged potential energy surfaces (PES). For a number of problems, including the prediction of receptor–ligand complexes for drug development [28,29], this technique proved superior to competing stochastic optimization methods. The idea behind the method is to flatten the potential energy surface in all regions that lie significantly above the best estimate for the minimal energy (E_0). In STUN the dynamic process explores not the original, but a transformed PES,

$$E_{\mathrm{STUN}} = 1\mathrm{n}\left(x + \sqrt{x^2+1}\right) \tag{2}$$

which dynamically adapts and simplifies during the simulation. Here $x = \gamma(E - E_0)$, where E is the energy, and E_0 the best energy found so far. The problem-dependent transformation parameter [23] γ controls the steepness of the transformation (we used. $\gamma = 0.5$ kcal/mol) The transformation in Eq. (2) ameliorates the difficulties associated with the original transformation [23], because $E_{\mathrm{STUN}} \propto 1\mathrm{n}(E/kT)$ continues to grow slowly for large energies. The ficticious temperature of STUN must be dynamically adjusted in order to accelerate convergence. STUN works best if its dynamical process alternates between low-temperature ‘local search’ and high-temperature ‘tunneling phases’. At finite temperature the dynamics of the system then becomes diffusive at energies $E >> E_0$ independent of the relative energy differences of the high-energy conformations involved. On the untransformed PES, STUN thus permits the simulation to 'tunnel' through energy barriers of arbitrary height.

17.1.2.2 Parallel tempering

The parallel (or simulated) tempering technique [26,27] was introduced to overcome difficulties in the evaluation of thermodynamic observables for models with very rugged potential energy surfaces and has been applied previously in several protein folding studies [30–32]. Low-temperature simulations on rugged potential energy surfaces are trapped for long times in similar metastable conformations because the energy barriers to structurally potentially competing different conformations are very high. The idea of PT is to perform several concurrent simulations of different replicas of the same system at different temperatures and to exchange replicas (or temperatures) between the simulations i and j with probability:

$$p = \min\left(1, \exp\left(-\left(\beta_j - \beta_i\right)\left(Ei - E_j\right)\right)\right) \tag{3}$$

where $\beta_i = 1/k_\beta T_i$ and E_i are the inverse temperatures and energies of the conformations, respectively. The temperature scale for the highest and lowest temperatures is determined by the requirement to efficiently explore the conformational space and to accurately resolve local minima, respectively. For proteins the temperatures must thus fall in a bracket between approximately 2–600 K. As described elsewhere [12] we have used an *adaptive temperature control* for the simulations: starting with an initial, ordered set of geometrically distributed temperatures we monitored the exchange rate between adjacent temperatures. If the exchange rate between temperature i and $i+1$ was below 0.5 per cent, then all temperatures above t_i were lowered by 10 per cent of $t_{i+1}-t_i$. If the exchange rate was above 2 per cent, then all all temperatures above t_i were increased by the same difference.

To further improve the computational efficiency of PT we also used a *replication step*, in which the best conformation replaces the conformation at the highest temperature every 250 000 simulation steps. This mechanism results in a rapid, large-scale exploration of the folding funnel around the best conformation found near the presently best conformation. The parallel tempering method was implemented in our program using the MPI communication library, which is available on most present day parallel computational architechtures with distributed memory. Since the communication effort is low (only the temperatures and energies need to be exchanged) and communication occurs only every few thousand steps, when replica exchange is attempted, this implementation scales very well with the number of processors.

17.1.2.3 Basin hopping method

An alternative approach to effectively eliminate high-energy transition states of the PES is used in the the basin hopping technique [24] (BHT), also known as Monte Carlo with minimization. This method simplifies the original potential energy surface 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 many cases the additional minimization effort to find an associated local minimum is more than compensated for by the increase of efficiency of the stochastic search on the simplified potential energy surface.

For the protein simulations we replace a single minimization step with a simulated annealing run [33]. Within each SA simulation, new configurations are accepted according to the Metropolis criterion. The temperature is decreased geometrically from its starting point to the final value, which must be chosen to be small compared with typical energy differences between competing metastable conformations, to ensure convergence to a local minimum (typically 2–5 K). Depending on the choice of the starting temperature, the SA search can deviate more or less significantly from its starting conformation. The individual relaxation step is thus parameterized completely by the starting (T_S), the final temperature and the number of steps. We investigate various choices for the numerical parameters of the method, but have always used a geometric cooling schedule.

At the end of one annealing step the new conformation was accepted if its energy difference to the current configuration was no higher than a given threshold energy ε_T, an approach recently proven optimal for certain optimization problems [34]. Throughout this study we use a threshold acceptance criterion of 1 kcal/mol.

17.1.2.4 Evolutionary strategies

While basin hopping and STUN are essentially sequential algorithms, PT provides some degree of inherent parallelism, which suits present day distributed architectures. In order to use even larger numbers of processors, we implemented an evolutionary strategy, in which the computational work is performed by many independent client computers that request tasks from a master computer. The master maintains a list of open tasks comprising the active conformations of the population. Each client performs an increasingly extensive energy minimization on the conformation it is given. When the client returns a new conformation after completing its task, it may replace an existing conformation following a scheme that balances the diversity of the population and the continued energetic improvement of its members. Specifically the client performs either a Monte Carlo (MC) or a simulated annealing (SA) [33] simulation of specified length on the conformation. Conformations are drawn randomly from the active population. When the client returns a new conformation after completing its task, the result is stored. Additionally, the new conformation replaces the energetically worst conformation in the active population, provided its energy is lower than the highest energy of the population and that it differs by at least 3Å backbone root mean square deviation (RMSB) from all members of the active population. If the new conformation has an RMSB of less than 3Å to some conformation of the population, it replaces this conformation if it is lower in energy.

This approach builds on the strength of the basin hopping technique, which is used for the individual steps. Since stochastic optimization methods never locate the global optimum with certainty, several independent simulations must be undertaken to obtain a relative degree of confidence in the convergence of the methods. In the evolutionary approach the progress a single simulation has made towards the optimum is exploited, because the resulting conformation will become a member of the active population. As such it may replace less optimized conformations and speed up the convergence of the overall population in comparison used a population of uncorrelated replicas.

17.2 RESULTS

17.2.1 The trp-cage protein

Using the PFF01 force field we simulated 20 independent replicas of the 20 amino-acid trp-cage protein [6,36] (pdb code 1L2Y) with a modified version of the stochastic tunneling method [9,23]. Six of 25 simulations reached an energy within 1 kcal/mol of the best energy, all of which correctly predicted the native experimental structure of the protein (see Fig. 17.1 (left)). We found a strong correlation between energy and RMSB deviation to the native structure for all simulations. The conformation with the lowest energy had a backbone root mean square deviation of 2.83 Å.

We also folded this protein with the parallel tempering method [12]. We found that the standard approach, which preserves the thermodynamic equilibrium of the simulated populations, did not reach very low energies even for the low-temperature replicas. We believe that the reason for this convergence failure was the insufficient exchange probability between replicas at different temperatures. We therefore introduced the adaptive temperature control described in the methods section. Fig. 17.2) shows the energies and corresponding temperatures for a simulation using 30 replicas. The temperature adjustment scheme results in a temperature distribution that permits frequent exchange of replicas and significantly speeds convergence. The best final structure associated with the lowest temperature in the simulation with 30 replicas had an RMSB deviation of 2.01 Å. We found convergence of the method using eight to thirty replicas. However, a minimal number of at least eight replicas appears to be required to fold the protein, for lower replica numbers it appears that even the adaptive temperature scheme fails to generate rapid replica exchange while spanning both high and low temperatures required for the speedy exploration of the free energy surface and the refinement of local minima, respectively.

Finally we have folded the trp-cage protein with the basin hopping technique. In comparison with the stochastic tunneling method we note that care must be taken with the parameterization of the basin hopping technique. First of all, very high starting temperatures, above 600 K, are required to permit a sufficient exploration of the free energy surface. This was also observed in the parallel tempering simulations (see Fig. 17.2) bottom

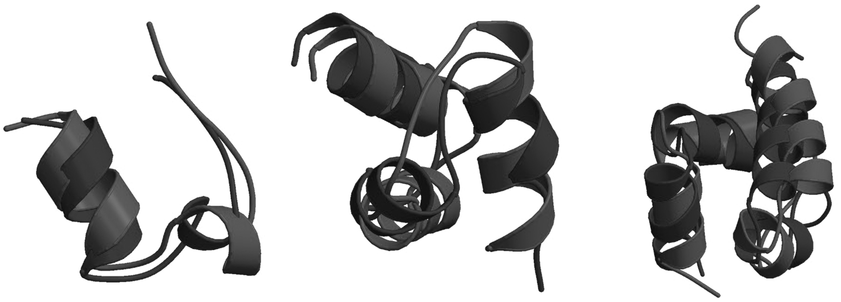

Fig. 17.1. Overlay of the native(red) and folded (blue) structures of trp-cage protein [35], the HIV accessory protein [13] and the bacterial ribosomal protein L20 [14] (see Color Plate 25).

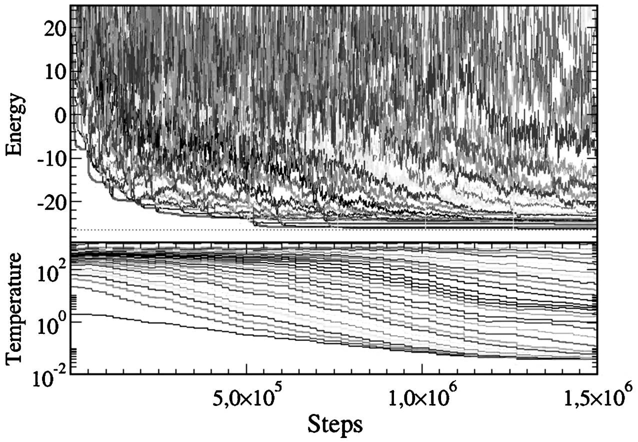

Fig. 17.2. Energies (upper panel) and temperatures (lower panel) of the 30 replica modified parallel tempering simulation of the trp-cage protein reported in the text. The dotted line in the upper panel corresponds to the estimate of the global optimum of the free energy (obtained independently). The lower panel demonstrates a rapid equilibration of the temperatures during the simulation. The upper panel demonstrates the convergence of the energy and the rapid exchange of information between the different replicas as discussed in the text (see Color Plate 26).

panel). As in PT, the lowest temperature had to be chosen in the range of 2–6 K to ensure that local minima are resolved well. In the stochastic tunneling method, the nonlinear transformation of the energy permits large-scale relaxations through thermodynamically forbidden regions, in basin hopping this effect can only be achieved by raising the temperature to unphysical values. We cannot rule out the possibility that basin hopping simulations with low starting temperatures would converge eventually, however, it appears that such an approach would not be computationally competitive. Furthermore, we note that the convergence of the basin hopping method is improved dramatically when the length of the relaxation run is moderately increased with the number of the basin hopping cycle. When we compared simulations comprised of basin hopping steps of constant length mid simulations where the length increased with the square root of the cycle number, we found much lower energies for the latter after investing the same total number of function evaluations in each run.

Using the basin hopping method with a starting temperature of $T_s = 800\ K$ and a final temperature of $T_f = 3\ K$ the lowest six of 20 simulations converged to the native structure. A total of 12 of these simulations approached the native conformation as its estimate of the optimum. The energies and RMSB deviations of all simulations are shown in Fig. 17.3). The plot indicates the existence of a set of structures with 2–3 Å RMSB deviation, which may correspond to the folding funnel, and a competing metastable conformation with about 5 Å RMSB. While all methods correctly identify the folding funnel, the basin hopping approach results in the lowest energies. Note that the second best simulation has an RMSB of only 1.8Å to the native conformation and loses in energy with less than 0.5 kcal/mol.

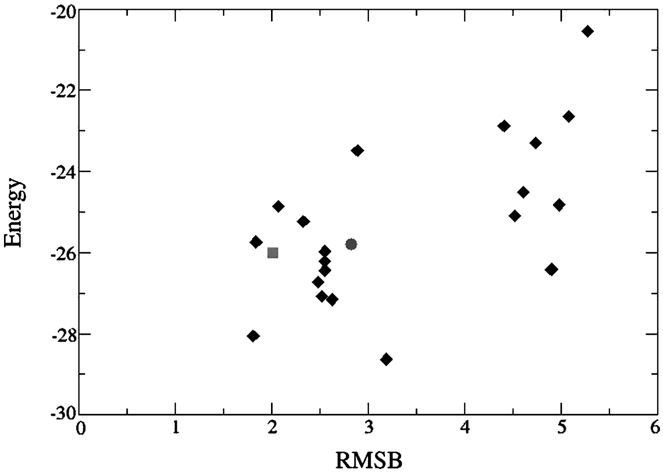

Fig. 17.3. Energy vs. RMSB plot for the final energies of the twenty basin hopping simulations described in the text (diamonds). For comparison we also indicate the best energy result for the STUN method (circle) and for the thirty processor PT simulation (square) (see Color Plate 27).

17.2.2 The HIV accessory protein

Encouraged by this result, we applied the modified basin hopping or Monte Carlo with minimization (MCM) strategy [8,25] to fold the structurally conserved 40 amino-acid headpiece of the HIV accessory protein [10]. We performed 20 independent simulations and found the lowest five to converge to the native structure (see tbldecoyhiv) [14]. The first non-native decoy appears in position six, with an energy deviation of 5 kcal/mol and a significant RMSB deviation. The table demonstrates that all low-energy structures have essentially the same secondary structure, i.e., position and length of the helices are always correctly predicted, even if the protein did not fold correctly.

AQ1

The good agreement between the folded and the experimental structure is also evident from Fig. 17.1 (center), which shows the secondary structure alignment of the native and the folded conformations. The good physical alignment of the helices illustrates the importance of hydrophobic contacts to correctly fold this protein. An independent measure to assess the quality of these contacts is to compare the C_β–C_β (which correspond to the NOE constraints of the NMR experiments that determine tertiary structure) in the folded structure to those of the native structure. We found that 66 per cent (80 per cent) of the C_β–C_β distances agree to within one (1.5) standard deviations of the experimental resolution.

We also performed a simulation of the HIV accessory protein using the adapted parallel tempering method [13]. We used 20 processors of an INTEL XEON PC cluster and ran the simulation for a total of 30×10^6 energy evaluations for each configuration, which corresponds to approximately 500 CPU hours on an 2.4 GHz INTEL XEON processor. All simulations were started with random conformations at high temperatures to allow for rapid, unbiased relaxation of the structures and the temperature distribution. The final conformation with the lowest energy/temperature converged to within 1.23/2.46 Å backbone root mean square (RMSB) deviation to the best known decoy/NMR structure of the HIV accessory protein. The overlay of the experimental and the converged structure (see Fig. 17.1) demonstrates the good agreement between the conformations, the difference in NOE constraints demonstrates that not only short-range, but also long-range distances are correctly predicted. Considering the ensemble of final conformations, we find many structures closely resembling the native conformation. The RMSB deviations of the next four lowest conformations (all within 1.5 kcal/mol of the minimal energy of XXX) have RMSB deviations of 3.14/2.23/3.78/3.00 Å, respectively, to the native decoy. AQ2

17.2.3 The bacterial ribosomal protein L20

In the course of the simulations on the HIV accessory protein we explored methods to share information between the independent basin hopping simulations in order to improve the overall convergence. For the 60 amino-acid bacterial ribosomal protein L20 (pdb-code 1GYZ) we thus experimented with the evolutionary technique described in the methods section. Starting from a seed population of random structures, we performed the folding simulation in three phases: (1) generation of starting structures of the population, (2) evolutionary improvement of the population, and (3) refinement of the best resulting structures to ensure convergence.

In phase (1) we performed high-temperature (500K) Monte Carlo simulations of 50 000 steps each. In these runs we reduced the strength of the solvent interactions (V_S) by 20 per cent to facilitate the rapid formation of secondary structures. It has been argued that hydrophobic collapse competes with secondary structure formation in protein folding. In the collapsed conformational ensemble, large-scale conformational changes, such as those required for secondary structure formation, occur only rarely. The goal of this simulation phase was the generation of a wide variety of competitive starting conformations for further refinement.

At the end of this simulation we had gathered in excess of 17 000 distinct decoys that were ranked according to their total energy as well as according to the individual energy terms of the force field (V_S, V_{LJ}, V_{HB}, V_C(Sidechain) and V_C(backbone). For each criterion we selected the best 50 conformations and eliminated duplicates to arrive at a population of 266 starting structures for phase (2) of the procedure. This population was relaxed in 14 000 SA simulations as described in the methods section. At the end of this step we selected the 50 conformations best in total energy for further refinement. In phase (3) we performed 5 500 SA simulations on this subpopulation. The length of the individual relaxation simulations was gradually increased from 10^5 steps per simulation to 2.3×10^6.

The best conformation had approached the native conformation to about 4.6 Å RMSB deviation. In total, six of the lowest ten conformations approach the native structure, while four others misfolded. Note that the selection criterion for the active population (see methods section) precludes the occurrence of the same configuration to within 3 Å RMSB; this dominance of near native conformations of the total ensemble is particularly encouraging.

In order to quantify the overall improvement of native content during the simulation, we defined the native content of the simulated ensemble as a weighted average of the deviations of the population and the native conformation: for a population of size N we add $100(N-R+1)/N$ for each near-native decoy (RMSB less than 4Å) ranked at position R by energy to the total native score of this population. A score of 100 thus corresponds to a native decoy placed at the top position, while a near-native decoy at the very bottom contributes just unity. Non-native conformations contribute nothing. Using this measure the final population contains in excess of 20 per cent of near-native conformations, its native score exceeds 800, increasing 60-fold during simulation phases (2) and (3).

17.3 CONCLUSION

Since the native structure dominates the low-energy conformations arising in all of these simulations, our results demonstrate the feasibility of all-atom protein tertiary structure prediction for three different proteins ranging from 20–60 amino acids in length with a variety of different optimization methods. The free energy approach thus emerges as a viable trade-off between predictivity and computational feasibility. While sacrificing the folding dynamics, a reliable prediction of its terminus, the native conformation – which is central to most biological questions – can be achieved.

The computational advantage of the optimization approach stems from the possibility of visiting unphysical intermediate conformations with high energy during the search. This goal is realized with different mechanisms in all of the employed stochastic optimization methods. In the stochastic tunneling method, the nonlinear transformation of the PES permits the dynamical process to traverse abritrarily high energy barriers at low termperatures; in basin hopping and parallel tempering, simulation phases at very high temperatures accomplish the same objective.

Our results indicate that the simple basin hopping method is very efficient in the determination of the global optimum of the free energy surface of realistic all-atom protein models. It is encouraging that the same structure was also found for the trp-cage protein using the parallel tempering and the stochastic tunneling method and also for the HIV accessory protein using the parallel tempering method. This finding indicates that the result of the folding approach is not an artefact of the optimization strategy. In direct comparison, however, we find that the basin hopping technique gave the lowest energies. Since it is virtually parameter-free, and very simple to implement, it emerges as a natural work-horse for our approach.

Its one important distavantage is the fact that different basin hopping simulations are completely independent of one another. Because the underlying optimization problem in all-atom protein folding is very difficult, and the free energy surface is very rugged, several simulations must be undertaken to obtain a relative degree of confidence in the convergence of the approach. From this perspective it is desirable to design an optimization method in which different members of the simulated population can learn from one another. In the evolutionary approach we have explored, here is one particularly simple scheme to ensure that the total computational effort is concentrated on the best conformations that arise at intermediate stages of the simulations. So far it leads to the folding of the largest protein at all-atom resolution, but much work remains to be done to optimize this approach.

17.4 DISCUSSION

This review indicates that all-atom protein structure prediction with stochastic optimization methods becomes feasible with present day computational resources. The fact that three proteins were reproducibly folded with different optimization methods to near-native conformation increases the confidence in the parameterization of our all-atom protein force field PFF01. The presently available evidence indicates that the comparatively straightforward basin hopping routine is a good work-horse to evolve individual conformations. The resolution of several independent basin hopping simulations may be enhanced by the use of evolutionary algorithms, such as the one used for the bacterial ribosomal protein L20 . We note that the master–client model for this strategy is asynchronous and can be implemented outside the simulation program using the standard TCP/IP, FTP or HTTP protocols. Using generic libraries, such as MPI, it can also be easily implemented on MPP architectures. Since the amount of information that needs to be exchanged is very small, while the effort in a single basin hopping cycle is substantial, there is virtually no loss in an asynchronous parallel implementation. This makes the evolutionary approach investigated here suitable for present day GRID architectures. While the present results demonstrate proof of principle, much work remains to be done to arrive at an optimal strategy.

Protein structure prediction with stochastic optimization methods requires two separate key ingredients: an accurate force field and efficient optimization techniques. One cannot overemphasize the importance of the interplay of optimization methods and force field validation. Rational force field development mandates the ability to generate decoys that fully explore competing low-energy conformations to the native state. The success of different optimization strategies depends strongly on the structure of the potential energy surface. As a result, the development of efficient optimization techniques for an all-atom protein structure prediction depends on the availability of a force field that stabilizes native conformations of proteins with appreciable hydrophobic cores. For helical proteins the bottleneck in *ab initio* all-atom structure prediction now lies in the development of optimization strategies that significantly increase the system size and increase the reliability of the predictions. Based on the results reviewed here, it is sensible to investigate improvements of the evolutionary techniques based on the basin hopping method for this purpose.

The application of this methodology to a wide range of proteins will generate large decoy sets of metastable conformations that compete with the native structure of the protein. These decoy sets may in turn be used to improve the parameterization of the force field. By its very nature, the approximation of the free energy of the system mandates the use of implicit solvent models. This implies that interactions with the solvent and intramolecular electrostatic interactions must be parameterized in accurate, yet efficient effective models. Since both effects are highly nontrivial, the free-energy approach can only approximate, but not duplicate, the results of all-atom explicit water simulations. The present evidence indicates that the native conformation is reproduced to 3–4 Å resolution with PFF01, but the results of the free-energy approach could be refined in all-atom simulations that start from a set of low-energy decoys.

17.5 ACKNOWLEDGEMENTS

We acknowledge the BMBF, the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2, WE 1863/14-1) and the Kurt Eberhard Bode Stiftung for financial support. Parts of the simulations were performed at the KIST teraflop cluster.

17.6 REFERENCES

1 D. Baker and A. Sali, Science, 294 (2001) 93.
2 J. Schonbrunn, W.J. Wedemeyer and D. Baker, Curr. Op. Struc. Biol., 12 (2002) 348.
3 N. Go and H.A. Scheraga, Macromolecules, 9 (1976) 535.
4 P. Ulrich, W. Scott, W.W.F. van Gunsteren and A.E. Torda, Proteins, SF&G, 27 (1997) 367.
5 C.D. Snow, H. Nguyen, V.S. Pande and M. Gruebele, Nature, 420 (2002) 102.
6 C. Simmerling, B. Strockbine and A. Roitberg, J. Am. Chem. Soc., 124 (2002) 11258.
7 C.B. Anfinsen, Science, 181 (1973) 223.
8 Z. Li and H. Scheraga, Proc. Nat. Acad. Sci. USA, 84 (1987) 6611.
9 A. Schug, T. Herges and W. Wenzel, Phys. Rev. Lett., 91 (2003) 158102.
10 T. Herges and W. Wenzel, Phys. Rev. Lett., 94 (2004) 018101.
11 T. Herges and W. Wenzel, Biophys. J., 87 (5) (2004) 3100.
12 A. Schug, T. Herges and W. Wenzel, Europhyics Lett., 67 (2004) 307.
13 A. Schug, T. Herges and W. Wenzel, Proteins, 57 (2004) 792.
14 A. Schug, T. Herges and W. Wenzel, J. Am. Chem. Soc., 126 (2004) 16736.
15 H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata and I. Shimanda, Biochemistry, 40 (1992) 9665.
16 U. Mayor, N.R. Guydosh, C.M. Johnson, J.G. Grossmann, S. Sato, G.S. Jas, S.M.V. Freund, D.O.V. Alonso, V. Daggett and A. R. Fersht, Nature, 421 (2003) 863.
17 T. Herges, H. Merlitz and W. Wenzel, J. Ass. Lab. Autom, 7 (2002) 98.
18 R. Abagyan and M. Totrov, J. Molec. Biol., 235 (1994) 983.
19 T. Herges, A. Schug, B. Burghardt and W. Wenzel, Intl. J. Quant. Chem., 99 (2004) 854.
20 F. Avbelj and J. Moult, Biochemistry, 34 (1995) 755.
21 D. Eisenberg and A.D. McLachlan, Nature, 319 (1986) 199.
22 K.A. Sharp, A. Nicholls, R. Friedman and B. Honig, Biochemistry, 30 (1991) 9686.
23 W. Wenzel and K. Hamacher, Phys. Rev. Lett., 82 (1999) 3003.
24 A. Nayeem, J. Vila and H. Scheraga, J. Comp. Chem., 12 (5) (1991) 594.
25 J.P. Doye and D. Wales, J. Chem. Phys., 105 (1996) 8428.
26 G.J. Geyer, Stat. Sci., 7 (1992) 437.
27 K. Hukushima and K. Nemoto, J. Phys. Soci. Japan, 65 (1996) 1604.
28 H. Merlitz and W. Wenzel, Chem. Phys. Lett., 362 (2002) 271.
29 H. Merlitz, B. Burghardt and W. Wenzel, Chem. Phys. Lett., 370 (2003) 68.
30 U. Hansmann and Y. Okamoto, J. Comput. Chem., 18 (1997) 920.
31 U. Hansmann, Eur. Phys. J. B, 12 (1999) 607.
32 C. Lin, C. Hu and U. Hansmann, Proteins, 53 (2003) 436.
33 S. Kirkpatrick, C. Gelatt and M. Vecchi, Science, 220 (1983) 671.
34 J. Schneider, I. Morgenstern and J. Singer, Phys. Rev. E, 58 (1998) 5085.
35 A. Schug, A. Verma, T. Herges and W. Wenzel, Proteins, (2005) submitted.
36 J.W. Neidigh, R.M. Fesinmeyer and N.H. Anderson, Nature Struct. Biol., 9 (2002) 425.

AUTHOR QUERY FORM

ELSEVIER

Modern Methods for Theoretical Physical Chemistry of Biopolymers

JOURNAL TITLE:	**MMTP-STARIKOV**
ARTICLE NO:	**Ch017**

Queries and / or remarks

Query No	Details required	Author's response
AQ1	Please indicate what "tbldecoyhiv" means	
AQ2	Please explain what XXX stands for in the sentence "The RMSB......to the native decoy"	
AQ3	Is the book published in ref. [35]?	

AUTHOR QUERY FORM

ELSEVIER

Modern Methods for Theoretical Physical Chemistry of Biopolymers

JOURNAL TITLE:	**MMTP-STARIKOV**
ARTICLE NO:	**Ch017**

Queries and / or remarks