Comparison of Stochastic Optimization Methods for All-Atom Folding of the Trp-Cage Protein

Alexander Schug,[a] Thomas Herges,[a] Abhinav Verma,[b] Kyu Hwan Lee,[c] and Wolfgang Wenzel*[a]

The performances of three different stochastic optimization methods for all-atom protein structure prediction are investigated and compared. We use the recently developed all-atom free-energy force field (PFF01), which was demonstrated to correctly predict the native conformation of several proteins as the global optimum of the free energy surface. The trp-cage protein (PDB-code 1L2Y) is folded with the stochastic tunneling method, a modified parallel tempering method, and the basin-hopping technique. All the methods correctly identify the native conformation, and their relative efficiency is discussed.

1. Introduction

Ab initio protein tertiary structure prediction (PSP) and the elucidation of the mechanism of the folding process are among the important outstanding problems of biophysical chemistry.[1,2] The many complementary proposals for PSP span a wide range of representations of the protein conformation, which range from coarse grained models to atomic resolution. The choice of representation often correlates with the methodology employed in structure prediction, which in turn determines the computational cost of the approach.

We have recently demonstrated a feasible strategy for all-atom protein structure prediction[3–5] 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.[5] We have already demonstrated the reproducible and predictive folding of three proteins: the 20 amino acid trp-cage protein (1L2Y),[3,6] the structurally conserved headpiece of the 40 amino acid HIV accessory protein (1F4I),[4,7] and the 60 amino acid bacterial ribosomal protein L20.[8] In addition, we could show that PFF01 stabilizes the native conformations of other proteins, for example, the 52 amino acid protein A[9,10] and the engrailed homeodomain (1ENH) from *Drosophila melanogaster*[11] as the global optimum of the free energy model.

All-atom methods, even with implicit solvent, are clearly among the computationally most demanding strategies for protein structure prediction/folding. Significant computational resources are required for this approach, and therefore it is important to compare the efficiency of different optimization strategies. One important advantage of optimization-based techniques results from their ability to quickly locate the native conformation without recourse to the physical folding dynamics or pathway. This implies that the simulation may pass through unphysical conformations or jump large distances in the conformational space. Little is presently known about the efficiency of different optimization methods for all-atom protein folding. Herein, we investigate three different optimization techniques and compare them with respect to their efficiency. Using all three techniques we have reproducibly folded the 20 amino acid trp-cage protein,[12] one of the fastest folding proteins known. All methods converge to near-native conformations, thus increasing confidence in the reliability of the underlying force field PFF01. However, we find that the techniques differ in their ability to really resolve the low-energy region of the free energy surface, where a modified version of the basin-hopping approach performs best. Since "in silico" (computer-simulated) protein folding has rightfully been compared with the search for a needle in the proverbial haystack, small differences in energy resolution can significantly influence the reliability of the prediction.

2. Methods

2.1. 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.[4,5,13] The force field, which strictly

[a] *Dr. A. Schug, Dr. T. Herges, Dr. W. Wenzel*
Forschungszentrum Karlsruhe, Institut für Nanotechnologie
P.O. Box 3640, 76021 Karlsruhe (Germany)
Fax: (+49) 7247 82 6434
E mail: wenzel@int.fzk.de

[b] *A. Verma*
Forschungszentrum Karlsruhe, Institut für wissenschaftliches Rechnen
P.O. Box 3640, 76021 Karlsruhe (Germany)

[c] *Dr. K. H. Lee*
Supercomputational Materials Lab
Korean Institute of Science and Technology, Seoul (Korea)

speaking parameterizes the internal free energy of the protein excluding backbone entropy, is parameterized with the following nonbonded interactions [Eq. (1)]:

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

where r_{ij} denotes the distance between atoms i and j and $g(i)$ the type of 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.[13–15] The nontrivial electrostatic interactions in proteins are represented by 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.[16] Interactions with the solvent were first fitted in a minimal solvent-accessible surface model[17] parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides.[18] 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 atoms along the bond, and the angle between the CO and NH axis.[5]

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 side-chain and backbone dihedral angles, which are attempted with 30 and 70% probability, respectively. The moves for the side-chain 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 are generated from a move library that was designed to reflect the natural amino-acid-dependent bias toward 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.[19] While driving the simulation toward the formation of a secondary structure, it contains no bias toward helical or sheet structures beyond that encountered in nature. Notably, the large-scale moves generated are likely to be accepted only at very high temperatures or at the very start of the simulation. At low temperature their acceptance probability falls to zero.

The energy evaluations in our model are significantly faster than those in most molecular dynamics (MD) simulations. In particular for side-chain moves, only very few atoms change position in many attempted moves. Our simulation package takes advantage of this fact by evaluating only those interactions that have changed. In addition, the full energy evaluation is omitted for moves that result in clashing conformations. Since over 60% of the computational effort is spent in the calculation of the solvent term, this also significantly speeds up the energy evaluation.

2.2. Optimization Methods

The low-energy part of the free energy landscape of proteins is extremely rugged due to the comparatively close packing of the atoms in the native structure. Rugged potential energy surfaces (PESs) are characterized by the existence of many low-lying minima, which are separated by high-energy barriers. For this reason, the global optimum of such a surface is difficult to obtain computationally. The presently available evidence indicates that optimization-based protein structure prediction falls into this class of problems, because simple optimization methods, such as steepest descent or simulated annealing, are almost always trapped in metastable conformations. Suitable optimization methods must therefore be able to speed up the simulation by avoiding high-energy transition states, by adapting large-scale moves wherever possible or by accepting unphysical intermediates. Here we investigate three different optimization methods: the stochastic tunneling method,[20] the basin-hopping technique,[21,22] and the parallel tempering method.[23,24] 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 is an inherently parallel optimization strategy that is well-suited to the 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 optimization methods that are capable of exploiting such architectures; that is, a high degree of parallelism with very little and optimally asynchronous communication is desirable.

2.2.1. Basin-Hopping Method

One of the simplest ideas to effectively eliminate high-energy transition states of the potential or free energy surface is employed in the basin-hopping technique[21] (BHT), also known as Monte Carlo with minimization. This method simplifies the original PES by replacing the energy of each conformation with the energy of a nearby local minimum (see Figure 1). 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 by the increase in efficiency of the stochastic search on the simplified PES. The basin-hopping technique and derivatives[14] have been used previously to study the PES of model proteins[25] and polyalanines using all-atom models.[26,27] In contrast to this work, we use a simulated annealing process for the minimization step, because analytical gradients for the SASA implicit solvent model of our force field are computationally very difficult to obtain.

For the protein simulations we replace a single minimization step with a simulated annealing (SA) run.[28] Within each SA simulation, new configurations are accepted according to the

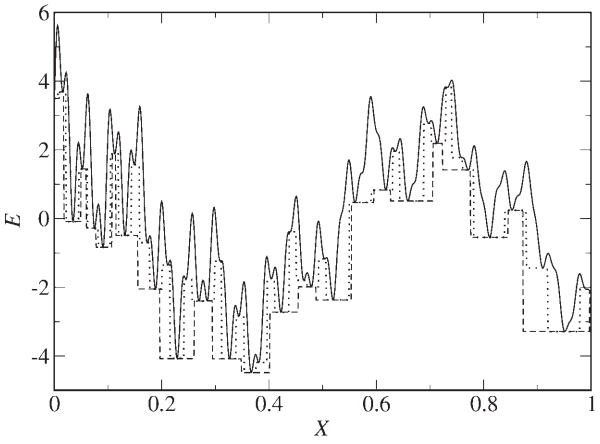

Figure 1. Schematic one dimensional potential energy surface ( ) and its mapped surfaces in basin hopping with local minimization (•••••) and basin hopping with minimization by simulated annealing ( ). The dotted curve is obtained by mapping each point of the original potential to its closest local minimum. The dashed curve is obtained by permitting the search to overcome intervening barriers of an average height that corresponds to kT_{av}, where T_{av} is a suitably averaged temperature of the simulated annealing run.

Metropolis criterion. The temperature is decreased geometrically from its starting value to the final value, which must be chosen as small compared to typical energy differences between competing metastable conformations, to ensure convergence to a local minimum (typically 2–5 K). Depending on the choice of 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 temperature (T_s), the final temperature, and the number of steps. We investigated various choices for the numerical parameters of the method, but have always used a geometric cooling schedule.

Each SA run is typically much more expensive than local minimization using gradient-based techniques, but it can nevertheless be competitive for very rugged PESs, or when the computation of the gradient of the potential is prohibitive. In a very rugged PES, such as that illustrated in Figure 1, strict local minimization changes the conformations only little, while SA-based minimization results in a significant further reduction of the complexity of the PES. In our model the computation of the gradient is much more expensive than the computation of the energy, because the SASA term involves the numerical integration of the atomic surfaces. To evaluate the gradient accurately, the number of integration points must be increased significantly.

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 proven optimal for certain optimization problems.[29] Throughout this study we use a threshold acceptance criterion of 1 kcal mol^{-1}.

2.2.2. Stochastic Tunneling Method

The stochastic tunneling technique (STUN)[20] was proposed as a generic global optimization method for complex rugged PESs. For a number of problems, including the prediction of receptor–ligand complexes for drug development,[30,31] this technique proved superior to competing stochastic optimization methods. The idea behind the method is to flatten the PES in all regions that lie significantly above the best estimate for the minimal energy (E_0). In STUN the dynamical process explores not the original, but a transformed PES [Eq. (2)],

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

which dynamically adapts and simplifies during the simulation (see Figure 2). 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[20] γ controls the steepness of the transformation [we used $\gamma = 0.5$ (kcal/mol)$^{-1}$]. The transformation in Equation (2) ameliorates the difficulties associated with the original transformation,[20] because $E_{\mathrm{STUN}} \propto \ln(E/kT)$ continues to grow slowly for large energies. The fictitious temperature of STUN must be adjusted to accelerate convergence.[3] 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 $E \gg E_0$ (see Figure 2), 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. In comparison to the basin-hopping approach, there is no need for extensive local minimization, but the nonlinear transformation tends to make the high-energy dynamics diffusive.

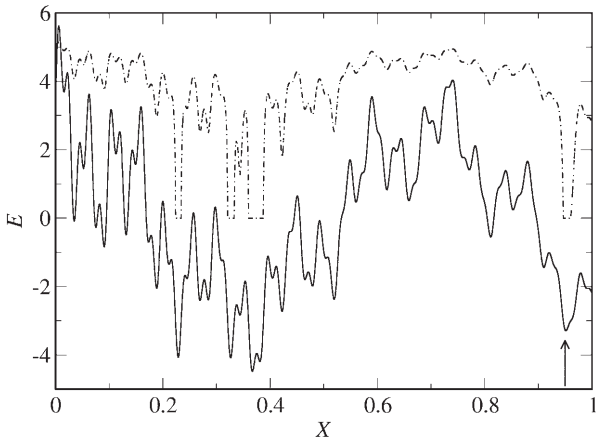

Figure 2. Schematic one dimensional potential energy surface ( ) (same as in Figure 1) and the effective potential energy surface after the STUN transformation, assuming that the local minimum indicated by the arrow has already been found. The effective potential energy surface is truncated at zero. The remaining relevant minima for the search are still very pro nounced, but the high energy features of the PES are significantly smooth ed. The diffusion time to cross the barrier from $x \approx 0.95$ to the global mini mum is significantly reduced by the near elimination of the many interven ing local minima in the effective potential energy surface.

2.2.3. Parallel Tempering

The parallel (or simulated) tempering (PT) technique[23,24] was introduced to overcome difficulties in the evaluation of thermodynamic observables for models with very rugged PESs and was applied previously in several protein-folding studies.[32–34] Low-temperature simulations on rugged PESs 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 [Eq. (3)]:

$$p = \min\{1, \exp[\quad(\beta_j \quad \beta_i)(E_i \quad E_j)]\} \tag{3}$$

where $\beta_i = 1/k_B T_i$ 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. Thus, for proteins the temperatures must fall in a bracket of approximately 2–1000 K. As described elsewhere,[6] 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%, then all temperatures above t_i were lowered by 10% of $t_{i+1} \quad t_i$. If the exchange rate was above 2%, then all temperatures above t_i were increased by the same difference. These exchange rates are very small compared with standard MD implementations for protein folding,[35] which results from the large range of temperatures that must be spanned by the optimization approach in comparison to MD.

To further improve the computational efficiency of PT we also use 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 PT method was implemented in our program using the MPI communication library, which is available on most present-day parallel computational architectures 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.

3. Results

First we investigated the folding of the 20 amino acid trp-cage protein[12,36] (PDB code 1L2Y) with the basin-hopping technique. We noted that very high starting temperatures are required to permit a sufficient exploration of the free energy surface. The lowest temperature had to be chosen in the range of 2–6 K to ensure that local minima were well-resolved. 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. For all simulations reported here we used a starting temperature of $T_s = 800$ K and a final temperature of $T_f = 3$ K. All simulations were started from a sticklike unfolded conformation with no secondary structure and a root-mean-square backbone (RMSB) deviation of 12.94 Å. 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. We performed 20 independent simulations comprising basin-hopping steps of constant length ($N = 10\,000$, 1000 cycles) and 20 independent simulations where the length of the individual step increased with the square root of the cycle number m ($N = 10\,000 \times \sqrt{m}$, 150 cycles). Figure 3 demonstrates that the overall structure of the free energy surface is explored well in both sets of simulations, but that the simulations with increasing cycle length reached much lower energies.

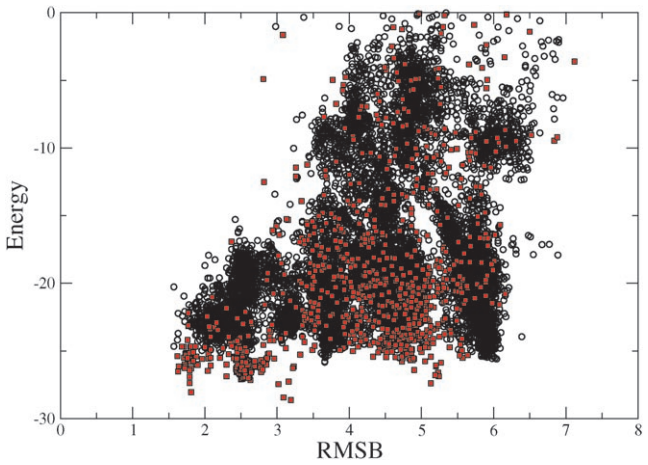

Figure 3. RMSB versus energy plot of all accepted conformations in the energy window between zero and 30 kcal mol ^{1} for the basin hopping simulations with constant length (○) and increasing length (red square). The total number of function evaluations was the same, so there are fewer observations in the latter simulation, which nevertheless reaches lower ener gies.

A total of 12 of these simulations approached the native conformation as its estimate of the optimum. The energies and (RMSB) deviations of these conformations are shown in Table 1. The best conformation had an energy of 28.63 kcal mol^{-1} and a RMSB deviation of 3.19 Å to the native conformation. Its overlay with the native structure is illustrated in Figure 4. The second-best configuration has a RMSB value of only 1.81 Å and loses energy by only about 0.6 kcal mol^{-1}. Figure 3 illustrates that there are at least four to five distinct families of metastable conformations. The plot indicates the existence of a set of structures with 1.5–3 Å RMSB deviation, which may correspond to the folding funnel, and a competing metastable conformation with about 5 Å RMSB. This competing conformation appears seventh in the decoy table, with an

Table 1. Energies, RMSB deviations, and secondary structure content of the decoys for the trp cage protein generated in 20 independent basin hopping simulations with increasing length per cycle. Note that the best and the second best decoys differ only in the position of the turn sepa rating the two helices, which completely destroys the tertiary structure.

Energy	RMSB	Three state secondary structure
28.631	3.19	CHHHHHHTTHHHHHTCCSCC
28.051	1.81	CHHHHHHHHTHHHHTCCSCC
27.159	2.63	CHHHHHHHHTHHHHTCTTTC
27.073	2.52	CHHHHHHHHTHHHHTCTTTC
26.727	2.48	CHHHHHHHHTHHHHTCTTTC
26.437	2.55	CHHHHHHHHTHHHHTCTTTC
26.413	4.90	CHHHHHTCTTHHHHHCTTTC
26.205	2.55	CHHHHHHHHTHHHHTCTTTC
25.969	2.55	CHHHHHHHHTHHHHTCTTTC
25.738	1.84	CHHHHHHHHTHHHHTCCSCC
25.240	2.33	CHHHHHHHHHHHHHCCSSCC
25.091	4.52	CHHHHHHHHTCSSTTSTTTC
24.865	2.07	CHHHHHHHHTHHHHTCCSCC
24.824	4.98	CHHHHHHHHHSSSTTSCSCC
24.514	4.61	CHHHHHHHHTSCCTTCTTTC
23.477	2.89	CHHHHHHHHSHHHHHCTTTC
23.290	4.74	CHHHHHHHHTCSSSSSTTTC
22.874	4.41	CHHHHHHHHTCSCTTCSSCC
22.649	5.08	CHHHHHHHHTCSSCCCTTTC
20.548	5.28	CCBSSSCBSSHHHHTCTTTC

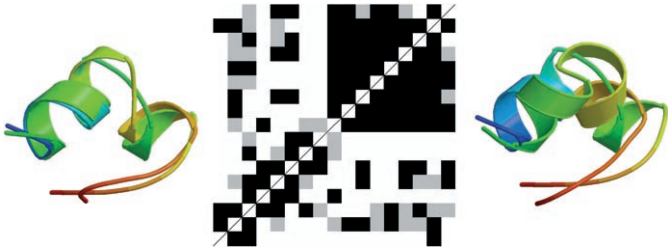

Figure 4. Overlay of the native and folded structures of trp cage protein (left) and the corresponding C_β C_β matrix (center). A pixel in row *i* and column *j* of the color coded distance map indicates the difference in the C_β C_β distances of the native and the folded structure. Black (gray) squares indi cate that the C_β C_β distances of the native and the other structure differ by less than 1.5 (2.25) Å, respectively. White squares indicate larger deviations. The right panel shows the misfolded conformation associated with the sev enth decoy in Table 1. The experimental conformation has the less pro nounced second helix (blue terminus).

energy difference of less than 2 kcal mol^{-1} to the native conformation. In secondary structure it differs from the native conformation only in the position of the turn between the first and second helix. This misfolded conformation is shown in Figure 4 (right).

We also performed 25 independent simulations of the same protein with a modified version of the stochastic tunneling method.[3,20] The length of each simulation was 1.2×10^7 steps, comparable to that of the basin-hopping simulation with increasing cycle length reported above. Six of 25 simulations reached an energy within 1 kcal mol^{-1} of the best energy of 25.73 kcal mol^{-1}, all of which correctly predicted the native experimental structure of the protein. There was a strong correlation between energy and RMSB deviation to the native structure for all simulations. The conformation with the lowest energy had a RMSB deviation of 2.83 Å. Both tunneling phases and local-search phases, corresponding to small and large values of the transformed energy, respectively, are required to converge the simulations.

This protein was also folded with the parallel tempering method.[6] 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. Convergence of the method was found using eight to 30 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 the high and low temperatures required for the speedy exploration of the free energy surface and the refinement of local minima, respectively. The total numerical effort of the parallel tempering method is the product of the replica number and the steps per replica. Therefore, this method can only be competitive if the high-temperature replicas, which never generate good low-energy decoys, significantly speed the search of the PES.

Figure 5 shows the energies and corresponding temperatures for a representative simulation using ten replicas. The temperature adjustment scheme results in a temperature dis-

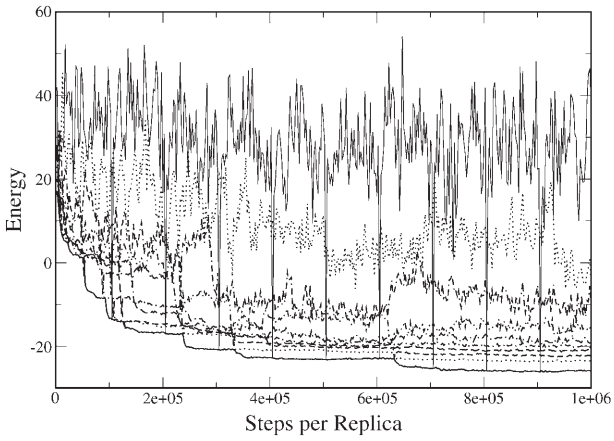

Figure 5. Energy versus step number diagram of a representative ten replica modified parallel tempering simulation of the trp cage protein. The data demonstrate the convergence of the energy and the rapid exchange of in formation between the different replicas as discussed in the text.

tribution that permits frequent exchange of replicas and significantly speeds convergence. Here we performed ten independent parallel-tempering simulations with 10^6 function evaluations for each of the ten replicas used. Figure 5 demonstrates the convergence of the method; frequent exchange of the replicas leads to the generation of new optimal conformations, as is evident from the crossing of the lines between steps 4–6×

10^5, which result in the generation of a new best structure after about 6.4×10^5 steps. Figure 5 nicely illustrates that significantly new minima are never generated from the lowest-temperature replica, but require the exchange mechanism. The near-vertical peaks of the highest-temperature simulation result from the replication step discussed in the Methods section. This mechanism generated a new best structure after the replication step with 3.5×10^5 steps, cascading down through the replicas and reaching the lowest temperature replica at step 3.7×10^5. The best final structure associated with the lowest temperature had an energy of -25.3 kcal mol^{-1} and a RMSB deviation of 3.3 Å.

4. Conclusions

In agreement with previous studies,[25,26] 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 by using the parallel tempering and stochastic tunneling methods. This finding indicates that the result of the folding approach is not an artifact of the optimization strategy. In direct comparison, however, we found 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 workhorse for our approach.

The energies of the best conformation and its RMSB deviations for all simulations are shown in Figure 6. Figure 6 clearly

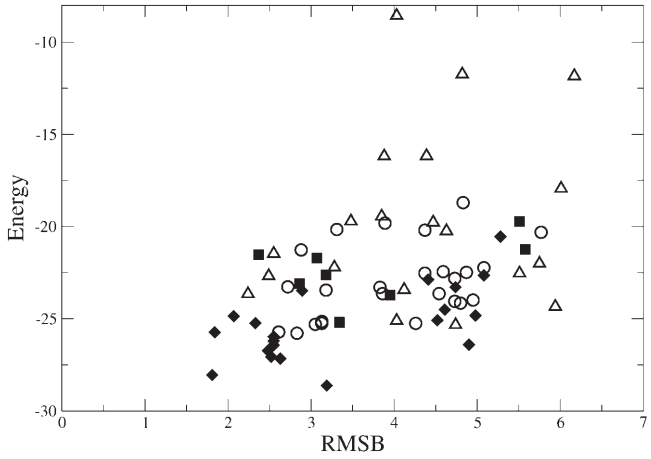

Figure 6. Energy versus RMSB plot for the final energies of the 20 basin hop ping simulations with and without increasing cycle length (◆ and △, respec tively). For comparison we also indicate the best energy result for the STUN method (○) and for the 30 processor PT simulation (■).

demonstrates that the basin-hopping approach is most efficient, but only when the version with increasing cycle length is being used. Figure 6 illustrates nicely that searching for the global optimum of an all-atom protein energy landscape can be compared to the search for a needle in a haystack. The accuracy and reliability of the predictions depend strongly on the availability of efficient optimization methods to explore the landscape.

The rough hierarchy of the parallel tempering, stochastic tunneling, and basin-hopping techniques that emerges from this study can be rationalized by a comparison of the underlying methods. We noted in the Introduction that optimization-based strategies for protein folding/structure prediction have an advantage in comparison with simulations of the folding pathway because unphysical conformations can be visited in the search. In the parallel tempering method, such unphysical conformations (at the physiological temperature) are generated in the high-temperature replicas, which is a comparatively mild relaxation of the energetic constraints selecting the native conformation. The stochastic tunneling method, by comparison, applies a very nonlinear transformation to the underlying free energy surface, and as a result the simulation may pass through even more unphysical regions of the conformation space. The relative success of this approach (in comparison to parallel tempering) arises because the gradient of the transformed PES appears to be still sufficiently strong to guide the simulation back into the physical realm.

The basin-hopping approach, just as with parallel tempering, generates unphysical conformations at very high temperatures. Basin-hopping simulations that heat only moderately do not explore the conformation space well. Therefore, it appears at first sight surprising that basin-hopping methods outperform the stochastic tunneling technique (at least for some parameterizations). The basin-hopping approach brings a new feature to the family of optimization methods investigated in this study: the simulation time is no longer continuous. The threshold acceptance criterion of the basin-hopping approach introduces a new element that is never possible in a physical simulation: the possibility of discarding a part of the trajectory and restarting at some earlier point in time. In contrast to the other techniques, basin-hopping simulations remember the (typically) best configuration attained so far and thus have the ability to discard search processes that have gone astray (about 60% of the cycles are rejected). To demonstrate that this feature significantly contributes to the success of this method we performed a set of simulations comprising a single simulated annealing run in the same temperature bracket with the same total number of function evaluations, none of which achieved energies even below -20 kcal mol^{-1}.

5. Summary

The native structure dominates the low-energy conformations arising in all of these simulations, and thus our results demonstrate that the trp-cage protein is folded to about 3 Å RMSB resolution in the PFF01 force field. This resolution is comparable to other implicit solvent simulations for the same protein,[36] but could be significantly improved in all-atom simulations. Nevertheless, the free energy approach 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. Different mechanisms realize this principle in the different optimization methods: in the stochastic tunneling method, the nonlinear transformation of the PES permits the dynamical process to traverse arbitrarily high-energy barriers at low temperatures; in basin hopping, the PES is simplified through the mapping to a smoother surface; and in parallel tempering, simulation phases at very high temperatures accomplish the same objective.

The data indicate that the comparably straightforward basin-hopping routine is a good workhorse for evolving individual conformations. Similar results were obtained in a recent study[37] on short $\alpha\alpha$, $\beta\beta$, and mixed $\alpha\beta$ peptides using a modified free energy model based on the ECEPP3[38] potential. The threshold acceptance step can be performed outside the simulation program that runs the basin-hopping cycle and distributed among any number of nodes using standard communication protocols. This makes the basin-hopping technique suitable for GRID-type architectures, which presently deliver high computational power at superior price/performance ratios.

The future area of application of all-atom protein structure prediction with optimization methods in a free energy model depends on the availability of efficient optimization methods to perform the underlying simulations. With this investigation we have contributed to a much-needed systematic study of the suitability of different optimization methods for this very important problem. The results need to be confirmed and compared with work on other proteins and optimization strategies. The findings provide suitable benchmarks for a realistic and widely studied system that can be investigated with comparatively modest computational effort. In comparison to the numerical effort of our techniques prior to this study, the average in silico folding time was cut by about one order of magnitude through the systematic investigation of different optimization strategies, which indicates that significant progress can still be made on this computationally intensive problem.

Acknowledgements

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

Keywords: conformation analysis · optimization · protein folding · protein structures · stochastic processes

[1] D. Baker, A. Sali, *Science* **2001**, *294*, 93 96.
[2] J. Schonbrunn, W. J. Wedemeyer, D. Baker, *Curr. Opin. Struct. Biol.* **2002**, *12*, 348 352.
[3] A. Schug, T. Herges, W. Wenzel, *Phys. Rev. Lett.* **2003**, *91*, 158 102.
[4] T. Herges, W. Wenzel, *Phys. Rev. Lett.* **2005**, *94*, 018 101.
[5] T. Herges, W. Wenzel, *Biophys. J.* **2004**, *87*, 3100 3109.
[6] A. Schug, W. Wenzel, *Europhys. Lett.* **2004**, *67*, 307 313.
[7] A. Schug, T. Herges, W. Wenzel, *Proteins* **2004**, *57*, 792 798.
[8] A. Schug, T. Herges, W. Wenzel, *J. Am. Chem. Soc.* **2004**, *126*, 16 736 16 737.
[9] D. Snow, H. Nguyen, V. S. Pande, M. Gruebele, *Nature* **2002**, *420*, 102 106.
[10] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, I. Shimanda, *Biochem istry* **1992**, *40*, 9665 9672.
[11] 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, A. R. Fersht, *Nature* **2003**, *421*, 863 867.
[12] J. W. Neidigh, R. M. Fesinmeyer, N. H. Anderson, *Nat. Struct. Biol.* **2002**, *9*, 425 430.
[13] T. Herges, H. Merlitz, W. Wenzel, *J. Ass. Lab. Autom.* **2002**, *7*, 98 104.
[14] R. Abagyan, M. Totrov, *J. Mol. Biol.* **1994**, *235*, 983 1002.
[15] T. Herges, A. Schug, W. Wenzel, *Int. J. Quantum Chem.* **2004**, *99*, 854 893.
[16] F. Avbelj, J. Moult, *Biochemistry* **1995**, *34*, 755 764.
[17] D. Eisenberg, A. D. McLachlan, *Nature* **1986**, *319*, 199 203.
[18] K. A. Sharp, A. Nicholls, R. Friedman, B. Honig, *Biochemistry* **1991**, *30*, 9686 9697.
[19] J. T. Pedersen, J. Moult, *J. Mol. Biol.* **1997**, *269*, 240.
[20] W. Wenzel, K. Hamacher, *Phys. Rev. Lett.* **1999**, *82*, 3003 3007.
[21] A. Nayeem, J. Vila, H. A. Scheraga, *J. Comput. Chem.* **1991**, *12*, 594 605
[22] J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111 5116.
[23] G. J. Geyer, *Stat. Sci.* **1992**, *7*, 437 483.
[24] K. Hukushima, K. Nemoto, *J. Phys. Soc. Jpn.* **1996**, *65*, 1604 1608.
[25] J. Wales, P. E. J. Dewbury, *J. Chem. Phys.* **2004**, *121*, 10 284 10 290.
[26] P. N. Mortenson, D. J. Wales, *J. Chem. Phys.* **2001**, *114*, 6443 6454.
[27] P. N. Mortenson, D. A. Evans, D. J. Wales, *J. Chem. Phys.* **2002**, *117*, 1363 1376.
[28] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **1983**, *220*, 671 680.
[29] J. Schneider, I. Morgenstern, J. M. Singer, *Phys. Rev. E* **1998**, *58*, 5085 5095.
[30] H. Merlitz, W. Wenzel, *Chem. Phys. Lett.* **2002**, *362*, 271 277.
[31] H. Merlitz, B. Burghardt, W. Wenzel, *Chem. Phys. Lett.* **2003**, *370*, 68 73.
[32] U. H. E. Hansmann, Y. Okamoto, *J. Comput. Chem.* **1997**, *18*, 920 933.
[33] U. H. E. Hansmann, *Eur. Phys. J. B* **1999**, *12*, 607 612.
[34] C. Y. Lin, C. K. Hu, U. H. Hansmann, *Proteins* **2003**, *53*, 436 445.
[35] A. E. Garcia, N. Onuchic, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 13 898 13 903.
[36] C. Simmerling, B. Strockbine, A. Roitberg, *J. Am. Chem. Soc.* **2002**, *124*, 11 258 11 259.
[37] R. A. Abagyan, M. Totrov, *J. Comput. Phys.* **1999**, *151*, 402 421.
[38] G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paternali, A. Zagari, S. Rumsey, H. A. Scheraga, *J. Phys. Chem.* **1992**, *96*, 6472 6484.