Investigation of the parallel tempering method for protein folding

Alexander Schug, Thomas Herges, Abhinav Verma and Wolfgang Wenzel[1]

Forschungszentrum Karlsruhe, Institut für Nanotechnologie, PO Box 3640, 76021 Karlsruhe, Germany

E-mail: wenzel@int.fzk.de

Abstract
We investigate the suitability and efficiency of an adapted version of the parallel tempering method for all-atom protein folding. We have recently developed an all-atom free energy force field (PFF01) for protein structure prediction with stochastic optimization methods. Here we report reproducible folding of the 20-amino-acid trp-cage protein and the conserved 40-amino-acid three-helix HIV accessory protein with an adapted parallel tempering method. We find that the native state, for both proteins, is correctly predicted to 2 Å backbone root mean square deviation and analyse the efficiency of the simulation approach.

(Some figures in this article are in colour only in the electronic version)

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 which has confined all-atom protein structure prediction to just 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

[1] Author to whom any correspondence should be addressed.

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 parametrization of the free energy of the protein in its environment exists and that 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], the villin headpiece [13] and the 60-amino-acid bacterial ribosomal protein L20 [14]. In addition we were able to show 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 *Drosophila melanogaster* [16].

However, currently little is known about the suitability and relative efficiency of various stochastic optimization methods for all-atom protein folding with free energy force fields. Because all-atom protein folding requires substantial computational resources it is important to develop techniques that can exploit the most powerful computational architecture currently available, i.e. massively parallel computers with distributed memory. For this reason we have investigated the parallel tempering method [17, 18] (PT) as a possible method for all-atom protein folding. We have applied this technique both to the 20-amino-acid trp-cage protein [19, 6] and to the conserved 40-amino-acid headpiece of the HIV accessory protein [10]. For both proteins we demonstrate predictive folding in the PT method, using 4–30 replicas.

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 [20, 10]. 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 parametrized with the following non-bonded interactions:

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

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 as a set of 138 proteins of the PDB database [21, 20, 22]. The non-trivial electrostatic interactions in proteins are represented via group-specific dielectric constants ($\epsilon_{g(i),g(j)}$ depending on the amino acid to which atom i belongs). The partial charges q_i and the dielectric constants were previously derived in a potential-of-mean-force approach [23]. Interactions with the solvent were first fitted in a minimal solvent accessible surface model [24] parametrized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly–X–Gly family of peptides [25]. 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–NH) which depends on the OH distance, the angle between N, H and O along the bond and the angle between the CO and the NH axis [11]. PFF01 was specifically optimized to fold one helical protein, the villin headpiece, which has only 12% sequence homology to the HIV accessory protein investigated here. So far PFF01 has only been applied to fold α-helical proteins, which are much easier to treat than β-sheet structures. Efforts to extend the methodology to treat such systems are currently under investigation.

The only degrees of freedom considered in the simulation are rotations about the main chain and side-chain dihedral angles. Such moves are attempted with the relative probability of 70% and 30% respectively. The side-chain moves are random changes of the dihedral angle by up to 5°; such moves are also attempted for half of the attempted backbone moves. The remaining backbone moves are drawn from a library [26] and set the backbone dihedral angle to a new angle drawn from a probability distribution that was generated for the particular amino acid from conformations of the PDB database. Such moves significantly speed the simulation as they generate secondary structure, but do not bias the move towards either helices or beta sheets above and beyond their natural probability.

2.2. Parallel tempering

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 the simulation by avoiding high energy transition states, adapting large scale moves or accepting unphysical intermediates.

The parallel tempering technique [17, 18] was introduced to overcome difficulties in the evaluation of thermodynamic observables for models with very rugged potential energy surfaces and applied previously in several protein folding studies [27–29]. 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

$$p = \min(1, \exp(-(\beta_j - \beta_i)(E_i - E_j))), \tag{2}$$

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 requirements to efficiently explore the conformational space and to accurately resolve local minima, respectively. For proteins the temperatures must thus fall in a range between approximately 2 and 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 rate of exchange between adjacent temperatures. If the rate of exchange between temperature i and $i+1$ was below 0.5%, then all temperatures above t_i were lowered by 10% of $t_{i+1} - t_i$. If the exchange rate was above 2%, then all temperatures above t_i were increased by the same difference. To further improve the computational efficiency of PT we also use introduced 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 currently best conformation.

3. Results

Figure 1 shows the energy versus time plot of typical adapted PT simulations for the trp-cage protein (pdb-code: 1L2Y) and the conserved 40-amino-acid headpiece of the HIV accessory protein (pdb-code: 1F4I; sequence: QEKEAIERLK ALGFEESLVI QAYFACEKNE NLAANFLLSQ) respectively. The trp-cage protein is among the fastest folding proteins known and has a compact two-helix native structure. The HIV accessory protein folds into a three-helix bundle with an appreciable hydrophobic core. Both proteins folded with the modified parallel tempering method [12] and attained energies comparable to those obtained independently with other methods. The best energy of the trp-cage protein was -26.25 kcal mol^{-1}, slightly improving on the best previous estimate of -25.8 kcal mol^{-1} obtained with the stochastic tunnelling methods [9], while the best energy of 1F4I was -117 kcal mol^{-1}, compared to -119.5 kcal mol^{-1} obtained with an adapted basin hopping scheme [10].

The accuracy of the predicted structures is illustrated in figure 2, which visualizes close correspondence of the folded and the experimentally observed structures for both proteins in the top panel. The bottom panel shows the C_β–C_β distance maps for both proteins, which correspond to distance constraints similar to those observed in NMR experiments.

For the trp-cage protein the lowest conformation had a backbone root mean square deviation (RMSB) of 2.01 Å at the end of this simulation. Considering the ensemble of final conformations, we find many structures closely resembling the native conformation. The four next lowest conformations (in energy) had RMSB of 2.56, 1.81, 2.91, 3.08 Å. For 1F4I the final conformation with the lowest energy/temperature had 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 RMSB deviations of the next four lowest conformations (all within 1.5 kcal mol^{-1} of the minimal energy) have RMSB deviations of 3.14/2.23/3.78/3.00 Å respectively from the native decoy.

Figure 1 indicates many exchanges of configurations as a function of step number. According to equation (2), the probability for two conformations to exchange rises rapidly when their energy difference shrinks. As a result, rapid replica exchange takes place between simulations at nearby energies. On the other hand, the temperature distribution of the simulations must span the range from a few kelvins, where local optimization takes place, to 600–1000 K where a rapid exploration of the free energy surface can occur. The temperature adjustment scheme described in the methods section serves to equilibrate the temperatures to satisfy these conflicting objectives as well as is possible; in particular the simulation of 1F4I, which was started at somewhat unreasonable temperatures, managed to equilibrate quickly to a nearly stable temperature set, as can be seen in figure 3. We note, however, that the simulations fail to converge if the number of temperatures drops too low; for the trp-cage protein, three of four simulations with only four temperatures failed to reach the NMR structure. In figure 4 we analyse the distribution of the final temperatures, which equilibrates to a near geometric distribution for $T > 1$ K.

The replication step introduced above led to a significant improvement of the relaxation rate of the overall simulation and introduced an element of true parallelism into the optimization scheme (which in the original PT method is mediated only indirectly by the temperature exchange). The use of such a replication step, while focusing the computational effort on the best available structure, has posed a risk of narrowing the search to a small part of the conformational space. Figure 5 illustrates that this has not occurred in the present simulation, where the high temperature simulations continue to generate a large degree of structural diversity in the sampling space.

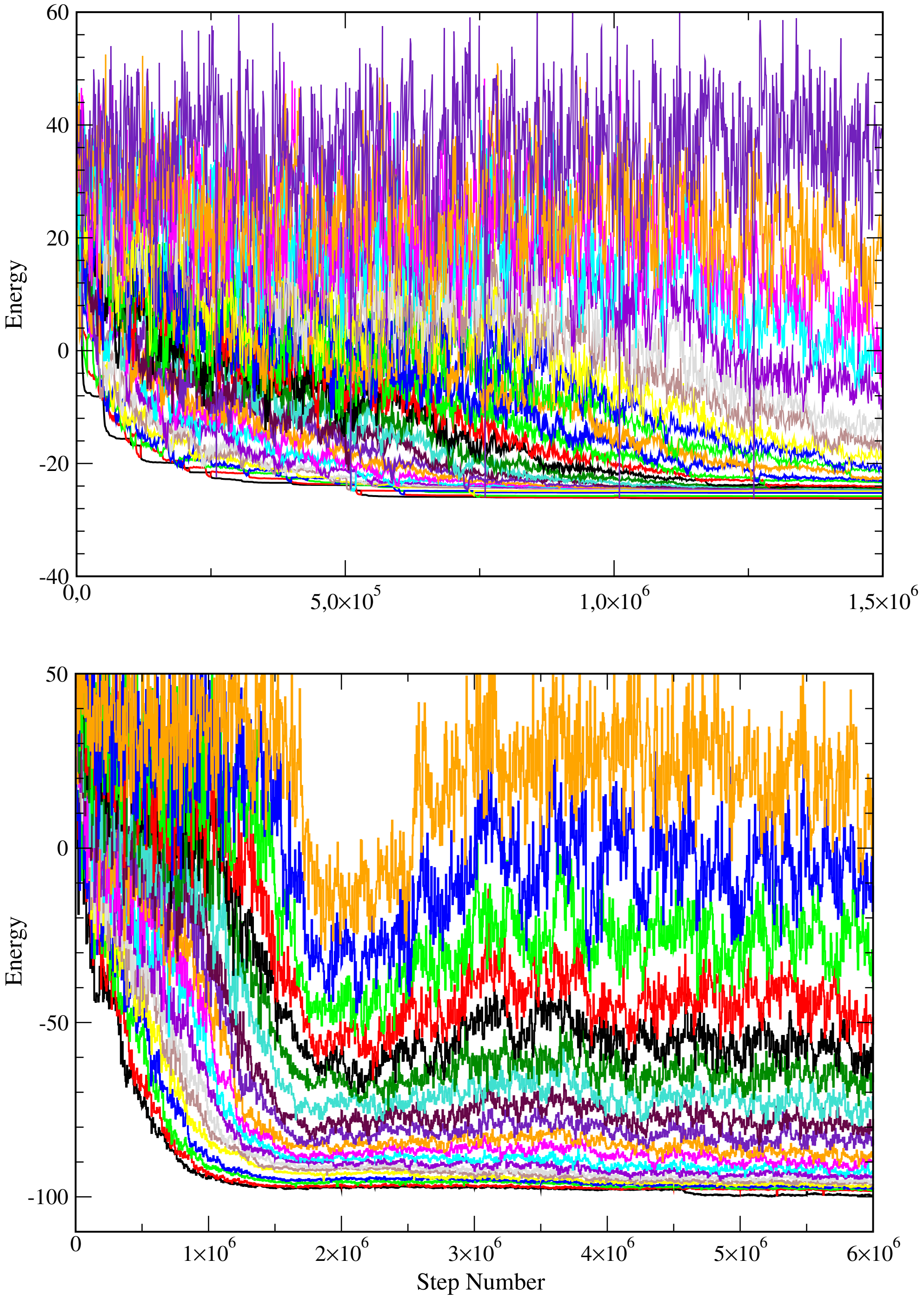

Figure 1. Energies of parallel tempering simulation of the trp-cage protein (top) and the 40-amino-acid headpiece of the HIV accessory protein (bottom) as a function of the number of steps. For the trp-cage and the HIV accessory protein simulation, 30 and 20 replicas respectively were used. For the latter simulation, data for only the first 6×10^6 steps are shown, but the simulation was continued for a total of 3×10^7 steps to ensure that convergence was achieved.

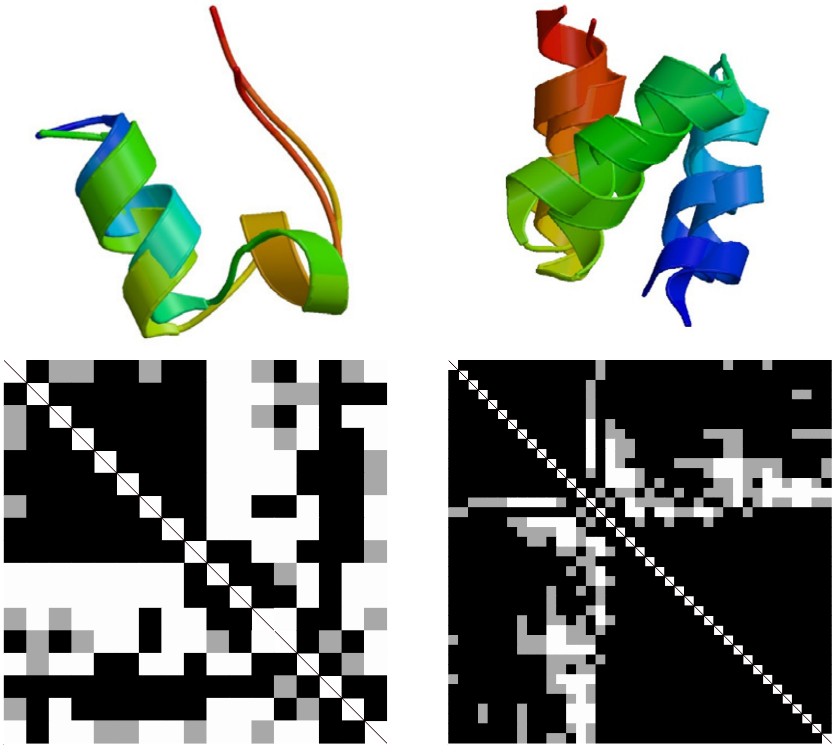

Figure 2. The top panel shows the overlay of the secondary structure elements in the tertiary structure of the experimental and the folded structures of the trp-cage (left) protein and the HIV accessory protein (right). The bottom panel shows the C_β–C_β maps of the distance between the folded and the experimental structures for both proteins. Each square in the C_β–C_β distance map illustrates the deviation between the C_β–C_β distance of two amino acids in the NMR and the C_β–C_β distance of the same amino acids in the folded structure. Black (grey) squares indicate a deviation of less than 1.50 Å (2.25 Å). White squares indicate larger deviations.

4. Discussion

Since the native structure dominates the low energy conformations arising in all of these simulation, these results demonstrate the feasibility of all-atom protein tertiary structure prediction with the adapted version of the parallel tempering method. We note that both dynamic temperature adjustment as well as replication contribute significantly to the convergence properties of the method. While these measures speed up convergence, the thermodynamic equilibrium of the conformations is naturally lost. The free energy approach emerges as a viable trade-off between predictivity and computational feasibility. The computational efficiency of the optimization approach stems from the possibility of visiting unphysical intermediate high energy conformations during the search. While sacrificing the folding dynamics, a reliable prediction of its terminus, the native conformation—which is central to most biological questions—can be achieved.

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

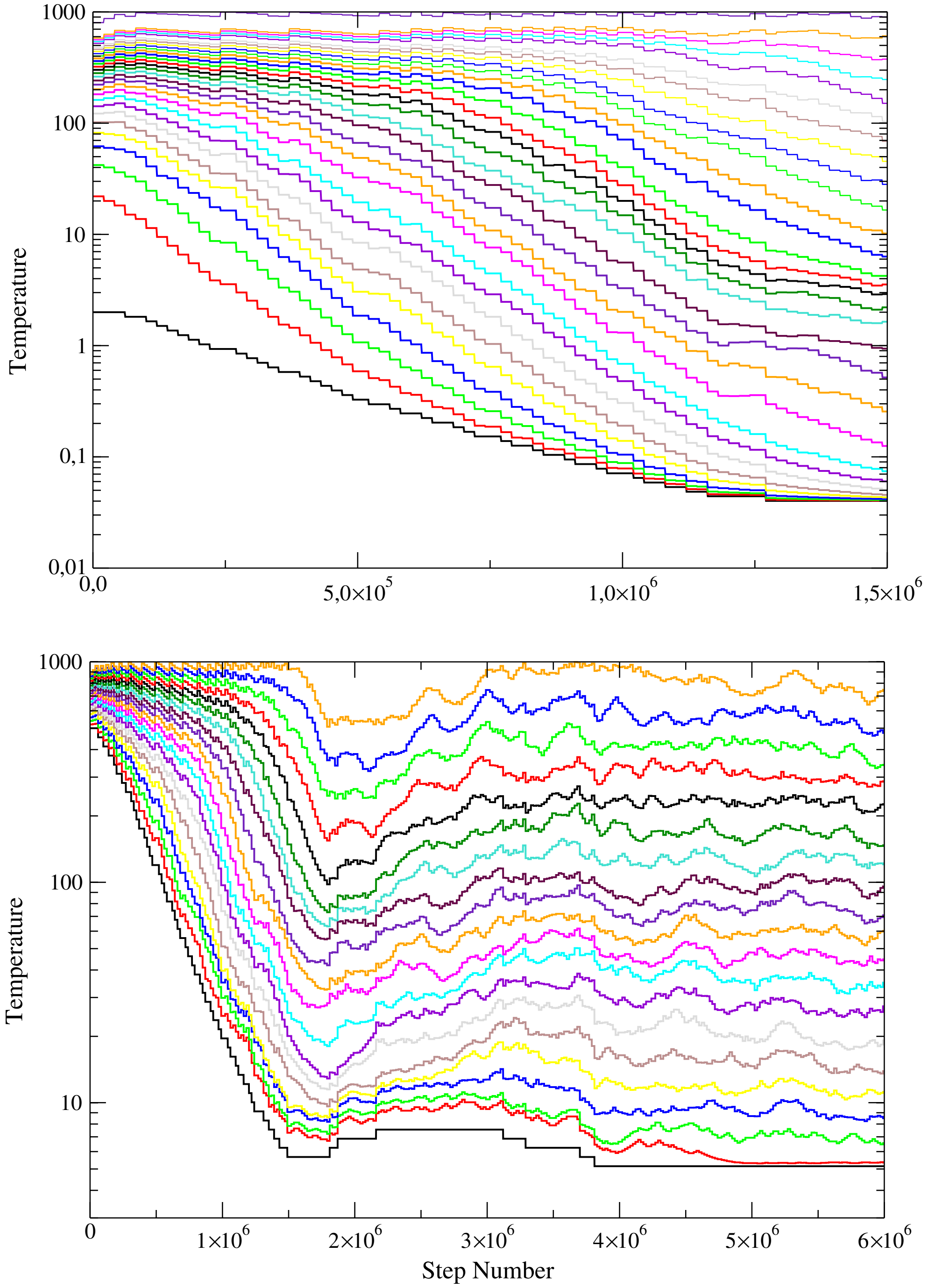

Figure 3. Self-adjusting temperatures of the adapted parallel tempering simulation of the trp-cage protein (top) and the 40-amino-acid headpiece of the HIV accessory protein (bottom) as a function of the number of steps. For the trp-cage and the HIV accessory protein simulation, 30 and 20 replicas respectively were used.

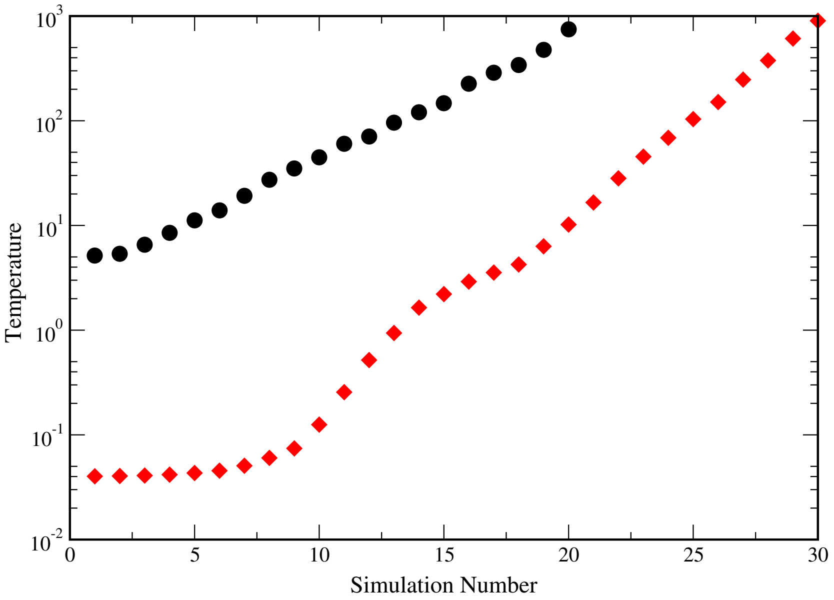

Figure 4. Distribution of the final temperatures in the adapted parallel tempering algorithm in application to the trp-cage protein (diamonds) and the HIV accessory protein (circles). Note that the final temperatures obey a near geometric distribution above 1 K.

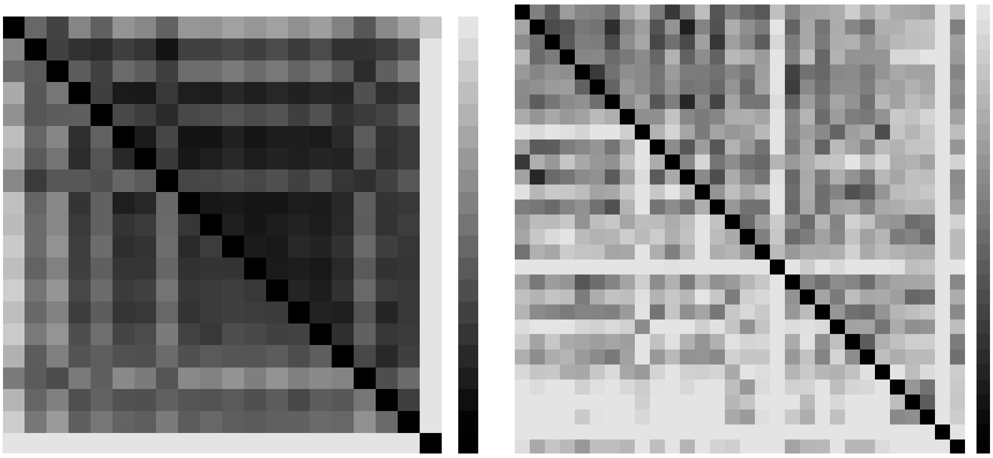

Figure 5. An illustration of the degree of similarity between the terminal configurations of the modified PT simulation of the trp-cage (left) and HIV accessory protein (right). Each row represents a different structure (in order of increasing energy from the top). Colour codes (the scale on the left, ranging from zero (black) to more than 4 Å (white)) indicate the similarity to the other structures. The upper triangle measures the backbone RMS; the lower triangle indicates the heavy atom RMSD. The red bars right and bottom indicate that the simulation at the highest temperature is very different from all other conformations. At intermediate temperatures there is a set of simulations which are similar among themselves, but still different from the native conformation.

different optimization strategies depends strongly on the structure of the potential energy surface. As a result, the development of efficient optimization techniques for all-atom protein structure prediction depends on the availability of a force field that folds 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 that can be treated with present day computational resources. We note that PSP on the basis of force field optimization fits the computational paradigm of globally distributed grid computing even better than protein folding using a molecular dynamics approach.

Acknowledgments

We thank the Fond der Chemischen Industrie, the BMBF, 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.

References

[1] Baker D and Sali A 2001 Protein structure prediction and structural genomics *Science* **294** 93–6

[2] Schonbrunn J, Wedemeyer W J and Baker D 2002 Protein structure prediction in 2002 *Curr. Opin. Struct. Biol.* **12** 348–52

[3] Go N and Scheraga H A 1976 On the use of classical statistical mechanics in the treatment of polymer chain conformation *Macromolecules* **9** 535–42

[4] Ulrich P, Scott W, van Gunsteren W F and Torda A E 1997 Protein structure prediction force fields: Parametrization with quasi-Newtonian dynamics *Proteins, SF&G* **27** 367–84

[5] Snow C D, Nguyen H, Pande V S and Gruebele M 2002 Absolute comparison of simulated and experimental protein folding dynamics *Nature* **420** 102–6

[6] Simmerling C, Strockbine B and Roitberg A 2002 All-atom structure prediction and folding simulations of a stable protein *J. Am. Chem. Soc.* **124** 11258–9

[7] Anfinsen C B 1973 Principles that govern the folding of protein chains *Science* **181** 223–30

[8] Li Z and Scheraga H A 1987 Monte Carlo minimization approach to the multiple minima problem in protein folding *Proc. Natl Acad. Sci. USA* **84** 6611–5

[9] Schug A, Herges T and Wenzel W 2003 Reproducible protein folding with the stochastic tunneling method *Phys. Rev. Lett.* **91** 158102

[10] Herges T and Wenzel W 2004 Reproducible in-silico folding of a three-helix protein in a transferable all-atom forcefield *Preprint* physics/0310146

[11] Herges T and Wenzel W 2004 Development of an all-atom forcefield for tertiary structure prediction of helical proteins *Biophysical J.* **87** 3100–9

[12] Schug A, Herges T and Wenzel W 2004 All-atom folding of the trp-cage protein in an all-atom forcefield *Europhys. Lett.* **67** 307–13

[13] Herges T, Schug A and Wenzel W 2004 *Protein Structure Prediction with Stochastic Optimization Methods: Folding and Misfolding the Villin Headpiece* (*Lecture Notes in Computer Science* vol 3045) pp 454–64

[14] Schug A, Herges T and Wenzel W 2004 Reproducible folding of a four helix protein in an all-atom forcefield *J. Am. Chem. Soc.* submitted

[15] Gouda H, Torigoe H, Saito A, Sato M, Arata Y and Shimanda I 1992 Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures *Biochemistry* **40** 9665–72

[16] 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 and Fersht A R 2003 The complete folding pathway of a protein from nanoseconds to microseconds *Nature* **421** 863–7

[17] Geyer G J 1992 *Stat. Sci.* **7** 437

[18] Hukushima K and Nemoto K 1996 Exchange Monte Carlo method and application to spin glass simulations *J. Phys. Soc. Japan* **65** 1604–8

[19] Neidigh J W, Fesinmeyer R M and Anderson N H 2002 Designing a 20-residue protein *Nat. Struct. Biol.* **9** 425–30

[20] Herges T, Merlitz H and Wenzel W 2002 Stochastic optimisation methods for biomolecular structure prediction *J. Ass. Lab. Autom.* **7** 98–104

[21] Abagyan R and Totrov M 1994 Biased probability Monte Carlo conformation searches and electrostatic calculations for peptides and proteins *J. Mol. Biol.* **235** 983–1002

[22] Herges T, Schug A, Burghardt B and Wenzel W 2004 Exploration of the free energy surface of a three helix peptide with stochastic optimization methods *Int. J. Quantum Chem.* **99** 854–93

[23] Avbelj F and Moult J 1995 Role of electrostatic screening in determining protein main chain conformational preferences *Biochemistry* **34** 755–64

[24] Eisenberg D and McLachlan A D 1986 Solvation energy in protein folding and binding *Nature* **319** 199–203

[25] Sharp K A, Nicholls A, Friedman R and Honig B 1991 Extracting hydrophobic free energies from experimental data:relationship to protein folding and theoretical models *Biochemistry* **30** 9686–97

[26] Pedersen J T and Moult J 1997 *J. Mol. Biol.* **269** 240

[27] Hansmann U H E and Okamoto Y 1997 Numerical comparison of three recently proposed algorithms in the protein folding problem *J. Comput. Chem.* **18** 920

[28] Hansmann U H E 1999 *Eur. Phys. J.* B **12** 607

[29] Lin C Y, Hu C K and Hansmann U H 2003 Parallel tempering simulations of hp-36 *Proteins* **53** 436–45