

All-atom folding of the trp-cage protein with an adaptive parallel tempering method

A. SCHUG AND W. WENZEL

*Forschungszentrum Karlsruhe, Institut für Nanotechnologie,
P.O. Box 3640, 76021 Karlsruhe, Germany*

PACS. 87.15.-v – Biomolecules: structure and physical properties.

PACS. 87.15.Aa – Theory and modeling; computer simulation.

PACS. 02.60.Pn – Numerical optimization.

Abstract. – Using the recently developed protein free-energy forcefield PFF01 we report the reproducible all-atom folding of the 20 amino-acid trp-cage protein to within 2.0 Å backbone RMS deviation to the experimental structure with modest computational resources. We used an adapted version of the parallel tempering method as an inherently parallel stochastic optimization method. We find that near native structures dominate the low-energy spectrum of the final conformations and investigate the efficiency of the method as a function of the number of replicas in application to *all-atom* protein structure prediction.

Protein structure prediction on the basis of the amino acid sequence alone remains one of the major outstanding challenges of theoretical biophysics [1–5]. In the post-genomic era, sequence information for proteins abounds, while structural and mechanistic information remains scarce. Theoretical methods for protein structure prediction may help to close the gap between the available sequences and structures and elucidate mechanisms of proteins that are difficult to handle experimentally (e.g. transmembrane proteins). With the development of reliable forcefields [6, 7] and robust simulation techniques [4, 5, 8], protein structure prediction may assist in the understanding and quantitative analysis of protein-protein or protein-ligand association [9, 10] at an atomistic level.

While homology based methods have demonstrated steady progress in the past decade [11], the assessment of atomistic de-novo prediction strategies has been less favorable [2, 3, 12]. Atomistic simulations of the folding process remain confined to small peptides due to their large computational cost [13–15]. Based on the thermodynamic paradigm of protein folding [16] free energy based methods describe the native structure of the protein as the global optimum of a suitable free-energy forcefield. This approach is potentially much faster and more predictive than the costly simulation of the folding pathway, but will obviously sacrifice dynamical information.

We have recently reported the rational development of a transferable all-atom free-energy forcefield (PFF01) [7] that correctly predicts the native structure of several proteins with 20–60 amino acids, as the global minimum of the free-energy surface (FES). Reproducible folding could be demonstrated for the 20-amino acid trp-cage protein [4], the 40 amino acid headgroup

of the HIV accessory protein [5] and the the 36 amino-acid headgroup of villin [8, 13, 17, 18]. An alternate free energy forcefield was recently used to predict the structure of the B domain of staphylococcal protein A from first principles [6].

Despite these advances in the modeling of the free energy surface, little is presently known about the efficiency of various optimization strategies to reliably determine the global optimum of the FES at the all-atom level. There is mounting evidence, that PFF01 or similar free-energy forcefields may correctly predict the native structure of a large family of proteins [7], but the lack of optimization methods to reliably locate the associated global minima of the forcefield emerges as the central bottleneck in all atom protein structure prediction.

Here we investigate an adapted version of the parallel tempering (PT) method as a candidate optimization method for all-atom protein structure prediction. We employ this method to reproducibly fold the native structure of the trp-cage protein [15, 19] (pdb code 1L2Y), one of the fastest folding proteins known [20], to within experimental resolution (backbone RMS deviation, 2.01Å). Our results provide additional evidence that the free-energy forcefield PFF01 stabilizes the native conformation of the trp-cage protein as its native conformation. We also present an analysis of performance of the method, in particular of the in-silico folding rate on the number of replicas used.

Methods. – We have recently developed an all-atom (with the exception of apolar CH_n groups) free-energy protein forcefield (PFF01) that models the low-energy conformations of proteins with minimal computational demand [5, 21]. In the folding process at physiological conditions the degrees of freedom of a peptide are confined to rotations about single bonds. The forcefield is parameterized with the following non-bonded interactions:

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

Here r_{ij} denotes the distance between atoms i and j and $g(i)$ the type of the amino acid i . The Lennard Jones parameters (V_{ij}, R_{ij} for potential depths and equilibrium distance) depend on the type of the atom pair and were adjusted to satisfy constraints derived from a set of 138 proteins of the PDB database [21–23]. 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 [24]. Interactions with the solvent were first fit in a minimal solvent accessible surface model [25] parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides [26]. A_i corresponds to the area of atom i that is in contact with a fictitious solvent. Hydrogen bonds are described via dipole-dipole interactions included in the electrostatic terms and an additional short range term for backbone-backbone hydrogen bonding (CO to NH) which depends on the OH distance, the angle between N,H and O along the bond and the angle between the CO and NH axis [7].

The parallel (or simulated) tempering technique [27, 28] 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 [29–31]. 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

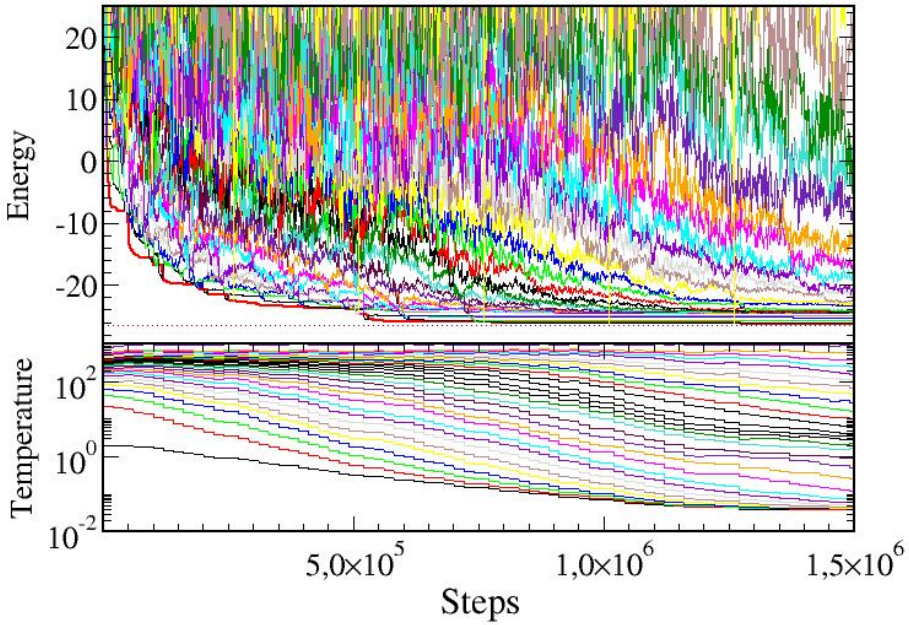


Fig. 1 – Energies (upper panel) and temperatures (lower panel) of the 28 replica modified parallel tempering simulation of the trp-cage protein reported in the text. The dotted line in the upper panel corresponds to the estimate of the global optimum of FES (obtained independently). The lower panel demonstrates 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.

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:

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

where $\beta_i = 1/k_B T_i$ and E_i are the inverse temperatures and energies of the conformations respectively. The choice of the exchange probability in equation (2) ensures that all simulations remain in thermodynamic equilibrium at their respective temperatures. The PT method thus offers the opportunity to simultaneously evaluate thermodynamic expectation values over a wide range of temperatures. The exchange mechanism improves the conformational averaging of the low-temperature simulations, because the exchange with high-temperature simulations provides a mechanism to overcome the high energy barriers between low-lying metastable conformations. Without loss of generality one may confine the exchange mechanism to simulations which are adjacent in temperature.

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 approx-

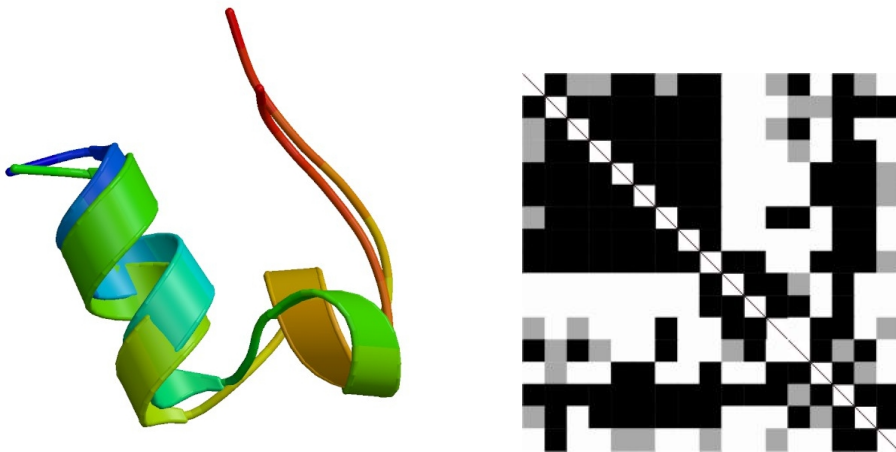


Fig. 2 – Overlay (left) and $C_{\beta} - C_{\beta}$ (right) distance map of the folded and the NMR structure and the predicted structure of the trp-cage protein. A pixel in row i and column j of the color coded distance map indicates the difference in the $C_{\beta} - C_{\beta}$ distances of the native and the folded structure. Black (grey) squares indicate that the $C_{\beta} - C_{\beta}$ distances of the native and the other structure differ by less than 1.5 (2.25) Å respectively. White squares indicate larger deviations.

imately 2-600 K. Note that this temperature scale refers to the fictitious temperature scale of the simulation, the physical temperature scale is always set to 300 K by the parameterization of the implicit solvent model. The energy distribution is often chosen to result in either geometric or constant differences of subsequent inverse temperatures.

Applied as an optimization technique, however the simulation associated with the lowest temperature will typically yield the estimate for the global optimum, while all others are required to generate different conformations. The computational effort of the method rises linearly with the number of temperatures, the computational efficiency of a PT based optimizer thus decreases when more than the minimally required number of simulations is used. Little is presently known about the optimal temperature distribution and the *in-silico* folding rates as a function of the number of temperatures in application to all-atom protein folding.

A recent investigation of PT to the HIV accessory protein [32] found that its standard implementation was very inefficient in comparison to the adapted basin hopping method [5,33]. We overcame this shortcoming of the original parallel tempering method with the introduction of an *adaptive temperature control* for all 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 $t_{i+1} - t_i$. If the exchange rate was above 2%, then all all temperatures above t_i were increased by the same difference.

Adaptive temperature control appears as an important ingredient, but not as the cure for the transition state problem in protein folding. We note that individual simulations of a traditional PT simulation do not exchange structural information. As a result, provided that all low temperature simulations should find near native conformations, they would have to do so individually. To improve the computational efficiency of PT we thus 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

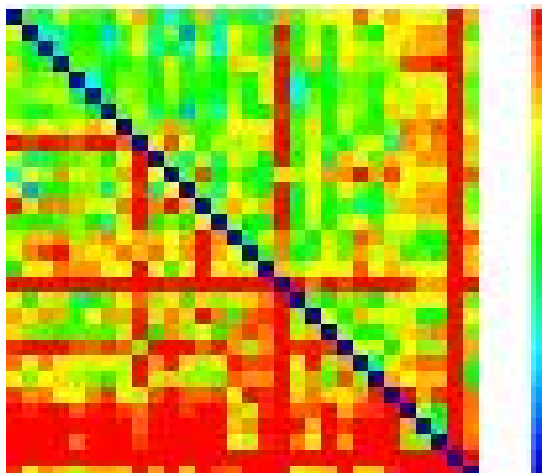


Fig. 3 – Illustration of the degree of similarity between the terminal configurations of the modified PT simulation of the trp-cage protein. Each row represents a different structure (ordered by increasing energy from the top). Color codes indicate the similarity to the other structures on a scale from 0-4 Å: blue squares represent close structures red squares different structures ($4 \text{ \AA} < \text{RMSB}$). The upper triangle measures backbone RMS, the lower the heavy atom RMS deviation.

exploration of the folding funnel around the best conformation found near the presently best conformation.

Results. – Using this method we performed simulations comprising up to 10^7 steps per temperature with 4,7,14 and 28 concurrent simulations. All simulations were started with random conformations at high temperatures to allow for rapid, unbiased relaxation of the structures and the temperature distribution. The convergence of the method and the adjustment of the temperatures is shown in Figure (1) for the simulation using 28 replicas. After an initial period of equilibration, the temperatures gradually relaxed into brackets, but kept fluctuating during the course of the simulation. The figure also demonstrates that the simulations achieve the large exchange rates required for efficient protein folding as evidenced by the touching bands of the energy distribution. At the end of this simulation, the lowest conformation had a backbone rms deviation (RMSB) of 2.01 Å. Considering the ensemble of final conformations, we find many structures closely resembling the native conformation. the next lowest conformations (in energy) had RMSB of 2.56, 1.81, 2.91, 3.08 Å etc, indicating predictive folding of the trp-cage protein. The overlay of the experimental and the converged structure (see Figure (2)) illustrates the good agreement between the conformations, the difference in NOE constraints demonstrates convergence of the long-range native contacts, in particular of the trp-cage region (off-diagonal black are in the right panel of the figure).

Analyzing the simulations with different numbers of replicas we found one of the simulations with four processors to converge to the native structures (RMSB 2.5 Å), all others froze at conformations 3-6 kcal/mol above the global optimum of the FES. Two of the simulations with eight processors converged to the native state (2.05/2.56 Å RMSB respectively), while the simulations with 14 and 28 processors both converged. Thus a minimal number of replicas appears to be required to ensure folding, a finding rationalized by the fact that the simulation must span the temperature range of approximately 2-500 K to ensure both local optimization

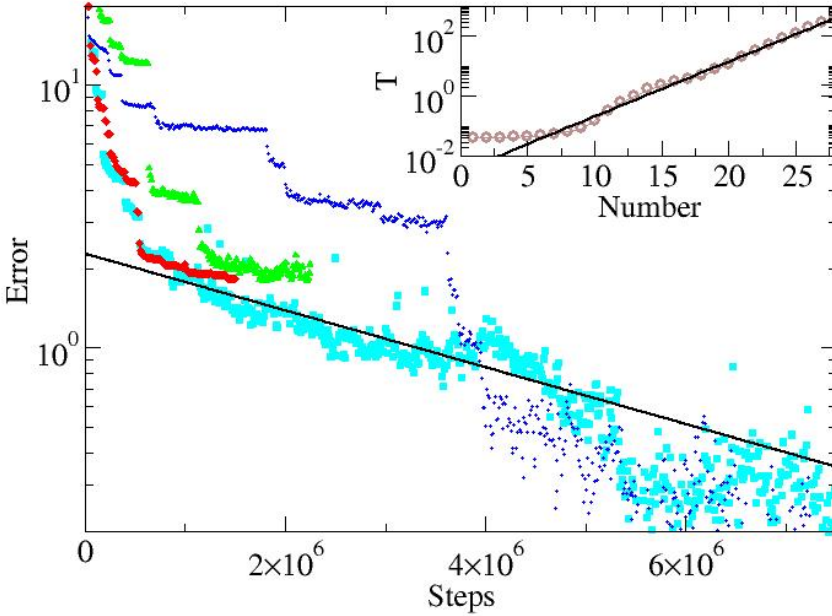


Fig. 4 – Log-normal plot of the errors in the energy associated with the lowest temperature (in kcal/mol) for selected simulations against the number of steps (blue/green/turquoise/red indicate 4,7,14,28 replicas respectively). The black line corresponds to an exponential fit to the $N_r = 14$ data indicating exponential of the simulation. The inset shows the final temperature distribution (averaged over the last 10^5 steps), the black line was fitted to a geometrical increase of the temperatures for $n_t > 7$.

and rapid relaxation.

Due to the lack of predictive all-atom folding studies, little is known about the performance of different optimization methods and their dependence on the number of replicas (processors). Figure (4) shows the dependence of the error of the simulation as a function of the number of steps. Encouragingly we find an *exponential convergence* of the method for all simulations that converged to the native state. For a small number of processors the convergence is dominated by rare events (see e.g. blue line) — it is clearly visible that the simulation is trapped in metastable states for a significant amount of time. This finding is commensurate with the observation that some simulations with few replicas do not converge within the allotted number of steps. For a larger number of replicas ($n_r = 8, 14, 28$) the error converges more smoothly, but appears to rapidly saturate as a function of n_r . The reason for this saturation can be traced to the equilibrated distribution of temperatures shown in the inset of the figure. We note that the first 10 simulations all run at temperatures below 1K, i.e. perform only local minimizations. Effectively they generate no new information, once good local minima have been found. The remaining temperatures appear to be distributed according to a geometric distribution ($T_n = T_7 q^{(n-7)}$ for $n \geq 10$). This observation validates the

temperature distribution schemes often used in PT simulations.

Conclusions. – In this investigation we reported the unbiased, reproducible folding of the 20-amino acid trp-cage protein from random initial condition to within experimental resolution with an adapted parallel tempering method. These results demonstrate the viability of the free-energy based approach to all-atom protein structure prediction and validate the PFF01 forcefield in particular for the trp-cage protein. This approach presented here complements MD simulations [14, 15] because it offers a rational criterion for unbiased *protein structure prediction*, whenever a particular structure occurs reproducibly in the low-energy spectrum of the simulations. The price paid for this relative certainty is the loss of direct insight into the kinetics of the folding process.

The development of reliable free energy forcefields [7] and efficient optimization techniques offers an increasingly viable route for protein structure prediction at the all-atom level. Presently, the availability of efficient optimization methods, rather than inaccuracies of the forcefield, appear to be the bottleneck towards the treatment of larger proteins, in particular for beta-sheet proteins, which are computationally much more demanding. The introduction of the modified PT method is thus an important step towards the development of parallel optimization techniques that can be efficiently implemented on massively parallel, distributed computational architectures that dominate the high-end of presently available computational resources.

We are grateful for many stimulating discussions with S. Gregurick and J. Moult, who contributed part of the code with which these simulations were performed. WW acknowledges support from the German National Science Foundation (We 1863/10-2), the KIST supercomputational materials lab, in particular Dr. K.-H. Lee, where some of the simulations reported here were performed, and the BMWF.

REFERENCES

- [1] D. Baker and A. Sali, *Science* **294**, 93 (2001).
- [2] J. Pillardy, C. Czaplowski, A. Liwo, J. L. and D. R. Ripoll, R. Kamierkiewicz, S. O. and W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye, and H. A. Scheraga, *Proc. Natl. Acad. Sci.(USA)* **98**, 2329 (2001).
- [3] J. Schonbrunn, W. J. Wedemeyer, and D. Baker, *Curr. Op. Struc. Biol.* **12**, 348 (2002).
- [4] A. Schug, T. Herges, and W. Wenzel, *Phys. Rev. Letters* **91**, 158102 (2003).
- [5] T. Herges and W. Wenzel, *Reproducible in-silico folding of a three-helix protein in a transferable all-atom forcefield*, <http://www.arXiv.org: physics/0310146> (2004).
- [6] J. Vila, D. Ripoll, and H. Scheraga, *Proc. Natl. Acad. Sci.(USA)* **100**, 14812 (2004).
- [7] T. Herges and W. Wenzel, *Development of an all-atom forcefield for tertiary structure prediction of helical proteins*, (submitted to *Proteins*) (2004).
- [8] U. H. E. Hansmann, *Phys. Rev. Letters* **88**, 068105 (2002).
- [9] J. Drews, *Science* **287**, 1960 (2000).
- [10] H. Merlitz, B. Burghardt, and W. Wenzel, *Chem. Phys. Lett.* **370**, 68 (2003).
- [11] E. Lattman, *Proteins* **44**, 399 (2001).
- [12] R. Bonneau, J. Tsui, I. Ruczinski, D. Chivian, C. M. E. Strauss, and D. Baker, *Proteins* **45**, 119 (2001).
- [13] Y. Duan and P. A. Kollman, *Science* **282**, 740 (1998).
- [14] C. D. Snow, H. Nguyen, V. S. Panda, and M. Gruebele, *Nature* **420**, 102 (2002).
- [15] C. Simmerling, B. Strockbine, and A. Roitberg, *J. Am. Chem. Soc.* **124**, 11258 (2002).
- [16] C. B. Anfinsen, *Science* **181**, 223 (1973).
- [17] T. Herges, A. Schug, and W. Wenzel, *Protein Structure Prediction with Stochastic Optimiza-*

- tion Methods: Folding and Misfolding the Villin Headpiece* (2004), (accepted for publication in Lecture Notes in Computational Science,(Springer, New York)).
- [18] T. Herges, A. Schug, and W. Wenzel, *All atom folding and misfolding of the villin headpiece in a free-energy forcefield* (2004), (submitted).
 - [19] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Anderson, *Nature Struct. Biol.* **9**, 425 (2002).
 - [20] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, *J. Am. Chem. Soc.* **124**, 12952 (2002).
 - [21] T. Herges, H. Merlitz, and W. Wenzel, *J. Ass. Lab. Autom.* **7**, 98 (2002).
 - [22] R. Abagyan and M. Totrov, *J. Molec. Biol.* **235**, 983 (1994).
 - [23] T. Herges, A. Schug, B. Burghardt, and W. Wenzel, *Low Energy Conformations of a Three Helix Peptide in an All-Atom Biomolecular Forcefield* (2004), accepted in *Intl. J. Quantum Chem.*
 - [24] F. Avbelj and J. Moult, *Biochemistry* **34**, 755 (1995).
 - [25] D. Eisenberg and A. D. McLachlan, *Nature* **319**, 199 (1986).
 - [26] K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig, *Biochemistry* **30**, 9686 (1991).
 - [27] E. Marinari and G. Parisi, *Europ. Phys. Letters* **451**, 1992 (19).
 - [28] A. P. Lyubartsev, A. A. Martinovski, S. V. Shevkunov, and P. Vorontsov-Velyaminov, *J. Chem. Phys.* **96**, 1776 (1992).
 - [29] U. Hansmann and Y. Okamoto, *J. Comput. Chem* **18**, 920 (1997).
 - [30] U. Hansmann, *Eur. Phys. J. B* **12**, 607 (1999).
 - [31] C. Lin, C. Hu, and U. Hansmann, *Proteins* **53**, 436 (2003).
 - [32] A. Schug, T. Herges, and W. Wenzel, *An adaptive parallel tempering method for protein folding* (2004), (submitted).
 - [33] A. Nayeem, J. Vila, and H. Scheraga, *J. Comp. Chem.* **12**(5), 594 (1991).