Reproducible Protein Folding with the Stochastic Tunneling Method

A. Schug,[1] T. Herges,[1] and W. Wenzel[1]

[1]*Forschungszentrum Karlsruhe, Institut für Nanotechnologie, 76021 Karlsruhe, Germany*

(Received 27 February 2003; published 6 October 2003)

We report the reproducible folding of the 20 amino-acid protein trp cage using a novel version of the stochastic tunneling method and a recently developed all-atom protein free-energy force field. 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, in total eight simulations converged to the native structure. We find a strong correlation between energy and root-mean-square deviation to the native structure for all simulations.

DOI: 10.1103/PhysRevLett.91.158102 PACS numbers: 87.15.Cc, 02.60.Pn, 02.70.Ns

Protein structure prediction on the basis of the amino-acid sequence alone remains one of the grand outstanding challenges of theoretical biophysics [1–3]. In the postgenomic era, sequence information for proteins abounds, while structural and mechanistic information remains scarce. While the sequences for hundreds of thousands of proteins are known, the PDB database contains just over 20 000 entries [4]. Theoretical models may help to close this gap, and elucidate mechanisms of proteins that are difficult to handle experimentally (e.g., transmembrane proteins). With the development of reliable force fields [5] and robust simulation techniques [6], protein structure prediction may assist in the understanding and quantitative analysis of protein-protein or protein-ligand association [7] at an atomistic level. Ultimately questions regarding the dynamics of biological function may be addressed.

While homology based methods have demonstrated steady progress in the past decade, the assessment of atomistic *de novo* prediction strategies has been less favorable [2,3]. *De novo* prediction strategies at the all-atom level are presently rare, in part because of their enormous computational cost [8,9]. The prediction of protein tertiary structure with free-energy force fields may significantly reduce this cost, because the native structure can be determined with global optimization methods [2,6,10], without recourse to the folding dynamics, orders of magnitude faster than with direct simulation methods.

We have recently reported the rational development of a transferable all-atom free-energy force field (PFF01) [5] that correctly predicts the native structure of two three-helix proteins, the 36 amino-acid headgroup of villin (pdb-code:1VII) [6,8] and the 40 amino-acid headgroup of the HIV accessory protein (pdb-code:1F4I), as the global minimum of the free-energy surface (FES). The FES of both proteins has several deep and complex folding funnels, with several nontrivial branching points. Optimization methods to efficiently and reliably locate the low-energy minima of such complex, rugged FES are still lacking.

In this investigation we show that an adapted version of the stochastic tunneling algorithm is an efficient and reliable method for reproducible structure prediction of trp cage [11,12](pdb-code 1L2Y), one of the fastest folding proteins [13]. Our results demonstrate that the PFF01 force field is transferable from the three-helix peptides to systems with significantly less helical content.

Model.—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 [14]. 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 nonbonded interactions:

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

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 as a set of 138 proteins of the PDB database [14–16]. The nontrivial electrostatic interactions in proteins are represented via group-specific dielectric constants (ϵ_i depending on the amino acid to which atom i belongs). The partial charges q_i and ϵ_i were previously derived in a potential-of-mean-force approach [17]. Interactions with the solvent were first fit in a minimal solvent accessible surface model [18] parametrized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides [19]. A_i corresponds to the area of atom i that is in contact with a fictitious solvent. The σ_i were adjusted to stabilize the native state of the 36-amino acid headgroup of villin (pdb-code 1VII) as the global minimum of the force field [20]. 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 takes the form: $V_{hb}(\mathrm{CO}_i, \mathrm{NH}_j) = R(\tilde{r}_{ij})\Gamma(\phi_{ij}, \theta_{ij})$ where $\tilde{r}_{ij}$, ϕ_{ij}, and θ_{ij} designate the OH distance, ϕ is the angle between N, H, and O along the bond, and θ is the angle between the CO and NH axis. R and Γ were fitted as a corrective potential of mean force to the same set of proteins described above [15].

Optimization Technique.—The stochastic tunneling technique (STUN) [21] 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 [22,23], this technique proved superior to competing stochastic optimization methods. In STUN the dynamical process explores not the original, but a transformed PES, which dynamically adapts and simplifies during the simulation. 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). Even at finite temperature the dynamics of the system then becomes diffusive at energies above $E \gg E_0$ 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.

The original transformation proposed for STUN [21] is helpful for many PES, but leads to failure for peptide simulations, where the first pair of atoms clashing converts the transformed PES to a "golf-course" landscape. Further clashes then cost no extra energy and the entire clashing conformational space is transformed into a single featureless plateau with no guiding force to a nonclashing conformation. Applied to proteins, this deficiency limited the applicability of the original STUN to relatively short peptides.

For the peptide simulations reported here we replace the original transformation [21] with:

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

with $x = \gamma(E - E_0)$, where E is the energy, E_0 the best energy found so far. The problem-dependent transformation parameter [21] γ controls the steepness of the transformation [we used $\gamma = 0.5(\mathrm{kcal/mol})^{-1}$]. The transformation in Eq. (2) ameliorates the difficulties associated with the original transformation, because $E_{\mathrm{STUN}} \propto \ln(E/kT)$ continues to grow slowly for large energies. Even so, the fictitious 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.

Results.—We performed 25 simulations of 12.5×10^6 energy evaluations each of the 1L2Y peptide starting from randomized dihedral angle conditions. 1L2Y is a designed 20-residue Trp-cage protein [11] that folds spontaneously into a globular fold which exhibits a stable secondary/tertiary structure not often found in proteins of this size [12]. Its secondary structure contains both an α helix extending from residue 2-8 and an 3-10 helix in residues 11-14. The presence of three proline residues near the C terminus of the peptide prohibits the existence of secondary structure in this region. In experiment, the protein autonomously folds into a tertiary structure in which the proline residues pack against Tyr-3 and Trp-6 near the N terminus. Each of the starting conformations had a large backbone root mean square deviation (RMSB) from the native configuration and a large positive energy. We have independently confirmed that the native state of the protein is close to a local minimum of the force field (RMSB = 2.61 Å, throughout this Letter the first of 20 reported NMR structures are used [11]). This deviation sets the scale on which the similarity of the resulting structures can be judged. The minimal energy and RMSB to the original/relaxed NMR structure are shown in Table I. The estimate of the best energies still spans a wide energy window, emphasizing the fact that presently little can be learned about the global minimum of such complex potential energy surfaces from a single isolated

TABLE I. Energy (in kcal/mol) and RMSB deviation to the relaxed NMR (RMSB-1) and the NMR (RMSB-2) structure of 1L2Y for the twenty-five simulations described in the text. Also shown is the secondary structure assignment (DSSP) for the lowest energy conformation of each simulation (C = coil, H = helix, T = turn, S = sheet). The first two lines give the data for the original/relaxed NMR structure, respectively.

Energy	RMSB-1	RMSB-2	Secondary structure
19.29	2.61	0.00	CHHHHHHCCCCTTHHHHTC
−25.73	0.00	2.61	CHHHHHHHHTHHHHTSCCCC
−25.79	1.81	2.83	CHHHHHHHHTHHHHTCCSCC
−25.31	2.52	3.05	CHHHHHHHHTHHHHTCTTTC
−25.25	2.55	3.13	CHHHHHHHHTHHHHTCTTTC
−25.25	3.30	4.26	CHHHHHHHHTCHHHHCTTTC
−25.24	2.56	3.13	CHHHHHHHHTHHHHTCTTTC
−25.15	2.57	3.13	CHHHHHHHHHHHHHHTCTTTC
−24.15	3.98	4.80	CHHHHHHHHTCSCTTSSSCC
−24.06	4.43	4.73	CHHHHHHHHTSSCSSSTTTC
−23.99	4.50	4.95	CHHHHHHHHTCSSTTSTTTC
−23.64	3.50	3.86	CHHHHHCTTHHHHHTCTTTC
−23.64	3.70	4.54	CHHHHHHHHHCTTTSSCSCC
−23.45	2.58	3.18	CHHHHHHHHTHHHHTCTTTC
−23.30	2.96	3.83	CHHHHHHHHTCHHHHCTTTC
−23.27	2.51	2.72	CHHHHHHHHTHHHHTCTTTC
−22.82	4.67	4.73	CHHHHHHHHSSCCTTCTTTC
−22.53	3.66	4.37	CHHHHHHTTSHHHHHCCSCC
−22.49	5.10	4.87	CHHHHHHHHTCCSSSSCCCC
−22.45	4.01	4.59	CHHHHHHHHHCSSTTCTTTC
−22.23	4.68	5.08	CHHHHHHHHTSSSTTSTTTC
−21.27	3.43	2.88	CHHHHHHHHTHHHHTCSCCC
−20.31	5.63	5.77	CHHHHHHHHHSSCTTCSSCC
−20.20	3.57	4.37	CHHHHHHHHHCSSTTSSSCC
−20.16	3.22	3.31	CHHHHHHHHSCHHHHCTTTC
−19.82	3.78	3.89	CHHHHHHHHHCTTTCCCSCC
−18.70	4.64	4.83	CHHHHHHSSSHHHHTCCSCC

simulation. The lowest six simulations resulted in estimates of the optimal energy that were within 1 kcal/mol of the best energy found, which differed by 0.06 kcal/mol from the energy of the (relaxed) NMR structure. All of these structures differed by less than 3 Å RMSB deviation from the relaxed NMR structure, the best structure deviated by only 1.81 Å RMSB. An analysis of the secondary structure content revealed a close similarity of all low-energy structures. Almost all structures correctly predict the first helix (residues 2–10) and the five lowest structures also correctly predict the second helix (residues 12–15). The six lowest structures differ among one another by between 1.01 and at most 2.91 Å in the RMSB deviation, which is similar to the deviation between NMR structure and the presumed global optimum of the PES. Figure 1(a) demonstrates the good agreement between the best structure and the NMR structure of the protein. Both helical fragments and the trp cage are correctly reproduced in the simulation. The majority of the RMSB deviation results from the difference in the conformation of the unstructured tail of the protein.

The *a posteriori* analysis of all structures reveals that NMR-like conformations occur also three times at higher energies. Collectively these data demonstrate that STUN predictively identified the global optimum of the force field as the native structure of 1L2Y with a success rate of over 30%. The best structure to significantly deviate in energy (the seventh simulation from the top), differs significantly in helical content and misses the second helix as illustrated in Fig. 1(b). Note that these two structures differ in energy by only 1.65 kcal/mol. We note that the resolution obtained in our calculations is of similar quality when compared to MD simulations with an implicit solvent model [12], but worse in comparison to MD simulations with explicit water. The latter have no equivalent in an optimization approach, as the entropic contributions of the solvent must always be parametrized in a free-energy force field. The use of a simple area based implicit solvent model may be the resolution limiting factor in simulations of this kind.

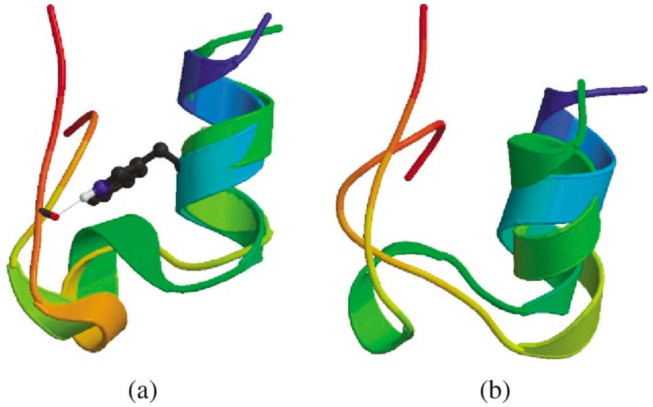

FIG. 1 (color). Secondary structure illustration of the NMR structure of 1L2Y comparing with the best structure (left) found in the simulations. On the right we show the first misfolded structure [#7 in Table I] which lacks both secondary helix and trp cage.

The total and transformed energy of the model for the simulation which resulted in the lowest energy are shown in Fig. 2. This figure clearly illustrates the alteration between local search cycles and tunneling phases characteristic of STUN. Note that successive local minima differ by energies of the order of only 1–2 kcal/mol while the energy barriers separating them are of the order $O(100\text{ kcal/mol})$. This behavior, resulting from the need to either partially unfold or to traverse Pauli-forbidden regions between successive local minima, is an indication of the typical ruggedness of realistic protein free-energy landscapes. A successful exploration of the low-energy part of the free-energy surface thus requires an efficient mechanism to escape deep local minima. In STUN a new minimum has been found, whenever the transformed energy in the lower panel of the figure reaches zero. Note that not each cooling cycle results in a new optimum.

Figure 3 illustrates STUNs ability to tunnel between very different conformations of similar energy in the low part of the PES. The figure encodes the degree of similarity of the local minima obtained in the simulation shown in Fig. 2, with the native structure shown on top. Dark squares indicate very different conformations, light squares very similar ones. Tunneling phases occur between each of the local minima of the STUN simulation. As is evident from the figure STUN has similar probabilities of generating structures that are reasonably similar and structures that are quite different in its tunneling

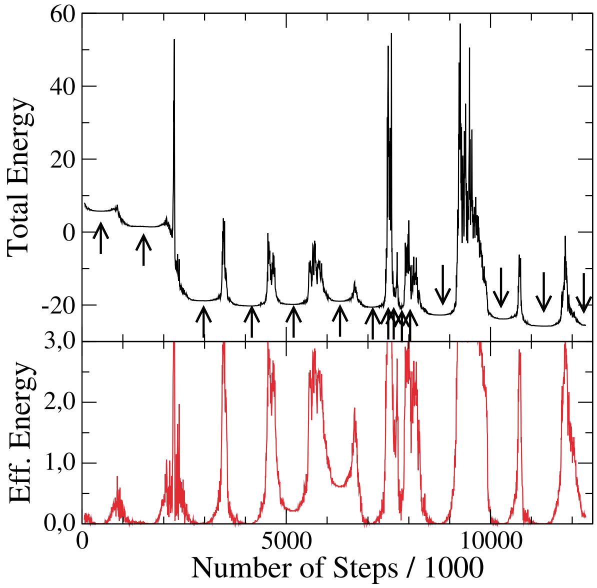

FIG. 2 (color). Total (top) and effective (bottom) energy of the STUN simulation resulting in the estimate of the global optimum of 1L2Y in PFF01. The horizontal axis gives the number of steps in 1000 s, the total energy is given in kcal/mol, the effective energy is dimensionless. The arrows indicate the configurations analyzed in Fig. 3.

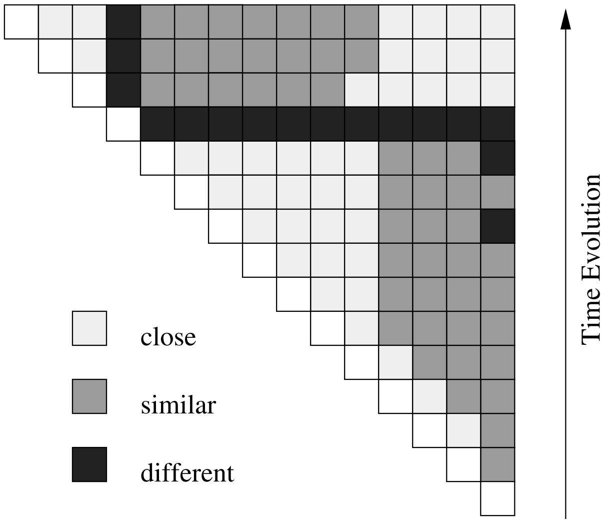

FIG. 3. Illustration of the degree of similarity of successive local minima in a STUN simulation of 1L2Y. Each row represents a different structure along the simulation pathway: the bottom/top rows represent the initial/final structures, with all local minima [defined by the end of a cooling cycle, see arrows in Fig. 2] in between. Color codes indicate the similarity to all previous structures: white squares represent identity, light shade close structures (RMSB < 1 Å), medium shade similar structures (1 Å < RMSB < 2.5 Å), and dark shade relatively different structures (2.5 Å < RMSB). Dark areas in the off diagonal indicate that the simulation explored very different regions of phase space, lighter areas (e.g., top right corner) indicate that the simulation returned to an area of phase space it had visited before.

process. The nonlinear transformation ensures that the energy scale of the present best estimate of the global optimum is not lost. In the fourth last structure a major tunneling event has taken place, which generated a metastable conformation that was completely different from all structures previously seen in the simulation. In the next tunneling phase, the simulation returned to the "folding" path, but not quite to the original structure (as illustrated by the block of medium gray squares) in the three top rows. Finally the simulation settled into a local minimum that was not very different from the local minima that occurred at the very beginning of the simulation (light gray squares at the end of the three top rows of the figures). This analysis demonstrates both STUN's ability to efficiently explore the low-energy structures (global moves) and its robustness in continuously revisiting good candidates for the global optimum.

The approach presented here complements MD simulations [9,12] because it offers a rational criterion for unbiased *protein structure prediction*, whenever a particular structure occurs reproducibly with the best energy in several simulations. The price paid for this relative certainty is the loss of direct insight into the kinetics of the folding process. Here we have demonstrated the reproducible folding of the trp-cage protein using an all-atom biomolecular free-energy force field with a success rate exceeding 30%. We were able to demonstrate the efficiency and robustness of an adapted version of the stochastic tunneling method for this purpose. We believe that further analysis of the STUN trajectories permits an elucidation of the folding pathway, this dynamical analysis of the FES will be reported elsewhere.

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. W.W. acknowledges support from the German National Science Foundation (We 1863/11-1), the Bode Foundation, and the BMWF.

[1] D. Baker and A. Sali, Science **294**, 93 (2001).
[2] J. Pillardy *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 2329 (2001).
[3] J. Schonbrunn, W. J. Wedemeyer, and D. Baker, Curr. Opin. Struct. Biol. **12**, 348 (2002).
[4] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. Bourne, Nucleic Acids Res. **28**, 235 (2000); http://www.rcsb.org/pdb.
[5] T. Herges and W. Wenzel (to be published).
[6] H. Hansmann, Phys. Rev. Lett. **88**, 068105 (2002).
[7] *Virtual Screening for Bioactive Molecules*, edited by H. J. Böhm and G. Schneider (Wiley, New York, 2001).
[8] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[9] C. D. Snow, H. Nguyen, V. S. Pand, and M. Gruebele, Nature (London) **420**, 102 (2002).
[10] U. Hansmann and Y. Okamoto, J. Phys. Chem. B **103**, 1595 (1999).
[11] J.W. Neidigh, R. M. Fesinmeyer, and N. H. Anderson, Nat. Struct. Biol. **9**, 425 (2002).
[12] C. Simmerling, B. Strockbine, and A. Roitberg, J. Am. Chem. Soc. **124**, 11 258 (2002).
[13] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. **124**, 12 952 (2002).
[14] T. Herges, H. Merlitz, and W. Wenzel, J. Assoc. Lab. Autom. **7**, 98 (2002).
[15] R. Abagyan and M. Totrov, J. Mol. Biol. **235**, 983 (1994).
[16] T. Herges, A. Schug, B. Burghardt, and W. Wenzel (to be published).
[17] F. Avbelj and J. Moult, Biochemistry **34**, 755 (1995).
[18] D. Eisenberg and A. D. McLachlan, Nature (London) **319**, 199 (1986).
[19] K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig, Biochemistry **30**, 9686 (1991).
[20] A. Schug, H. Merlitz, and W. Wenzel, Nanotechnology **14**, 1161 (2003).
[21] W. Wenzel and K. Hamacher, Phys. Rev. Lett. **82**, 3003 (1999).
[22] H. Merlitz and W. Wenzel, Chem. Phys. Lett. **362**, 271 (2002).
[23] H. Merlitz, B. Burghardt, and W. Wenzel, Chem. Phys. Lett. **370**, 68 (2003).