In Silico Folding of a Three Helix Protein and Characterization of Its Free-Energy Landscape in an All-Atom Force Field

T. Herges and W. Wenzel*

Forschungszentrum Karlsruhe, Institut für Nanotechnologie, 76021 Karlsruhe, Germany

(Received 29 October 2003; published 5 January 2005)

We report the reproducible first-principles folding of the 40 amino-acid, three-helix headpiece of the HIV accessory protein in a recently developed all-atom free-energy force field. Six of 20 simulations using an adapted basin-hopping method converged to better than 3 Å backbone rms deviation to the experimental structure. Using over 60 000 low-energy conformations of this protein, we constructed a decoy tree that completely characterizes its folding funnel.

DOI: 10.1103/PhysRevLett.94.018101 PACS numbers: 87.15.Cc, 02.60.Pn, 02.70.Ns

Available genomic and sequence information for proteins contains a wealth of biomedical information that becomes accessible when translated into a three-dimensional structure [1]. While theoretical models for protein structure prediction [2,3] that partially rely on experimental information have shown consistent progress [4], the assessment of *de novo* strategies that rely on sequence information alone has been much less favorable [5]. The development of such techniques [6,7], in particular, of transferable first-principles all-atom folding methods, would significantly benefit the understanding of protein families where little experimental information is available, the prediction of novel folds, as well as the investigation of protein association and dynamics which are presently difficult to probe experimentally. Recent progress for small peptides [3,8–10] documents both the feasibility of this approach as well as its limitations [11,12], in particular, those associated with the direct simulation of the folding process through molecular dynamics [13].

Our approach is based on the thermodynamic hypothesis [14] that many proteins are in thermodynamic equilibrium with their environment: their native state thus corresponds to the global minimum of their free-energy landscape [15,16]. The free energy of the system is accessible either indirectly by explicit ensemble averaging of the combined internal energy of protein and solvent, or directly in a free-energy force field where an implicit solvation model approximates direct interactions with the solvent as well as most of the entropic contributions. We developed an all-atom protein force field (PFF01) [10,17,18] with an area-based implicit solvent model that approximates the free energy of peptide conformations in the natural solvent. Here we investigate the structurally conserved 40 amino-acid headpiece of the autonomously folding HIV accessory protein (1F4I-40) [19] with a modified basin-hopping technique [20,21]. Out of 20 simulations, the five energetically lowest correctly reproduced the NMR structure of this three-helix protein with a backbone rms deviation of less than 3 Å. The combination of decoy based model development [22] for the free energy with efficient stochastic optimization methods suggests a viable route for protein structure prediction at the all-atom level with present-day computational resources.

Model.—We have recently developed an all-atom protein force field (PFF01), which was used to reproducibly fold the 20 amino-acid trp-cage protein [10]. PFF01 [18] comprises an atomically resolved electrostatic model with group specific dielectric constants and a Lennard-Jones parametrization that was adapted to the experimental distance distributions from crystal structures of 138 proteins [23,24]. Interaction with the fictitious solvent are modeled in a simple solvent accessible surface approach [25], where the solvation free energies per unit surface were fitted to the enthalpies of solvation of the Gly-X-Gly series of peptides [26]. The only low-energy degrees of freedom available to the peptide during the folding process are rotations of the dihedral angles of the backbone and the side chains; these are the only moves considered during the simulation. There are two move classes, the small random rotations about a single angle and the library moves, which set a particular backbone dihedral to a permitted value in the Ramachandran plot.

If an accurate model for the free energy of the protein in its environment is available, stochastic optimization methods can be used to locate the global optimum of the free-energy landscape orders of magnitude faster than traditional simulation techniques. We adapted the basin-hopping technique [20] for protein simulations by replacing a single minimization step with a standard simulated annealing run [27] with self-adapting cooling cycle and length [28]. At the end of one annealing step, the new conformation was accepted if its energy difference to the current configuration was no higher than a given threshold energy ε_T, an approach recently proven optimal for certain optimization problems [29]. Each simulation was started from a nonclashing conformation with random backbone dihedral angels and performed in three separate steps: First, we used a high temperature bracket of 800/300 K ($\varepsilon_T = 15$ K) for the annealing window and reduced the strength of the solvent terms in the force field by 20%. The second step started from the final configurations of the first run, used the same annealing window, but the full solvent

TABLE I. Table of the ten lowest energy decoys (of 20, remainder available by request from the authors) with backbone rms deviation to the NMR structure and secondary structure content. The first row designates the secondary structure content of the NMR structure.

Name	RMSB	Energy	Secondary structure content
N	0.00		ccHHHHHHHHHclcbHHHHHHHHHHclcccHHHHHHHHHc
D01	2.34	−119.54	cHHHHHHHHHHHlcbcHHHHHHHHHHHHbHHHHHHHHHHc
D02	2.41	−117.52	cHHHHHHHHHHHlcbHHHHHHHHHHHHHbHHHHHHHHHHc
D03	2.76	−116.25	cHHHHHHHHHHHlcbHHHHHHHHHHHHHbHHHHHHHHHHc
D04	2.40	−115.85	cHHHHHHHHHHHlbbHHHHHHHHHHHHHbHHHHHHHHHHc
D05	2.43	−114.67	cHHHHHHHHHHHlcbHHHHHHHHHHHcbHHHHHHHHHHHc
D06	6.48	−114.06	cHHHHHHHHHHHcccbHHHHHHHHHHHHbHHHHHHHHHHc
D07	2.57	−113.65	cHHHHHHHHHHHlbbcHHHHHHHHHHHHbHHHHHHHHHHc
D08	4.61	−107.72	cHHHHHHHHHcclccHHHHHHHHHHHHHlclHHHHHHHHc
D09	4.14	−106.29	cHHHHHHHHHHHcbcbHHHHHHHHHbblcHHHHHHHHHHc
D10	5.92	−103.88	cHHHHHHHHHHHlcHHHHHHHHHbcbcclbHHHHHHHHHc

interactions. In the third step the resulting structures were further annealed in a low temperature bracket of 600/3 K ($\varepsilon_T = 1$ K). Within each annealing run the temperature was geometrically decreased; also the number of steps per annealing run was gradually increased to ensure better convergence. In total, each simulation comprised 10^7 energy evaluations with a computational effort roughly corresponding to a 10 ns molecular dynamics simulation (in vacuum with 1 fs time step). After this time no significant energy fluctuations occurred in the simulations, indicating that each had settled into a metastable configuration.

Results.—Using PFF01 we performed 20 independent modified basin-hopping simulations of the structurally conserved 40 amino-acid headpiece of the HIV accessory protein (pdb-code 1F4I, with sequence QEKEAIERLK ALGFEESLVI QAYFACEKNE NLAANFLLSQ). Although both fold into three-helix bundles, the HIV accessory protein is very different from the villin headpiece for which the force field was developed in both secondary and tertiary structures. The two proteins have less than 12% sequence similarity [30] and differ in their secondary structure content. The best structures found in each run were ranked according to their energy, and the rms backbone (RMSB) deviation to the NMR structure was computed. Table I demonstrates that the five lowest structures had to good accuracy converged to the NMR structure of the protein. The first non-native decoy appears in position six, with an energy deviation of 5 kcal/mol (in our model) and a significant RMSB deviation. The table demonstrates that all low-energy structures have essentially the same secondary structure; i.e., the position and length of the helices are always correctly predicted, even if the protein did not fold correctly. The degree of secondary structure content and similarity decreases for the decoys with higher energy (data not shown), in good correlation with their energy.

The good agreement between the folded and the experimental structure is also evident from Fig. 1(a), which shows the secondary structure alignment of the native and the folded conformations. The good physical alignment of the helices illustrates the importance of hydrophobic contacts to correctly fold this protein. An independent measure to assess the quality of these contacts is to compare the C_β-C_β distances (which correspond to the nuclear Overhauser effect constraints of the NMR experiments that determine tertiary structure) in the folded structure to those of the native structure. The color-coded C_β-C_β distance in Fig. 2 demonstrates a 66% (80%) coincidence of the C_β-C_β distance distances to within 1.5 Å (2.25 Å), respectively. The dark diagonal block indicates intrahelical contacts, which are, perhaps not too surprisingly, resolved to very good accuracy. The off-diagonal dark blocks, however, indicate that also a large fraction of long-range native contacts is reproduced correctly.

Starting from intermediate structures of the folding simulations, we generated over 60 000 low-energy conformations (decoys). Decoys with a root mean square deviation of the backbone of less than 3 Å were grouped into families with free-energy brackets of 2 kcal/mol. We then

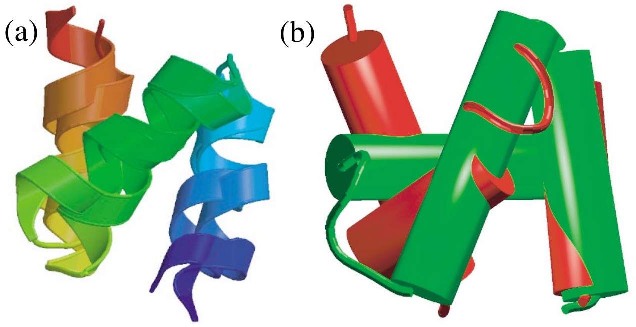

FIG. 1 (color). (a) Overlay of the secondary structure elements of the native configuration and the folded structure of 1F4I-40. (b) Overlays of the secondary structure elements of the native (green) configuration and the lowest misfolded decoy (red) of 1F4I.

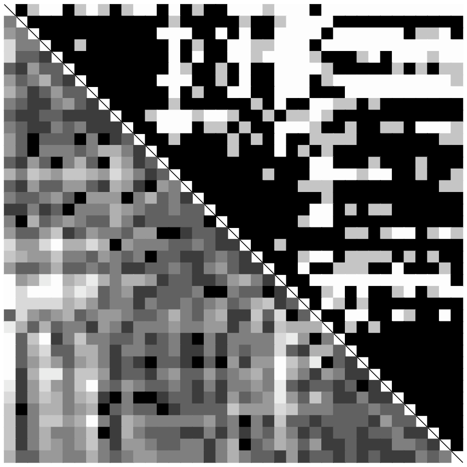

FIG. 2. C_β-C_β distance error map for the folded structure 1F4I in comparison to the NMR structure. The upper triangle shows absolute and the lower relative errors, respectively. Each square encodes the deviation between the C_β-C_β distance of two amino acids in the NMR to the C_β-C_β distance of the same amino acids in folded structure. Black (gray) squares indicate a deviation of less than 1.50 Å (2.25 Å). White squares indicate large deviations.

resolved the topological hierarchy [21,31,32] of the associated potential energy surface through the construction of a decoy tree (Fig. 3) that illustrates the low-energy structure of the free-energy surface. Beginning from the best conformation, we draw a vertical line for each decoy family in this window. Moving upward in energy, the number of decoys in each family grows almost exponentially in the low-energy region which we can resolve well. As a result, the diversity of each family grows until different families unite. Family membership is associative; i.e., as soon as two decoys in different branches have an RMSB deviation of less than 3 Å, all members of both families belong to one superfamily. Pictorially, this representation results in an inverted treelike structure that characterizes the topology of the metastable states of the free-energy surface.

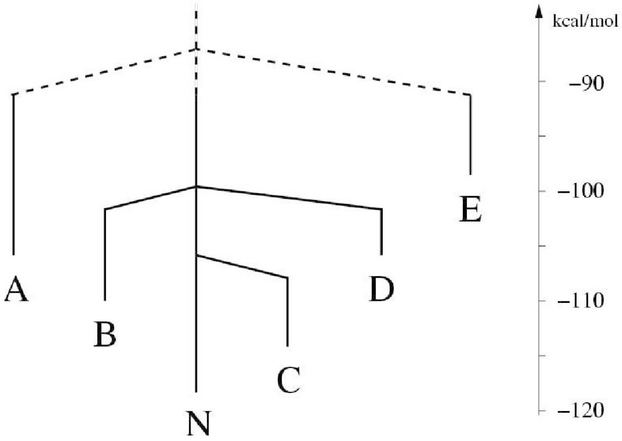

FIG. 3. Decoy tree of the low-energy configurations of the 40 amino-acid headpiece of the HIV accessory protein. The horizontal axis depicts the total energy, the chart on the right the total number of decoys in the primary and secondary funnels.

Trees with very short stems and many low-energy branches are characteristic of glassy potential energy surfaces, which are associated with Levinthals paradox [33,34] in the context of protein folding. Well structured trees with few terminal branches suggest the existence of a folding funnel [15], consistent with the "new" paradigm for protein folding [35]. From this perspective the structure in Fig. 3 is consistent with the existence of a very broad folding funnel [15] with pronounced competing secondary metastable conformations, which are depicted as the nonnative terminal minima of the tree. The discretization of the energy axis in intervals of 2 kcal/mol starting from the native conformation results in a smoothing of the free-energy surface. Each line in the figure corresponds to a family containing hundreds to thousands of structures, which are all associated with the same low-lying metastable conformation (the terminal point of the branch). Simulations trapped in such a metastable state must overcome a potential energy barrier of the order of the energy difference to the next highest branching point of the tree to visit another structure. The branching points of the tree were constructed only from structures of the decoy set and not through independent transition state search among the terminal structures. In addition, main-chain entropy is neglected in this analysis, which results in an overestimation of the barrier. Further investigations to determine more accurately the transition states are presently under way.

The lowest competing terminal branch (branch *C*), associated with decoy D06 in Table I, is less than 5 kcal/mol above the best native decoy. Decoy D06 has comparable energy to competing decoys but much larger RMSB deviation and has few long-range native contacts. This structure [see Fig. 1(b)] has also three helices of comparable length and sequence location, but differs from the native structure in the relative alignment of the helices with respect to each. The RMSB deviation of the low-lying terminal structures to the NMR structures is large (i.e., comparable with the RMSB deviation to unfolded structures), indicating that conserved secondary structure elements, rather than distance constraints, characterize the folding funnel. The low-lying local minima are thus characterized by varying spatial arrangements of similar secondary structure elements, a property that is ill represented by either RMSB deviation or the number of native contacts.

Discussion.—This work demonstrates reproducible folding for a three-helix protein in an all-atom free-energy force field. Investigations of the protein folding mechanism and methods for protein structure prediction will benefit from the availability of accurate physical models that are amenable to present-day computational resources [1]. This investigation brings together an all-atom free-energy model that is entirely determined by the amino-acid sequence alone and a simulation approach that permits predictive folding with moderate computational effort. It thus helps to close the gap between simplified protein models [3] and explicit-water molecular dynamics studies [13]. The results of this study are thus free from the influence of simplifying assumptions incurred by the use of either coarse grained models or artificial interactions designed to stabilize the native structure. The HIV accessory protein has a substantial hydrophobic core. Our analysis thus demonstrates that the PFF01 is capable of capturing the important competing physical interactions that lead to secondary and tertiary structure formation.

We have also developed a simulation approach that permits predictive folding, removing sampling uncertainties typical of unfolding [36] or replica exchange simulations [37]. The simulation methodology used here can succeed only in a free-energy approach, which is based on the thermodynamic hypothesis [14] that proteins are in thermodynamic equilibrium with the environment under physiological conditions. Following this hypothesis, the native structure of the protein can be predicted using stochastic optimization method orders of magnitude faster than by direct simulation. Our results demonstrate that the important influence of the solvent can be modeled with a relatively simple solvent accessible surface approach.

The analysis of the free-energy surface supports the funnel paradigm of protein folding for a nontrivial protein with a significantly larger hydrophobic core than was previously possible. The relatively small number of terminal branches of the decoy tree offers the first glimpse of the experimentally inaccessible structure of the folding funnel. It is commensurate with, but does not in itself prove, the existence of a broad folding funnel with well defined secondary metastable states which may constitute important folding intermediates. The free-energy optimization approach used here permits the characterization of these low-lying states, which surprisingly share very similar secondary structure with the native configurations. Further investigations must show whether this pattern persists also for other proteins.

We are grateful to S. Gregurick, J. Moult, and J. Pedersen for discussions and portions of the code used in these simulations. This work was supported by the Deutsche Forschungsgemeinschaft, the Bode Foundation, and the BMWF.

*Electronic address: wenzel@int.fzk.de

[1] D. Baker and A. Sali, Science **294**, 93 (2001).

[2] J. Schonbrunn, W.J. Wedemeyer, and D. Baker, Curr. Opin. Struct. Biol. **12**, 348 (2002).

[3] A. Liwo *et al.*, Proc. Natl. Acad. Sci. U.S.A. **99**, 1937 (2002).

[4] J. Moult *et al.*, Proteins: Struct., Funct., Genet. **45**, 2 (2001).

[5] R. Bonneau *et al.*, Proteins: Struct., Funct., Genet. **45**, 119 (2001).

[6] G.M. Crippen, J. Mol. Biol. **260**, 467 (1996).

[7] J. Pillardy *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 2329 (2001).

[8] C.D. Snow *et al.*, Nature (London) **420**, 102 (2002).

[9] C. Simmerling, B. Strockbine, and A. Roitberg, J. Am. Chem. Soc. **124**, 11 258 (2002).

[10] A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102 (2003).

[11] U.H.E. Hansmann, Phys. Rev. Lett. **88**, 068105 (2002).

[12] C. Lin, C. Hu, and U. Hansmann, Proteins: Struct., Funct., Genet. **53**, 436 (2003).

[13] Y. Duan and P.A. Kollman, Science **282**, 740 (1998).

[14] C.B. Anfinsen, Science **181**, 223 (1973).

[15] J.N. Onuchic, Z. Luthey-Schulten, and P. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[16] C. Hardin *et al.*, J. Comput. Chem. **23**, 138 (2003).

[17] T. Herges, A. Schug, H. Merlitz, and W. Wenzel, Nanotechnology **14**, 1161 (2003).

[18] T. Herges and W. Wenzel, Biophys. J. **87**, 3100 (2004).

[19] E.S. Withers-Ward *et al.*, Biochemistry **39**, 14 103 (2000).

[20] A. Nayeem, J. Vila, and H. Scheraga, J. Comput. Chem. **12**, 594 (1991).

[21] J.P. Doye and D. Wales, J. Chem. Phys. **105**, 8428 (1996).

[22] M. Chhajer and G.M. Crippen, BMC Struct. Biol. **6**, 4 (2002).

[23] M.J. Sippl, G. Nemethy, and H.A. Scheraga, J. Phys. Chem. **88**, 6231 (1984).

[24] R. Abagyan and M. Totrov, J. Mol. Biol. **235**, 983 (1994).

[25] D. Eisenberg and A.D. McLachlan, Nature (London) **319**, 199 (1986).

[26] K.A. Sharp *et al.*, Biochemistry **30**, 9686 (1991).

[27] S. Kirkpatrick, C. Gelatt, and M. Vecchi, Science **220**, 671 (1983).

[28] T.H.A. Verma, A. Schug, and W. Wenzel, "Comparision of Stochastic Optimization Methods for All-Atom Protein Folding" (to be published).

[29] J. Schneider, I. Morgenstern, and J. Singer, Phys. Rev. E **58**, 5085 (1998).

[30] C. Gille *et al.*, Bioinformatics **12**, 2489 (2003).

[31] O. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).

[32] C.L. Brooks, J.N. Onuchic, and D.J. Wales, Science **293**, 612 (2001).

[33] C. Levinthal, J. Chim. Phys. Physicochim. Biol. **65**, 44 (1968).

[34] B. Honig, J. Mol. Biol. **293**, 283 (1999).

[35] A. Šali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).

[36] U. Mayor *et al.*, Nature (London) **421**, 863 (2003).

[37] A.E. Garcia and N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **100**, 13 898 (2003).