Free-Energy Landscape of the Villin Headpiece in an All-Atom Force Field

Thomas Herges and Wolfgang Wenzel*
Forschungszentrum Karlsruhe
Institut für Nanotechnologie
Postfach 3640
D-76021 Karlsruhe
Germany

Summary

We investigate the landscape of the internal free-energy of the 36 amino acid villin headpiece with a modified basin hopping method in the all-atom force field PFF01, which was previously used to predictively fold several helical proteins with atomic resolution. We identify near native conformations of the protein as the global optimum of the force field. More than half of the twenty best simulations started from random initial conditions converge to the folding funnel of the native conformation, but several competing low-energy metastable conformations were observed. From 76,000 independently generated conformations we derived a decoy tree which illustrates the topological structure of the entire low-energy part of the free-energy landscape and characterizes the ensemble of metastable conformations. These emerge as similar in secondary content, but differ in tertiary arrangement.

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 (Baker and Sali, 2001; Moult et al., 2001; Schonbrunn et al., 2002). Investigations of the protein landscape may offer insights into the folding funnel and help elucidate folding mechanism and kinetics. A complete characterization of the low-energy landscape for proteins remains a difficult task because of their complexity.

One of the central paradigms of protein folding is that many proteins in their native conformation are in thermodynamic equilibrium with their environment (Anfinsen, 1973). Exploiting this characteristic, researchers have been able to predict the structure of the protein by locating the global minimum of its free-energy surface (Li and Scheraga, 1987; Liwo et al., 2002; Onuchic et al., 1997) without recourse to the folding dynamics, a process that is potentially much more efficient than the simulation of the folding process. For questions only regarding the folded structure of the protein, PSP based on global optimization of the free-energy may offer a viable alternative, provided that suitable parameterization of the free-energy of the protein in its environment exists and that the global optimum of this free-energy surface can be found with sufficient accuracy (Li and Scheraga, 1987).

We have recently demonstrated a feasible strategy for all-atom protein structure prediction (Schug et al., 2003; Herges and Wenzel, 2004a, 2004b) in a minimal thermodynamic approach. We developed an all-atom force field for proteins (PFF01), which parameterizes the internal free-energy (Snow et al., 2004) of a protein on the basis of physical interactions (Herges and Wenzel, 2004a). Using various stochastic optimization methods (Schug et al., submitted), we have already demonstrated the reproducible and predictive folding of several proteins in PFF01: the 40 amino acid HIV accessory protein (1F4I) (Herges and Wenzel, 2004b), the 20 amino acid trp-cage protein (1L2Y) (Schug et al., 2003), the 60 amino acid four-helix bacterial ribosomal protein L20 (1GYZ) (Raibaud et al., 2002; Schug et al., 2004b), and the DNA binding domain of the human centromere protein B with 56 amino acids (unpublished data). In addition, we could show that PFF01 stabilizes the native conformations of other helical proteins, e.g., a 45 amino acid headpiece of protein A (Snow et al., 2002; Gouda et al., 1992; Zhou and Karplus, 1999; Vila et al., 2004), and the engrailed homeodomain (1ENH) from *Drosophilia melangaster* (Clarke et al., 1994; Mayor et al., 2003).

Here we investigate tertiary structure formation of the autonomously folding 36 amino acid headpiece of the villin protein (PDB code, 1VII). The villin headpiece did not fold in a landmark explicit-water all-atom simulations using the AMBER (Duan and Kollman, 1998) force field. A recent optimization-based approach using the ECEPP/2 force fields with an implicit solvent model (Hansmann, 2002) found a two-helix structure as the estimate of the global optimum of the free-energy surface. Three-helix structures were also observed, indicating that the difficulty of folding this protein in silico may originate from both deficiencies of the force field and/or limitations in the sampling of the conformational space.

Here we performed 50 independent simulations using modified basin hopping technique (comprising 1.67 × 10^8 energy evaluations in total) from random initial conditions. Half of the runs found conformations associated with the folding funnel leading toward the native state, the lowest energy simulation associated with the native folding funnel resulted in a structure with a 3.3 Å backbone rms deviation (rmsd) to the NMR conformation. We also found one low-lying metastable conformation within 1 kcal/mol of the independently obtained estimate of the global optimum of the free-energy surface. We find that in silico folding of the villin headpiece requires at least ten times the numerical effort that was invested in the larger 40 amino acid HIV accessory protein, a three-helix bundle, which we recently folded reproducibly using the same force field and the same optimization strategy.

In order to rationalize these differences, we characterized the low-energy portion of its free-energy surface (FES) of the protein using a decoy tree approach

*Correspondence: wenzel@int.fzk.de

(Becker and Karplus, 1997) generated from over 76,000 decoys that grouped into 14,000 families. This analysis demonstrates that the villin headpiece has several metastable three-helix conformations, which have almost the same energy and secondary structure content as the native state, but which differ significantly in tertiary structure from the NMR conformation. This characterization of the relevant part of the free-energy surface of the protein is presently difficult to accomplish by other means, but may contribute significant insight into the folding properties of proteins both experimentally and in silico.

Results

We first performed 500 independent Monte-Carlo simulations at 500 K on random initial conformations of the villin headpiece (PDB code, 1VII; sequence, MLSDE DFKAV FGMTR SAFAN LPLWK QQNLK KEKGLF), each comprising 2 × 10^5 energy evaluations. The initial conformations were generated by setting the dihedral angles of the protein to random values and selecting nonclashing conformations; they have no secondary structre. The 50 conformations with the lowest energies that differed by at least 4 Å in their backbone rmsd from one another were selected as starting conformations for the optimization runs. These conformations span a wide variety of possible structures, but we nevertheless found that they already contain much of the correct secondary structure. During this phase of the simulation, the helical content of these structures increased from zero to about 45% (the native conformation has 53% helical content). In this context it is important to note that the simulation temperature (500 K) is largely unrelated to the "system temperature" that is determined by the parameterization of the free-energy force field (300 K), because the force field has no kinetic energy term.

Starting from these conformations, we performed 50 independent basin hopping runs, as described in the methods section, for 100 cycles each, resulting in 67 × 10^6 energy evaluations for each decoy. The dependence of the energies as a function of the basin hopping cycle for 20 of these runs that resulted in the lowest energies is shown in Figure 1. The efficiency of the basin hopping technique to generate new, nontrivial structures with better energies is shown in the inset of the figure, which shows a histogram of the energy gain for a basin hopping step (i.e., the energy difference between the starting conformation and the conformation at the end of the SA run). The first 20 cycles were discarded for the construction of the histogram to eliminate the influence of equilibration effects in the beginning of the simulation. The basin hopping method used an threshold acceptance criterion, according to which all basin hopping steps with $\Delta E < 3$ kcal/mol (shown blue in the histogram) were accepted (acceptance ratio, 35%). Length and temperature bracket of the underlying SA simulation must be adapted to ensure a significant acceptance ratio to permit nontrivial dynamics in the simulations. When this is achieved, the energy of each individual run fluctuates significantly, as is evident in the figure. Some simulations visit the global minimum (e.g., red, pink, and violet curves), while others (e.g., dotted red curve) never even come close within the allotted time. In order to obtain an estimate of the energy of the global optimum of the FES, we performed 20 basin-hopping simulations starting from the NMR conformation. For these simulations the threshold acceptance criterion ensures that the vicinity of the NMR conformation is extensively probed. We therefore estimate the global optimum of the force field at −86.59 kcal/mol with a near native conformation (rmsd 3.6 Å).

The features of the best conformations of the 20 simulations that reached the lowest energies are summarized in Table 1, along with their rmsd to the NMR conformation and their secondary structure content. We note that best energies of these low-energy decoys are very close to one another and that a nonnative conformation emerges as the lowest decoy (energy difference, 0.69 kcal/mol). We also note that even the best conformation (decoy D01) fails to reach the energy of the NMR decoy of the decoy tree (−86.59 kcal/mol; blue line in Figure 1). Given the fluctuations in the energies of the basin-hopping procedure (see Figure 1), the present optimization strategy is inadequate to resolve energy differences on this scale. Should the FES have several minima that are energetically closer than the scale of the threshold criterion (3 K), it is a natural (and desirable) feature to visit these structures with nearly equal probability. The resolution of the available optimization methods is presently insufficient to ensure that all simulations converge to the native conformation.

In order to ensure the validity of this conclusion, we selected the best conformations from the five lowest energy runs. For each of these, we performed an additional simulation of 40 basin hopping cycles, comprising 50 × 10^6 energy evaluations. These demonstrated that up to 2 kcal/mol can be gained in the total energy of some structures, but we found no conformations with energies lower than the NMR decoy. The simulations associated with D01 and D02 never produced conformations that had a lower energy than the starting conformation and the energetic order of the other conformations was not significantly changed. These results support the conclusion that the present implementation of the modified basin hopping technique had reached the limit of its energy resolution.

In comparison to earlier studies (Daura et al., 1998; Lin et al., 2003), it is encouraging that three-helix structures with the right secondary structure dominate the low-energy decoys. The fraction of correct native structure increased further from the end of the high-temperature runs to the end of the basin-hopping runs. The rmsd values in the table also demonstrate that obtaining the correct secondary structure is a necessary, but not a sufficient condition for proper folding of the protein.

This investigation of the villin headpiece had consumed more than the 10-fold computational effort that was required to reproducibly and predictably fold the structurally conserved 40 amino acid headpiece of the HIV accessory protein (Herges and Wenzel, 2004b; Schug et al., 2004a). The two proteins have less than 12% sequence homology (Gille et al., 2003) and differ in their secondary structure content, but both fold into compact three-helix bundles. Although the HIV acces-

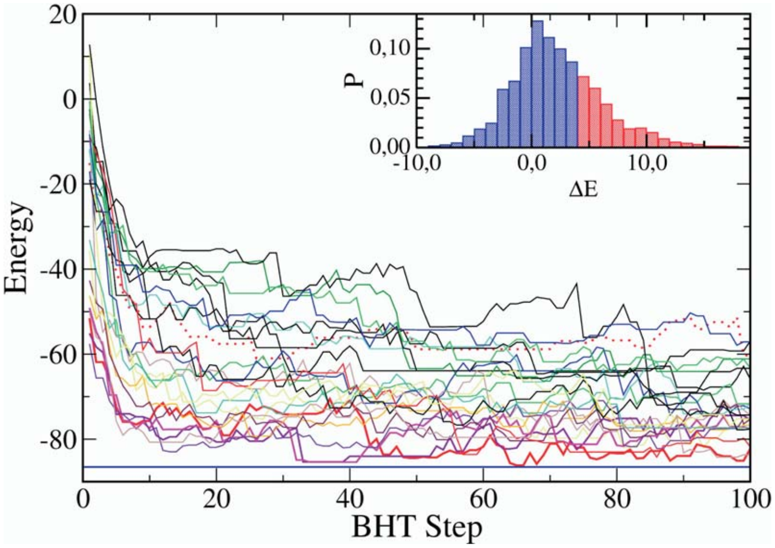

Figure 1. Accepted Energy as a Function of Step Number for the Best Twenty of Fifty Basin-Hopping Simulations Described in the Text

The bold red curve indicates the simulation that visited the best conformation; note that significant energy fluctuations are required to explore the conformational space. On the energy scale of the simulation, the best energies obtained are indistinguishable from the estimate for the globally best energy that was obtained independently (see text). Inset: Histogram of the energy difference between the starting and the final conformation in a single basin hopping step. Blue/red areas indicate accepted and rejected moves. Data was accumulated after the equilibration period of the first 20 steps to reflect equilibrium properties of the simulation.

sory protein is slightly larger, 8 of 20 basin-hopping simulations found the NMR structure and no competing low-energy decoys emerged. This finding illustrates the significant differences in the in silico folding rates of even the family of naturally occurring three-helix bundles.

Differences in the in silico folding rates of for protein models of equal lengths have often been traced to various measures of the complexity of their sequence dependent potential energy surfaces, but investigations into realistic proteins have been hampered by the lack of adequate all-atom models for which folding to the native state could be observed. In order to analyze the difference in the folding behavior, we have attempted to characterize the low-energy portion of the free-energy landscape of the villin headpiece and compare it with the landscape of the HIV accessory protein. Starting from randomized initial conformations, we generated in excess of 76,000 structures with the modified basin-hopping method. These were grouped in to 14,000 decoy families, according to the procedure outlined in the methods section and a decoy tree was generated. The low-energy portion of the 1VII free-energy surface is illustrated in Figure 3, it features seven almost isoenergetic nonnative terminal branches. The decoy associated with the terminal branches of the tree are shown in Figure 2. There are five three-helix and two two-helix

Table 1. Top Twenty Decoys after the Initial Simulations of the Villin Headpiece in PFF01

Code	Energy	Rmsd	Branch	Dist	Secondary Structure
NMR	–86.59				CHHHHHTTSSSCHHHHTTSCHHHHHHHHHHHTTCC
D01	–86.23	7.57	C	(3.31)	CHHHHHHHHHHCHHHHHHSHHHHHHHHHHHHHTCC
D02	–85.51	4.56	N	(3.27)	CHHHHHHHTSCHHHHHHCHHHHHHHHHHHHHTCC
D03	–85.35	5.80	N	(4.86)	CSHHHHHHHHHCHHHHHHCHHHHHHHHHHTTTCCC
D04	–85.11	4.30	N	(3.23)	CHHHHHHHTSCSHHHHHCHHHHHHHHHHHHHHTCC
D05	–84.08	4.13	N	(3.35)	CHHHHHHHTSCHHHHHHSHHHHHHHHHHHHHTCC
D06	–83.46	7.70	C	(3.14)	CHHHHHHHHHHCSSCSSCHHHHHHHHHHHHHTCC
D07	–82.97	4.62	N	(3.37)	CHHHHHHHCHHHHHHHHHSHHHHHHHHHHHHHSCC
D08	–82.55	5.93	N	(4.65)	CHHHHHHHHHHHHCHHHHHSCTTTCHHHHHHHHHC
D09	–82.03	6.50	A	(4.42)	CCCHHHHHHHTCSCHHHHHHSSSHHHHHHHHHHT
D10	–81.07	6.76	E	(4.08)	CCCHHHHHHHTSSCCSSCSSHHHHHHHHHHHHT
D11	–80.96	7.42	A	(4.00)	CHHHHHHHTCSCHHHHHHSCSHHHHHHHHHHHTCC
D12	–80.38	7.11	X		CHHHHHHHHHHHCCSHHHHCSSHHHHHHHHHHHTCC
D13	–80.36	7.56	C	(3.49)	CHHHHHHHHHHHHTTCCSSSHHHHHHHHHHHHHTCC
D14	–79.94	8.29	X		CHHHHHHHHHTTTSSCSCSSHHHHHCHHHHHHHT
D15	–79.85	3.86	N	(4.54)	CHHHHHHHHHTCSCHHHHHHSCHHHHHHHHHHHTS
D16	–79.25	5.03	N	(3.48)	CHHHHHHHHHCSSCHHHHHSHHHHHHHHHHHHHT
D17	–78.74	2.66	N	(2.75)	CHHHHHHHTSCCHHHHHHSCHHHHHHHHHHHHTCC
D18	–78.39	7.58	C	(3.59)	CHHHHHHHHHCCCHHHHHSHHHHHHHHHHHHHTCC
D19	–78.35	3.18	N	(3.49)	CHHHHHHHTSCCHHHHHHSCHHHHHHHHHHHHHHC
D20	–78.21	7.30	B	(4.00)	CHHHHHHHHHHHHHHHHHHSHHHHHHHHHHHHHTCC

Top 20 decoys after the initial simulations of the villin headpiece in PFF01 (see text) with their energy (in kcal/mol), the rmsd to the native structure, the closest terminal branch of the tree (see text), and the rmsd between the terminal branch and the decoy. Next we show the label of the conformation of the closest branch in the decoy tree (Figure 3) and the rmsd between the two conformations. An X donates a conformation that is larger than 5 Å rmsd from all branches of the tree. The last column shows the secondary structure content (computed with DSSP).

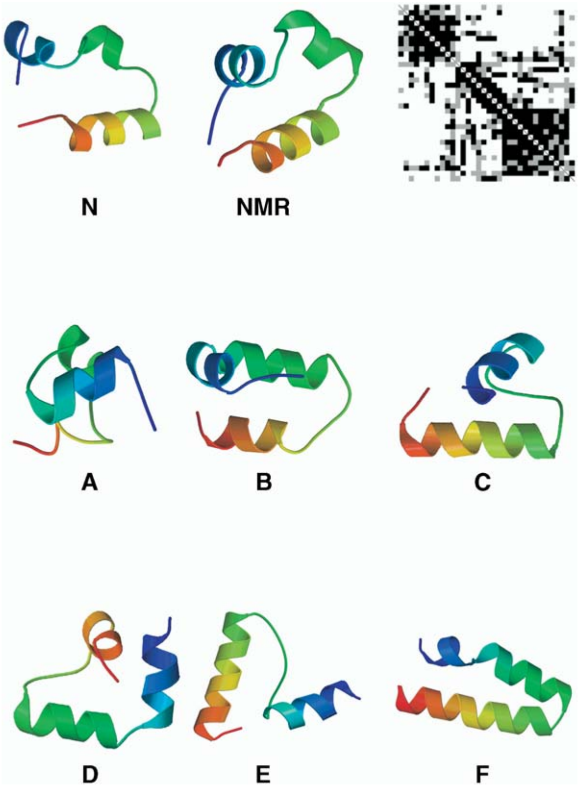

Figure 2. Native Structure and the Conformations of the Terminal Branches of the Decoy Tree of the Villin Headpiece

The labels refer to the branches in Figure 3 and the numbers indicate the rmsd deviations to the native structure (labeled NMR). The color-coded distance map in the top right compares the C_β and $-C_\beta$ distances of the native structure with that of the NMR decoy. The density of each square corresponds to the degree of similarity of the experimental and the model C(beta)-C(beta) distances for each pair of amino acids, which are enumerated along the protein sequence along the horizontal and vertical axis of the figure. Black (gray) squares indicate that the C_β and $-C_\beta$ distances of the native and the other structure differ by less than 1.5 and 2.25 Å, respectively. White squares indicate larger deviations.

structures at the bottom of their respective branches of the tree. One notes a distinct propensity for the formation of helices in those regions where the NMR structure forms helices 1 and 3. The central helix occurs less often and leads to misfolded decoys reminiscent of the two-helix structures found in (Lin et al., 2003). The energy spectrum shows a gap of less than 1 kcal/mol between the NMR decoy and the next competing structure and becomes nearly continuous with increasing energy. An analysis of the rmsd matrix of all low energy decoys suggests that many distinct metastable conformations were probed in the search.

In order to further analyze the results of the folding runs, we have computed the rmsd distance matrix between the conformations associated with the terminal branches of the decoy tree (Figure 3), as depicted in Figure 2 and the conformations D01–D20. Table 1 shows the label of the closest branch for each decoy and its rmsd deviation to the bottom structure. Out of the 20 simulations, the native branch (N) was visited 11 times, branch C was visited 4 times, branch A 2 times, and branches C and B 1 time. Two conformations (labeled X) could not be associated with any terminal branch, but their energies are very high. This finding crossvalidates the independent generation of the decoy tree and the simulation runs. The data from the construction of the decoy tree indicates that the folding runs did efficiently explore the low-energy conformational space, even if their energy resolutions is too low to ultimately distinguish between the conformations associated with the native and nonnative low-energy conformations. Similarly, it is encouraging that all low-energy decoys of the folding simulations could be associated with families already present in the tree, indicating that the decoy tree may contain a nearly complete characterization of the low-energy surface of the villin headpiece. It is also encouraging that 50% of the simulation runs are at least associated with the native branch.

Discussion

Based on this data, it is possible to compare the decoy trees of the villin headpiece and the HIV accessory protein (Herges and Wenzel, 2004a), see Figure 4. The latter has a much less branched structure and all competing low-energy branches terminate at a much higher relative energy. To illustrate the features of the FES surface associated with a given tree, we have schematically indicated two one-dimensional surfaces in the figure that are commensurate with the respective trees. Note that the distribution of branching points permits a qualitative comparison of the energy barriers required to bridge tow families. A comparison of these two surfaces thus indicates that folding simulations for the villin headpiece have much more opportunity to go astray when compared to those of 1F4I and will require much longer time to escape from metastable conformations.

Long-lived metastable conformations, such as those associated with terminal branches of the decoy tree are presently difficult to characterize experimentally or by direct simulation. Free-energy optimization strategies thus offer a unique opportunity to characterize the low-energy portion of the FES, even though they are incapable of characterizing the extended ensembles of folding intermediates that are characterized by large contributions of backbone conformational entropy (Garcia and Onuchic, 2003). We have therefore computed measures of native content for all decoy and generated trees in which the width of the branch illustrates similarity with the native structure for the chosen criterion. Figure 5 shows four such trees that indicate the presence of helices 1, 2, and 3 and the fraction of long-range native content in the decoy set respectively.

In addition to the topological information present in the tree, this information permits a characterization of the physical properties of the folding funnel and its metastable branches. By definition, all native characteristics are fully present in the terminus of the native branch. Helix 1 (amino acids: 4–11), for example, is also present in branches C, E, and F, but its content actually decreases with decreasing energy in branch B. Helix 2

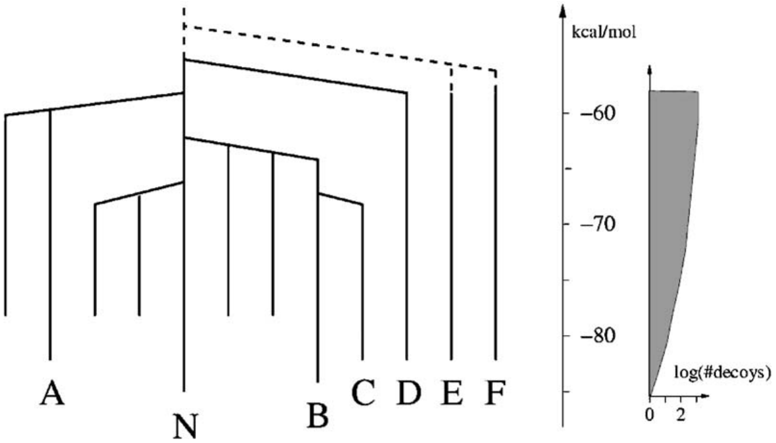

Figure 3. Decoy Tree for the Villin Headpiece in the Optimized Force Field

The structures corresponding to the terminal branches of the tree are shown in Figure 2. The tree was constructed from a set of 76,000 structures grouped in 14,000 families, as discussed in the Experimental Procedures.

(15–20), by comparison, is present in A, B, and D, but not in C. Helix 3 (24–34) is present only in branch C. Provided that the characterization of the free-energy surface in the tree is complete, this analysis suggests that none of the long-lived folding intermediates have the complete secondary structure. As a result, it is unlikely that a sequential folding scenario along a single path, reminiscent of Levinthals paradigm (Levinthal, 1968), is realized in this protein.

Instead, different folding paths will generate conformations in either the native or one of the nonnative families, which differ in their secondary structure content. In processes moving from one nonnative family to the next, part of the original secondary content is lost, and secondary structure in a different region of the protein is formed instead. Branches B and C apparently have too much content of either helix 2 or 3, respectively, to generate close by conformations of similar energy. Only transitions leading to the native branch lead to an increase in secondary structure. Note that the topological information in the tree does not imply that a folding path that has visited C must visit B before reaching N, as the trees represent highly simplified projections of complex multidimensional surfaces.

Of particular interest is the bottom panel of the figure, which illustrates the fraction of native contacts. It clearly demonstrates that essentially all long-range native contacts are formed only at the bottom of the native folding funnel. Few long-range contacts are already present in the metastable funnels or near the branching points. This indicates that the formation of nearly the entire secondary structure is required for the formation of stabilizing long-range contacts for the villin headpiece. Note that while all helices are prevalent already at comparatively high energies in the native funnel, the formation of the long-range native contacts occurs in a very small fraction of the conformation space, that appears accessible only after all secondary structure has formed. Even though the free-energy optimization strategy followed here yields no direct dynamical information, this finding is commensurate with a folding scenario in which the long-range native contacts are formed only a the very end of the folding process, after the transition state. In the absence of the characterization of the transition states separating the metastable conformations, however, no immediate conclusions regarding the experimental folding rate of the protein can be drawn from the decoy trees.

Conclusions

Recent investigations of all-atom protein structure prediction with free-energy models (Schug et al., 2003; Vila et al., 2004; Herges and Wenzel, 2004a, 2004b) provide increasing evidence that the native tertiary structure can be predicted as the equilibrium conformation of suitable free-energy force field, a finding commensurate with the free-energy paradigm of protein folding (Anfinsen, 1973). We find that PFF01 stabilizes the native conformation of a family of nonhomologous helical proteins as its global minimum. The all-atom representation in PFF01 or similar force fields permits the parameterization of the free-energy landscape on the ba-

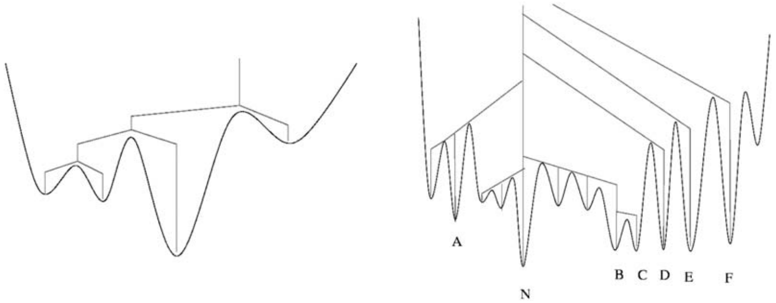

Figure 4. Schematic Representations of One-Dimensional Potential Energy Surfaces Compatible with the Decoy Trees Obtained for the All-Atom Models of the HIV Accessory Protein and the Villin Headpiece

HIV accessory protein, left; villin headpiece, right.

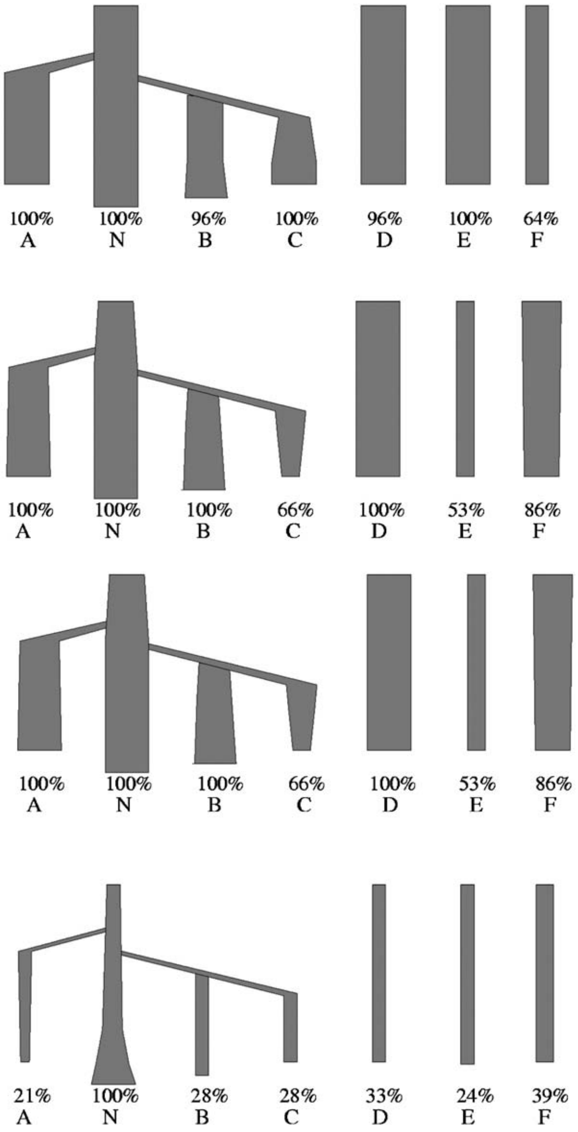

Figure 5. Illustration of the Secondary and Tertiary Structure Content of the Low-Energy Portion of the Free-Energy Surface of the Villin Headpiece

Each figure shows the decoy tree of the villin headpiece, where the width of the branch is proportional to some characteristic of the decoy family at the given energy. From top to bottom the characteristics are: the fraction of helical content in the regions where the native conformation has helices 1, 2, and 3, respectively. The bottom panel corresponds to the fraction of long-range native contacts, defined as the fraction of C_β and $-C_\beta$ distances that differed less than 2 Å from their native counterparts from residues 3–14 with residues 23–33.

sis of physical interactions that are well understood for smaller systems. Such parameterization promise better transferability in comparison to knowledge based methods, but can only be exploited when efficient optimization methods are available to accurately and reliably determine the global optimum of the FES at the all-atom level.

The use of physical interactions in all-atom representations incurs a large computational cost when compared to more coarse-grained or homology-based models. In an optimization approach, this increase in cost is partially compensated by the efficiency of the conformational search. Here we demonstrated that the native conformation of the villin headpiece and its entire low-energy FES can be characterized with atomic resolution with present-day computational resources. The total computational effort invested in this study corresponds to an MD trajectory of approximately 1.2 μs, comparable to the effort invested in a landmark explicit water simulation MD (Duan and Kollman, 1998) of the same system. This simulation failed to fold into the native state, while other recent MD simulations indicate that the native conformation is visited only rarely on this timescale. Our results suggest that protein structure prediction using stochastic optimization methods may become a viable alternative to explicit water MD for small proteins. In comparison to folding by molecular dynamics or replica exchange methods (Garcia and Onuchic, 2003), the disadvantage of the optimization strategies is the loss of dynamical information. We also note that the presently available evidence indicates that explicit water simulation models are superior to implicit solvent models with regard to the accuracy in which the native conformation is resolved (Simmerling et al., 2002).

In comparison to earlier implicit solvent free-energy models (Lin et al., 2003), a force field that correctly predicts the native conformation of several nontrivial proteins with 20–60 amino acids as its global optimum is now available. There is presently no evidence to suggest that this force field would not fold larger proteins, but optimization methods to perform such experiments with presently available computational resources are still lacking. Progress toward the study of larger proteins depends on further improvements of the energy resolution of the optimization method. Successful techniques must bridge the gap between efficient exploration of the overall structure of the FES and the very accurate resolution of competing local minima. Even comparatively simple three-helix proteins appear to pose significant challenges for which the villin headpiece provides an interesting example: while we were able to fully characterize the low-energy free-energy surface of the protein, reproducible folding, which could be achieved for the trp-cage protein (Schug et al., 2003) and the HIV accessory protein (Herges and Wenzel, 2004b), presently overtaxes the energy resolution of the basin-hopping method. The close proximity of the different branches of decoy tree may provide a rationalization of the difficulties encountered in prior folding studies, which also failed to converge to the native structure (Daura et al., 1998; Lin et al., 2003).

The construction of the decoy tree and its comparison for different proteins permits a qualitative rationalization of the folding dynamics that is presently inaccessible by other means. The number of branches of the decoy tree, their complexity, and energy distribution offers a straightforward characterization of the folding funnel of the protein and augments dynamic characterizations of the transition state ensemble (Garcia and Onuchic, 2003).

Experimental Procedures

Model

We have recently developed an all-atom (with the exception of apolar CH$_n$ groups) force field (PFF01) for the internal free-energy (Snow et al., 2004) of proteins, excluding backbone entropy, based on physical interactions (Herges et al., 2002; Herges and Wenzel, 2004a). Keeping bond-length and non-diehedral bond angles fixed, the force field parameterizes the nonbonded interactions:

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

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 as a set of 138 proteins of the PDB database (Abagyan and Totrov, 1994; Herges et al., 2002, 2004). 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 ϵ_i were previously derived in a potential-of-mean-force approach (Avbelj and Moult, 1995). Interactions with the solvent were first fit in a minimal solvent-accessible surface model (Eisenberg and McLachlan, 1986) parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly-X-Gly family of peptides (Sharp et al., 1991). 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 (Herges et al., 2003). 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) that takes the form:

$$V_{hb}(CO_i, NH_j) = R(\bar{r}_{ij})\Gamma(\phi,\theta_{ij}), \quad (2)$$

where $\bar{r}_{ij}$,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 potentials of mean force to the same set of proteins described above (Abagyan and Totrov, 1994; Herges and Wenzel, 2004a).

Move Classes

In the folding process, at 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 (a single) side chain or backbone dihedral angle, 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; the remainder is 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 (Pedersen and Moult, 1997). While driving the simulation toward the formation of secondary structure, it contains no bias toward helical or sheet structures beyond that encountered in nature. We note that 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 force field and the optimization methods are implemented in our program package POEM (Protein Optimization with Energy Methods), which was used to carry out all simulations.

Optimization Method

Monte Carlo with minimization (MCM), also known as basin-hopping approach, has been used to locate the global minima of many complex potential energy surfaces (Li and Scheraga, 1987; Doye and Wales, 1996; Wales and Doye, 1997; Herges and Wenzel, 2004a, 2004b). The minimization step simplifies the potential energy surface by mapping each conformation to a nearby local minimum. The increase of efficiency of MCM in comparison to the Monte Carlo method (Metropolis et al., 1953) on the original potential energy surface strongly depends on the average energy gain in the minimization procedure. For very rugged potential energy surfaces, such as those encountered in protein folding, local minimization yields comparatively little improvement. We have therefore replaced the local minimization by a simulated annealing (Kirkpatrick et al., 1983) run, starting at 660 K and then cooled with a geometric cooling cycle to 1 K. The number of steps in the cooling cycle is gradually increased according to $N_c = 10^5\sqrt{n_m}$, with the number of the minimization cycle n_m. The resulting configuration replaces the starting configuration according to a threshold acceptance criterion with a threshold of 3 kcal/mol. Other than in most other Monte Carlo schemes, the new conformation is accepted with probability one, if the energy is less than the threshold; otherwise, it is rejected.

Free-Energy Surface Topology

The topology of the low-energy part of the FES was analyzed in a decoy tree (Becker and Karplus, 1997; Brooks et al., 2001) that groups conformations in a given energy range into families, as a distance measure we used backbone rmsd, obtained as a rigid-body least-squares fit of two conformations. The tree was constructed from all decoys (local minima that differ by at least 1 Å rmsd from all other decoys) encountered in the simulations for sequence of equidistant energies E_0, E_1..., starting with the energy of the best conformation. A decoy with energy below E_n that has less than 3 Å rmsd to the decoy of just one family at the next lower energy level E_{n-1} is included into that family. If a decoy is associated with more than one family, the corresponding families are united, if it belongs to no existing family a new family containing just this decoy is created. For each family we by draw a vertical line in the energy window between E_{n-1} and E_n and merge the lines for the energy E_n where the families are united. This analysis results in an inverted tree-like structure that illustrates the energetic order and degree of structural similarity of conformations via their family association. For a force field that stabilizes the native structure the native family is represented by the branch of the tree that extends the furthest downward. In a force field that stabilize nonnative structures, perturbations in the parameters rebalance the lower portion of the tree.

Acknowledgments

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. We thank the Fond der Chemischen Industrie, the BMBF, the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2 and WE 1863/14-1), and the Kurt Eberhard Bode Stiftung for financial support. We thank Dr. K.-H. Lee and the KIST supercomputational materials lab, where some of the simulations reported here were performed.

References

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.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223–230.

Avbelj, F., and Moult, J. (1995). Role of electrostatic screening in determining protein main chain conformational preferences. Biochemistry *34*, 755–764.

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. Science *294*, 93–96.

Becker, O., and Karplus, M. (1997). The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. J. Chem. Phys. *106*, 1495–1517.

Brooks, C.L., Onuchic, J.N., and Wales, D.J. (2001). Taking a walk on a landscape. Science *293*, 612–613.

Clarke, N., Kissinger, C., Desjarlais, J., Gilliland, G., and Pabo, C. (1994). Structural studies of the engrailed homedomain. Protein Sci. *3*, 1779–1787.

Daura, X., Jaun, B., Seebach, D., van Gunsteren, W., Daura, A.E.M., Juan, B., Seebach, D., van Gunsteren, W., and Mark, A.E. (1998). Reversible peptide folding in solution by molecular dynamics simulation. J. Mol. Biol. *280*, 925–932.

Doye, J.P., and Wales, D. (1996). On potential energy surfaces and relaxation to the global minimum. J. Chem. Phys. *105*, 8428.

Duan, Y., and Kollman, P.A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science *282*, 740–744.

Eisenberg, D., and McLachlan, A.D. (1986). Solvation energy in protein folding and binding. Nature *319*, 199–203.

Garcia, A.E., and Onuchic, N. (2003). Folding a protein in a computer: an atomic description of the folding/unfolding of protein a. Proc. Natl. Acad. Sci. USA *100*, 13898–13903.

Gille, C., Lorenzen, S., Michalsky, E., and Frömmel, C. (2003). Kiss for strap: user extensions for a protein alignment editor. Bioinformatics *12*, 2489–2491.

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*, 9665–9672.

Hansmann, U.H.E. (2002). Global optimization by energy landscape paving. Phys. Rev. Lett. *88*, 68105.

Herges, T., and Wenzel, W. (2004a). An all-atom force field for tertiary structure prediction of helical proteins. Biophys. J. *870*, 3100–3109.

Herges, T., and Wenzel, W. (2004b). Reproducible in-silico folding of a three-helix protein in a transferable all-atom forcefield. Phys. Rev. Lett. *94*, 18101.

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

Herges, T., Schug, A., Merlitz, H., and Wenzel, W. (2003). Stochastic optimization methods for structure prediction of biomolecular nanoscale systems. Nanotechnology *14*, 1161–1167.

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. Intl. J. Quant. Chem. *99*, 854–893.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. Science *220*, 671–680.

Levinthal, C. (1968). Are there pathways for protein folding? J. Chim. Phys. *65*, 44–45.

Li, Z., and Scheraga, H. (1987). Monte carlo minimization approach to the multiple minima problem in protein folding. Proc. Natl. Acad. Sci. USA *84*, 6611–6615.

Lin, C., Hu, C., and Hansmann, U. (2003). Parallel tempering simulations of hp-36. Proteins *53*, 436–445.

Liwo, A., Arlukowicz, P., Czaplewski, C., Oldizeij, S., Pillardy, J., and Scheraga, H. (2002). A method for optimising potential energy functions by a hierarchichal design of the potential energy landscape. Proc. Natl. Acad. Sci. USA *99*, 1937–1942.

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 micorseconds. Nature *421*, 863–867.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. *210*, 1087–1092.

Moult, J., Fidelis, K., Zemia, A., and Hubbard, T. (2001). Critical assessment of methods of protein structure (casp): round iv. Proteins *45*, 2–7.

Onuchic, J.N., Luthey-Schulten, Z., and Wolynes, P. (1997). Theory of protein folding: the energy landscape perspective. Annu. Rev. Phys. Chem. *48*, 545–600.

Pedersen, J.T., and Moult, J. (1997). J. Mol. Biol. *269*, 240–259.

Raibaud, S., Lebars, I., Guillier, M., Chiaruttini, C., Bontems, F., Rak, A., Garber, M., Allemand, F., Springer, M., and Dardel, F. (2002). Nmr structure of bacterial ribosomal protein l20: implications for ribosome assembly and translational control. J. Mol. Biol. *323*, 143–151.

Schonbrunn, J., Wedemeyer, W.J., and Baker, D. (2002). Protein structure prediction in 2002. Curr. Opin. Struct. Biol. *12*, 348–352.

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

Schug, A., Herges, T., and Wenzel, W. (2004a). All-atom folding of the three-helix hiv accessory protein with an adaptive parallel tempering method. Proteins *57*, 792–798.

Schug, A., Herges, T., and Wenzel, W. (2004b). Predictive in-silico all-atom folding of a four helix protein with a free-energy model. J. Am. Chem. Soc. *126*, 16736–16737.

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*, 9686–9697.

Simmerling, C., Strockbine, B., and Roitberg, A. (2002). All-atom strucutre prediction and folding simulations of a stable protein. J. Am. Chem. Soc. *124*, 11258–11259.

Snow, C.D., Nguyen, H., Pande, V.S., and Gruebele, M. (2002). Absolute comparison of simulated and experimental protein folding dynamics. Nature *420*, 102–106.

Snow, C.D., Qiu, L., Du, D., Gai, F., Hagen, S.J., and Pande, V.S. (2004). Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. Proc. Natl. Acad. Sci. USA *101*, 4077–4082.

Vila, J., Ripoll, D., and Scheraga, H. (2004). Atomically detailed folding simulation of the b domain of staphylococcal protein a from random structures. Proc. Natl. Acad. Sci. USA *100*, 14812–14816.

Wales, D.J., and Doye, J.P. (1997). Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. J. Phys. Chem. *101*, 5111.

Zhou, Y., and Karplus, M. (1999). Folding a model three helix bundle protein: thermodynamic and kinetic analysis. J. Mol. Biol. *293*, 917–951.