J|A|C|S
ARTICLES
Published on Web 03/25/2006

Role of Topology, Nonadditivity, and Water-Mediated Interactions in Predicting the Structures of α/β Proteins

Chenghang Zong,*[,†,‡] Garegin A. Papoian,[§] Johan Ulander,[||] and Peter G. Wolynes[†,‡,⊥]

Contribution from the Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093-0371, Center for Theoretical Biological Physics, University of California, San Diego, La Jolla, CA 92093-0371, Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, AstraZeneca R&D Mlndal, Medicinal Chemistry, S-42183 Mlndal, Sweden, and Department of Physics, University of California, San Diego, La Jolla, CA 92093-0371

Received December 19, 2005; E-mail: czong@ucsd.edu

Abstract: The folding of α/β proteins involves most of the commonly known structural and dynamic complexities of the protein energy landscapes. Thus, the interplay among different structural components, taking into account the cooperative interactions, is important in determining the success of protein structure prediction. In this work we present further developments of our knowledge-based force field for α/β proteins, introducing more realistic modeling of many-body interactions governing the folding of β-sheets. The model's innovations highlight both specific topological characteristics of secondary structures and the generic nonadditive interactions that are mediated by water. We also investigate how a coarse biasing of the protein morphology can be used to understand the role of heterogeneity in protein collapse. Analysis of the simulation results for three test α/β proteins indicates that the addition of the topological and many-body ingredients to the model helps to greatly reduce the roughness in the energy landscape. Consequently, high quality candidate structures for α/β proteins can be generated from simulated annealing runs, using very modest amounts of computer time.

1. Introduction

Trying to understand how the nearly unique (but averaged!) protein structures are encoded from sequences and thereby to reproduce this process in silico has been a longstanding pursuit in theoretical chemistry. Energy landscape theory explains the basic physics of how Levinthal's paradox is overcome, but progress in structure prediction requires also attention to chemical and biological detail. While refinement of low resolution structures to the level of X-ray structures remains difficult, the generation of low resolution models has progressed greatly in recent years.[1−3] Nevertheless reliably searching even the low resolution conformation space for some protein topologies is still the fundamental step in making successful tertiary structure predictions.

The folding of a β-sheet in a protein, with its hydrophobic core having an extensive hydrogen bonding network, is a highly cooperative process.[4] In this paper we demonstrate that predictions of low resolution structure for α/β proteins are significantly improved when cooperative effects in β-sheet formation are taken into account even at the coarse grained level. Long-range water-mediated interactions, coupled with cooperative β-strand formation potentials, help to overcome the intrinsic topological frustration in the folding of α/β proteins which often makes their folding slow even in the laboratory. It has long been recognized that the formation of secondary structure elements may help to reduce the conformation space of the system. On the other hand, incorrect packing between such elements, if formed permanently, may result in topological frustration for the protein chain. When two protein segments approach each other early in the folding process, alignment of interstrand hydrogen bonds will promote β-sheet formation. The component β-strands then become straight and stiff. Similar straightening of β-strands is induced by interactions with the α-helices in α/β proteins. These processes promote topological frustration, where slow anisotropic rearrangements dominate folding dynamics. Nonadditive water-mediated interactions on the contrary may diminish the effect of such topological frustration by facilitating escape from incorrectly formed β-strand folds.

The pairwise-additive knowledge-based potentials, used traditionally in protein structure prediction, have recently been supplemented by many-body potentials that mimic the structural

† Department of Chemistry and Biochemistry, University of California, San Diego.
‡ Center for Theoretical Biological Physics, University of California, San Diego.
§ University of North Carolina at Chapel Hill.
|| AstraZeneca R&D Mlndal.
⊥ Department of Physics, University of California, San Diego.

(1) Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D. *Science* **2005**, *310*, 638.
(2) Chivian, D. et al. *Proteins: Structure, Function, and Bioinformatics* **2005**, *61*, 157.
(3) Fujitsuka, Y.; Takada, G. C. S. *Proteins: Structure, Function, and Bioinformatics* **2006**, *62*, 381.
(4) Guo, C.; Levine, H.; Kessler, P. *PRL* **2000**, *84*, 3490.

10.1021/ja058589v CCC: $33.50 © xxxx American Chemical Society

effects of water.[5] We have found the addition of specific water-mediated interactions significantly improves the quality of structure prediction for large α-proteins.[5] Since α/β proteins are characterized by more complex topologies than all α-helical proteins, the water-induced nonadditive forces are expected to strongly affect the folding process. When strands are close in sequence, as happens, in the formation of local β-hairpins, chain diffusion is an efficient mechanism for bringing two segments together. The interactions between segments are nevertheless still required to be cooperative in order to stabilize the final closure. On the other hand, when forming β-sheets, which contain strands that are far apart in sequence, long-range interactions play a crucial role. The formation of β-sheets is thus strongly coupled to the formation of the correct globular fold.

The starting point for our structure prediction potential development is the Associative Memory Hamiltonian (AMH) introduced by Friedrichs and Wolynes.[6] The mathematical form of the minimal frustration principle based on the ratio of folding and glass transition temperature is then used to optimize the model parameters. In this way, structure prediction potentials that were pairwise additive were developed for α and α/β proteins.[7,8] For the α-proteins, the traditional pairwise additive potentials have recently been supplemented by adding many-body water-mediated potentials accounting for interstitial water.[5] The water-mediated potential is context sensitive, depending on the local protein density around a pair of residues. In particular, two residues interact differently through water on the protein surface from how they do through the protein, i.e., when they are buried in the core. The propensity for various amino acids to be buried, yet another nonadditive effect of water, was also taken into account through a local-density based burial profile potential. In this work we have added similar water-mediated potentials to the α/β protein structure prediction Hamiltonian to investigate whether the improved treatment of nonadditive interactions yields more nativelike β-sheet formation.

One of the important questions in folding α/β proteins is the interplay between formation of α-helices and β-strands. In general, helices are relatively local in sequence and form much faster than the β-strands. These, possibly transient helical structures, in turn may promote the directional collapse of the β-strand regions. Formation of the helices helps to align the β-strands, which in turn promotes the β-sheet nucleation process. Consistent with these speculations, a recent experimental study on folding of a single-chain monellin indicated that α-helical content appears significantly prior to the chain collapse into an oblate shape.[9] 80% of the β-content forms only after the helix becomes well folded. These observations prompted us to examine in more detail the role of helices in the collapse and alignment of the β-strands.

Yet other important components of α/β protein structural architecture are the turns and the loops that connect α-helices and β-strands. The folding of turn regions in β-hairpins has been found in some cases to couple strongly to the formation of the segments they connect.[10] Some loops exhibit strong internal stability. Ω-Loops are notable examples. Turn regions are much harder to predict using knowledge-based approaches than are α-helices and β-strands. Nevertheless, β-hairpins with less than eight residues can be reliably predicted using knowledge-based approaches.[11] Loops spanning a larger number of residues are harder for bioinformatic approaches. Predicting the structure of large loops is significantly hindered by their noncompact nature. In such conformations, exposed to water, water-mediated interactions play an important role for stability.

We also explore in this paper how topological frustration may be alleviated by introducing a morphological bias into simulations. Large scale morphology changes are slow even in the laboratory. Since protein shapes usually deviate from spherical, knowing their final dimension and biasing their global morphology help to decrease the large configurational entropy of the unfolded state and allow escape from early topological traps. Such a morphological constraint bias may be analogous to the cage effect thought to occur within chaperones.[12] Low-resolution information about the protein shape can often be obtained from X-ray crystallography or cyro-EM experiments, even in cases when an atomic level resolution protein structure has not been solved.[14] By carrying out simulations with different spherical and nonspherical morphological bias potentials, we can study systematically which specific protein conformations are preferred under the constraints of a particular shape.

In this paper we demonstrate that adding water-mediated interaction potentials and cooperative β-strand formation potentials to the structure prediction Hamiltonian for α/β proteins leads to significantly improved prediction results. In particular, we find that topological frustration in nonlocal β-sheet formation during collapse is alleviated by the water interactions which make escape easier from topological traps. We discuss in detail how the nonadditivity effects and the morphological bias affect the chain topology in protein folding. The article is organized in the following way. First, we describe various terms of the structure prediction Hamiltonian, including the cooperative β-strand potentials, emphasizing the novelties introduced in the present work. Next, the parameter optimization procedure is briefly outlined. Finally, some specific structure prediction results are discussed in detail for three α/β proteins. These test proteins are not homologous to the training proteins that were used in optimizing potential parameters, providing an objective evaluation of the Hamiltonian's performance.

2. Modeling

Our structure prediction efforts are based on the Associative Memory Hamiltonian (AMH), which has been extensively documented in prior works.[7,8,15] In this section we briefly outline the further features introduced in the paper. The AMH is intrinsically a coarse-grained model, where each residue is

(5) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3352.
(6) Friedrichs, M. S.; Wolynes, P. G. *Science* **1989**, *246*, 371.
(7) Hardin, C.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 14235.
(8) Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 1679.
(9) Kimura, T. et al. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2748.
(10) Munoz, V.; Thompson, P.; Hofrichter, J.; Eaton, W. *Nature* **1997**, *390*, 196.
(11) de la Cruz, X.; Hutchinson, E.; Shepherd, A.; Thornton, J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11157.
(12) Takagi, F.; Koga, N.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *14*, 892.
(13) Chikenji, G.; Fujitsuka, Y.; Takada, S. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3145.
(14) Wu, Y.; Chen, M.; Lu, M.; Wang, Q.; Ma, J. *J. Mol. Biol.* **2005**, *350*, 571.
(15) Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. *IBM J. Res. Dev.* **2001**, *45*, 475.

represented by C_α, C_β, and O atoms. The Hamiltonian contains two major components: (i) sequence-independent polymer physics terms to describe the backbone interactions; (ii) sequence-dependent knowledge-based potentials optimized to achieve folding of a number of training proteins. The backbone interactions include chain-connectivity, excluded-volume, Ramachandran, and chirality potentials. The sequence-dependent interactions involve only $C_\alpha-C_\alpha$, $C_\alpha-C_\beta$, and $C_\beta-C_\beta$ pairs. These interactions are grouped into three proximity classes according to the sequence distance between the interacting residues, as follows: short range ($3 \leq |i - j| < 5$), medium range ($5 \leq |i - j| \leq 8$), and long range ($|i - j| > 8$). For the short and medium classes, a pairwise interaction in the target protein is associated with a corresponding pairwise interaction in memory proteins whose structures are already known. The short and medium range interactions are based on preliminary alignments of sequence to the memories. In this level they play a guiding role analogous to the choice of fragments in fragment assembly methods.[2,13] In our evaluation studies here, homologous proteins are rigorously excluded from the memory set. Further details are given in the Appendix.

For the long-range proximity class of interactions, a simple square well potential is used, unrelated to the memory proteins. The terms of this function are partitioned into two wells, based on the physical distance. The first well covers the 4.5 Å to 6.5 Å interval, representing a simple contact between two residues. The second well covers the 6.5 Å to 9.5 Å interval, representing protein-mediated or water-mediated interactions. To determine whether an interaction is protein- or water-mediated, the local density around each pair of residues is computed.[5] When both residues are not surrounded by many other residues, they will instead be surrounded by water which mediates the inter-residue interactions at an appropriate distance of 6.5 Å to 9.5 Å. On the other hand, when one or both of the residues in a pair are buried, then water mediation is switched smoothly to protein mediation, using a many-body switching function.[5] In addition to the water-mediated potential between residue pairs, an additional interaction term, based on the burial profile of each amino acid, was introduced to describe the propensity of amino acids to partition between water and the protein interior.[5] The burial profile potential includes three wells that characterize the likelihood for a particular amino acid to be in low, medium, or high local density. It is important to point out here that owing to the context dependence both the second well potential in the long-range proximity class and the burial potential are nonadditive.

The hydrogen bonding interactions were modified to include additional geometrical constraints for the β-strands. Since a β-strand has to be quite extended in order to effectively form the hydrogen bonding network, we added a constraint term to allow only small curvature of the strand in the β-sheet formation. Furthermore, we set three sequence-separation-based proximity classes for hydrogen bonding potentials: for the first class the sequence distance for a pair of interacting residues is less than 19; for the second class it is between 19 and 45; for the third class it is larger than 45. The hydrogen bonding potentials include three terms to represent pairwise interactions, parallel nonadditivity, and antiparallel nonadditivity, respectively (please see Appendix for further details). When both pairwise and nonadditive interactions are present, the hydrogen bonds sometimes become too difficult to break, once they are formed. To avoid strong local collapse of β-strands, we only turned on two of the nonadditive terms (Λ_2 and Λ_3 terms in the Appendix) when the interacting residues are both predicted to be in a β-strand from a secondary structure prediction server JPRED.[16] Here we hypothesize that these residues are the ones giving the most energetic stabilization to the β-sheets. A similar conjecture about the stabilization of β-strands has also been discussed for the folding mechanism of β-hairpins.[10]

Several proteins containing β-components have been shown to undergo specific collapse.[18] These include, for example, protein L and cold shock protein. The observations indicate that the early stage of forming β-sheets leads to a heterogeneous collapse of the protein chain. In the AMH, we treat this collapse heterogeneity with the so-called "liquid-crystal" potential in analogy to the nematic phase in liquid crystals.[19] Since this collapse involves forming either antiparallel or parallel strands, we include two corresponding terms in the "liquid-crystal" potential. In addition to the antiparallel and parallel strands with large sequence separation among the strands, β-hairpins are prominent in the architecture of the β-proteins. Since β-hairpins are relatively local in sequence, it is possible to identify possible β-hairpin candidates using secondary structure prediction results. We employed the following criterion to determine possible β-hairpin candidates. In a segment of 18 consecutive residues, (i) if two β-strands of similar size and with an intervening small loop region (less than eight) are predicted in the secondary structure prediction, and (ii) if the loop region is predicted with high probability (larger than 5 on the scale of 10), then we conjecture that a β-hairpin is likely to form in this region. We introduced a pairwise interaction for residues in this region to bias the formation of a β-hairpin conformation (details are given in the Appendix). Tuning these pairwise potentials changes the turning tendency of the polypeptide backbone. Although the protein chain becomes preferentially bent in specific places by such a β-hairpin potential, the hairpins cannot form unless the hydrogen bond interactions further stabilize the formation of β sheets.

Finally, we introduced a potential to explicitly control the collapse. The corresponding collapse potential biases the gyration radius of the protein chain. A harmonic constraint on the gyration radius favors a spherical protein shape. However, a nonspherical shape is common for many native protein structures. Therefore, an asymmetrical collapse potential can be quite helpful. Here, we introduce such an asymmetrical collapse potential, which independently constrains the gyration along *x*-, *y*-, and *z*-axes. Using different bias parameters for each axis controls the general shape of the collapsed protein. Thus, experimentally derived information on the protein morphology, available from many low-resolution techniques, may be incorporated into the prediction scheme by such an asymmetrical collapse potential. Alternatively several values of these weak constraints can be scanned in order to cover all possibilities in sampling.

(16) Cuff, A. J.; Clamp, E. M.; Siddiqui, S. A.; Finlay, M.; Barton, G. J. *Bioinformatics* **1998**, *14*, 892.
(17) Zemla, A. *Nucleic Acids Research* **2003**, *31*, 3370.
(18) Chahine, J.; Nymeyer, H.; Socci, V. L. N.; Onuchic, J. *PRL* **2002**, *88*, 168101.
(19) Luthey-Schulten, Z.; Ramirez, B. E.; Wolynes, P. G. *J. Phys. Chem.* **1995**, *99*, 2177.

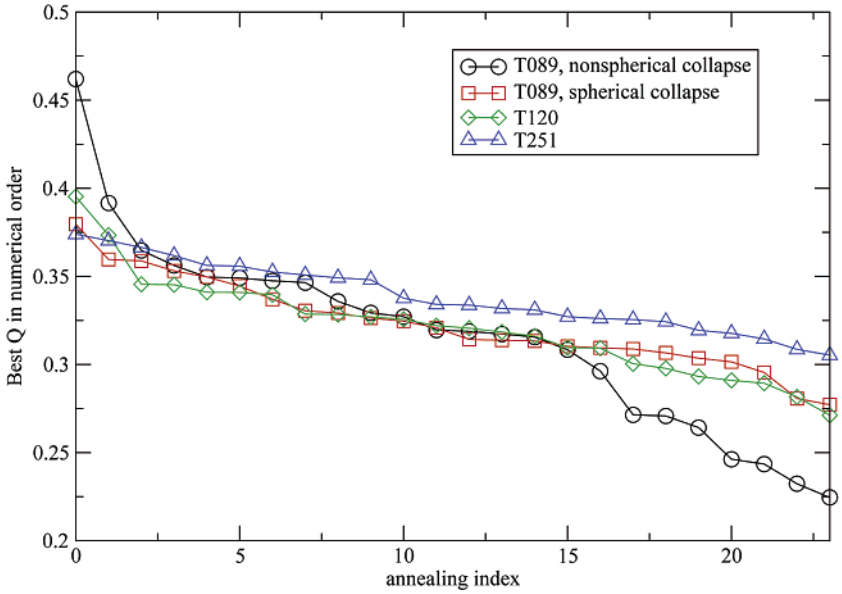

Figure 1. Best Q sampled in 24 annealing runs for proteins T089, T120, and T251.

3. Constrained Self-Consistent Optimization

A self-consistent optimization scheme was used to tune the various interaction strengths in the Hamiltonian. The optimization is based on the minimum frustration principle.[20] The energetic stabilization in the folding process is described by the energy gap, δE, between the molten-globule states and the nativelike states. At the folding temperature T_f, the energy gain, δE, is balanced by the loss of configurational entropy S_c. Thus, the folding temperature T_f is expressed as $\delta E/S_c$.[20] In this optimization, an important statistical characteristic of the folding energy surface is the roughness of the energy landscape, quantitatively described by the energy variance of molten-globule states, $\sqrt{\Delta E^2}$. The ratio of this variance to the entropy of the molten-globule states, $\sqrt{\Delta E^2/S_{mg}}$, provides an estimate of the polypeptide chain's glass transition temperature T_g. Maximizing the ratio of the folding temperature over the glass transition temperature, T_f/T_g, provides a quantitative procedure to minimize the frustration presented in a knowledge-based Hamiltonian for a training set of proteins.

In this optimization scheme, additional constraints are imposed upon the mean and the variance of the molten-globule structures for each proximity class. Thus, the optimization preserves the energy balance between different proximity classes. We used 14 α/β proteins to "train" the Hamiltonian. Decoy structures imitating the molten globule were self-consistently generated from sampling at high temperature, $1.2T_f$. A nativelike ensemble of structures was also generated from biasing sampling to the native region. A Lagrangian functional, containing the constraints on the mean and variance, was minimized for each proximity class.[7]

4. Results and Discussion

We carried out molecular dynamics simulations with temperature quenching to search for low energy conformations (more details about MD simulation are given in the Appendix). Three α/β proteins, which were dissimilar to all of the training proteins, were used to test our current model. For each protein, 24 simulated annealing runs were carried out. The best Q score for each run is reported in Figure 1. The Q score is defined as $2\sum_{i<j-2}\exp[-(r_{ij}-r_{ij}^N)^2/2\sigma_{ij}^2]/(N-1)(N-2)$, which describes a critical assessment of all pairwise distances within structures. The structure with $Q = 1.0$ corresponds to the native structure, while the conformations having Q values near 0.4 are typically characterized by an ~6 Å RMSD fit to the native structure. We used yet another similarity measure to compare conformations, the Z score calculated with the combinatorial extension (CE) algorithm.[21] This score identifies a general topological similarity disregarding the sequence information. In general, a Z-score of 3.5 indicates significant structural similarity, while strong structural similarity is achieved for Z-scores larger than 4.0.

Three test proteins were CASP targets, with indices T089, T120, and T251. The crystal structure analysis for these proteins indicates they have diverse topologies. For instance, the β-sheets in the test proteins are quite different in both their shapes and their locations. T089 is a single domain from protein 1E4F (a CASP4 target), taken from residues 86 to 166 in the PDB. In the T089 native structure, a long three-strand β-sheet is formed around the α-helix. The second test protein, T120, having 115 residues, is an N-terminal domain of human XRCC4DNA repair protein 1FU1 (a CASP4 target). The native structure of this protein is comprised of two sandwichlike β-sheets with two helices connecting them. The third test protein, T251, which contains 99 residues, was taken from protein 1XG8 (a CASP6 target). This protein is comprised from three outer helices, in addition to a four-strand β-sheet mainly located in the core of the protein. An interesting aspect of the T089 and T251 topology is the nonlocal nature of β-sheets, with β-strands separated far apart in sequence. This makes these proteins challenging targets for structure prediction.

Prior to starting the simulations, we located possible β-hairpin regions by using secondary structure prediction results[16] for the test proteins. However, for large loop regions, like the Ω loop

(20) Goldstein, R. A.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918.

(21) Shindyalov, I.; Bourne, P. *Protein Eng.* **1998**, *11*, 739.

Figure 2. A predicted structure for protein T089 with the $Q = 0.38$ (CE: $Z = 3.9$) (left), the native structure (middle) and the contact map (right). The prediction was generated with the spherical collapse potential.

in T089, we did not find any specific structural information from secondary structure prediction. Moreover, long loop conformation may strongly depend on the folding of other segments in a protein. Therefore, we did not add any biasing to the residues in the long loop regions.

The structure prediction results for three test proteins, evaluated using the Q score, are summarized in Figure 1. The Q scores are plotted in the sorted order of numerical values. For each of the three test proteins, we reached conformations with a Q score greater than 0.35 within 24 short annealing runs. Using the CE score measure, we found that about 10 annealing runs for each protein sampled structures with a Z larger than 3.7, corresponding to rather native topologies. When the nonspherical collapse potential was added to constrain the overall topology of one of the targets, T089, the prediction results were significantly improved. The best Q score reached a high value of 0.45, exhibiting very strong similarity to the native structure. Samplings of the best predicted structures for three test proteins are presented in Figures 2, 3, 4, and 5. The native structures and the contact maps are also shown for comparison. Both structural drawings and the corresponding contact maps indicate that the predicted structures are very similar to the native structures, with some discrepancy in the packing of secondary structure elements. The global RMSD value for the predicted structures in Figures 2, 3, 4, and 5 are 10.5 Å, 6.3 Å, 6.0 Å, and 12.4 Å respectively. The larger values reflect bad relative placement of correct large substructures. A measure of the substructure quality is provided by LGA (Local Global Alignment[17]). We used the LGA server to analyze this aspect of our prediction results. We list the results as follows: For the predicted structure in Figure 2, there are 49 out of 81 residues within 5 Å and the RMSD for the residues under this distance cutoff is 3.02 Å. For the predicted structure in Figure 3, there are 56 out of 81 residues within 5 Å with an RMSD of 3.23 Å. For the predicted structure in Figure 4, there are 83 out of 115 residues within 5 Å having an RMSD of 2.88 Å. And for the predicted structure in Figure 5, there are 45 out of 99 residues within 5 Å with an RMSD of 2.29 Å.

To quantify further the predictive power of the improved AMH, we use an umbrella sampling algorithm (setup details are given in the Appendix) to compute the free energy profile and the thermodynamic energy profile for T089 along the reaction coordinate, Q (Figure 6). The average energy decreases with increasing Q, indicating an energy landscape funneled toward the native state. Furthermore, the asymmetrical collapse

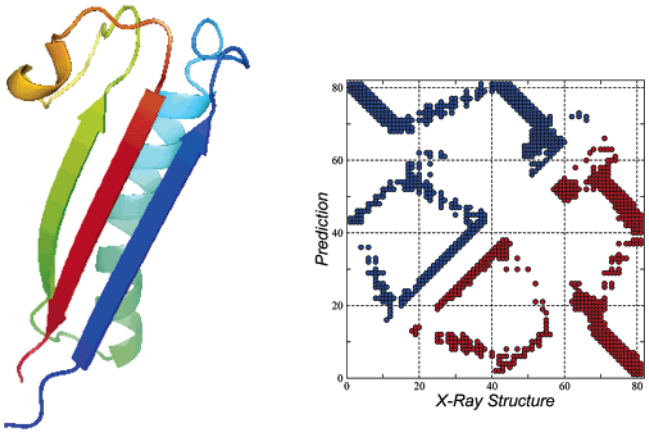

Figure 3. A predicted structure for protein T089 with the $Q = 0.46$ (CE: $Z = 4.1$) (left) and the contact map (right). The prediction was generated with the nonspherical collapse potential.

potential improves the funneling for protein T089, compared with the landscape having the isotropic collapse potential. Although the energy decreases as the conformations become more nativelike, the entropy loss reverses this trend and shifts the minimum of the free energy to a relatively low Q value. In Figure 6, for the asymmetrical collapse potential, the minimum of the free energy curve is located at $Q > 0.3$ at $T = 1.00$. For the isotropic collapse potential, the broad minimum of the free energy profile at $T = 1.00$ is located around the molten-globule region, with Q spanning from 0.2 to 0.3.

Thermodynamically, the average energy decrease strongly funnels the configurations toward the protein native state. On the other hand, the ruggedness of the energy landscape also critically affects the folding dynamics. The energy ruggedness leads to a glass transition at the low temperatures needed to stabilize the native structure, making search via simulated annealing difficult. To quantify the emergence of glassy behavior, while lowering the temperature, we evaluated Q-autocorrelation functions (Figure 7). These correlations provide dynamic information on the ruggedness of the energy landscape at any given temperature. With decreasing temperature, the valleys of the energy landscape become too deep for the protein chain to overcome by simple thermal fluctuations in the simulation time allowed, leading to trapping in low energy conformations. For protein T089, when T is lowered below 0.9, the system no longer efficiently explores the configurational space on the simulation time scale. We did find that the glass

Figure 4. A predicted structure for protein T120 with the $Q = 0.39$ (CE: $Z = 4.7$) (left), the native structure (middle), and the contact map (right).

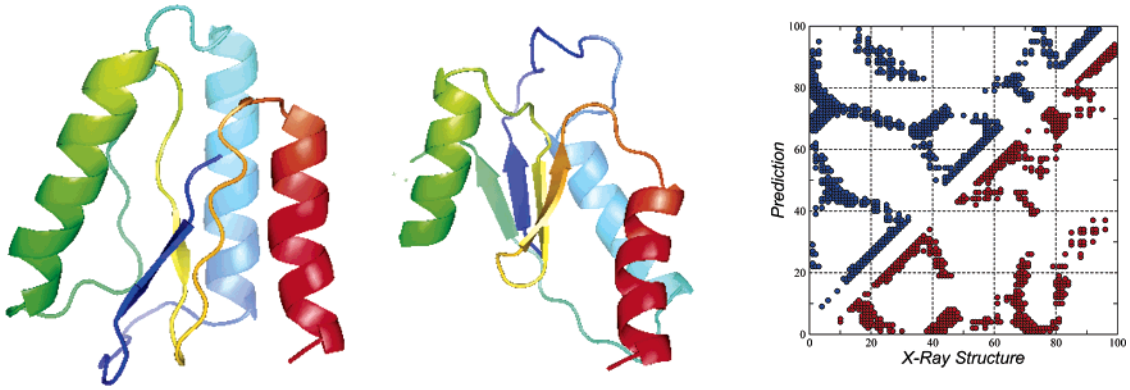

Figure 5. A predicted structure for protein T251 with the $Q = 0.37$ (CE: $Z = 3.6$) (left), the native structure (middle), and the contact map (right).

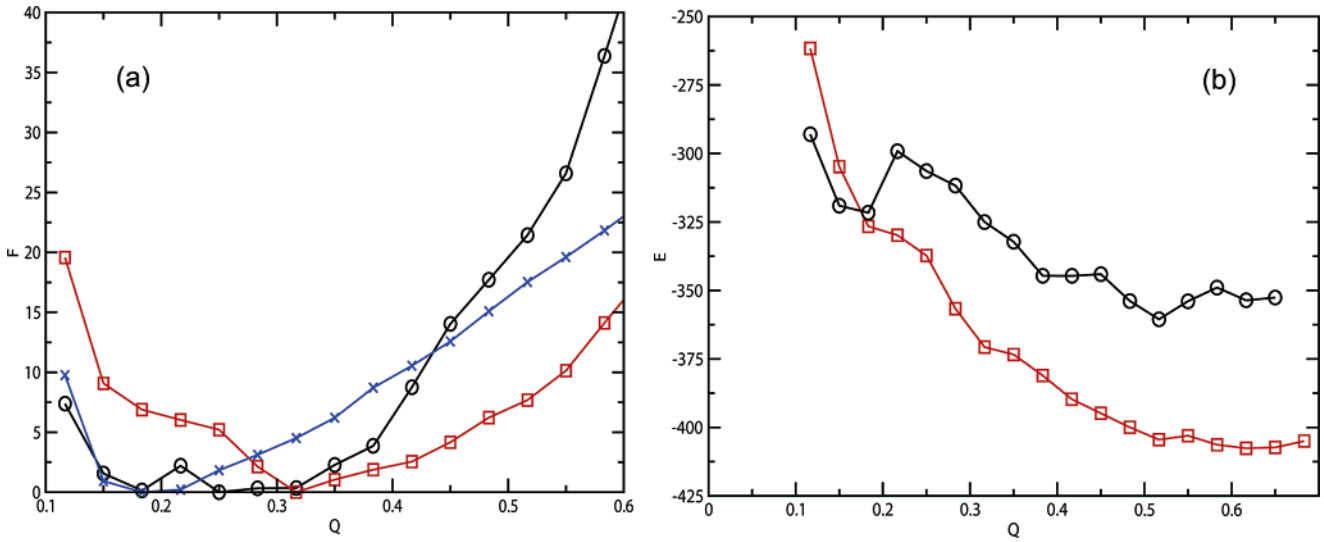

Figure 6. Free energy calculations for protein T089. (a) Free energy as a function of Q at $T = 1.0$. (b) Energy as a function of Q at $T = 1.0$. The line with circle symbols corresponds to the spherical collapse potential; the line with square symbols corresponds to the nonspherical collapse potential; the line with cross symbols corresponds to the result from the previous model without water-mediated potentials and that did not contain the modifications in β-potentials.

temperature for T089 with these innovations is significantly lower than the results from the previous AMH study on this system.[8] We attribute this lowering to the water-mediated interactions and the adjusted β-potentials that, in turn, help to decrease the energy ruggedness of the landscape of the molten-globule states, resulting in a glass transition below $T = 0.9$. Reducing the energy ruggedness allows efficient sampling of nativelike structures at lower temperatures, since the energy gradient favors native structures, while the entropy plays a lesser role when the temperature is low. Therefore, the depression of the glass temperature explains why good quality structures are sampled in our simulated annealing runs, even though the minimum of the free energy for T089 is around the molten-globule states at $T = 1.00$.

Use of the asymmetrical collapse potential decreases the number of available topologies compared with the isotropic collapse potential. The resulting entropy reduction for the molten-globule states shifts the minimum of the free energy profile for T089 to a more nativelike region. For the isotropic collapse potential the main free energy well is located around $0.2 < Q < 0.3$ with CE scores reaching 3.5 (Figure 8), while for the asymmetrical collapse potential the main free energy

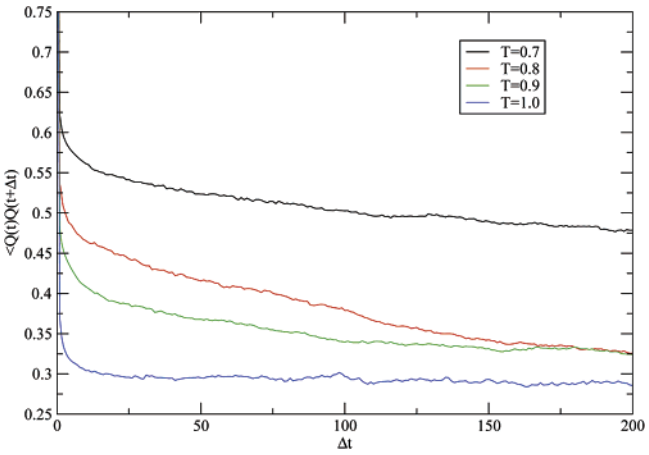

Figure 7. Q autocorrelation function at different temperatures for protein T089 (the collapse potential is spherical). The x-axis is time measured in units that correspond to roughly 250 ns of laboratory time.

well is located around $0.3 < Q < 0.4$ with CE scores reaching as high as 4.0. The free energy cost to reach $Q = 0.4$ is only about $3kT$ for the asymmetrical collapse potential, suggesting that in any given simulation run we expect about $e^{-3} \approx 5\%$ chance to sample $Q > 0.4$ structures per relaxation time. Furthermore, from a statistical viewpoint, we can sample regions of higher free energy by carrying out additional annealing runs. For instance, we estimate that $Q \approx 0.55$ (RMSD ≈ 4 Å) structures for T089 may be generated using 30 CPU-day computational resources.

Both the folding of the secondary structure elements and the interplay between forming α-helices and β-sheets are important for determining the overall tertiary architecture. In Figure 9, we plot the folding progress of individual secondary structure elements for simulated annealing runs that generated the highest Q score. The prediction of individual secondary structure segments is measured by $Q_{\text{seg}} = \sum_{\{i,j\}\in\text{seg}}\Theta(r_c - r_{ij}^N) \exp[-(r_{ij} - r_{ij}^N)^2/2\sigma_{ij}^2]/\sum_{\{i,j\}\in\text{seg}}\Theta(r_c - r_{ij}^N)$. From the prediction of each secondary structure segment of T089 and T120, we observe that α-helices form much earlier than β-sheets. Only when the α-helices reach $Q \approx 0.6$, does the nativeness of β sheets start to appreciably increase. In addition, we find that the α-helices exhibit residual structure signals even at high temperatures. On the other hand, the β-sheets do not exhibit any residual structures at high temperatures, when the overall Q values are around 0.1. The early folding of α-helices provides a nucleus of hydrophobic surface that helps to align the β-strands.

Moreover, we provide a sequence of folding snapshots in Figure 10 to illustrate the progression of conformations in the

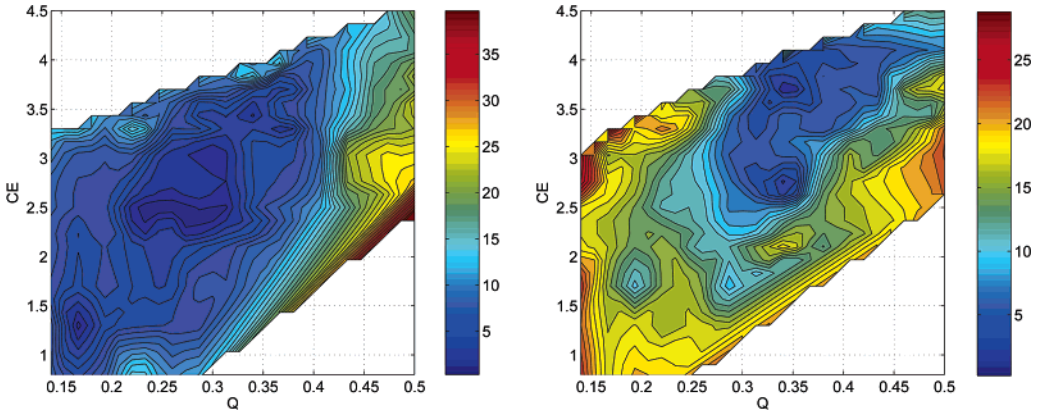

Figure 8. Free energy for protein T089 as a function of Q (x axis) and CE score (y axis) at $T = 0.9$. (Left panel) Using the spherical collapse potential. (Right panel) Using the nonspherical collapse potential.

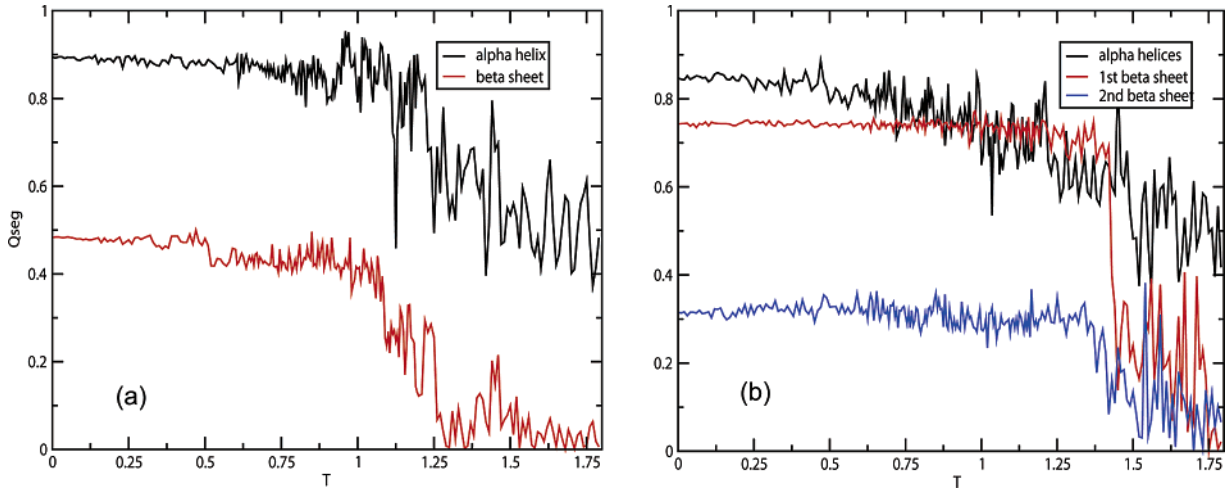

Figure 9. The folding of secondary segments of protein T089 (a) and protein T120 (b) in five annealing runs which generate a Q score larger than 0.35. For protein T089: the α-helical region covers residues 23−39; the β-sheet region covers residues 4−13, 43−54, and 70−80. For protein T120: the α-helical region covers residues 49−58 and 63−74; the first β-sheet region covers residues 2−8, 18−24, 31−37, and 41−48; the second β sheet region covers residues 84−88, 94−101, and 104−112. For protein T089, the spherical collapse potential was used.

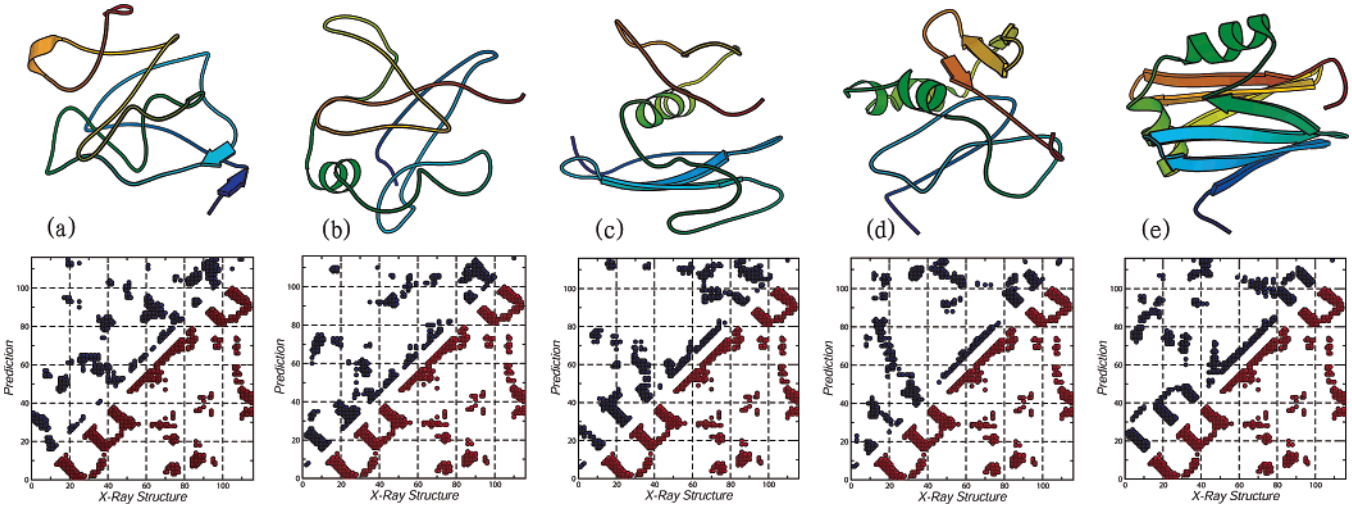

Figure 10. A sequence of snapshots was taken from a simulated annealing trajectory for protein T120. The contact map is drawn below each structure. This trajectory eventually sampled a $Q = 0.39$ structure. Snapshots (a−e) are sampled at $T = 1.75$, 1.5, 1.48, 1.44, and 1.17, respectively.

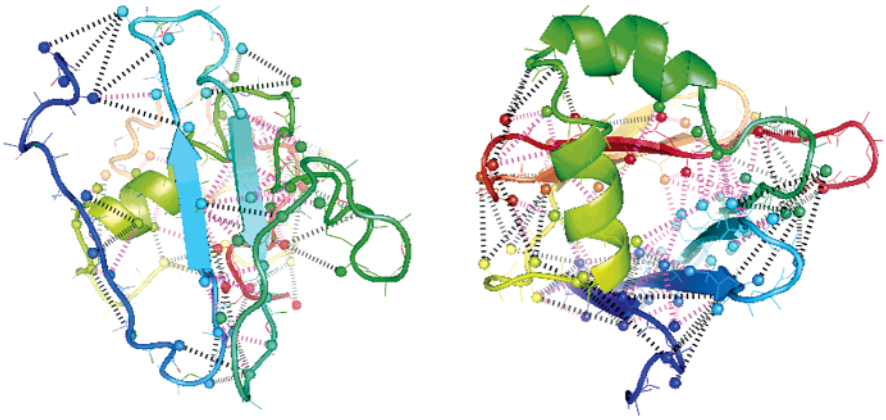

Figure 11. Long range interactions in a predicted structure for protein T120. (Left) One structure at the early stage of folding with $Q = 0.25$. (Right) The well folded structure with $Q = 0.39$. Spheres represent β-atoms. Dashed lines with magenta color present the contacts with distances between 4.5 Å and 6.5 Å. Dashed lines with black color represent the water-mediated interactions with distances between 6.5 Å and 9.5 Å.

β-sandwichlike protein T120. An initial collapsed conformation is shown in Figure 10a. As the temperature is decreased, the helix starts to form with some alignment of β-strands (Figure 10b). Partial formation of the N-terminal and C-terminal sheets were observed (Figure 10c and d). As temperature is further decreased, the full hydrogen bonding network is formed, producing a very nativelike conformation.

A prominent role of the β-sheets is apparently to shield the hydrophobic core from the hydrophilic surface. With the C_β atoms considered as the interaction centers and with the above-mentioned switching mechanism, the water-mediated potential differentiates between the hydrophobic and the hydrophilic sides of a β-sheet. Furthermore, before two segments reach each other within 6 Å and start to form hydrogen bonds, the water-mediated interactions encourage the approach of the β-strands in the range from 6.5 Å to 9.5 Å. There is "water-mediated" closure of β-strands. One snapshot of closing β-strands showing such water-mediated interactions is presented in the left panel in Figure 11. The closing of the strand in blue by the water-mediated interactions is shown as dashed lines. In the right panel of Figure 11, we present the predicted structure from Figure 5, highlighting the water-mediated interactions. We observe that the shell filled with water-mediated interactions covers well the hydrophobic core region. Overall, the predicted structures that are nativelike contain a favorable network of water-mediated interactions surrounding the protein core.

5. Conclusion

In summary, our work demonstrates that correct modeling of cooperativity that is largely mediated by water is crucial for obtaining accurate structure predictions of β-sheets in α/β proteins. First, many weakly specific tertiary interactions are involved in forming β-sheets. During the early events of protein folding, these tertiary interactions may occur prior to locking of hydrogen bonds between β-strands. On the other hand, in the case of helices, the local hydrogen bonds appear prior to forming those tertiary interactions. Water-mediated interactions are important for the recognition between β-strands. They help to reduce the topological frustration, which in turn leads to more efficient sampling of structures having nativelike packing. In α/β proteins, the early folding of α-helices provides patches of hydrophobic surface to nucleate the alignment of β-strands. This mechanism can be incorporated into a general capillary picture as discussed by Wolynes.[22] The exact timing of events between nucleation processes and the formation of secondary structures regulates the collapse of proteins. In the early stages of protein

(22) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 6170.

folding, the collapse can be either nonspecific or specific. Our comparative analysis indicates that potentials that favor early specific collapse over early nonspecific collapse significantly improve structure prediction. Water-mediated potentials may be combined with higher resolution models that include more details of the side chains that take into account efficient packing of nativelike protein structures.

Acknowledgment. We especially thank Michael Prentiss for useful discussions; we also want to thank all the previous P.G.W. group members for their cumulative efforts in the development of the AMH (Associative Memory Hamiltonian) code. The efforts of the P.G.W. group in protein folding are supported through NIH Grant No. 5R01GM44557.

Supporting Information Available: The Appendix is available in the Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

JA058589V