Folding Transition-State and Denatured-State Ensembles of FSD-1 from Folding and Unfolding Simulations

Hongxing Lei, Shubhra Ghosh Dastidar, and Yong Duan*

UC Davis Genome Center and Department of Applied Science, University of California, Davis, California 95616

Received: June 14, 2006; In Final Form: August 18, 2006

Characterization of the folding transition-state ensemble and the denatured-state ensemble is an important step toward a full elucidation of protein folding mechanisms. We report herein an investigation of the free-energy landscape of FSD-1 protein by a total of four sets of folding and unfolding molecular dynamics simulations with explicit solvent. The transition-state ensemble was initially identified from unfolding simulations at 500 K and was verified by simulations at 300 K starting from the ensemble structures. The denatured-state ensemble and the early-stage folding were studied by a combination of unfolding simulations at 500 K and folding simulations at 300 K starting from the extended conformation. A common feature of the transition-state ensemble was the substantial formation of the native secondary structures, including both the α-helix and β-sheet, with partial exposure of the hydrophobic core in the solvent. Both the native and non-native secondary structures were observed in the denatured-state ensemble and early-stage folding, consistent with the smooth experimental melting curve. Interestingly, the contact orders of the transition-state ensemble structures were similar to that of the native structure and were notably lower than those of the compact structures found in early-stage folding, implying that chain and topological entropy might play significant roles in protein folding. Implications for FSD-1 folding mechanisms and the rate-limiting step are discussed. Analyses further revealed interesting non-native interactions in the denatured-state ensemble and early-stage folding and the possibility that destabilization of these interactions could help to enhance the stability and folding rate of the protein.

Introduction

A detailed description of the transition-state ensemble (TSE) is an important step toward full elucidation of a protein folding mechanism. The TSE is a set of non-native structures that collectively form the highest free-energy barrier that a protein has to cross during its folding process. The functional significance of the TSE lies in its strategic location on the free-energy landscape of protein folding such that crossing of the TSE should lead to rapid progress toward the native structure during the folding reaction. It is generally recognized that the TSE might contain the key tertiary and secondary contacts that are mostly responsible for both the protein's stability[1,2] and its folding processes. Consequently, attempts have been made both theoretically[3−7] and experimentally[8,9] to identify the protein folding/unfolding TSEs. Some of the examples include the seminal work of Fersht and co-workers who characterized the TSEs of several prototypical small proteins according to Φ-value analyses.[10−14] In a recent collaboration, Fersht and Daggett and their co-workers[15] combined Φ-value analyses from experiments and unfolding simulations to elucidate detailed information on the folding transition states. In this work, we extend such an effort by combining folding and unfolding simulations to investigate three key areas on the free-energy landscape.

Because the TSE forms the barrier in the free-energy landscape and the associated structures are unstable and prone to go toward either side of the reaction coordinate, it is difficult to obtain high-resolution structures of the TSE from direct experimental measurements. The inherent heterogeneity of the TSE implies that there might be wide variety of structures with common features. Therefore, experimental observations are averaged over the ensemble, and the experimentally identified TS is taken to represent the average features of the TSE. Nevertheless, knowledge of the structural variation within the TSE can help to clarify the mechanism of narrowing the conformational space toward the native state. Because of their high spatial and temporal resolution, molecular dynamics (MD) simulations are advantageous for identifying these transient structures.

One major problem in computational studies of the protein folding process is the short time scale achievable from the simulations, which limits the studies to within several microseconds.[16] Therefore, unfolding simulations were performed to allow thorough sampling of the conformational space. An implicit assumption of this approach based on unfolding simulations is that the folding and unfolding processes follow the same or similar pathways in the reverse directions. However, because unfolding simulations are mostly carried out under high-temperature conditions to allow fast unfolding to occur within the short simulation time, the free-energy landscape might have been altered slightly.[17] Because of these potential complications, identification of the TSE from unfolding simulations directly is not a straightforward exercise. In this article, we report our efforts to combine both folding and unfolding simulations to identify the TSE of folding of the mini-protein FSD-1.

The denatured-state ensembles of proteins were once thought to have been structureless random coils. However, NMR

* Corresponding author. Telephone: (530) 754-7632. Fax: (530) 754-9648. E-mail: duan@ucdavis.edu.

10.1021/jp063716a CCC: $33.50 © 2006 American Chemical Society
Published on Web 10/07/2006

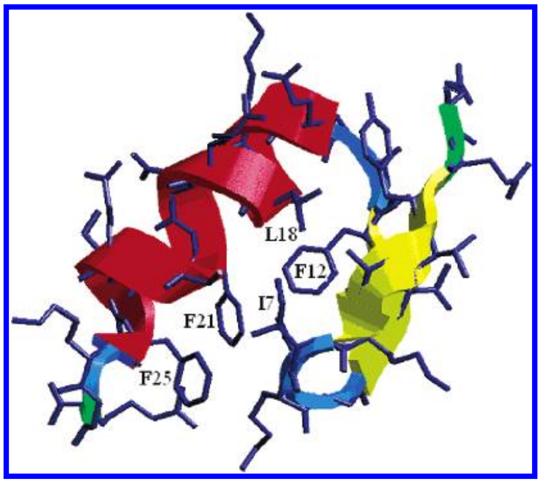

Figure 1. Native structure of FSD-1. The helix is in red, and the β-hairpin is in yellow. The side chains at the helix/sheet interface are labeled with a single letter code and residue index.

experiments[18] indicated substantial residual structures in the denatured state. Since then, efforts have been made to characterize the denatured-state residual structures by both experimental[19-22] and computational[23] approaches. Unlike the native structure, the denatured ensembles are highly heterogeneous and are difficult to study experimentally. In this work, we combine both unfolding simulations at 500 K starting from the native structure and folding simulations at 300 K starting from the extended conformation to characterize the denatured-structure ensemble and early-stage folding. Analyses revealed detailed information on the interactions that were partially responsible for stabilizing the non-native states. Further comparison between the structures of the denatured- and transition-state ensembles allows for the proposal of a possible folding pathway.

FSD-1 is a small α/β protein of 28 residues (QQYTAKIKGR$_{10}$-TFRNEKELRD$_{20}$FIEKFKGR). It was designed based on the backbone structure of a zinc-finger protein domain using a full-sequence-design protocol[24] by dead-end elimination[25] with low sequence identity to the template. The NMR structure (Figure 1) of FSD-1 shows that residues 3−13 form a β-hairpin and residues 15−25 form a helix under its native conditions. The α/β interface contains mostly hydrophobic residues including Ala$_5$, Ile$_7$, Phe$_{12}$, and Leu$_{18}$, Phe$_{21}$, Ile$_{22}$, Phe$_{25}$. The terminal residues were not well resolved in the NMR experiment.

Methods

The initial structure for the unfolding simulations was taken from the NMR ensemble (PDB code1FSD). Models 1−10 were chosen for the simulation. The initial structure was solvated using a truncated octahedral box of water molecules represented according to the TIP3P model,[26] ensuring that the edge of the solvent box was at least 9 Å away from the solute. This required a box with sides of length 50 Å and a total of ~11 000 atoms. The system was minimized and equilibrated via a constant-pressure, constant-temperature simulation. After the equilibration phase at each trajectory, constant-volume, constant-temperature simulations were performed, and the coordinates were saved every 20 ps. The MD simulations were conducted with the AMBER simulation package,[27] and the protein was represented using the Duan et al. force field (AMBER ff03).[28] Particle Mesh Ewald (PME)[29] was applied to calculate long-range electrostatic interactions; SHAKE[30] was applied to freeze the vibrations of the bonds connecting hydrogen atoms. A 2.0-fs time step was used.

TABLE 1: Summary of the Simulations

set	starting point	temp (K)	length of each trajectory (ns)	number of independent trajectories	description
UTRAJ	native	500	10	10	unfolding
NTRAJ	native	300	10	10	native
FTRAJ	unfolded	300	200	5	early folding
TSTRAJ	unfolding TS	300	10	10	folding/unfolding

The unfolding temperature was set to 500 K, and 10 independent trajectories were run, with each extending to 10.0 ns starting from the native structure. This set of trajectories is labeled UTRAJ. For comparison, 10 trajectories were run at 300 K for 10 ns each starting from the native state, and this set is denoted NTRAJ. Five independent simulations at 300.0 K were performed, with each running to 200.0 ns, to investigate the early folding process. In this set of simulations, the initial structure was the fully extended state. After an initial collapsing process modeled by a short simulation in the Generalized-Born[31,32] solvent model, the root-mean-square deviation (RMSD) reached ~8 Å. Five extended structures were selected from which the folding simulations continued using the same protocol and solvent models as used in the UTRAJ and NTRAJ sets. This trajectory set is labeled FTRAJ.

An initial estimate of the TSE was obtained from the unfolding simulations at 500 K by analyses of the free-energy landscape, which allowed identification of an area as defined by the conditions RMSD = 4.0 ± 0.2 Å and radius of gyration (R_g) = 9.1 ± 0.2 Å. There were a total of 42 snapshots in the defined area. Ten conformations were selected from this set of 42 structures for approximately 2 frames per each of the 5 trajectories that were structurally dissimilar from one another by visual inspection. Using these 10 structures as the starting points, 10 different trajectories were run for 10.0 ns at 300 K. This trajectory set is referred to as TSTRAJ. A summary of the simulations is provided in Table 1.

Results and Discussions

Native-State Ensemble. We first examined the stability of FSD-1 at room temperature. Consistent with our earlier observations[33] and experimental findings,[24] FSD-1 was marginally stable at room temperature. The backbone root-mean-square deviation (RMSD) from that of the native structure in the 10 NTRAJ trajectories at 300 K remained mostly within ~2.0 Å, although it also reached ~3 Å from time to time. In other trajectories, although the RMSD transiently reached ~5 Å in some trajectories, it came back quickly to below 3 Å. This level of RMSD is higher than the typical RMSDs observed in simulations of other stable proteins. The radius of gyration (R_g), which measures the size and compactness of the protein, was relatively stable and fluctuated between 9 and 10.5 Å, indicating that the protein remained roughly similar in size to the native structure.

A two-dimensional contour map of the population density was generated with the data from the NTRAJ trajectories using WHAM[34,35] (Figure 2) with the RMSD (y axis) and the radius of gyration (R_g, x axis) as the reaction coordinates. The figure shows that sampling in these 10 trajectories was mainly around the native state, consistent with our earlier observation. The most populated region was around RMSD ≈ 1.5−2 Å, which is the native-structure basin. Extension to the region with RMSD ≈ 3.5 Å was also observed. The broad native basin enabled the protein to sample a wide range of conformational space. More importantly, it suggests that the folding transition-state ensemble lies beyond an RMSD of 3.5 Å. It also indicates that the protein

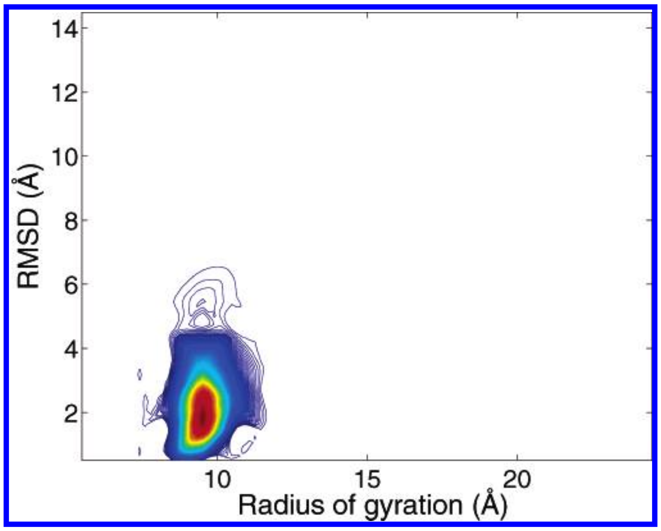

Figure 2. Two-dimensional population distribution around the native state (NTRAJ) at 300 K. The RMSD and R_g are the reaction coordinates. The population is represented by the color gradient, where red is the most populated area.

has a rather low tendency to cross the region of RMSD ≈ 4 Å, suggesting the presence of an (free) energy barrier. Overall, these results are consistent with the observations that FSD-1 is a marginally stable protein[24,33] and its native basin appears to be relatively broad.

General Features of the Unfolding Trajectories. The details of the unfolding process of FSD-1 at relatively lower temperatures were reported in a previous work.[33] Here, the unfolding was conducted at 500 K, which is higher than the previously reported temperatures for unfolding.[33] The elevated temperature allows better sampling of the unfolded state. Figure 3 shows the RMSD from the native structure of two of the total of 10 unfolding trajectories (UTRAJ). The RMSD reached more than 4.0 Å within 1.0 ns at 500 K and 8−10 Å within 5 ns. During this time, the unfolding of the β-hairpin took place first, whereas most of the helix retained its structure and denatured slowly. After that, the RMSD fell back to below 5.0 Å and resumed an increasing trend after 5.5 ns, eventually reaching ~10 Å at the end of the trajectory (10 ns). The R_g value remained close to that of the native protein (~9.5 Å) until 1.5 ns, then started to increase, and reached ~15 Å within 5 ns when the RMSD reached ~8−10 Å. A rapid collapse was observed between 5 and 5.5 ns when R_g fell rapidly back to ~9.5 Å, close to the R_g value of the native state. However, this was a completely denatured state because the corresponding RMSD was greater than 6 Å. The reduction in R_g is indicative of compact structures even in the denatured states. R_g then fluctuated in a narrow band (between 10 and 12 Å) until 8 ns, and started to increase again to 17 Å after 8 ns until the end of the trajectory.

Denatured-State Ensemble and Early Folding Processes. Figure 4 shows the two-dimensional contour map of the population density obtained from the 10 unfolding trajectories using the RMSD and R_g as the reaction coordinates. Apart from the high population around the native structures (RMSD ≈ 2.5, R_g ≈ 9.5), population in the denatured state is also high. In fact, at 500 K, the most populated region is around RMSD ≈ 7 Å and R_g ≈ 10 Å, which is a fully denatured state and represents the denatured-state ensemble. Interestingly, although the denatured-state ensemble is structurally very different from the native state, as judged by the large (~7-Å) RMSD, the R_g values are quite similar; the native-state R_g value of ~9.5 Å is comparable to the ~10-Å value of the most populated denatured state. This implies that the denatured-state ensemble is dominated by compact structures.

We conducted five folding simulations (FTRAJ) at 300 K starting from the extended conformations to examine the early-stage folding. Figure 5 shows the RMSD from two of the five folding trajectories. The RMSD decreased to ~6 Å within 100.0 ns and fluctuated around this value until 200.0 ns, when the trajectory was stopped. During the slow decrease of the RMSD, an occasional sudden jump of the RMSD was also observed, e.g., between 85 and 100.0 ns in one trajectory and between 130 and 140.0 ns in the other, indicative of unfolding events. In some cases, these unfolding events were accompanied by increases in R_g, indicating that the protein moved toward an extended state transiently. For example, R_g transiently reached above 17 Å at around 85 ns in one trajectory when the RMSD also increased. However, these transient events soon dissipated, and the previous trend of the RMSD was resumed. In these

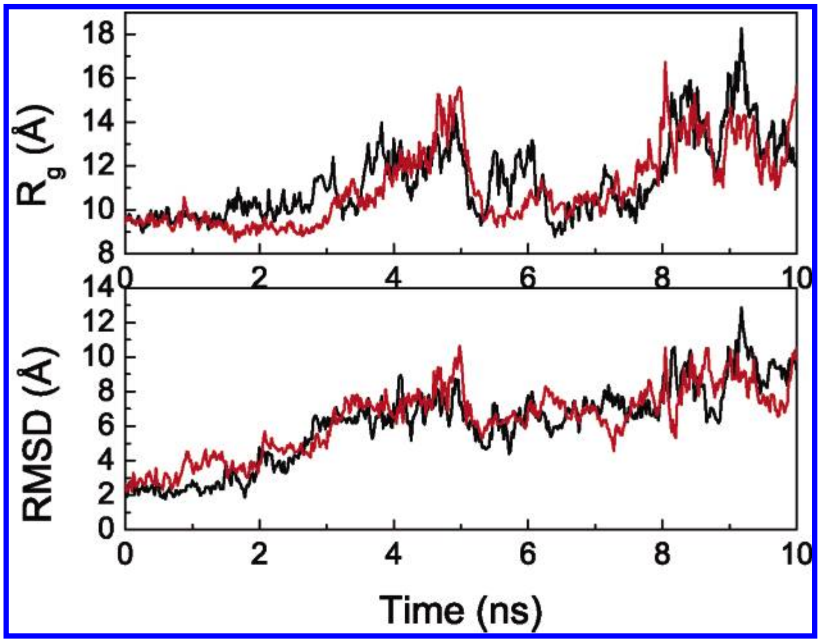

Figure 3. R_g and RMSD from two representative unfolding trajectories of the UTRAJ set at 500 K.

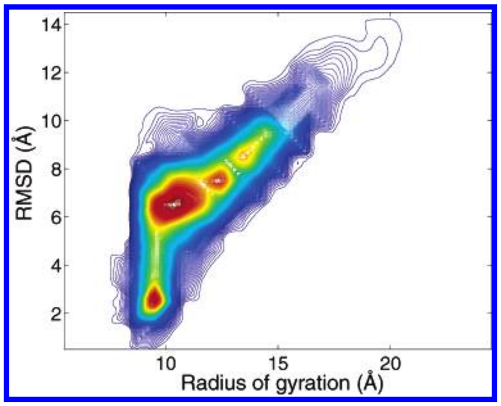

Figure 4. Two-dimensional population contour from the unfolding trajectories (UTRAJ) at 500 K. The coloring scheme is the same as in Figure 2.

two trajectories, the lowest RMSD value was around ~5 Å. Similar events were also observed in other trajectories (data not shown).

A two-dimensional distribution map, shown in Figure 6, was generated by combining the data from all of the folding trajectories (FTRAJ) using the weighted histogram analysis method (WHAM).[34,35] The most populated region was around RMSD ≈ 5 Å and R_g ≈ 9 Å; both values are notably smaller than RMSD ≈ 7 Å and R_g ≈ 10 Å observed in the unfolding simulations. The difference suggests a shift toward the compact and native state. Presumably, such a shift was due to the early folding process. Interestingly, there were additional populated regions at RMSD ≈ 7 Å and R_g ≈ 9.0 Å and RMSD ≈ 9.0 Å and R_g ≈ 9.5 Å. These regions were not observed in the 500 K unfolding simulations. The difference suggests that the free-energy landscape is notably smoother at higher temperature.

$C_\alpha - C_\alpha$ contact maps were calculated for both the unfolding (UTRAJ) and folding (FTRAJ) simulations and are shown in Figure 7 for comparison. The unfolding contact map was calculated for the snapshots within the general basin of the denatured state (RMSD > 5.5 Å), whereas the folding map was obtained from the second half of the trajectories (100−200 ns). A rather interesting observation was the residual helical secondary structures in the denatured state, including the native helix. As for the folding map, the pattern of secondary structures resembled that of the denatured state. A notable difference was the partial formation of the non-native contacts including long-range hydrophobic contacts between F12 and F21, I22 that partially stabilized a transient β-hairpin of fragment $F_{12}RNE_{15}$-$KELRD_{20}FI$. These long-range contacts were responsible for the increased contact order observed in the FTRAJ simulations (discussed later).

The conformations evolved in the five folding trajectories were examined by a clustering analysis. The structures whose main-chain RMSD values were within 2.5 Å of each other were assigned to the same cluster. The representative structures (taken from the center of the cluster) of the most populated clusters from the folding trajectories are shown in Figure 8. These structures were all reasonably compact and had partial formation of the secondary structural elements, including both the native and non-native secondary structures.

We further examined the secondary structures in the early-stage folding. Figure 9 shows the secondary structures averaged over the second half (100−200 ns) of the FTRAJ trajectories. The native secondary structures are also shown in the figure as green and red triangles for comparison. In comparison to the native secondary structures, the second β-strand, the loop region ($R_{10}TFRN$), and the C-terminal portion of the helix ($F_{21}IEKFK$) were mostly in their respective native conformations during the simulations. These fragments were also in their respective native secondary structures in the UTRAJ simulations (i.e., contact maps in Figure 7). Thus, residual (native) secondary structures can exist in the denatured-state ensemble and are perhaps the early folding nucleus. However, the N-terminal β-strand (Y_3TAK) stayed mostly in the helical region in the early stage of folding (FTRAJ), forming a non-native helix. This was probably due to the relatively high helical propensity of Ala_5 and Lys_6. According to the Chou−Fasman[36] scale, the helix propensities of Ala and Lys are, respectively, 1.4 and 1.2, notably higher than their β-sheet propensities (0.83 and 0.74, respectively). Thus, A_5K_6 facilitated helix nucleation in the denatured-state ensemble.

On the other hand, although most residues of the C-terminal helix ($E_{15}KELRDFIEKF_{25}$) had a high helix population in early

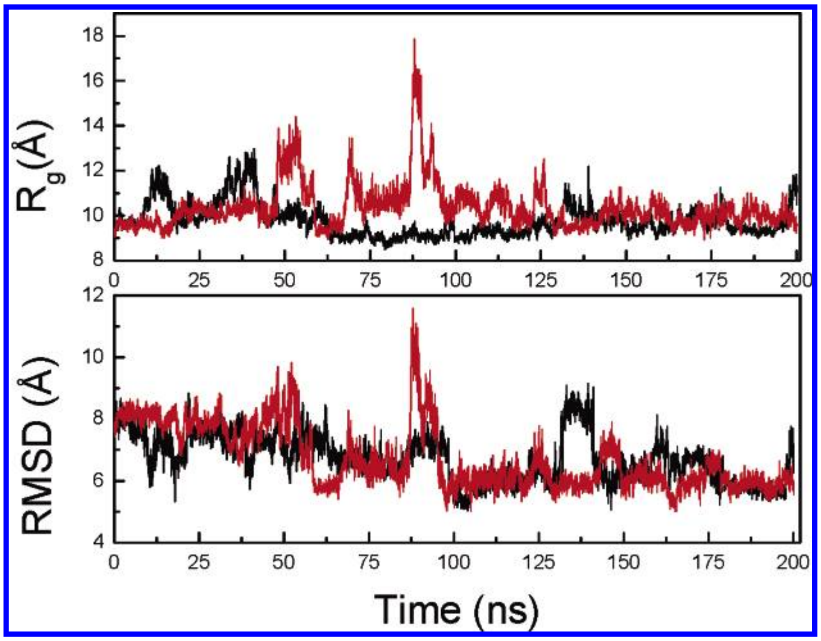

Figure 5. R_g and RMSD of two representative trajectories of the folding simulations (FTRAJ) at 300 K.

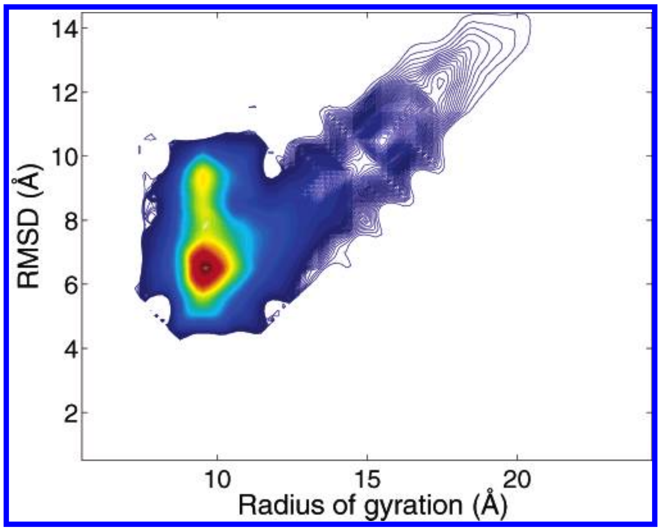

Figure 6. Two-dimensional population distribution from the folding trajectories (FTRAJ) at 300 K. The coloring scheme is the same as in Figure 2.

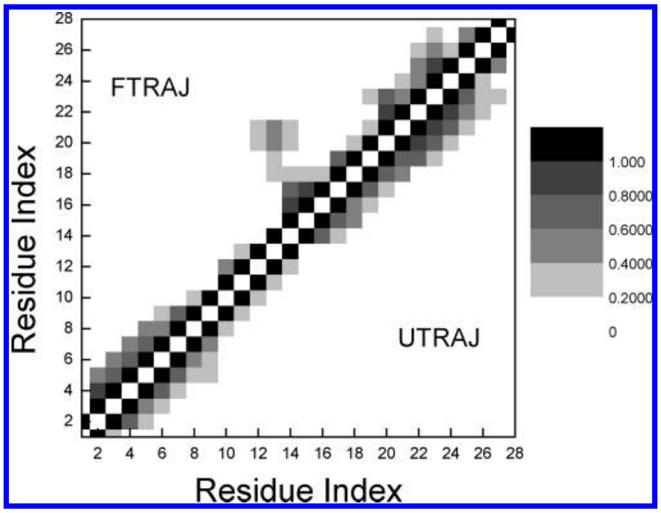

Figure 7. Comparison of $C_\alpha-C_\alpha$ contact maps calculated for the unfolding (UTRAJ, lower-right triangle) and folding (FTRAJ, upper-left triangle) simulations. The gray scale indicates the fractional occupancy. The cutoff distance is 6 Å.

folding (FTRAJ), the helix was broken in the middle primarily because of the lack of helix formation in three residues, E_{17}, R_{19}, and D_{20}. Judging from the strong helical populations of E_{15} (84%) and E_{23} (61%), the lack of helix population of E_{17} was likely due to local (non-native) interactions. Indeed, a non-native salt bridge was formed between E_{17} and R_{19}. This salt bridge was quite stable during the simulation with an occupancy rate of more than 58% when averaged over 100−200 ns of the FTRAJ simulations, which was the highest occupancy rate among all salt bridges found in the same period. The observed high stability was partially due to their close proximity. Such local attractive forces facilitated the formation of short-range salt bridges, as observed in many high-resolution protein structures.[37] Because E_{17} and R_{19} are next to each other when the local sequence assumes the β-sheet conformation, the non-native salt bridge "locks" the local fragment into the non-native β-sheet conformation and reduces the folding rate. In summary, the early-stage folding events and the denatured-state ensemble included the formation of both native and non-native secondary structures. Evidentially, the non-native structures would have to dissipate in the subsequent folding processes and could have a negative impact on both folding kinetics and stability.

In an attempt to enhance the stability of FSD-1, Sarisky and Mayo examined the relevant sequences[38] using energetic analyses of the FSD-1 native structure. Here, we propose that enhancement of the stability and folding rate could arise by substituting the key residues that help to stabilize the non-native secondary structures and salt bridges. These proposed changes are based on analyses of the denatured-state ensemble. Thus, our approach is complementary to the work of Sarisky and Mayo. Two examples are residues Ala_5 and Lys_6, which are part of the first β-strand. Because they have relatively high helical propensities, as discussed earlier, they likely facilitate the formation of the non-native helix in the denatured-state ensemble. A possible substitution is K6R because Arg has almost equal propensities for helix and sheet formation according to the Chou−Fasman scale, whereas Lys has a much stronger helical propensity according to the same scale. One might also contemplate substituting Ala_5 for a less helical residue (e.g., Ile). Other possible changes to stabilize the helix include E_{17}, R_{19}, and D_{20}. Some of the likely beneficial substitutions include R19K and D20E, both of which increase the overall helical propensity. Another useful strategy might be destabilization of the $E_{17}-R_{19}$ non-native salt bridge.

Folding Transition-State Ensemble. The transition-state ensemble is characterized by its instability because it resides on a peak of the free-energy landscape. Therefore, in simulations, the population around the TSE should be much lower than in the native- and unfolded-state ensembles. Hence, the TSE can be identified from the unfolding simulations,[3−7] although caution should be applied because the unfolding TS and the folding TS might be slightly different from each other. In our simulations, an interesting observation from the unfolding (500 K) simulations was that the native-state and denatured-state ensemble were separated by a less populated region around RMSD $\approx$ 3.5−5.5 Å and $R_g \approx$ 9.0−10.0 Å, as shown in Figure 4. The low population is indicative of the presence of a barrier in the free-energy landscape. This barrier is present like a crest between the highly populated troughs in the free-energy landscape. The low-density region corresponds to a high energy barrier that stands in the path from the denatured state to the native state and the reverse. On the other hand, the folding

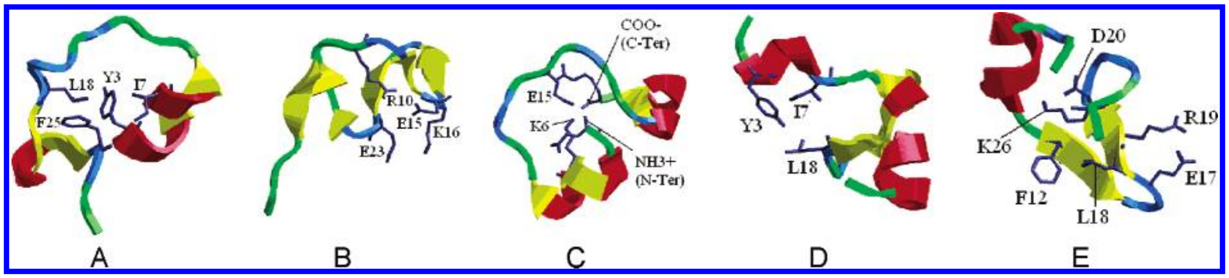

Figure 8. Representative structures of the most populated clusters in the folding trajectories (FTRAJ).

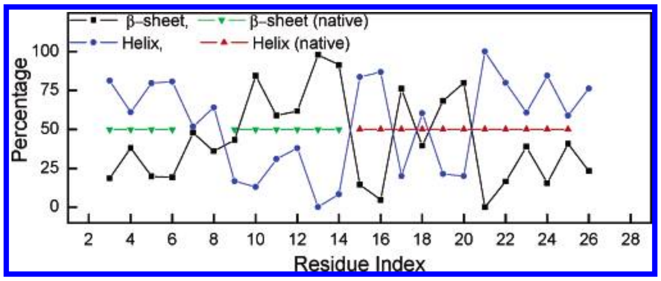

Figure 9. Average percentages of helix (blue) and β-sheet (black) from the five folding trajectories (FTRAJ). The secondary structures of the native structures are shown in green (β-sheet) and red (helix).

simulations at 300 K sampled the region RMSD $\approx$ 5 Å and $R_g \approx$ 9 Å, and the folding process met resistance at around RMSD $\approx$ 4.5 Å and $R_g \approx$ 9.5 Å. Thus, the region also showed the characteristics of a high (free) energy barrier at the folding temperature (300 K) and was close to the transition-state ensemble identified from the unfolding simulations.

Because the position of the TSE is a maximum on the free-energy surface, it is expected that the process can go in either direction (i.e., toward either the native or the denatured state) if the simulations start from the TSE. Thus, a possible validation of the proposed TSE is to conduct a series of simulations from the TSE structures. Ten simulations were performed (TSTRAJ) starting from different conformations with RMSD $\approx$ 3.5−5.5 Å and $R_g \approx$ 9.0−10.0 Å selected randomly from the unfolding trajectories. Indeed, as expected, four trajectories demonstrated various degrees of decreasing RMSD (Figure 10), indicating that the protein moved toward the native structure in the trajectories, whereas others demonstrated increasing RMSD (data not shown) with the structures moving toward the denatured state. In particular, one trajectory demonstrated an almost complete folding process, and its RMSD started from ~4.8 Å and decreased to 2.5 Å by the end of 10.0 ns. Thus, the structure reached the general basin of the native-structure ensemble. Such rapid folding is indicative of a downhill process. Three other trajectories also demonstrated various degrees of decreasing RMSD (~4 Å). In the remaining six trajectories (data not shown), some structures unfolded completely and moved toward higher values of the RMSD. In all of these trajectories, the variation of R_g was small, and the value fluctuated within the range 9−10 Å, which is similar to the native R_g (~9.5 Å).

Representative structures of the TSE are shown in Figure 11. A common feature of these structures is the substantial formation of native secondary structures. In all cases, the native helix was almost complete, and the β-hairpin opened up, partially exposing the hydrophobic core to solvent. Notable variations of the structures were observed around the β-hairpin. In most cases, the β-hairpin was partially formed, and the overall topology was close to the native structure. These observations were confirmed by the residue contacts formed during the simulations. The average $C_\alpha-C_\alpha$ contacts of the TSE structures are shown in Figure 12 and are compared to the native map. In addition to the near completion of the native helix, the β-hairpin also started to form, starting from the turn region. The contact map also shows that the turn was the nucleation site of the β-hairpin. Thus, improvement in the turn is likely beneficial to the overall stability and folding of the protein. Among the non-native contacts, the N-terminal β-strand showed signs of a transient helix, similar to that observed in the denatured state.

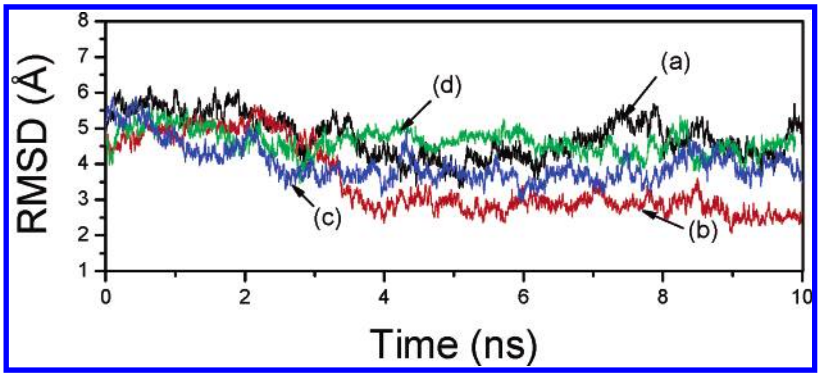

Figure 10. RMSDs of four trajectories at 300 K started from the unfolding TSE (TSTRAJ). Labels a−d indicate the corresponding starting structures shown in Figure 11.

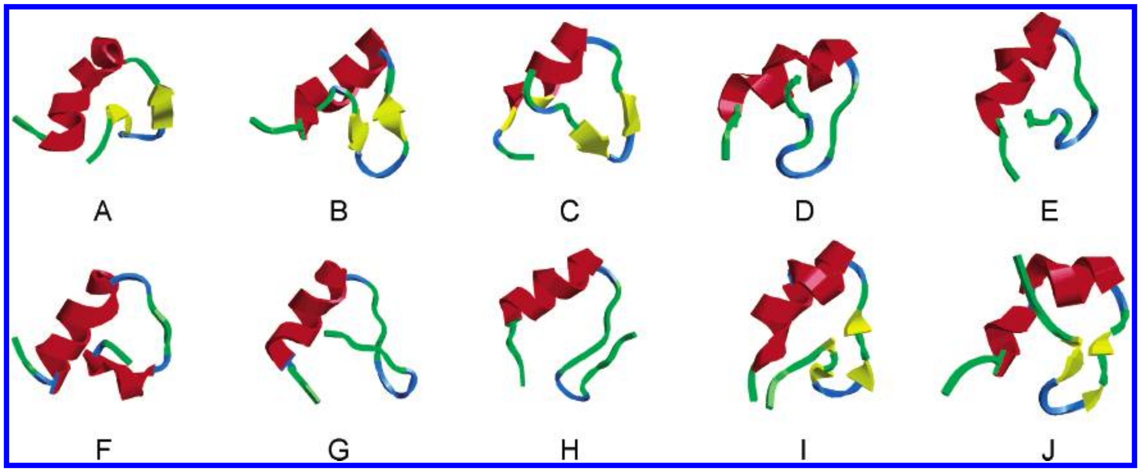

Figure 11. Representative structures of the transition-state ensemble that were used as the starting structures in the TSTRAJ simulations. The close resemblance to the native secondary structures is readily apparent.

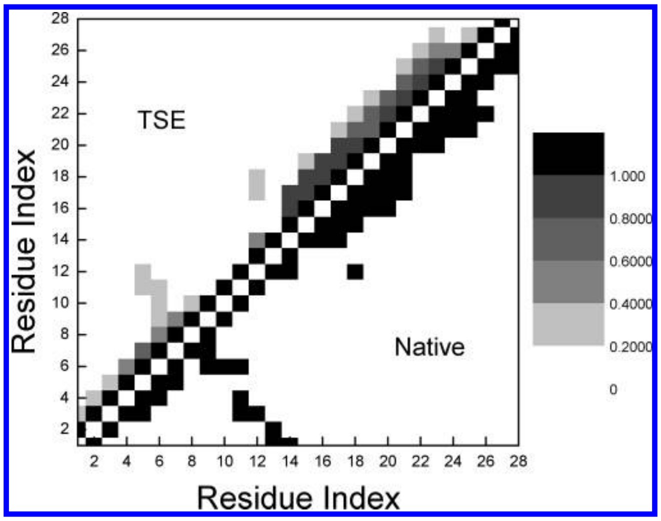

Figure 12. $C_\alpha-C_\alpha$ contact map in the native NMR structure (lower-right) and that averaged over the TSE structures (upper-left). The gray scale indicates the fractional occupancy. The cutoff distance is 6 Å.

Discussion

The simulations reported herein were performed in both directions (folding/unfolding) of the folding reaction coordinate and were started from different points on the conformational space and at different temperatures. The major aim of this work was to characterize the TSE of the folding process by combining all of the information gathered from the simulations. We identified the high (free) energy barrier that separates the native state from the unfolded conformations.

We investigated three key areas of the free-energy landscape of FSD-1. In the denatured-state ensemble, there was a considerable amount of residual secondary structure, albeit both the native and non-native secondary structures exist. Thus, when measured by the overall secondary structure population, the transition (e.g., thermal melting) between the native- and denatured-state ensembles is expected to be smooth. This is consistent with the experimental observation that FSD-1 has a rather smooth melting curve[24,38] when monitored using circular dichroism. In fact, experimentally, the transition, as measured by the CD signal, is marked by a wide range from ~4.0 to 80 °C with the middle point close to ~40 °C. The wide range of the transition also suggests a somewhat flexible native structure and a broad native free-energy basin. Indeed, this was observed in the simulations.

On the other hand, the structures of the transition-state ensemble were characterized by near-native secondary structures and an overall near-native topology. A consistent observation was the partial formation of the β-hairpin and partial unfolding of the native hydrophobic core. This suggests that completion of the native structure is triggered by the simultaneous formation of both the β-hairpin and the native core in a cooperative manner. Furthermore, the folding of FSD-1 is initiated by the substantial formation of native (helical) secondary structures that lead to tertiary structure formation toward the TSE. This is consistent with the framework models.[39−41]

A notable difference between the denatured-state structures and the TSE structures is the lack of formation of the overall topology in the former. Although these denatured structures are compact and, on average, have (transient) native secondary structures, the overall topology of these structures does not resemble that of the native structure or the structures of the TSE. On the basis of this observation, we propose that the rate-limiting step in the folding of FSD-1 is the formation of the correct topology that leads to the TSE structures.

We calculated the relative contact orders[42−44] of the representative structures observed in the simulations to obtain a qualitative assessment of the topological entropy.[43−45] The relative contact orders of the early-stage clusters (Figure 8) ranged from 0.134 to 0.175, and those of the TSE structures (Figure 11) were between 0.117 and 0.163. The FSD-1 native structure has a relative contact order of 0.139, which is very close to the average relative contact order of the TSE structures (0.135) and notably lower than that of the early-stage structures (0.150). The higher contact orders in the early-stage structures imply that there was substantial formation of non-native long-range contacts in these structures and that the native structure has favorable chain (topology) entropy.

We note that the denatured states were identified from our unfolding simulations at 500 K. This temperature is notably higher than the typical experimental unfolding temperatures. Thus, the structures identified from the simulations could be even "more denatured" than the ones in typical thermal denaturation experiments. For the denatured-state ensemble, as expected, the average contact order was 0.100, the lowest of all states, because of the lack of formation of any long-range contacts. An interesting observation was the substantial increase in the contact order in the early stage of folding in comparison to the (fully) denatured state, indicative of long-range contacts and compact structures. This was largely due to the nonspecific hydrophobic collapse. For example, the persistent long-range contacts among F_{12}, F_{21}, and I_{22} observed in early-stage folding were stabilized by the hydrophobic force. However, as some of these long-range contacts were non-native, they had to dissipate in the subsequent folding, which led to lower contact order. The increase in chain (topology) entropy was a favorable direction that drove the protein toward the native structures. Thus, chain (topology) entropy appears to play an important role in protein folding. Interestingly, we found that the chain (or topological) entropy favored the native state in comparison to those structures found in early-stage folding and was one of the driving forces to unfold some of the early-stage compact structures.

Furthermore, the similarity in contact orders of the native and TSE structures is consistent with the observation of similar topological structures in these two states. Because of this similarity, one might be able to use the native structure to estimate the contact orders of the transition-state ensemble from which the folding rates can be estimated.

Accurate identification of the folding TSE is an important step toward understanding folding mechanisms. In this work, we combined unfolding and folding simulations with a set of simulations that started from a small set of selected perspective TSE structures. Although the results were consistent with the notion that these structures were likely representative of the TSE, some cautionary notes are clearly warranted. Most notable is the small set of selected structures in the "refolding" (TSTRAJ) simulations. Obviously, the 10 simulations, regardless of where they started, were insufficient to provide solid statistics for a rigorous identification of TSE. Thus, the conclusions based on these 10 simulations are rather qualitative. Fortunately, these conclusions are also consistent with the observations based on other sets of simulations, including both folding and stability simulations. Thus, we are cautiously optimistic that the identified structures capture the main features of the TSE.

Conclusion

Three key areas of the free-energy landscape of the FSD-1 protein, i.e., native, transition, and denatured, and their respective structural ensembles have been investigated by combined folding and unfolding molecular dynamics simulations with explicit solvent. The native ensemble of FSD-1 rests on a relatively flat free-energy basin marked by the high flexibility of the protein. The TSE was initially identified from unfolding simulations at 500 K and was examined by 10 folding simulations starting from selected structures of the ensemble. Among these, four trajectories moved closer to the native structure as judged by the main-chain RMSD, and one moved into the native ensemble within 10.0 ns. Judging from the main-chain RMSD, the TSE is about 4.5 Å from the native structure and is characterized by the substantial formation of native secondary structures, including both the α-helix and *β*-sheet, with partial exposure of the hydrophobic core to the solvent. Residual secondary structures were observed in the denatured-state ensemble obtained from the unfolding simulations. These secondary structures were also present in the early stage of folding. Thus, they are likely the folding nucleus of early-stage folding. The presence of non-native secondary structures in the denatured state is consistent with the smooth melting curve observed using circular dichroism. Taken together, the results suggest that the rate-limiting step of FSD-1 folding is the development of tertiary structures and the key fragments of secondary structures. This is followed by a cooperative step in which completion of the secondary structures and packing of the hydrophobic core take place simultaneously. Analyses indicated that non-native interactions involving the pairs Ala_5, Lys_6 and Glu_{17}, Arg_{19} were partially responsible for stabilizing the non-native structures in the denatured-state ensemble. We propose that destabilization of these interactions could help to enhance the stability and folding rate of the protein.

Acknowledgment. We thank Dr. Kevin Plaxco for providing programs to calculate the contact orders. This work was supported by research grants from the NIH (Grants GM64458 and GM67168 to Y.D.). Use of the Pymol, VMD, and Rasmol graphics packages is gratefully acknowledged. H.L. and S.G.D. contributed equally to this work.

References and Notes

(1) Lindorff-Larsen, K.; Rogen, P.; Paci, E.; Vendruscolo, M.; Dobson, C. M. *Trends Biochem. Sci*. **2005**, *30*, 13.

(2) Vendruscolo, M.; Dokholyan, N. V.; Paci, E.; Karplus, M. *Phys. Rev. E* **2002**, *65*, 061910.

(3) Day, R.; Bennion, B. J.; Ham, S.; Daggett, V. *J. Mol. Biol*. **2002**, *322*, 189.

(4) Li, A.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 10430.

(5) Li, A.; Daggett, V. *J. Mol. Biol*. **1996**, *257*, 412.

(6) Levitt, M. *J. Mol. Biol*. **1983**, *168*, 621.

(7) Dastidar, S. G.; Mukhopadhyay, C. *Phys. Rev. E* **2005**, *72*, 051928.

(8) Sosnick, T. R.; Dothager, R. S.; Krantz, B. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17377.

(9) Anil, B.; Sato, S.; Cho, J. H.; Raleigh, D. P. *J. Mol. Biol.* **2005**, *354*, 693.

(10) Jemth, P.; Day, R.; Gianni, S.; Khan, F.; Allen, M.; Daggett, V.; Fersht, A. R. *J. Mol. Biol.* **2005**, *350*, 363.

(11) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1525.

(12) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17327.

(13) Fersht, A. R.; Sato, S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7976.

(14) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14338.

(15) 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.; Fersht, A. R. *Nature* **2003**, *421*, 863.

(16) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.

(17) Dinner, A. R.; Karplus, M. *J. Mol. Biol.* **1999**, *292*, 403.

(18) Neri, D.; Billeter, M.; Wider, G.; Wuthrich, K. *Science* **1992**, *257*, 1559.

(19) Zhang, O.; Formankay, J. D. *Biochemistry* **1995**, *34*, 6784.

(20) Farrow, N. A.; Zhang, O. W.; Formankay, J. D.; Kay, L. E. *Biochemistry* **1995**, *34*, 868.

(21) Zhang, O. W.; FormanKay, J. D. *Biochemistry* **1997**, *36*, 3959.

(22) Kortemme, T.; Kelly, M. J. S.; Kay, L. E.; Forman-Kay, J.; Serrano, L. *J. Mol. Biol.* **2000**, *297*, 1217.

(23) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153.

(24) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82.

(25) Dahiyat, B. I.; Sarisky, C. A.; Mayo, S. L. *J. Mol. Biol.* **1997**, *273*, 789.

(26) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, W. R.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(27) Case, D. A.; Darden, T. A.; T. E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, 2004.

(28) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem*. **2003**, *24*, 1999.

(29) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.

(30) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comput. Phys.* **1977**, *23*, 327.

(31) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297.

(32) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

(33) Lei, H.; Duan, Y. *J. Chem. Phys.* **2004**, *121*, 12104.

(34) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.

(35) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339.

(36) Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 211.

(37) Sarakatsannis, J. N.; Duan, Y. *Proteins* **2005**, *60*, 732.

(38) Sarisky, C. A.; Mayo, S. L. *J. Mol. Biol.* **2001**, *307*, 1411.

(39) Kim, P. S.; Baldwin, R. L. *Annu. Rev. Biochem.* **1982**, *59*, 631.

(40) Ptitsyn, O. B. *J. Protein Chem.* **1987**, *6*, 273.

(41) Kim, P. S.; Baldwin, R. L. *Annu. Rev. Biochem.* **1990**, *59*, 631.

(42) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985.

(43) Makarov, D. E.; Keller, C. A.; Plaxco, K. W.; Metiu, H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3535.

(44) Makarov, D. E.; Plaxco, K. W. *Protein Sci.* **2003**, *12*, 17.

(45) Weikl, T. R.; Dill, K. A. *J. Mol. Biol.* **2003**, *329*, 585.