PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A simple simulation model can reproduce the thermodynamic folding intermediate of apoflavodoxin

María Larriva,[1] Lidia Prieto,[1] Pierpaolo Bruscolini,[2] and Antonio Rey[1,2]*

[1] Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain
[2] Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Corona de Aragón, 42, E-50009 Zaragoza, Spain

ABSTRACT

Flavodoxins are single domain proteins with an α/β structure, whose function and folding have been well studied. Detailed experiments have shown that several members of this protein family present a stable intermediate, which accumulates along the folding process. In this work, we use a coarse-grained model for protein folding, whose interactions are based on the topology of the native state, to analyze the thermodynamic characteristics of the folding of *Anabaena* apoflavodoxin. Our model shows evidence for the existence of a thermodynamic folding intermediate, which reaches a significant population along the thermal transition. According to our simulation results, the intermediate is compact, well packed, and involves distortions of the native structure similar to those experimentally found. These mainly affect the long loop in the protein surface comprising residues 120–139. Although the agreement between simulation and experiment is not perfect, something impossible for a crude model, our results show that the topology of the native state is able to dictate a folding process which includes the presence of an intermediate for this protein.

Proteins 2010; 78:73–82.

Key words: protein folding; topology-based models; free energy surface; populations; Monte Carlo.

INTRODUCTION

The protein folding process constitutes a very interesting and important problem which is nowadays being thoroughly studied by the use of both experimental and computer simulation techniques[1,2].

Many single domain proteins fold following a two-state process,[3] in which only the folded and the unfolded states are significantly populated at the transition temperature. The two-state model has been therefore highly used to study both the thermodynamics of the folding transition and its kinetic features.[4] However, some exceptions to this universally accepted behavior have been found in recent years in small or medium-sized proteins. They include the so-called downhill or barrierless folding process,[5,6] in which a continuous transition is found in some fast folding proteins; and folding processes which have been characterized as three-state (or higher).[7] These correspond to the presence of at least one intermediate species, with a population high enough to be detected by different experimental techniques, between the native and denatured states. For proteins showing three-state folding, the two-state model cannot properly fit the experimental data (from calorimetry, for example),[8] and different techniques usually yield different results for the unfolding process. In recent years, proteins with intermediates showing partly unfolded conformations are being the subject of intense research, because these intermediates may in some occasions be important for protein function or for aggregation processes related to folding diseases.[9,10]

Among these, the flavodoxin protein family has been one of the most analyzed.[11] These proteins have an important function as electronic transporters in redox processes of bacteria. Although this function requires the presence of a flavin mononucleotide cofactor (FMN) bound to the protein, the apo form of the flavodoxin from *Anabaena* is stable, and its structure has been determined through X-ray diffraction.[12] Moreover, the folding of this protein has been experimentally determined to have a stable thermodynamic interme-

The authors dedicate this article to the memory of Angel R. Ortiz.
The authors state no conflict of interest.

Grant sponsor: Spanish Ministerio de Educación y Ciencia; Grant numbers: FIS2006-12781-C02, CSD2007-00010; Grant sponsor: Universidad Complutense de Madrid/Banco Santander Central Hispano; Grant number: GR58/08.

Lidia Prieto's current address is Department of Chemistry, City College of New York, Convent Avenue at 138th St., New York, NY 10031.
*Correspondence to: Antonio Rey, Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain. E-mail: jsbach@quim.ucm.es

Received 2 March 2009; Revised 16 June 2009; Accepted 17 June 2009
Published online 6 July 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22521

diate in thermal unfolding experiments, although chemical denaturation experiments have found the usual two-state unfolding behavior.[8,13–15] The stable intermediate has been also found in flavodoxins from different bacteria,[16–20] although not in all the proteins from this family.[21,22]

In the last decades, it has been proposed that the main characteristics of the folding process for proteins may be highly related to the topology of the native structure.[23,24] Therefore, the folding process could be adequately represented by taking into account only the attractive interactions among residues which are in contact in the native state. This approach has been the basis of an extensive use of topology-based simulation models (originally termed Gō-type models[25]) to analyze the characteristics of the folding pathways, mainly kinetic but also thermodynamic.[25–42] A very recent review of the different types of models belonging to this category can be found in Ref. 43. The use of these models does not mean that the topology of the native state is always the only relevant information needed to understand the folding characteristics of a given protein, with the amino acid sequence playing a minor role, as recent experiments have shown (see, for e.g., Refs. 44 and 45 on the folding of proteins with similar structures). These systems clearly indicate that the study of the influence of the native interactions alone on the characteristics of the folding process has yet to be checked in different interesting cases, being the thermodynamic intermediate of apoflavodoxin from *Anabaena* a rather appealing one.

In the past years, some of us have developed a simple, coarse-grained computer model to simulate the thermodynamic characteristics of the folding/unfolding process for proteins, using a topology-based interaction energy.[46] Given the reduction in complexity of the system brought by the simplified description of the protein geometry and energetics, we have been able to use this model to study the thermodynamic characteristics, including the free energy surface, for the thermal folding/unfolding equilibrium of several proteins,[42] as well as to analyze some methodological implications of this type of simulation models.[41,47,48] In this present manuscript, we apply the method to search for the thermodynamic intermediate of the folding of apoflavodoxin from *Anabaena*. The structure of this protein, together with the map which shows the native contacts which define the interaction potential for our simulations (see Materials and Methods) is shown in Figure 1(a), and it has been taken from the Protein Data Bank[49] (PDB) file 1FTG.[12]

MATERIALS AND METHODS

We use a coarse-grained model for the representation of both the protein geometry and the interaction energy. The model has been already described and tested in previous works from our group,[41,42,46–48] and only its most relevant features will be summarized here.

We use a single-bead representation for every amino acid, centered at the corresponding α-carbon position. The α-carbon trace of the native protein is sketched in Figure 1(b). Neighbor beads along the sequence are kept at a constant distance of 3.8 Å (corresponding to a trans peptide bond). A repulsive core prevents overlapping of non-neighbor beads. The native structure, as corresponding to the geometrical model used in this work, is depicted in Figure 1(c).

The attractive energy uses a topology-based potential, which employs as equilibrium conditions the distances among amino acid pairs found in the experimental struc-

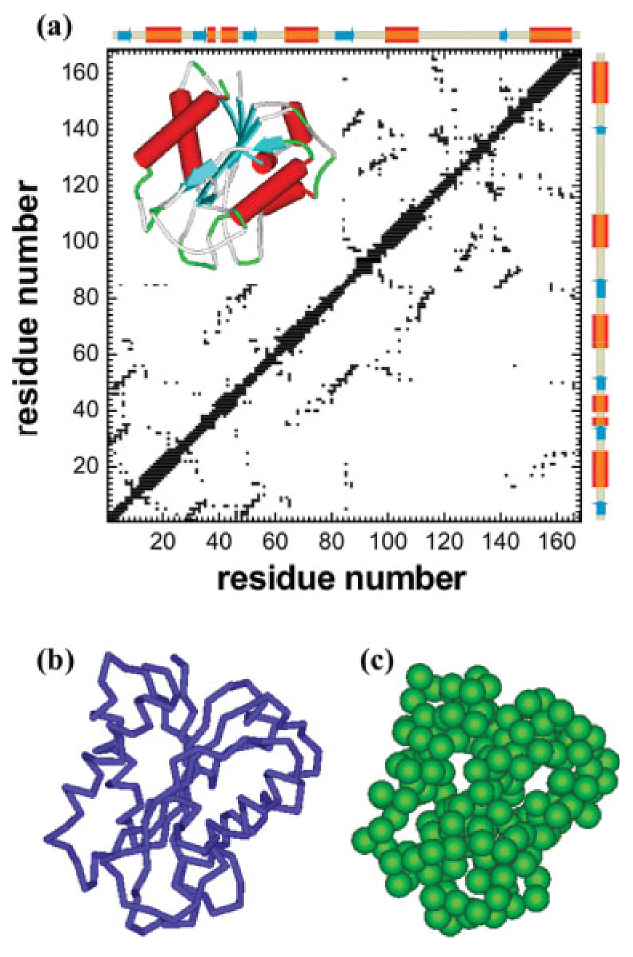

Figure 1

(**a**) Contact map for the apoflavodoxin structure, computed from the PDB file 1FTG. Part of the upper triangle contacts are hidden by a cartoon representation of the protein structure. The secondary structure elements, sketched along the axes, are taken from the header of the PDB file. (**b**) α-carbon trace of the native structure. (**c**) Native structure represented as it is used in the simulation model, with a bead per residue, centered at its α-carbon, depicted at a size corresponding to its repulsive core. The model does not differentiate between elements of secondary structure, and therefore no specific coloring method has been used in panels (b) and (c).

ture. Local interactions correspond to distances between units $i - i + 2$ and $i - i + 3$ (the latter with an added chirality criterium) along the sequence. For the remaining, nonlocal interactions, we only consider the interactions among residue pairs which are in contact in the native structure. We define that two residues are in contact when any pair of heavy atoms belonging to every one of them are at a distance closer than 4.5 Å. The mathematical part of the attractive interactions is defined as a truncated harmonic potential,[46] only acting when the distance d_{ij} between a pair of residues which are in contact in the native structure at a distance d_{ij}^{nat} fulfills the criterium $|d_{ij} - d_{ij}^{\mathrm{nat}}| \leq 0.7$ Å. This cut-off value, and the functional form of the potential, have proved to render adequate values for the thermodynamic characteristics of the folding process for different proteins in previous works from our group.[42,46,48]

Our definition of native contacts creates a contact map, that for the particular case of flavodoxin is shown in Figure 1(a). It has been calculated from the file 1FTG in the protein data bank (PDB).[49] This protein has 169 residues, in a structure with a β-sheet packed against two different layers of α-helices.[12] The coordinates of the first residue of the sequence are not included in the PDB file, and therefore the model used in this work only considers 168 residues. The structure is shown as a cartoon representation in the upper triangle of the contact map, superimposed to the contacts in this triangle, which are however symmetrical to those in the lower triangle. The map contains a total of 331 local-contacts and 421 nonlocal native contacts, according to our definition. In spite of the model simplicity, it is already a large protein, difficult to simulate when one is interested in the full folding/unfolding transition, which has probably made that only very few simulations of this protein have been reported up to know, without any mention to the thermodynamic intermediate.[50–52]

To properly sample the conformational space for the modeled polypeptide chain at different temperatures, we use a replica exchange (or parallel tempering) Monte Carlo simulation procedure,[53,54] which uses an adequate combination of Monte Carlo moves.[46] Given the complexity of the folding transition (see Results), with a rather large free energy barrier between the folded and unfolded states, it has been compulsory to use a large number of replicas at quite close temperatures in the transition region. Otherwise, the acceptance ratio for the replica exchange trials is not adequate to warrant a correct computation of equilibrium properties. Every simulation encloses then 45 different temperatures, from well above to well beyond the transition region, and takes about 200 CPU hours in current single processor machines. The results presented in this work correspond to the average values from 15 independent parallel-tempering simulations. All them start from a fully extended chain.

The analysis of the simulation results includes the calculation of energy fluctuations to compute the heat capacity of the system at every temperature, according to the expression

$$C_{\mathrm{v}}^{*} = \frac{\langle E^{*2} \rangle - \langle E^{*} \rangle^{2}}{T^{*2}} \tag{1}$$

Along this work, the temperature and the energetic quantities are used in reduced units: $T^{*} = T/T_{\mathrm{ref}}$; $E^{*} = E/(k_{\mathrm{B}}T_{\mathrm{ref}})$, where k_{B} is the Boltzmann constant.

We also carry out a histogram analysis of different energetic and structural properties, and a more complex Weighted Histogram Analysis Method (WHAM) to obtain free energy estimations from a complete and simultaneous consideration of the conformational space simulated at all the temperatures.[55–58]

The experimental characterization of the thermodynamic intermediate in reference[14] was based on the "integrity ϕ values", defined as

$$\phi = \frac{\Delta\Delta G_{\mathrm{I-D}}}{\Delta\Delta G_{\mathrm{N-I}} + \Delta\Delta G_{\mathrm{I-D}}} \tag{2}$$

in terms of the changes in the stability of the native (N), denatured (D), and intermediate (I) states upon mutations in the protein sequence; notice that this definition is different from the standard experimental ϕ values, where the kinetic transition state of the folding process is used instead of the intermediate state.[59]

For a direct comparison with the former ϕ values, we have calculated for every residue i a ϕ_i value, defined as

$$\phi_i = \left\langle \frac{\mathrm{NNC}_i \text{ in a conformation}}{\mathrm{NNC}_i \text{ in the folded conformation}} \right\rangle \tag{3}$$

where NNC_i refers to the number of nonlocal native contacts for residue i, with a 20% tolerance from the native distances. This mechanistic definition follows that used previously by other authors.[60,61] The denominator in Eqn (3) is computed from the PDB file. The average corresponds to all the conformations recorded from the simulations at a given temperature which belong to the intermediate state.

The main equivalence between the definitions of ϕ values in Eqs.(3) and (2) is that ϕ_i takes a value of 1 if residue i is in a fully native environment at the considered conformations, and a value of 0 if it has lost all its native contacts, therefore corresponding to an unfolded structure, at least at a local level. Notice however that the earlier numerical definition is a rather crude approximation[62] for the experimentally used ϕ values and, as a matter of fact, it is highly influenced by the number of native contacts of a given residue. This can be understood, for instance, by considering a pair of residues (i,j) that loose in the intermediate one nonlocal native contact between them. If residue i has n_i nonlocal native contacts and residue j has n_j, with $n_i \ll n_j$, the computed ϕ values for these two residues are completely different, even though they have both lost one and the same contact.

Thus, a quantitative agreement between experimental ϕ values and those calculated in this work according to Eq. (3) is not intended here. Therefore, we use the computed ϕ values from the simulations just as a qualitative estimation of the native structural elements which become distorted in the intermediate, and we complete this information with the analysis of individual native contacts, collected in frequency maps (see below), to better characterize the structural features of the intermediate resulting from our simulations.

In this work, we want to focus on the folding characteristics of the protein considered. Of course, the details of the model can influence the simulation results, as we have checked for our model in other protein structures,[41,46] and other research groups have shown as well.[63,64] In this work, however, we have not fitted any of the model parameters in order to get the experimentally observed intermediate. On the opposite, we have kept the same contact definition and cut-off distance, and identical model features, as those we have been using in our recent works on different proteins.[42,48] We should mention that, in these works, and other currently under development with larger proteins, we have found the model to correctly reproduce the thermodynamic characteristics of the folding transition experimentally found for different proteins, both in two-state and downhill folding processes. In none of these cases we have found the least evidence of thermodynamic intermediates induced by the simulation model. Therefore, we are convinced that the results shown in the next section have a real biophysical meaning, and are not an artifact of any specifically tuned model.

RESULTS AND DISCUSSION

As a first result of our simulations, the energy fluctuations at every temperature allow us to readily compute the heat capacity for the system, according to Eqn. (1). This is plotted in Figure 2 as a function of temperature, for the transition region. The figure shows a rather sharp peak, as it also happens in experimental calorimetry results for this protein.[8] As a matter of fact, the narrow temperature interval for the full peak is one of the interesting features of our simulation model. According to the peak maximum, the transition temperature would be around T^*_m = 0.645, and the full transition happens in a temperature interval of ±5% around this value. The experimental heat capacity curve for this protein has its maximum at approximately T_m = 320 K, with the transition spanning a total interval of roughly 40 K (from ∼300 to ∼340 K), thus representing an interval of ±6% around the maximum.[8] This agreement between experiment and simulation gives us a first clue that our model, although rather simple, yet does not create very broad and unspecific transitions, as it usually happens with other simple models in this family.[65,66] On the opposite, the first crude characteristics of the thermodynamic behavior for the folding/unfolding transition are properly represented.

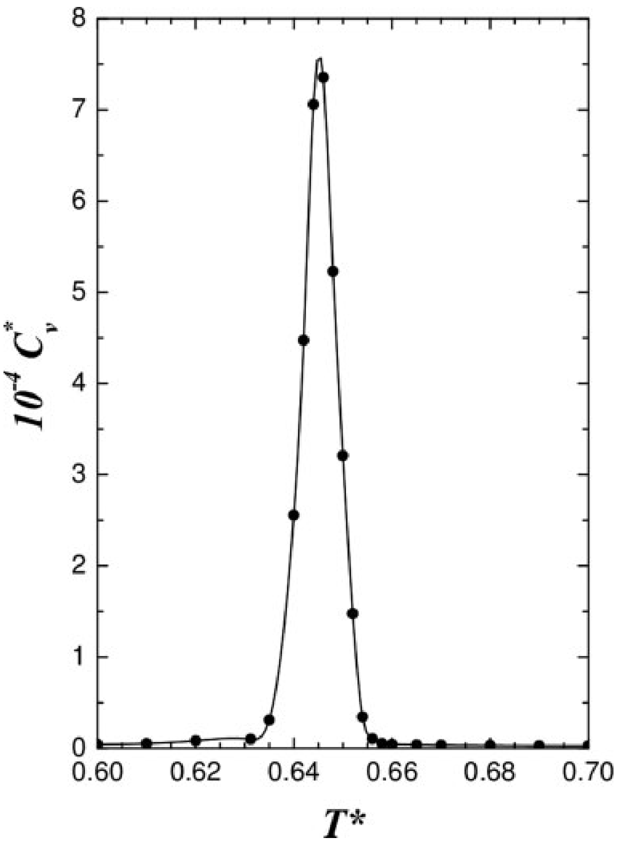

Figure 2

Heat capacity for the simulated system as a function of temperature at the transition region. Both axes are in reduced units. The solid line is drawn just as an eye-guide.

To study the type of transition, experimental studies have to fit the heat capacity curve to different kinds of models (two- or different types of three-state models, for example),[13] and/or to analyze the unfolding characteristics of the thermal transition through different experimental techniques.[8] Computer simulation, on the other hand, may provide a more direct view of the situation, since we have a plethora of conformations along the simulated trajectories which directly allow us to analyze the thermodynamic and structural features of the transition. For the type of study we are considering in this work, a very simple yet highly informative analysis of the numerical data is to calculate the energy histograms from the simulations at different temperatures. For some of these temperatures, in the transition region, these energy histograms are shown in Figure 3(a). At the lowest temperature shown, $T^* = 0.60$, below the transition region, the system shows a single narrow peak at low energies, corresponding to the folded structures present in this regime. Similarly, at the highest temperature shown, $T^* = 0.66$, above the folding/unfolding transition, the system energy shows again a single peak, which appears now at smaller (in absolute value) energies, and corresponds to sporadic native contacts (or native local conformations) in a chain which has become unfolded. At intermediate temperatures, close to the equilibrium unfolding temperature, a standard two-state transition would show the lowest energy peak becoming progressively less intense, while the high energy peak would continuously grow.[4] In our simulations, however, the native peak splits into two clearly separated maxima at the central temperatures shown in Figure 3(a), indicating the presence of an intermediate state which reaches a significant population in the transition region, and accumulates in an amount high enough to appear in the energy histogram as an independent peak. As a matter of fact, a detailed analysis of the profiles from these energy histograms has allowed us to integrate the different curves, in order to extract quantitative information about the population of the different states present along the transition. The results of this analysis are reported in Figure 3(b). Here, we can clearly appreciate that the population of the intermediate state (I) is equivalent to those of the native (N) and denatured (D) states in a narrow temperature range. We should mention that these results do not show the intermediate state to be the species of dominant population at any temperature, as it apparently happens in the three-state interpretation of the experimental data (see Fig. 3 in Ref. 8). However, apart from this small quantitative difference, the agreement between our simulation results and the experiment is quite remarkable, specially if we have into account that no sequence information at all is considered in the simulation model, and only the topology of the folded structure is creating the characteristics of the transition.

To get a deeper insight onto these characteristics, in Figure 4 we show the normalized (unit area) histograms for the different energy contributions in the simulation model. They are computed from the simulations at $T^* = 0.644$, at the maximum of the heat capacity curve. Specifically, we present the histograms for the number of native contacts, separated as nonlocal and local. We have considered a native contact ij to be present in a given conformation when the difference between the distance d_{ij} in it and the corresponding distance d_{ij}^{nat} in the folded conformation is less than 20% of the latter, a condition more stringent than that used in other works.[67,68] It can be seen that, while the curve for the local contacts shows two broad peaks, corresponding to compact and open conformations, the histogram for the nonlocal native contacts is able to discriminate two different peaks in the region of compact conformations, which should correspond to the folded and intermediate states. Let us recall that the number of native nonlocal global contacts for the structure of apoflavodoxin considered in this model is 421 (see Materials and Methods). Thus, the narrow and intense peak around 400 contacts corresponds to small distortions of the native state, typical of the thermal fluctuations shown by off-lattice simulation models at the transition temperature, while the broader and shorter peak centered around 360 contacts has to represent the intermediate state.

Although the analysis of the results at a single temperature is useful, more accurate and significant results can be obtained by simultaneously considering the simulation data from all the temperatures, because this provides a whole picture of the configurational space available for

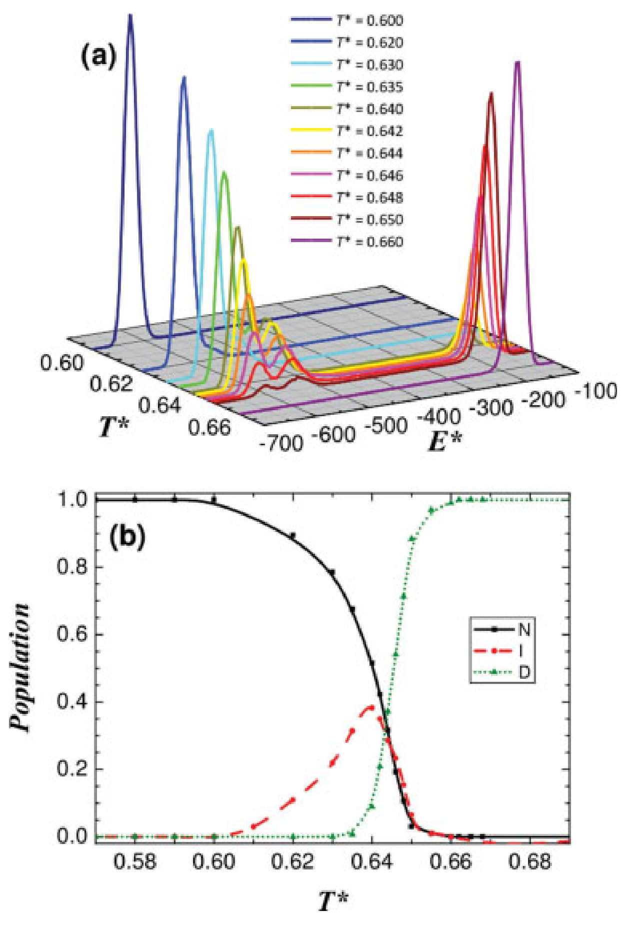

Figure 3

(**a**) Energy histograms at several temperatures in the transition region. Every line corresponds to the average over all the independent simulations computed. (**b**) Populations of the native (N), intermediate (I) and denatured (D) states as a function of reduced temperature.

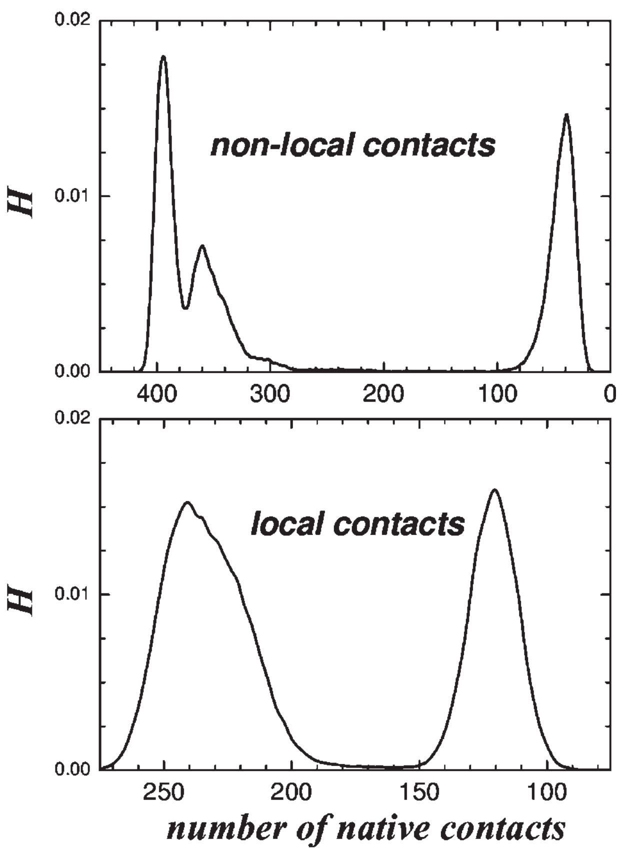

Figure 4

Histograms H of the energy components for the simulated system at a reduced temperature $T^* = 0.644$. Upper panel: number of nonlocal ($|i - j| \geq 4$) native contacts. Lower panel: number of local ($|i - j| = 2$ or 3) native contacts.

the polypeptide model. This is possible thanks to the WHAM method.[55–58] In Figure 5, we show the results for the free energy surface, projected onto the radius of gyration and the number of nonlocal native contacts of the conformations (upper map) and onto the root mean square deviation with respect to native (rmsd) and the number of nonlocal native contacts of the conformations (lower map). Although all the temperatures have been used to accumulate the histograms, the results presented correspond to the reduced temperature $T^* = 0.640$ where, according to our calculations (see Figure 3), the intermediate shows its maximum population. The insets in both maps show a closer view of the free energy surface in the region of compact conformations.

In the full maps of Figure 5, we can observe a narrow region corresponding to the compact structures, with a large number of nonlocal native contacts and a small radius of gyration or rmsd value (the native state in the PDB file has a radius of gyration $R_g = 14.37$ Å, computed from the α-carbon atoms), and a very wide and swallow peak, centered around $R_g = 32$ Å and rmsd > 15 Å, which corresponds to open conformations in the denatured state. Between compact and denatured states there is, according to our topology-based interaction model, a rather high-free energy barrier. The close up views in the figure insets provide, on the other hand, a much better description of the presence of a thermodynamic intermediate with a significant population in the thermal folding/unfolding process of this protein. In the upper map, we can see two free energy minima, with different numbers of long-range native contacts formed, but

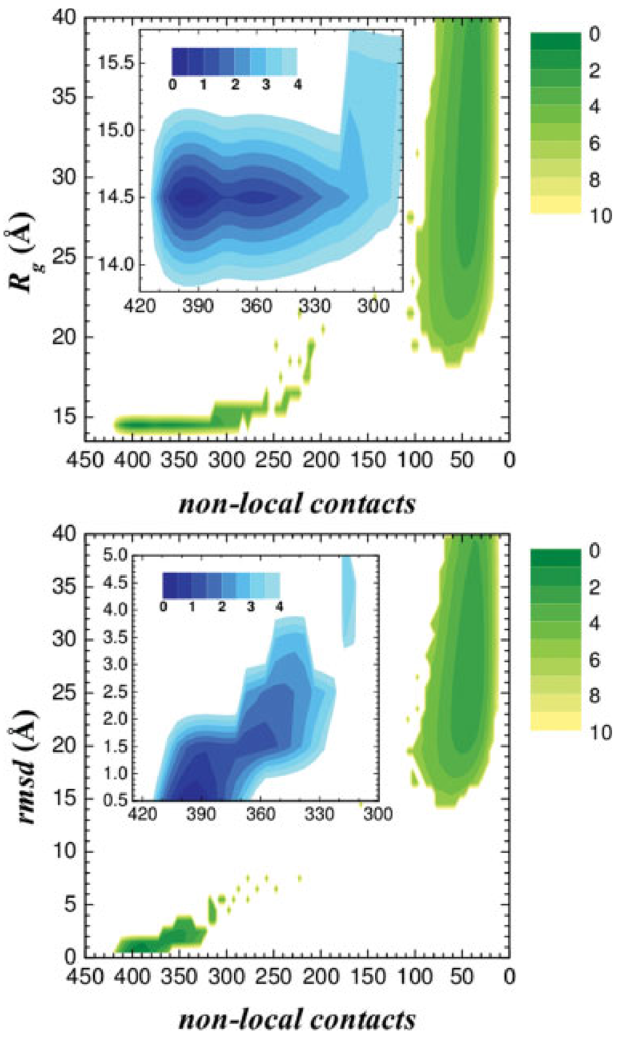

Figure 5

Free energy surfaces computed from the WHAM analysis of our simulation results. In the upper panel, the surface is projected onto the number of nonlocal native contacts and the radius of gyration of the structures. In the lower panel, the surface is projected onto the number of nonlocal native contacts and the root mean square deviation (rmsd) of the structures with respect to the PDB native conformation. In every panel, the inset shows a detailed view of the region corresponding to the compact conformations, to highlight the minima corresponding to the folded and the intermediate states. The free energy scale, as indicated in the legends, is in reduced units.

with essentially the same value for the radius of gyration (at least, at the resolution of the numerical analysis carried out). This implies that the intermediate is as compact as the native state, as it has been also pointed out by the experimental results.[14] The inset in the bottom map shows that, as expected, the intermediate exhibits a certain distortion with respect to the native conformation, because the rmsd values for the native state simulated are below 1.5 Å, whereas the intermediate state shows conformations with rmsd values ranging from 1.5 to about 3.5 Å. According to the simulations, there is a small free energy barrier between the native and the intermediate states at the transition temperature, although quantitatively it seems to be smaller than that obtained from the analysis of experimental data.[14] However, the free energy barriers we appreciate in these plots are projected onto arbitrarily chosen reaction coordinates (as the radius of gyration, the rmsd values or the number of nonlocal contacts) which are useful to discuss the simulation results, but have probably little relation to the "real reaction coordinate(s)" for the folding process. This fact, together with the pure thermodynamic character of our simulation methodology, preclude at this moment the possibility to use our barriers to even estimate any kinetic information on the folding process for apoflavodoxin.

The whole free energy surfaces in Figure 5 show other scattered minima which are less compact and much more unstructured than the thermodynamic intermediate we have already commented on. However, these minima have larger values of the free energy, implying populations negligible in comparison with the thermal intermediate we are considering. In addition, they do not become stabilized at other temperatures, and, therefore, we have discarded the possibility of additional significant intermediates in the folding/unfolding transition for this particular protein.

Finally, to evaluate all the possibilities of our interaction model, we have tried to characterize the structure of the thermodynamic intermediate, because this information has also been recently available from a fine analysis of the experimental data on different mutants of apoflavodoxin.[14] From our free energy maps, we can define a series of values for the total energy, the number of native contacts, the rmsd values, etc, which clearly allow us to discriminate if a given conformation recorded in a snapshot of our simulation results belongs to the intermediate state. As seen in Figure 3(b), these conformations add up to about one third of the total conformations at a certain temperature. This means that, for the analysis that follows, we have been able to use several thousand conformations, coming from different independent simulations at $T^* = 0.640$, which provide an adequate statistical significance for the results. These are shown in Figure 6. In part (a) of this Figure, we show with red circles the number of nonlocal native contacts that every residue shows in the folded conformation taken from the PDB file 1FTG, and also the ϕ values for the individual residues, according to the "mechanistic" definition used for this quantity (see Eq. (3) in Materials and Methods). The horizontal arrows indicate the disordered regions which have been experimentally characterized in the intermediate (the loops involving positions 90–100 and 120–139).[14] Just as a mostly qualitative estimation, it is possible to find a rough correlation between the experimental and the computed ϕ values, with a slope of 0.9 ± 0.1, but with a large scattering of the individual data. The general trend of the calculated and experimental ϕ values is

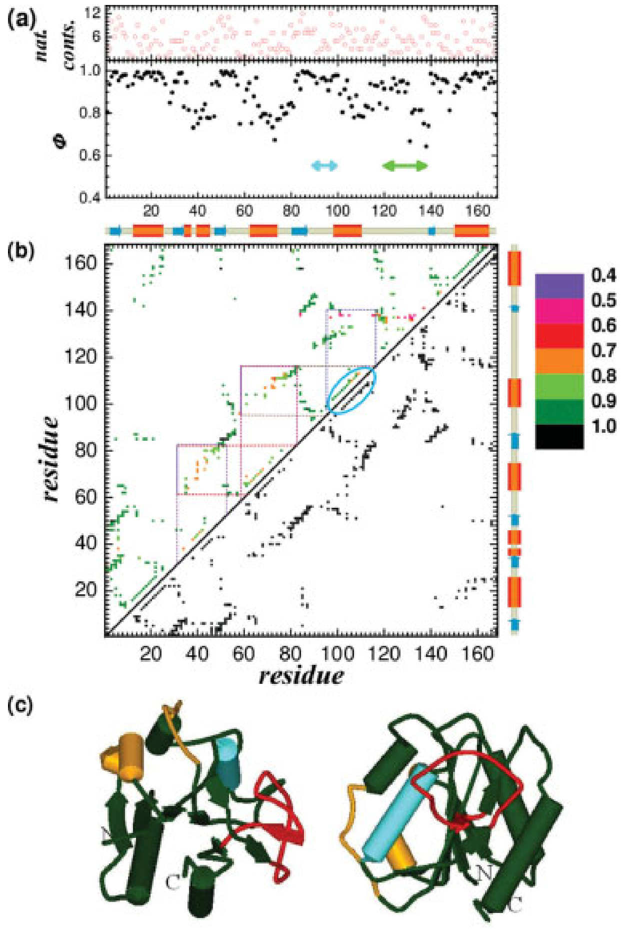

Figure 6

(**a**) ϕ values computed for the conformations detected as belonging to the intermediate at $T^* = 0.640$. The horizontal arrows show the regions which have been experimentally determined as more unstructured.[14] The red circles indicate the number of nonlocal native contacts in the PDB structure. (**b**) Frequency map for the native contacts in the conformations corresponding to the intermediate. The lower triangle shows the native contacts, for comparison. The color legend indicates the frequency of the individual contacts. (**c**) Two views of a cartoon representation of the native structure, colored according to the level of structural disorder in the intermediate, as detected by the simulation results. See text for details.

thus in a reasonable global agreement. Moreover, if we take the experimental ϕ values,[14] and classify them according to whether they show a residue as being in a native-like environment ($0.55 < \phi < 1$) or in a different one ($\phi \leq 0.55$ or $\phi > 1$), it is possible to find out a threshold value $\phi_t = 0.92$ in our numerical results so that the native–like or unfolded–like character is predicted in 82% of the experimentally investigated residues. These data suggest that the numerical ϕ values computed from the simulations convey a valuable coarse-grained information on the nature of the intermediate state. However, due to the different definitions of experimental and calculated ϕ values, mentioned in the previous section, we prefer to rely on the analysis of a frequency contact map, as shown in Figure 6(b), to get a more detailed information on the structure of the intermediate state. This is again a contact map for the protein. In this case, however, the upper triangle corresponds to a frequency map, that is, there is a colored scale which represents how often a given native contact appears in the conformations belonging to the intermediate state. Therefore, it provides a clear indication of the residual native structure present in a given set of conformations, in this case the snapshots characteristic of the intermediate state. Both in graphs (a) and (b), we can see that the intermediate shows a structure quite similar to the native state. This was expected, since we have already seen that the intermediate is as compact as the native state, and its structural deviations create rather modest values for the rmsd in the intermediate. Therefore, we find ϕ values above 0.8 for most of the residues, and contact frequencies above 0.8 for most of the native contacts. A few regions, however, show smaller values for these properties. The most important deviations from the native structure, with ϕ values close to 0.6 and contact frequencies of 0.5–0.6, appear at the loop which extends from positions 120–139 along the sequence, in agreement with the experimental results (green horizontal arrow in panel (a)).[14] The loss of native contacts for this loop implies also that the α-helix between residues 100 and 111 looses a fraction of its contacts, as indicated in the low values of the computed ϕ values for this region. The frequency contact map in Figure 6(b) shows, nevertheless, that this helix is formed with a high population in the intermediate (as shown by the $i - i + 4$ set of contacts parallel to the main diagonal and circled with a blue ellipse in the map), although it may have become slightly less well packed against the rest of the folded structure. This fact indicates that the ϕ values from simulation alone, as computed here and in other works, have to be considered with some caution, especially since they do not correspond to the definition of the experimental ϕ values, as it has been already mentioned in the Materials and Methods section. The frequency maps, however, may provide a better description of the structural features of the intermediate. On the other hand, these maps are not readily comparable with any experimental information available at this moment.

In other two regions of the sequence (involving positions 36–46 and 68–76), both the contact frequencies and the ϕ values are between 0.7 and 0.8. These two regions have not been reported as being unstructured in the experimental study of this protein intermediate.[14] Two residues mutated in these regions,[14] D43A and S71A, gave very small $\Delta\Delta G_{\text{N-D}}$ values and thus huge errors in the experimental ϕ results. Our simulations show a certain (though not severe) distortion in this region for the simulated intermediate state. The reason for this fact may be in the position of all these regions into the protein fold, which is indicated in Figure 6(c), where a schematic representation of the native structure is presented in two different views. The dark green regions are essentially the same in the intermediate and in the native structure. The most unstructured loop in our simulations, coincident with the experimental results, is shown in red. The affected helix which looses part of its contacts as a consequence is shown in cyan. The other two regions are shown in yellow. As it can be appreciated, they are also located in the protein surface, where the number of native contacts is less than in the protein core (something that can be somehow grasped from the number of native contacts at the top of Fig 6(a)). In addition, they form contacts with each other. In the contact frequency map, we have drawn a series of dotted lines indicating how the different elements which present partially disordered structures in the intermediate are related through contacts in the native state. Given the nature of the interaction potential used in this work, it is quite possible that a reduction in the number of attractive interactions for the helix circled in blue in the map has also affected the loop in positions 68–76, and this in turn has done the same with the broken helix in positions 36–46. The fact that all the contacts have the same weight in our simulation model, without any modulation due to the chemical nature of the amino acids involved in them, may be partly responsible of this spread in the disordering effects imposed by the loss of structure in the large loop shown in red in Figure 6(c), which is, according to our simulations, the structural element present in the intermediate whose structure more importantly deviates from the native one. For the shorter loop between residues 90–100, our model fails to predict any significant deviation from native in the structure of the intermediate state.

CONCLUSION

Proteins which show thermodynamic intermediates along their folding processes are common in large multi-domain proteins,[4] but only in a few cases they have been characterized for relatively small single-domain structures.[7] In this work, we have checked that the thermodynamic intermediate experimentally characterized for *Anabaena* apoflavodoxin[14] can be quite reasonably reproduced with a simple simulation model which only considers the contacts present in the native structure (a Gō-type or topology-based model). The model has not been tuned at all for the study of this particular protein. It uses the same characteristics and parameters we have found correct for the analysis of other proteins,[42,48] where no intermediates were found neither in the simulation, nor in the experiments.

The agreement between the results in this manuscript and experiment is not perfect in a few quantitative points, involving the specific population of the intermediate at the transition temperature and some minor discrepancies in the elements of the native structure that become unstructured in the intermediate. However, the model properly locates the long loop involving residues 120–139 as the portion of the protein whose structure is more severely distorted in the intermediate, while the fold is kept compact and with most of its hydrophobic cores practically unaffected. Although a more detailed study which takes into account the sequence of the protein may be desirable, the current model has made clear that, at least for this protein, the structure of the native state defines a folding process which includes the existence of the experimentally detected intermediate.

ACKNOWLEDGMENTS

M. L. and L. P. acknowledge a FPU Scholarship from Spanish Ministerio de Educación y Ciencia.

REFERENCES

1. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry and biology meet. Chem Rev 2006; 106:1559–1588.
2. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem. When will it be solved? Curr Opin Struct Biol 2007;17:342–346.
3. Jackson SE. How do small single-domain proteins fold? Proteins: structures and molecular properties. Fold Des 1998;3:R81–R91.
4. Creighton TE. Proteins: structures and molecular properties. New York: Freeman; 1993.
5. Garcia-Mira MM, Sadqi M, Fisher N, Sanchez-Ruiz JM, Muñoz V. Experimental identification of downhill protein folding. Science 2002;298:2191–2195.
6. Gruebele M. Downhill protein folding: evolution meets physics. CR Biol 2005;328:701–712.
7. Sancho J, Bueno M, Campos LA, Fernández-Recio J, Irún MP, López J, Machicado C, Pedroso I, Toja M. The 'relevant' stability of proteins with equilibrium intermediates. Scientific World J 2002; 2:1209–1215.
8. Irún MP, Garcia-Mira MM, Sanchez-Ruiz JM, Sancho J. Native hydrogen bonds in a molten globule: the apoflavodoxin thermal intermediate. J Mol Biol 2001;306:877–888.
9. Dobson CM. Protein folding and misfolding. Nature 2003;426:884–890.
10. Cremades N, Sancho J, Freire E. The native-state ensemble of proteins provides clues for folding, misfolding and function. Trends Biochem Sci 2006;31:494–496.
11. Sancho J. Flavodoxins: sequence, folding, binding, function and beyond. Cell Mol Life Sci 2006;63:855–864.
12. Genzor CG, Perales-Alcón A, Sancho J, Romero A. Closure of a tyrosine/tryptophan aromatic gate leads to a compact fold in apo flavodoxin. Nat Struct Biol 1996;3:329–332.
13. Fernández-Recio J, Genzor CG, Sancho J. Apoflavodoxin folding mechanism: an alpha/beta protein with an essentially off-pathway intermediate. Biochemistry 2001;40:15264–15245.
14. Campos LA, Bueno M, Lopez-Llano J, Jimenez MA, Sancho J. Structure of stable protein folding intermediates by equilibrium ϕ analysis: the apoflavodoxin thermal intermediate. J Mol Biol 2004; 344:239–255.
15. Campos LA, Garcia-Mira MM, Godoy-Ruiz R, Sanchez-Ruiz JM, Sancho J. Do proteins always benefit from a stability increase? Relevant and residual stabilisation in a three-state protein by charge optimisation. J Mol Biol 2004;344:223–237.
16. Steensma E, van Mierlo CPM. Structural characterization of apoflavodoxin shows that the location of the stable nucleus differes among proteins with a flavodoxin-like topology. J Mol Biol 1998;282:653–666.
17. Bollen YJM, Sánchez IE, van Mierlo CPM. Formation of on– and off–pathway intermediates in the folding kinetics of *Azotobacter vinelandii* apoflavodoxin. Biochemistry 2004;43:10475–10489.
18. Muralidhara BK, Wittung-Stafshede P. Thermal unfolding of apo and holo *Desulfovibrio desulfuricans* flavodoxin: cofactor stabilizes folded and intermediate states. Biochemistry 2004;43:12855–12864.
19. Cremades N, Bueno M, Neira JL, Velázquez-Campoy A, Sancho J. Conformational Stability of *Helicobacter pylori* flavodoxin. J Biol Chem 2008;283:2883–2895.
20. Cremades N, Sancho J. Molten globule and native state ensemble of *Helicobacter pylori* flavodoxin. Can crowding, osmolytes or cofactors stabilize the native conformation relative to the molten globule? Biophys J 2008;95:1913–1927.
21. López-Llano J, Maldonado S, Jain S, Lostao A, Godoy-Ruiz R, Sanchez-Ruiz JM, Cortijo M, Fernández-Recio J, Sancho J. The long and short flavodoxins. II. The role of the differentiating loop in apoflavodoxin stability and folding mechanism. J Biol Chem 2004;279:47184–47191.
22. Bollen YJM, van Mierlo CPM. Protein topology affects the appearance of intermediates during the folding of proteins with flavodoxin-like fold. Biophys Chem 2005;114:181–189.
23. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987; 84:7524–7528.
24. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
25. Gō N. Protein folding as a stochastic process. J Stat Phys 1983;30:413–423.
26. Taketomi H, Ueda Y, Gō N. Studies on protein folding, unfolding and fluctuations by computer simulation. Int J Pept Protein Res 1975;7:445–459.
27. Gō N, Taketomi H. Respective roles of short- and long-range interactions in protein folding. Proc Natl Acad Sci USA 1978;75:559–563.
28. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
29. Takada S, Wolynes PG. Microscopic theory of critical folding nuclei and reconfiguration activation barriers in folding proteins. J Chem Phys 1997;107:9585–9598.
30. Pande VS, Rokhsar DS. Is the molten globule a third phase of proteins? Proc Natl Acad Sci USA 1998;95:1490–1494.
31. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state

ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.

32. Clementi C, Jennings PA, Onuchic JN. How native state topology affects the folding of dihydrofolate reductase and interleukin-1β. Proc Natl Acad Sci USA 2000;97:5871–5876.
33. Hoang TX, Cieplak M. Sequencing of folding events in Gō-type proteins. J Chem Phys 2000;113:8319–8328.
34. Shimada J, Kussell EL, Shakhnovich EI. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. J Mol Biol 2001;308:79–95.
35. Erman B. Analysis of multiple folding routes of proteins by a coarse-grained dynamics model. Biophys J 2001;81:3534–3544.
36. Linhananta A, Zhou Y. The role of sidechain packing and native contact interactions in folding: discontinuous molecular dynamics folding simulations of an all-atom Gō model of fragment B of Staphylococcal protein A. J Chem Phys 2002;117:8983–8995.
37. Wang J, Wang W. Folding transition of model protein chains characterized by partition function zeros. J Chem Phys 2003;118:2952–2963.
38. Clementi C, García AE, Onuchic JN. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. J Mol Biol 2003;326:933–954.
39. Clementi C, Plotkin SS. The effects of nonnative interactions on protein folding rates: theory and simulation. Protein Sci 2004; 13:1750–1766.
40. Zuo G, Wang J, Wang W. Folding with downhill behavior and low cooperativity of proteins. Proteins 2006;63:165–173.
41. Prieto L, Rey A. Influence of the chain stiffness on the thermodynamics of a Gō-type model for protein folding. J Chem Phys 2007;126:165103-1–8.
42. Prieto L, Rey A. Influence of the native topology on the folding barrier for small proteins. J Chem Phys 2007;127:175101-1–11.
43. Sulkowska JI, Cieplak M. Selection of optimal variants of Gō-like models of proteins through studies of stretching. Biophys J 2008;95:3174–3191.
44. Gunasekaran K, Eyles SJ, Hagler AT, Gierasch LM. Keeping it in the family: folding studies of related proteins. Curr Opin Struct Biol 2001;11:83–93.
45. Zarrin-Afsar A, Larson SM, Davidson AR. The family feud: do proteins with similar structures fold via the same pathway? Curr Opin Struct Biol 2005;15:42–49.
46. Prieto L, de Sancho D, Rey A. Thermodynamics of Gō-type models for protein folding. J Chem Phys 2005;123:154903-1–8.
47. Prieto L, Rey A. Simulations of the protein folding process using topology-based models depend on the experimental structure. J Chem Phys 2008;129:115101-1–7.
48. Rey-Stolle MF, Enciso M, Rey A. Topology-based models and NMR structures in protein folding simulations. J Comput Chem 2009;30:1212–1219.
49. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
50. Leenders R, van Gunsteren WF, Berendsen HJC, Visser AJWG. Molecular dynamic simulations of oxidized and reduced *Clostridium beijrinckii* flavodoxin. Biophys J 1994;66:634–645.
51. Stagg L, Zhang SQ, Cheung MS, Wittung-Stafshede P. Molecular crowding enhances native structure and stability of α/β protein flavodoxin. Proc Natl Acad Sci USA 2007;104:18976–18981.
52. Martínez-Júlvez M, Cremades N, Bueno M, Pérez-Dorado I, Maya C, Cuesta-López S, Prada D, Falo F, Hermoso JA, Sancho J. Common conformational changes in flavodoxins induced by FMN and anion binding: The structure of Helicobacter pylori apoflavodoxin. Proteins 2007;69:581–594.
53. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 1997;281:140–150.
54. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 1999;314:141–151.
55. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. Phys Rev Lett 1988;61:2635–2638.
56. Ferrenberg AM, Swendsen RH. Optimized Monte Carlo data analysis. Phys Rev Lett 1989;63:1195–1198.
57. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 1992; 13:1011–1021.
58. Chodera JD, Swope WC, Pitera JW, Seok C, Dill KA. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. J Chem Theory Comput 2007; 3:26–41.
59. Fersht AR, Matouschek A, Serrano L. The folding of an enzime. I. Theory of protein engineering analysis of stability and pathway of protein folding. J Mol Biol 1992;224:771–782.
60. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature 2001;409:641–645.
61. Hubner IA, Shimada J, Shakhnovich EI. Commitment and nucleation in the protein G transition state. J Mol Biol 2004;336:745–761.
62. Allen LR, Paci E. Transition states for protein folding using molecular dynamics and experimental restraints. J Phys Condens Matter 2007;19:285211-1–15.
63. Kaya H, Chan HS. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? J Mol Biol 2003;326:911–931.
64. Liu Z, Chan HS. Solvation and desolvation effects in protein folding: native flexibility, kinetic cooperativity and enthalpic barriers under isostability conditions. Phys Biol 2005;2:S75–S85.
65. Lu D, Liu Z, Wu J. Structural transitions of confined model proteins: molecular dynamics simulation and experimental validation. Biophys J 2006;90:3224–3238.
66. Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. Proc Natl Acad Sci USA 2007;104: 12330–12335.
67. Finke JM, Onuchic JN. Equilibrium and kinetic folding pathways of a TIM barrel with a funneled energy landscape. Biophys J 2005; 89:488–505.
68. Patel B, Finke JM. Folding and unfolding of γTIM monomers and dimers. Biophys J 2007;93:2457–2471.