doi:10.1006/jmbi.2000.3693 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) **298**, 937–953

JMB

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

Cecilia Clementi*, Hugh Nymeyer and José Nelson Onuchic

Department of Physics
University of California at San Diego, La Jolla, CA 92093-0319, USA

**Corresponding author*

Recent experimental results suggest that the native fold, or topology, plays a primary role in determining the structure of the transition state ensemble, at least for small, fast-folding proteins. To investigate the extent of the topological control of the folding process, we studied the folding of simplified models of five small globular proteins constructed using a Gō-like potential to retain the information about the native structures but drastically reduce the energetic frustration and energetic heterogeneity among residue-residue native interactions. By comparing the structure of the transition state ensemble (experimentally determined by Φ-values) and of the intermediates with those obtained using our models, we show that these energetically unfrustrated models can reproduce the global experimentally known features of the transition state ensembles and "en-route" intermediates, at least for the analyzed proteins. This result clearly indicates that, as long as the protein sequence is sufficiently minimally frustrated, topology plays a central role in determining the folding mechanism.

Keywords: protein folding; transition state; folding intermediate; Φ-value analysis; molecular dynamics simulations

Introduction

Our understanding of the protein folding problem has been thoroughly changed by the new view that has emerged in the last decade. This new view, based on the energy landscape theory and funnel concept (Leopold *et al.*, 1992; Bryngelson *et al.*, 1995; Socci *et al.*, 1996; Onuchic *et al.*, 1997; Dill & Chan, 1997; Nymeyer *et al.*, 1998; Klimov & Thirumalai, 1996; Mirny *et al.*, 1996; Shea *et al.*, 1998), describes folding as the progressive evolution of an ensemble of partially folded structures through which the protein moves on its way to the native structure. The existence of a deep energy funnel in natural proteins and the relatively simple connectivity between most structurally close conformational states makes this description possible, even when only a few simple reaction coordinates that measure similarity to the native structure are used. The folding mechanism is controlled by both the shape of this free energy landscape and the roughness on it, which arises from the conflicts among interactions that stabilize the folded state and therefore can create non-native conformational traps (Bryngelson & Wolynes, 1987, 1989; Goldstein *et al.*, 1992).

The energetic roughness, however, is not the only limiting factor in determining a sequence's foldability. Even if the energetic roughness could be completely removed, the folding landscape would not be completely smooth. Theoretical (Wolynes, 1996; Nelson *et al.*, 1997; Nelson & Onuchi, 1998; Onuchic *et al.*, 1996; Socci *et al.*, 1997; Betancourt & Onuchic, 1995; Sheinerman & Brooks, 1998a; Micheletti *et al.*, 1999; Scheraga, 1992) and experimental (Grantcharova *et al.*, 1998; Martinez *et al.*, 1998) advances indicate that the final structure of the protein also plays a major role in determining a protein's foldability. Some particular folding motifs may be intrinsically more designable than others. To address this difference in foldability, which is not dependent on energetic frustration, we have introduced the concept of "topological frustration" (Nymeyer *et al.*, 2000; Onuchic *et al.*, 2000; Shea *et al.*, 1999).

E-mail address of the corresponding author:
cclementi@ucsd.edu

0022-2836/00/050937–17 $35.00/0

Let us imagine an ideal situation for which the order of native contact formation during folding is not biased. In this ''ideal'' situation, there are an enormously large number of equivalent folding pathways, and an analysis of the transition state ensemble would show that for this ensemble nearly all parts of the protein have a similar probability of participation. The structure in the transition ensemble has been estimated by analogy with minimalist lattice models made to reproduce the global landscape features of small, fast-folding proteins: similar Levinthal entropies, stabilities and energetic roughness, as gauged by the glass transition temperatures. These models show a transition state ensemble about halfway through the unfolded and folded states (Onuchic *et al*., 1995). In this ideal case, all the contacts in this transition ensemble would exist with the same probability.

Although the average amount of native formation in the transition ensemble is about 50 %, the lattice simulations show that, even when the sequence is designed to have substantially reduced energetic frustration, there are variations in the amount of nativeness of specific contacts in the transition state ensemble (Onuchic *et al*., 1996, 2000; Nymeyer *et al*., 2000). Real proteins display similar heterogeneity in contact formation. In systems with no energetic frustration and equal native interactions, these variations in the transition state ensemble are created solely by the folding motif and polymeric constraints that make certain contacts more geometrically accessible and stable than others. This variation in frequency that some contacts are made in the transition state ensemble generally reduces the entropy of the transition state and, when determined by the native motif, is a gauge of the amount of ''topological frustration'' in the system. Although this type of frustration can be modified by some design tricks (Plotkin & Onuchic, 2000), it cannot be completely eliminated: it reflects an intrinsic difficulty in folding to a particularly chosen shape. Minimalist models have shown how this heterogeneity leads to a transition ensemble that is a collection of diffuse nuclei with various levels of native contact participation (Onuchic *et al*., 1996). The minimalist models calibrated to real proteins show similar overall levels of contact heterogeneity as real proteins (Onuchic *et al*., 1996). This picture of a transition state composed of several diffuse nuclei has been confirmed by other lattice and off-lattice studies (Klimov & Thirumalai, 1998; Pande & Rokhsar, 1999). In addition to selecting sequences which have low levels of energetic frustration, evolution appears to have selected for a particular set of folding motifs which have reduced levels of ''topological frustration'', discarding other structures into which it is too difficult to fold (Betancourt & Onuchic, 1995; Wolynes *et al*., 1996; Nelson & Onuchic, 1998; Micheletti *et al*., 1999; Debe *et al*., 1999).

Guided by theoretical folding studies on lattice, off-lattice, and all-atom simulations (e.g. see Onuchic *et al*., 1995, 1996, 2000; Boczko & Brooks, 1995; Nymeyer *et al*., 2000; Shea *et al*., 1999), as well as recent experimental evidence (Grantcharova *et al*., 1998; Martinez *et al*., 1998; Chiti *et al*., 1999; Martinez & Serrano, 1999; Riddle *et al*., 1999), we suggest that real proteins, and especially small, fast-folding (sub-millisecond), two-state like proteins, have sequences with a sufficiently reduced level of energetic frustration that the experimentally observed ''structural polarization'' of the transition state ensemble (viz. the variation in the amount of local native structure) is primarily determined by topological constraints. That is, in well-designed sequences, the variations are more determined by the type of native fold than by differences in sequence, which leave the native fold relatively unchanged and the energetic frustration small.

The amount of native local structure in the transition state can be experimentally measured by using single and double point mutants as probes in the Φ-value technique (Fersht, 1994). If topology is a dominant source of heterogeneity in the transition state structure, then the majority of evolved sequences which fold to the same motif would exhibit similar local structure in the transition state ensemble. We provide evidence here that not only is this the case that much of the transition state ensemble is determined by the final folded form, but also for larger proteins that are not two-state folders, some ''en-route'' intermediates are determined by topological effects as well. Thus, it appears that the dominance of topology in folding extends even into some larger, slower-folding proteins with intermediates. This fact is consistent with some recent observations by Plaxco *et al*. (1998) that reveal a substantial correlation between the average sequence separation between contacting residues in the native structure and the folding rates for single domain proteins (Chan, 1998).

To ascertain the extent of topological control of the folding behavior, we created several simplified energetic models of small, globular proteins, using potentials created to minimize energetic frustration. We show that these energetically unfrustrated models reproduce nearly all the known global features of the transition states of the real proteins on whose native structures they are based, including the structure of folding intermediates. We directly compare the structure of the transition state ensemble experimentally determined by Φ-value measurements with the numerically determined value. The simulated transition state ensemble is inferred from structures sampled in equilibrium around the free energy barrier between the folded and unfolded states. This free energy is computed as a function of a single reaction coordinate that measures the fraction of formed native contacts. The validity of this method has been demonstrated by Onuchic *et al*. (1999) and Nymeyer *et al*. (2000).

Firstly we present in some detail the physical concepts underlying this work in the light of recent experimental results. Next we present results for a sample of five small, globular proteins, and compare these results against the available experimental data. The off-lattice model used in our study is presented in the Appendix. In order to investigate the relevance of the topology, we chose a model which reproduces the topological features of a given real protein and eliminates most of the energetic frustration and variations in the strength of native residue-residue contacts. The predicted transition state for these proteins is in good agreement with experimental evidence, supporting our hypothesis of the major role played by topology.

Checking the Folding Mechanism by Analyzing the Transition State Ensemble

How do we know what the folding transition state ensemble looks like? Experimental analysis of folding transition state ensembles has been largely performed using the Φ-value analysis technique introduced by Fersht and co-workers (Fersht, 1994). Φ-Values measure the effects that a mutation at a given position along the chain has on the folding rate and stability:

$$\Phi \equiv \frac{-RT\ln(k_{\mathrm{mut}}/k_{\mathrm{wt}})}{\Delta\Delta G^0} \qquad (1)$$

where k_{mut} and k_{wt} are the mutant and wild-type folding rates, respectively, R is the ideal gas constant, T is the absolute temperature, and $\Delta\Delta G^0$ is the difference in the total stability between the mutant and wild-type proteins in kcal/mol.

Because the folding event of small fast-folding proteins is well described as a diffusive process over a barrier determined by the free energy profile, the folding rate can be written as a Kramer-like equation (Socci *et al.*, 1996):

$$k = k_0 \exp(-\Delta G^{\ddagger}/RT) \qquad (2)$$

where k_0 is a factor depending on the barrier shape and the configurational diffusion coefficient of the system. If k_0 is insensitive to small sequence changes, as appears to be true for reasonably unfrustrated sequences (Onuchic *et al.*, 1996, 2000; Socci *et al.*, 1996; Nymeyer *et al.*, 2000; Shea *et al.*, 1999; Scalley & Baker, 1997; Munoz & Eaton, 1999), the Φ-value is then seen to be a ratio of free energy changes of the folding barrier to stability:

$$\Phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G^0} \qquad (3)$$

where $\Delta\Delta G^{\ddagger}$ is given by:

$$\Delta\Delta G^{\ddagger} = \Delta G^{\ddagger}_{\mathrm{mut}} - \Delta G^{\ddagger}_{\mathrm{wt}} = -RT\ln k_{\mathrm{mut}}/k_{\mathrm{wt}} \qquad (4)$$

When this relationship is valid and the mutation can be considered as a small perturbation, the Φ-value is a convenient measure of the fraction of native structure formed in the transition state ensemble around the site of the mutation. A Φ-value close to 1 means that the free energy change between the mutant and the wild-type is almost the same in the transition state and native state, indicating that native contacts involving the mutated residue are already formed at the transition state. Conversely, a Φ-value close to 0 means that the free energy change is the same in the transition state and unfolded states, so the local environment of the residue is probably unfolded-like. A detailed analysis of the mutation is needed to determine exactly which contacts are disrupted under mutation. Ideally, mutations are made which eliminate small hydrophobic side-groups. Studies using Φ-values with multiple same-site mutations generally support the accuracy of the Φ-value as a structural measurement of the transition ensemble (Matouschek *et al.*, 1995), although sizable changes in the transition state structure have been induced in at least one protein through a single point mutation (Burton *et al.*, 1997). In interpreting Φ-values, it is also important to remember that they measure only the relative change in structure, not the absolute amount of structure. This leads to the possibility that some mutants with low Φ-values may have nearly native local environments in the transition state, a possibility seen clearly in the experimental studies of procarboxypeptidase A2 (Villegas *et al.*, 1998).

The validity of Φ-values as structural measurements clearly supports the Kramer-like description of the folding rate and the fact that the Φ-value can be properly understood as a ratio of the free energy change of the transition ensemble over the change of the native ensemble (equation (3)). This equation is very convenient as a starting point for computing Φ-values. In several recent simulation papers for lattice and off-lattice protein models, we have investigated this issue at length (Nymeyer *et al.*, 2000; Onuchic *et al.*, 2000; Shea *et al.*, 1999). All these studies concluded that as long as the systems present a weak or moderate level of energetic frustration (such as the Gō-like models in this work), Φ-values determined from changes in the free energy barrier, in turn determined using a single simple reaction coordinate, yield quantitatively correct Φ-values. Therefore, all the calculations performed here were done utilizing equation (3); no actual kinetics were performed, only appropriate sampling of the protein configurational space (see Appendix and Socci & Onuchic (1995), Boczko & Brooks (1995), Onuchic *et al.* (1996) and Nymeyer *et al.* (2000), for example, for details). Technically, as long as the folding barriers are of a few k_BT or more and the displacement of the barrier position along this reaction coordinate under mutation is sufficiently small, the Φ-values can be computed using free energy perturbation:

$$\Phi = \frac{\Delta\Delta G^{\mathrm{TS}} - \Delta\Delta G^{\mathrm{U}}}{\Delta\Delta G^{\mathrm{F}} - \Delta\Delta G^{\mathrm{U}}} = \frac{\ln(e^{\Delta E/RT})_{TS} - \ln(e^{\Delta E/RT})_{U}}{\ln(e^{\Delta E/RT})_{F} - \ln(e^{\Delta E/RT})_{U}} \quad (5)$$

We used equation (5) to compute Φ-values for our protein models using fixed transition (TS), unfolded (U), and folded (F) regions identified by the free energy profile viewed using a single order parameter, Q, the fraction of native contacts formed in a given conformation.

What experimental evidence exists as to the role of topology in determining the average structure in the folding transition state ensemble? The clearest evidence to date of the role of topology comes from comparisons of the transition state structure of two homologues of the SH3 domain (src SH3 and α-spectrin SH3). These two homologs have only weak identity (≈30% identity with gaps), but the Φ-values at corresponding sequence positions are highly correlated (Grantcharova *et al.*, 1998; Martinez *et al.*, 1998), supporting the degeneracy in the folding behavior for these two sequences. Furthermore, one of these sequences has a strained Φ-Ψ conformation in the high Φ region of the distal turn. The fact that this strain does not detectably lower the Φ-values in the local neighborhood (Martinez *et al.*, 1998), suggests that the sequence details and local stability are less important for determining how structured a region is in the transition state ensemble than its location in the final folded conformation. Other evidence indicates that these results may be more generally applicable than simply for SH3 or β-sheet proteins. Sequence conservation has been shown not to correlate with Φ-values (Kim *et al.*, 1998), indicating that, in general, sequence changes at a given position in a protein weakly affect the Φ-value at that position.

Results for some small, fast-folding proteins (such as CI2 and the λ-repressor) suggest that the transition state is an expanded version of the native state, with a certain degree of additional heterogeneity over the structure (Itzhaki *et al.*, 1995; Burton *et al.*, 1997) (similar to the theoretical predications for small α-helical proteins (Onuchic *et al.*, 1995; Boczko & Brooks, 1995)), while results for other proteins (such as the β-sheet SH3 domain) show apparently greater structural heterogeneity in the transition state (Sheinerman & Brooks, 1998a,b). This difference in the degree of ''structural polarization'' emerging between small α-helix and β-sheet proteins suggests that the folding mechanism of a given protein is fundamentally tied to the type of secondary structural elements and their native arrangement. Current studies using the Φ-value technique have been made of src SH3 (Grantcharova *et al.*, 1998), α-spectrin SH3 (Martinez *et al.*, 1998), CI2 (Itzhaki *et al.*, 1995), barnase (Fersht *et al.*, 1992), barstar (Killick *et al.*, 1999), λ-repressor (Burton *et al.*, 1997), CheY (Lopez-Hernandez & Serrano, 1996), protein L (Kim *et al.*, 1998), procarboxypeptidase A2 (Villegas *et al.*, 1998), RNase H (Raschke *et al.*, 1999) and the tetrameric protein domain from tumor suppressor p53 (Mateu *et al.*, 1999).

Here, we analyze five proteins (SH3, CI2, barnase, RNase H and CheY) that have been extensively studied experimentally and for which, therefore, details of their transition state ensemble are quite well known. We generate sequences (and potentials) for simulating these different globular proteins. These sequences have the native backbone folds of real experimentally studied globular proteins but sequence and potential interactions designed to reduce drastically the energetic frustration and heterogeneity in residue-residue interactions. By comparing the transition state structures of these unfrustrated models with the experimental studies of their real protein cousins, we quantify the effects of the native topology. If topology completely determines how folding occurs, then the model and real proteins should have identical folding behavior and Φ-values. If energetic frustration and heterogeneity are critical for determining the folding mechanism, then the variations in Φ-values with position should bear greatly reduced similarity to those in the real proteins on which the computer homologues are based.

Two of the five studied proteins are simple two-state-like fast-folding proteins (SH3 and CI2), while the other three (barnase, RNase H and CheY) are known to fold through the formation of an intermediate state. We show not only that our simple models can reproduce most of the Φ-value structure, but also that models for barnase, RNase H and CheY correctly reproduce the folding intermediates of these proteins, suggesting that many of the ''en-route'' intermediates are also largely determined by the type of native fold.

We represent the five globular proteins using a simplified C^{α} model with a Gō-like (Ueda *et al.*, 1975) Hamiltonian as detailed in the Appendix. This potential is in its details unlike that of real proteins, which have residue-residue interactions with many components (Coulomb interactions, hydrogen bonding, solvent-mediated interactions, etc.). The crucial features of this potential are its low level of energetic frustration that characterizes good folders, and a native conformation equal to the real protein. The ability of this model to reproduce features of the real transition state ensemble and real folding intermediates is a strong indication that the retention of the topology is enough to determine the global features of their folding mechanism. Using these models, we simulate the dynamics of a protein starting from its native structure for several temperatures. To monitor the thermodynamics of the system, we group the configurations obtained during a simulation as a function of the reaction coordinate, Q, defined as the fraction of the native contacts formed in a conformation ($Q = 0$

at the fully unfolded state and $Q = 1$ at the folded state). The choice of Q as order parameter for the folding is motivated by the fact that in a funnel-like energy landscape, a well-designed sequence has the energy of its conformations reasonably correlated to degree of nativeness, and the parameter Q is a good measure of the degree of similarity with the native structure. Our Gō-like potential is minimally frustrated for the chosen native structure, and the prediction of transition state ensemble structures and folding rates for these Gō-like systems has been shown to be quite accurate (Socci *et al*., 1996; Shea *et al*., 1999; Nymeyer *et al*., 2000). From the free energy profile as a function of Q, it is easy to locate the unfolded, folded and transition state ensembles, as is shown below. Since these models consider totally unfrustrated sequences, they may not reproduce the precise energetics of the real proteins, such as the value of the barrier heights and the stability of the intermediates; nonetheless, they are able to determine the general structure of these ensembles.

In order to compare the folding process simulated using our model to the actual process for a given protein (as obtained from experimental Φ-values analysis), we need to choose a "mutation" protocol to compute Φ-values. Experimentally, the ideal mutation is typically one that removes a small hydrophobic side-group such as a methyl group that makes well-defined and identifiable residue-residue contacts in the native state. The Φ-value is then sensitive to this known contact. Our computational mutation is the removal of a single native bond, so our computer Φ-values are sensitive to the fractional formation of this bond Q_{ij} between residues i and j. We make these mutations because, as in most real mutations, they are sensitive to the formation of specific contacts, rather than being averages over interactions with many parts of the native structure. They mostly resemble the interaction Φ_{int}-value made by making double cycle mutants (Fersht *et al*., 1992). The Φ-values are computed from equation (5). In an ideal, perfectly smooth funnel-like energy landscape, all the Φ-values should be equal; in an energetically unfrustrated situation, Φ-values variations are due to the structure of the native conformation.

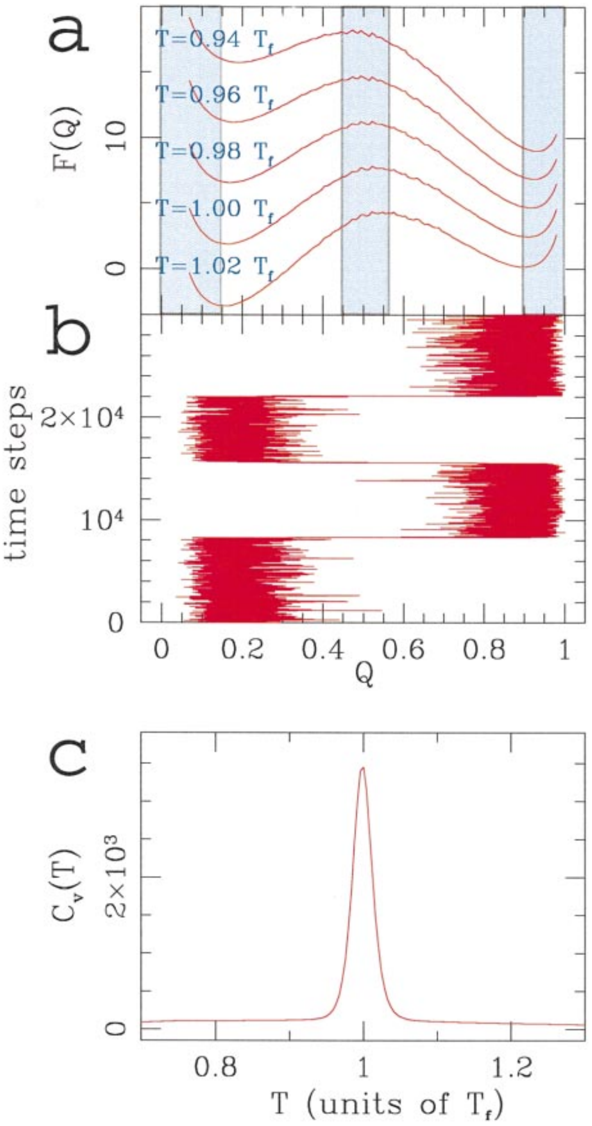

Figure 1. (a) Free energy $F(Q)$ as a function of the reaction coordinate Q around the folding temperature for the model of CI2. Free energies are measured in units of k_BT_f, where T_f is the folding temperature. The unfolded, folded and transition state regions are shown in the light blue shaded areas. (b) A typical sample simulation at a temperature around the folding temperature. The reaction coordinate Q as a function of time (measured in arbitrary units of molecular dynamics steps) is shown. The two-state behaviour is apparent from the data. The unfolded and folded states are equally populated at the folding temperature. (c) Heat capacity as a function of the temperature (units of folding temperature).

Determining the Transition State Ensemble of Small Globular Proteins

We have discussed the idea of "topological frustration" and its role in determining the structural heterogeneity of the transition state ensemble. We explore its role directly by creating protein models which drastically reduce the energetic frustration and energetic heterogeneity among residue-residue native interactions, leaving topology as the primary source of the residual frustration. Results obtained with these models, constructed using a C^{α} level of resolution with a Gō-like potential designed to fold to the native trace of chosen proteins, are then compared with the experimental data of those proteins. Five proteins with different folding motifs and different amounts of transition state heterogeneity (variation in Φ-values) and/or intermediates have been investigated.

We first analyze chymotrypsin inhibitor II (CI2), a mixed α-β protein with a broad distribution of Φ-values (nearly uniform from 0 to 1). Then we present an analysis for the src SH3 domain, a largely β-sheet protein with a more polarized transition state structure (a substantial number of large Φ-values). We then apply the same technique to barnase, RNase H and CheY, three other mixed α-β proteins which fold *via* a folding intermediate. Although these proteins are not two-state folding proteins, we demonstrate that topology is also the dominant determinant of their folding behavior. We show that the topology plays a major role, not only in the transition state ensemble, but it is also largely responsible for the existence and general structure of the folding intermediate. This result may be quite common for "en-route" folding intermediates and could provide a computational method for distinguishing between "on-pathway" and "off-pathway" structures which are inferred from experiments. To check the applicability of this method, the same approach presented here has been extended elsewhere (Clementi *et al*., 2000) to a pair of larger proteins (dihydrofolate reductase and interleukin-1β). Even for these very large proteins we found that the overall structure of the transition state and intermediate ensembles experimentally observed can be obtained utilizing similar simplified models.

Analysis of two-state folders: CI2 and SH3

CI2

The chymotrypsin inhibitor 2 (CI2) protein is a 64-residue protein, consisting of six β-sheets packed against an α-helix to form a hydrophobic core. Experimental studies (Jackson & Fersht, 1991a,b; Jackson *et al*., 1993a,b) have established that CI2 folding and unfolding can be modeled by simple two-state kinetics. The structure of the transition state for this protein has been extensively characterized by protein engineering (Itzhaki *et al*., 1995; Otzen & Fersht, 1995; Jackson & Fersht, 1991b), by free energy functional approaches (Shoemaker & Wolynes, 1999; Shoemaker *et al*., 1999), by a geometrical variational principle (Micheletti *et al*., 1999), and by all-atom molecular dynamics simulations (Li & Daggett, 1996; Kazmirski *et al*., 1999; Lazaridis & Karplus, 1997). These studies have shown the transition state has roughly half of the native interactions formed in the transition state ensemble and a broad distribution of Φ-values, in agreement with the general predictions of the energy landscape theory used with a law of corresponding states for small proteins (Onuchic *et al*., 1995, 1996). The broad distribution of Φ-values suggests that most hydrophobic contacts are represented at a level of about 50 % in the transition state ensemble.

We constructed a Gō-like C^{α} model of CI2 as described in the Appendix. Several fixed temperature simulations were made and combined using the WHAM algorithm (Swendsen, 1993) to generate a specific heat *versus* temperature profile and a plot of the potential of mean force as a function of

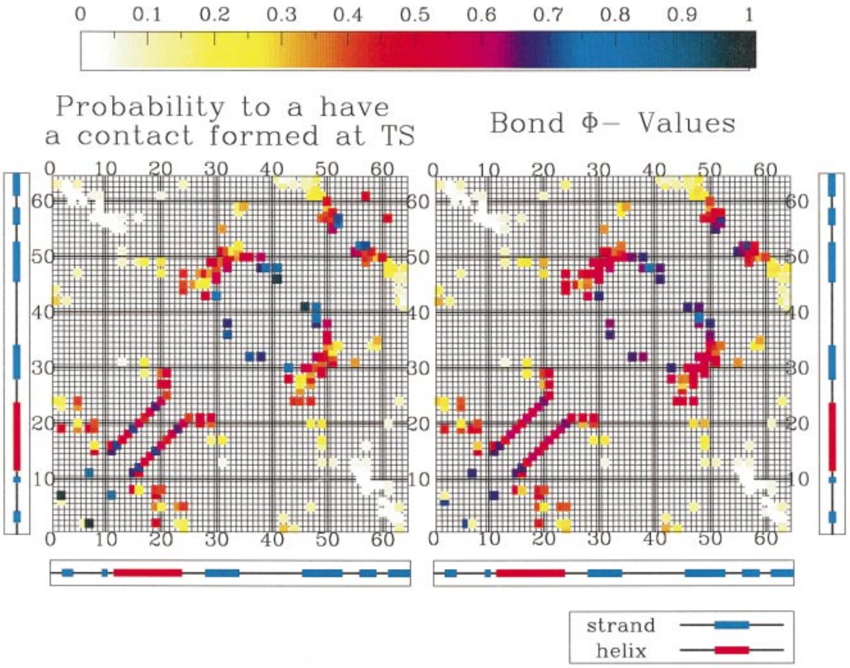

Figure 2. The results for the transition state structure from the simulations for CI2. The probability of native contact formation at the transition state (left panel), and bond Φ-values (right panel) are shown. Different colors indicate different values from 0 to 1, as quantified by the color scale. The α-helix, the interactions between the strands 4 and 5, and the mini-core (i.e. interactions between residues 32, 38 and 50) are the parts of the structure formed with the highest probability, although they are not fully formed. Overall, the transition state ensemble appears as an expanded version of the native state where most contacts have a similar probability of participation, but some interactions are less likely to occur. These results agree with the transition state structure experimentally obtained.

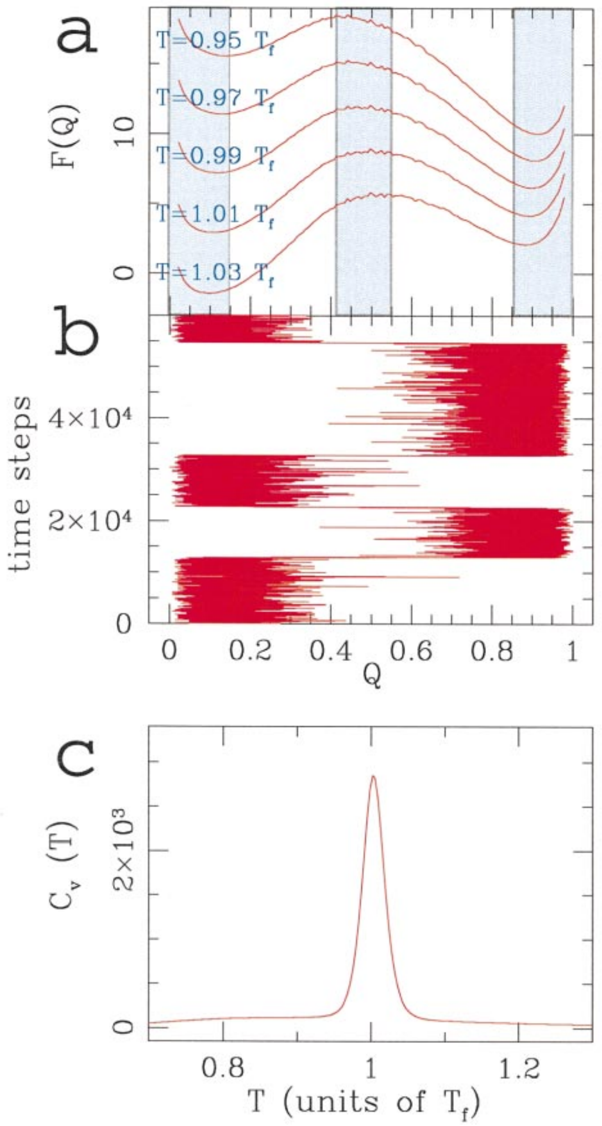

Figure 3. (a) Free energy $F(Q)$ as a function of the reaction coordinate Q for a set of temperatures around the folding temperature for the model of SH3. Free energies are measured in units of k_BT_f. The choices for the unfolded, folded and transition state regions are marked as shaded regions. (b) The reaction coordinate Q as a function of time (unit of molecular dynamics steps), from a typical sample simulation around the folding temperature. As in Figure 1, the two-state behaviour is apparent. At the transition temperature the model protein has an equal probability to be found in the unfolded or in the folded state. (c) Heat capacity as a function of the temperature, in units of folding temperature.

the folding order parameter Q (see Figure 1). From the free energy profile, we identified the dominant barrier, and used the thermal ensemble of states at its location to generate Φ-values from equation (5). The ranges of values of Q used to determine each of these ensembles is shaded in Figure 1. The mutations have been implemented by the removal of single attractive interactions (they are replaced with the same short-range repulsive interactions used between residues without native interactions). The values computed *via* this method are shown in Figure 2. Also shown in this Figure is the fractional formation of individual native contacts in the transition state. The small difference between these two Figures is primarily due to the fact that in the Φ calculations the native contact formations in the folded and unfolded states are also taken into account. Because of the higher concentration of contacts between residues nearby in sequence and the local conformational preferences, the unfolded state shows a high level of local structure. The inaccurate representation of local contacts in the unfolded state makes the short-range Φ-values less reliable as transition structure estimates than long-range Φ-values.

From the calculations, we detect three significant regions of large Φ-values: the α-helix, the mini-core defined by strands 3 and 4 and their connecting loop, and between the C terminus of strand 4 and the N terminus of strand 5. These regions generally have Φ-values in excess of 0.6. Slightly smaller values of about 0.5 exist for the short-range contacts between the N terminus of strand 3 and the C terminus of the α-helix and for contacts between strand 3 and strand 4. All other regions lack a consistent set of large Φ-values. Despite the large number of native contacts between strands 1 and 2 and the α-helix and between strands 5 and 6 and the α-helix, only low Φ-values are observed in this region (nearly all below 0.2). A comparison between these data and the exhaustive analysis by Fersht and colleagues (Otzen & Fersht, 1995) shows excellent overall agreement. They found that ''β-strands 1, 5 and 6 ... are not structured in the transition state....''. Strand 2 also shows a greatly reduced amount of structure. Furthermore, ''the central residues of β-strands 3 and 4 interact with the α-helix to form the major hydrophobic core of CI2.'' The hydrophobic mini-core in this region (defined as the cluster formed by side-chains of residues 32, 38, and 50) is detected by single mutant and double mutant Φ-values (Itzhaki *et al*., 1995) to be at least 30% formed in the transition ensemble. Similarly, they found the α-helix, particularly the N-capping region, to be highly ordered.

In summary, there is quite good overall agreement, except for a discrepancy in the short-range interactions in the loop region between strands 4 and 5. This protein shows generally higher Φ-values between interactions which are more local in sequence and lower Φ-values between interactions which are distant in sequence. The results are thus consistent with the picture of the transition state as a collection of non-specific and somewhat diffuse nuclei (Onuchic *et al*., 1995). This overall low level of frustration suggests a low level of ''topological frustration'' in this model as well and a particularly designable motif.

src SH3 domain

src SH3 is the 57-residue fragment of tyrosine-protein kinase that stretches from T84 to S140. It has five β-strands (and a short 3_{10} helix) in an antiparallel arrangement, forming a partial β sandwich. Experimental measurements have shown that the SH3 domain folds using a rapid, apparently two-state mechanism. A Φ-value analysis (Grantcharova *et al.*, 1998) reveals that the distal loop hairpin and diverging turn regions are both highly structured and docked together at the transition state; the hydrophobic interactions between the base of the hairpin and the strand following the diverging turn are partially formed, while other regions of src SH3 appear only weakly ordered in the transition state ensemble. The overall representation of the transition state structure of src SH3, having the distal loop and diverging turn largely formed and other regions weakly formed, agrees with studies of α-spectrin SH3, (Martinez *et al.*, 1998), which has a similar backbone structure but a dissimilar sequence (approx 30 % identity with gaps). This observed similarity along with evidence of a strained backbone conformation in the distal loop of the α-spectrin SH3 (Martinez *et al.*, 1998) supports the concept of ''topological'' dominance in folding (Grantcharova *et al.*, 1998).

Figure 3 shows the folding behavior as obtained from our dynamics simulations of the Gō-like analogous of the src SH3. The free energy barrier defining the transition state location is evident in the Figure. As above, we have computed Φ-values from equation (5) by mutating (removing) every native residue-residue attractive contact. The results of this calculation are shown in Figure 4. In addition to Φ-values, the contact formation probability at the transition state ensemble has been calculated. Our previous caveats concerning Φ-values for local interactions still apply. We observe the highest collection of off-diagonal (long-range) Φ-values is in the diverging turn, distal loop interaction exactly as seen from the experimental Φ-value measurements. We see very low values in the RT loop region, in accord with the two mutants in this loop. We also see medium-to-high values between the two β strands which are connected by the distal loop. The transition state structure of the SH3 presents a substantially larger degree of structural polarization than CI2, where the Φ-values are much more uniform. This suggests that SH3 has a backbone conformation which is intrinsically more difficult to fold, i.e. there is a greater level of ''topological frustration'' in this structure. Nevertheless the transition state composition is well reproduced for both the two proteins.

Analysis of three proteins which fold throughout the formation of an intermediate state: barnase, RNase H and CheY

Barnase, RNase H and CheY are three small αβ proteins (although larger than the previous two proteins): barnase is a 110 residue protein, composed by three α-helices (located in the first 45 residues) followed by five β-strands; RNase H consists of 155 residues which arrange themselves in five α-helices and five β-strands; CheY is 129 residues long, classic α/β parallel fold in which five β-strands are surrounded by five α-helices. Exper-

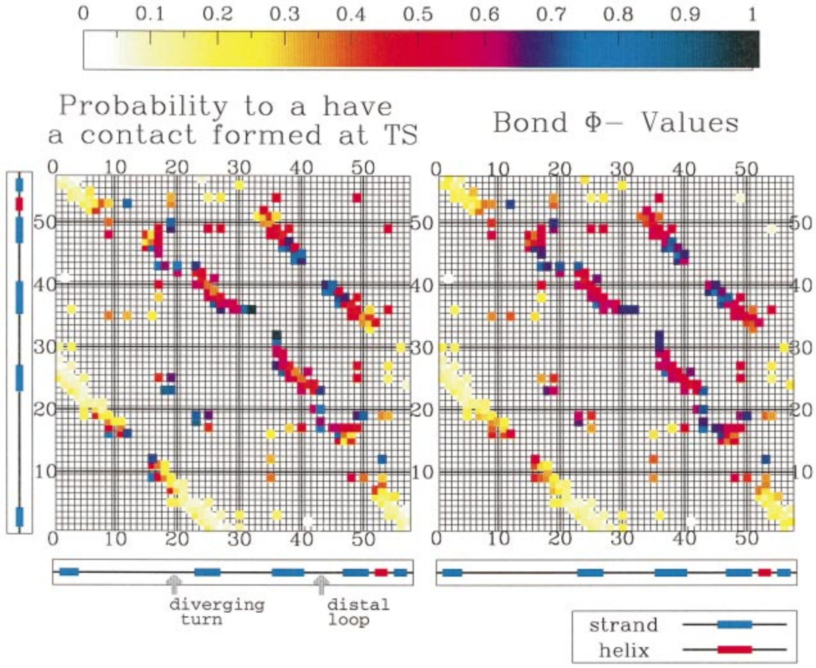

Figure 4. The transition state structure as obtained from the simulations for SH3. The panel on the left represents the probability for a native contact to be formed at the transition state, while the panel on the right shows the results for bond Φ-values. Different colors indicate different values from 0 to 1, as quantified by the color scale. The diverging turn and distal loop are marked on the contact map. The interactions within and between these two parts of the protein chain appear to be formed with high probability. The interactions between the two strands joined by the distal loop are partially formed, while the contacts involving the first 20 residues do not contribute to the transition state structure. This description of the transition state is in agreement with experimental results.

imental results show that these three proteins do not fold by following a simple two-state kinetics directly from the unfolded state to the native structure, but fold through the formation of a metastable intermediate which interconverts into the native state. This evokes an interesting question: is topology alone able to determine the presence of an intermediate in the folding process? In Figures 5, 7 and 9 we show evidence for the first time that such intermediates can be created solely from a Gō-like minimalist model which preserves the native topology. The presence of an intermediate during these protein's folding events is a requirement of the native protein motifs. The free energy changes upon mutations of a wild-type three-state protein are experimentally measured both for the intermediate and the transition state, to define two different sets of Φ-values for the protein:

$$\Phi_{\mathrm{I}} = \frac{\Delta\Delta G_{\mathrm{I}} - \Delta\Delta G_{\mathrm{U}}}{\Delta\Delta G_{\mathrm{F}} - \Delta\Delta G_{\mathrm{U}}}$$

$$\Phi_{\mathrm{TS}} = \frac{\Delta\Delta G_{\mathrm{TS}} - \Delta\Delta G_{\mathrm{U}}}{\Delta\Delta G_{\mathrm{F}} - \Delta\Delta G_{\mathrm{U}}} \qquad (6)$$

where Φ_{I} provides information about the structural composition of the intermediate state (I), and Φ_{TS} of the transition state (TS). Below we discuss in some details the results for the three proteins. Since, as for the first two proteins, the Φ-values and the native contact probabilities provide somewhat similar information, for simplicity, we show only the results obtained for the native contact probabilities (for safety we have checked the Φ-values and determined that similar information is recovered).

Barnase

The analysis of experimentally obtained Φ-values (Fersht *et al.*, 1992) for barnase shows that some relevant regions of the structure are fully unfolded in the intermediate while other regions are fully folded.

Figure 6 shows the intermediate and the transition state structure obtained from the Gō-like model. The intermediate shows substantial structural heterogeneity: there are very high probability values for interactions within the β-sheet region and its included loops, and very low values for interactions within the α-helices and their loops and between the α-helical and β-sheet regions. Some local short-range helical interactions are formed. The transition state ensemble structure shows the same structure as the intermediate with the addition of strong interactions within helices 2 and 3; between helix 2, helix 3, the first β-strand, and the intervening loops; and between the second β-strand and the second helix.

Comparing these simulation results with extensive mutagenesis studies of reference (Fersht *et al.*, 1992), we observe a good qualitative agreement. The β-sheet region is highly structured in the intermediate as it is the core region 3 (consisting of the packing of loop 3, that joins strands 1 and 2, and of loop 5, that joins strands 4 and 5, with the other side of the β-sheet). In agreement with experiments, the earliest formed part of the protein appears to be the β-sheet region. Also the core region 2 (formed by the hydrophobic residues from helix 2, helix 3, the first strand, and the first two loops) is found to be only weakly formed in the intermediate and the transition state.

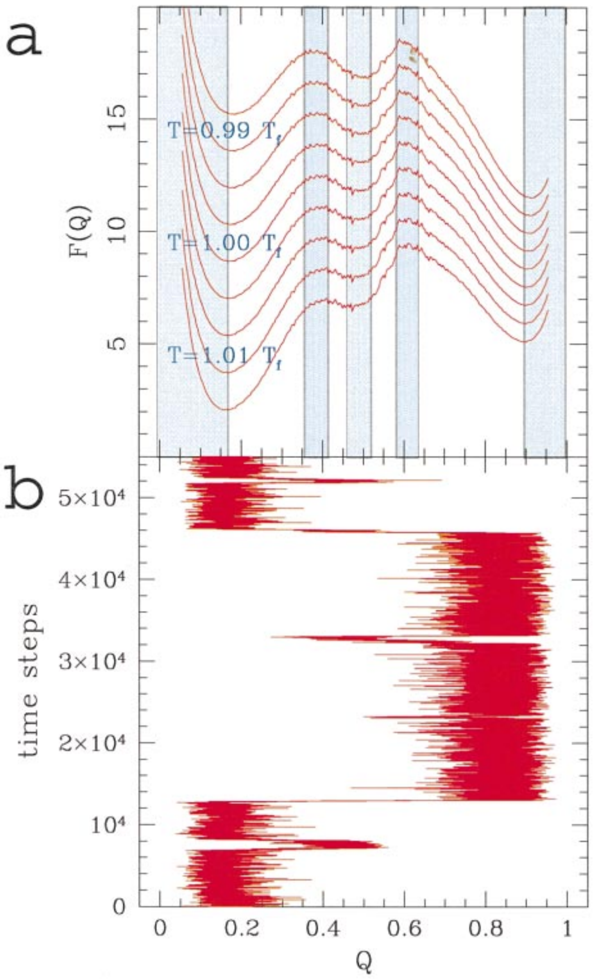

Figure 5. (a) Free energy $F(Q)$ of barnase protein as a function of the reaction coordinate Q around the folding temperature. Free energies are measured in units of $k_{\mathrm{B}}T_{\mathrm{f}}$. The unfolded, folded and intermediate state regions are green, while the tops of the two barriers are light blue. The local minimum in the free energy profile between the unfolded and folded minima locates the folding intermediate state. The presence of a folding intermediate state is also evident from (b), where the order parameter Q is plotted as a function of time for a typical molecular dynamics simulation around the folding temperature. In the interval $Q \in (0.4 - 0.5)$, the same state (i.e. with the same average structure) is visited both from the unfolded and folded structures.

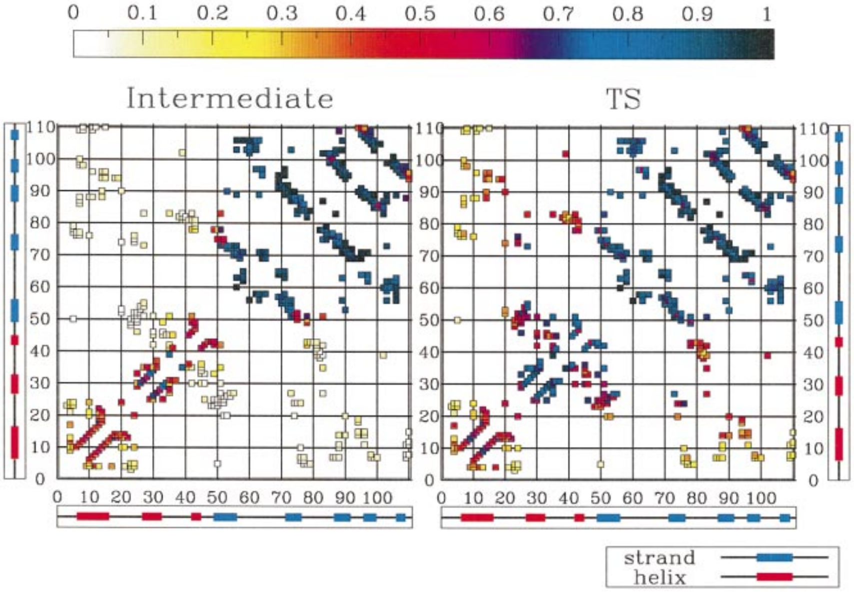

Figure 6. The probability of native contact formation for the intermediate (left panel) and transition state (right panel) structures as obtained from our simulations of barnase. Different colors indicate different values from 0 to 1, as quantified by the color scale. The earliest formed part of the protein appears to be the β-sheet region, in agreement with experimental results. The core 3 (formed by loops 3 and 5 to the β-sheet) is formed at the intermediate and transition state, while core 1 (the packing of helix 1 against the β-sheet) and core 2 (the interactions between the hydrophobic residues from helices 2 and 3, strand 1, and the first two loops) start to form only after the transition state. The formation of α-helix 1 occurs as a late event of the folding from our simulations, while from experimental results it seems to be already formed at the intermediate and transition state. The early formation of the α-helix is most probably due to energetic factors rather then from topology requirements (and then beyond the prediction possibility of this model), as detailed in the text.

There are two minor discrepancies between the barnase model and the experimental data. First, we slightly overestimate the formation of core region 2 in the transition state ensemble. Second, we underestimate the amount of structure in core region 1 (formed by the packing of the first helix against a side of the β-sheet) in both intermediate and transition ensemble. In particular, we under-represent the interaction between helix 1 and the β-sheet region. The experimentally observed early packing of helix 1 against the rest of the structure is not reproduced by our model. Clearly there are some important energetic factors which have been neglected by the simple model. These may be inferred from the barnase crystal structure. For example, one can see that helix 1 is largely solvent-exposed, with interactions between it and the remainder of the protein formed by only five of the 11 helix residues: 83% of the interactions reside on the hydrophobic residues Phe7, Ala11, Leu14 and Gln15, and 17% of the interactions are formed by the charged residues Asp8 and Asp12, while the solvent-exposed part of the helix is composed of polar residues. Large stabilizing interactions, other than tertiary (most hydrophobic) interactions, are neglected in the model, probably being responsible for the failure in predicting the formation of the structural parts involving helix 1. In this structural detail, it appears that the topological factors are not the leading determinant of the folding behavior.

Ribonuclease H

Kinetic studies of the wild-type RNase H have shown that an intermediate state is populated in the folding process, and the structure of this intermediate has been extensively investigated by circular dichroism, fluorescence and hydrogen exchange methods (Dabora & Marqusee, 1994; Yamasaki *et al.*, 1995; Dabora *et al.*, 1996; Chamberlain *et al.*, 1996; Raschke & Marqusee, 1997) and by protein engineering (Raschke *et al.*, 1999). Figure 7 shows that, consistent with the experimental evidence, we find an intermediate state in the folding process of the RNase H model. Experimental results indicate that the most stable region of the protein intermediate involves α-helix 1, strand 4, α-helix 4 and α-helix 2. Hydrogen exchange experiments have shown that α-helix 1 is the region of the protein most protected from exchange, suggesting that most of the interactions involving α-helix 1 are already significantly formed at the intermediate state of the folding process. Helix 4 and β-strand 4 are the next most protected regions, while α-helix 5 has a low-to-moderate level of protection. After the completion of the intermediate structure, the rate-limiting transition state involves the ordering of the β-sheet and α-helix 5. The packing of helix 5 across the sheet is found to be the latest folding event.

The results of the model for RNase H show good agreement with experimental evidence. As shown in Figure 8, the formation of contacts invol-

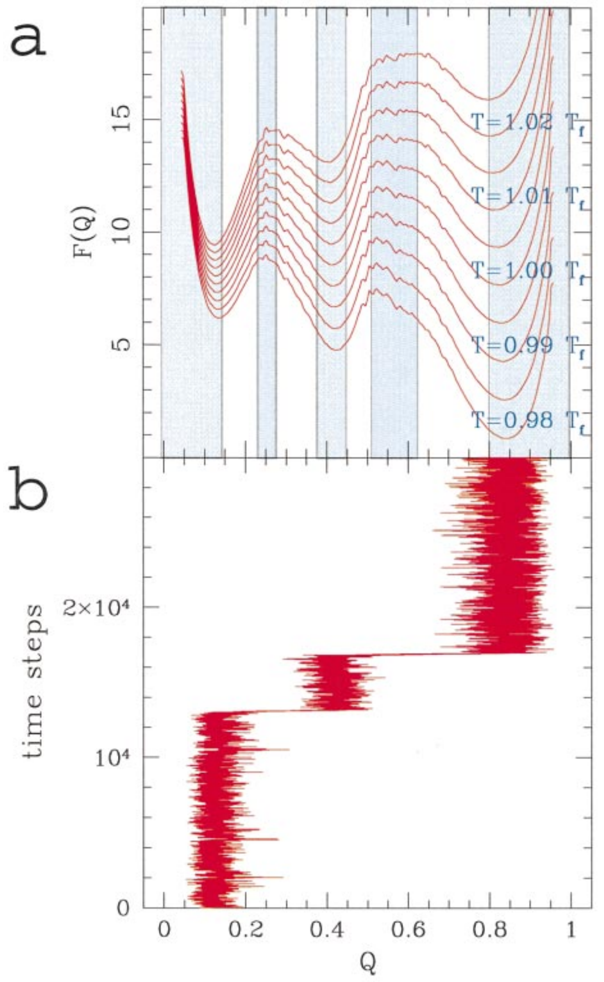

Figure 7. (a) Free energy $F(Q)$ of the model of RNase H as a function of the reaction coordinate Q around the folding temperature. Free energies are measured in units of k_BT_f. The regions corresponding to the unfolded, folded and intermediate state are green, while the top of the two barriers are light blue. A folding intermediate is detected as a local minimum in the free energy between the unfolded and folded minima. (b) The fraction of native contacts formed, Q, is plotted *versus* the simulation time for a sample of our simulations (at a temperature $T = 0.99T_f$) where the transition from unfolded to folded state is observed. The local minimum of (a) corresponds to a transiently populated intermediate (located at Q around 0.4) that later evolves to the fully folded state.

ving helix 1 is the earliest event in the folding process. Contacts arising from α-helix 4 and the β-strand 4 are then formed at the intermediate state and consolidated at the transition state. In agreement with the experimental results, we find that, at the transition state, interactions between α-helix 1, strand 4 and the rest of the protein are mostly formed; α-helix 4 is also well structured and interactions between helix 4 and the other parts of the protein are partly formed. Interactions among the strands are almost all formed, but the sheet is not yet docked to helix 5.

CheY

Utilizing protein engineering (Lopez-Hernandez & Serrano, 1996; López-Hernándex *et al*., 1997), the transition state of CheY has been characterized and can be described as a combination of two subdomains: the first half of the protein (subdomain 1), comprising the α-helices 1 and 2 and the β-strands 1-3, is substantially folded, whereas the second half (subdomain 2) is completely disorganized. Helix 1 seems to play the role of a nucleation site around which subdomain 1 begins to form. Moreover, an intermediate has been detected at the early stage of the folding process where all the five α-helices are rather structured. The last two helices, however, are very unstructured in the later occurring transition state. For this reason it has been suggested that a misfolded species is ‘‘visited’’ at the beginning of the folding process.

Our simple model detects two possible intermediates for this protein, one of them is an ‘‘en-route’’ intermediate that is short-lived and occurs just before the transition state ensemble (Q around 0.6 in Figure 9). Surprisingly, the unfrustrated model is also able to detect a ‘‘misfolded’’ trap in the folding of CheY. Since non-native interactions are not allowed in the model, this trap is a long-living partially folded state created by the topological constrains. There is no direct connection between this trap state and the fully folded state. The structure of this trap is shown in Figure 10 and it agrees with the experimental observation of all helices well structured. In contrast to the previously discussed proteins, the model of CheY seems to have a tendency to first form a ‘‘wrong’’ part of the protein and, when this happens, a partial unfolding must occur before the folding can be completed.

Finally, analyzing the transition state structure, we find good agreement with experimental data. As shown in Figure 10, the first part of the protein (subdomain 1) is almost fully folded at the transition state ensemble, while subdomain 2 is completely unfolded.

Conclusions

Recent theoretical studies and experimental results suggest that the folding mechanism for small, fast-folding proteins is strongly determined by the native state topology. The amount of energetic frustration, arising from the residual conflict among the amino acid interactions, appears largely reduced for these proteins so that topological constraints are important factors in governing the folding process. To explore this topological influence in real proteins, we analyzed the folding process of the Gō-like analogous of five real proteins. Since we have used Gō-like potentials, the energetic frustration is effectively removed from the system, while the native fold topology is taken into

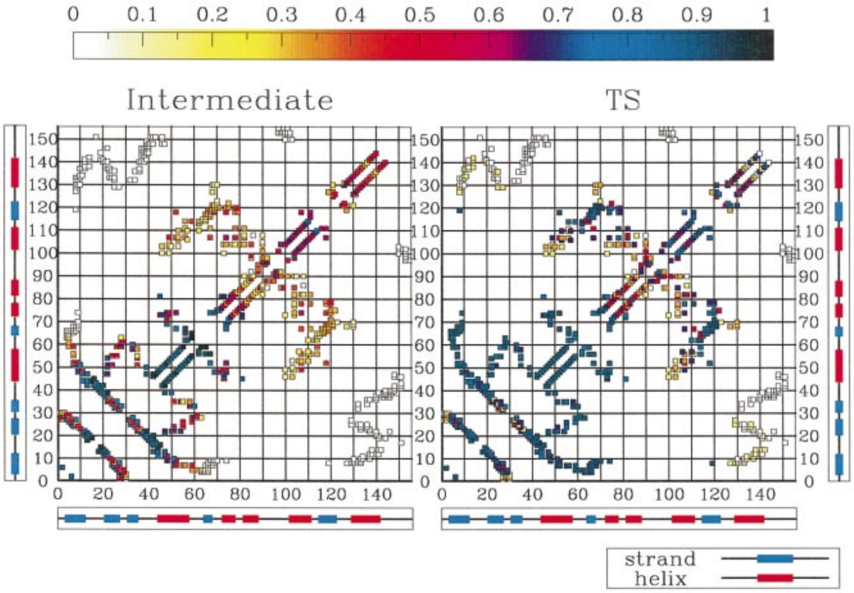

Figure 8. The probability of native contact formation at the intermediate (left panel) and transition state (right panel) structure, as observed for the RNase H model. Different colors indicate different values from 0 to 1, as quantified by the color scale. In agreement with experimental results, we found that interactions involving α-helix 1 are the first formed in the folding process. Contacts between α-helix 1 and strand 4 are highly probably formed at the intermediate. Also, α-helix 4 is well structured and the β-sheet is partly formed. These interactions strengthen at the transition state where also the β-sheet is almost completely formed, while the packing of helix 5 across the sheet is not yet accomplished.

account. It is important to highlight the fact that the results from such studies exhibit the overall topological features of the folding mechanism, although we do not expect the precise energetic values for barrier heights and intermediate state stabilities. For example, real proteins are not necessarily totally unfrustrated and they have only to minimize energetic frustration to a sufficiently reduced level in order to be good folders. Also, as long as energetic frustration is small enough, creating some heterogeneity at the native interactions may help to reduce topological frustration (Plotkin & Onuchic, 2000), and that will energetically favor some contacts over others.

The effective use of a small number of global order parameters as reaction coordinates in interpreting real data or studying more detailed protein folding models, depends critically on the degree of frustration present in real proteins (Nymeyer *et al.*, 2000). Since our results show that general structural features of the transition state ensemble in real proteins, at least for this class of fast-folding proteins, are reproducible by using a substantially unfrustrated potential, several different global order parameters should work to explain the folding mechanism. For this reason, it should not be a surprise that, utilizing energy landscape ideas and the funnel concept, some very simple models with approximate order parameters determined by a single or a few sequence approximations (Alm & Baker, 1999; Munoz & Eaton, 1999; Galzitskaya & Finkelstein, 1999) have been successful in predicting qualitative features of the transition state ensemble.

Again, we compared in details the structure of the transition state ensemble of the five proteins resulting from our simulations with experimental data. The agreement between our results and the experimental data supports the idea that energetic frustration is indeed sufficiently reduced and the protein folding mechanism, at least for small globular proteins, is strongly dependent on topological effects. The structure of the transition state ensemble of the CI2 presents a broad distribution of Φ-values, i.e. a reduced degree of structural polarization, in agreement with predictions based on the energy landscape theory (see Onuchic *et al.*, 1995, 1996). On the other hand, the structure of the SH3 transition state ensemble shows a higher degree of polarization. Nevertheless, by using our simplified Gō-like model, we reproduced the transition state composition for both proteins, demonstrating that topology is largely responsible for the observed experimental differences. The last three proteins we analyzed (barnase, RNase H and CheY) are known to fold through a three-state kinetics, involving the formation of an intermediate structure. Our Gō-like model of these proteins also folds with three-state kinetics, with intermediates that are analogous to those detected experimentally. This fact suggests that topology is also a dominant factor in determining the ‘‘en-route’’ intermediates.

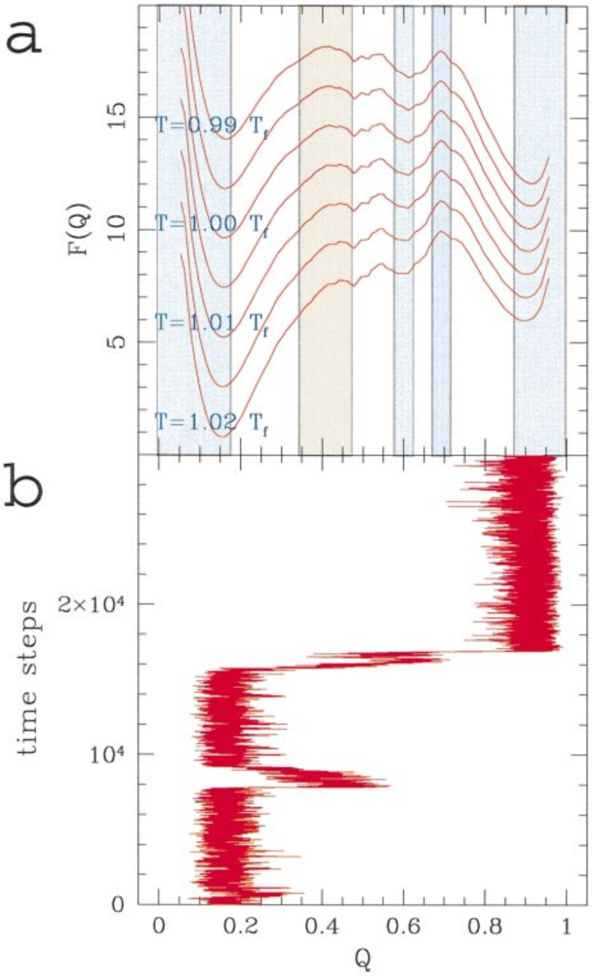

Figure 9. (a) Free energy $F(Q)$ profile for the model of CheY plotted as a function of the reaction coordinate Q for a set of temperatures around the folding temperature. Free energies are measured in units of k_BT_f. Unlike the corresponding Figures for barnase (Figure 5) and RNase H (Figure 7), two different structures are populated between the folded and unfolded states. In addition to the "en-route" intermediate state (green as the regions corresponding to the folded and unfolded states), a "misfolded" intermediate structure (marked in brown at Q around 0.4) is transiently visited from the unfolded state. The tops of the two barriers are light blue. In agreement with experimental results, we found that in this "misfolded" structure, all the five α-helices are rather structured while, in the later occurring "en-route" intermediate and transition state ensemble, helices 4 and 5 are completely unstructured (see Figure 10). (b) Typical sample of the simulation around the folding temperature, in a region where the folding occurs. The first transiently populated intermediate state corresponds to a structure where all the helices are formed. Before proceeding to the folded state, a partial unfolding occurs.

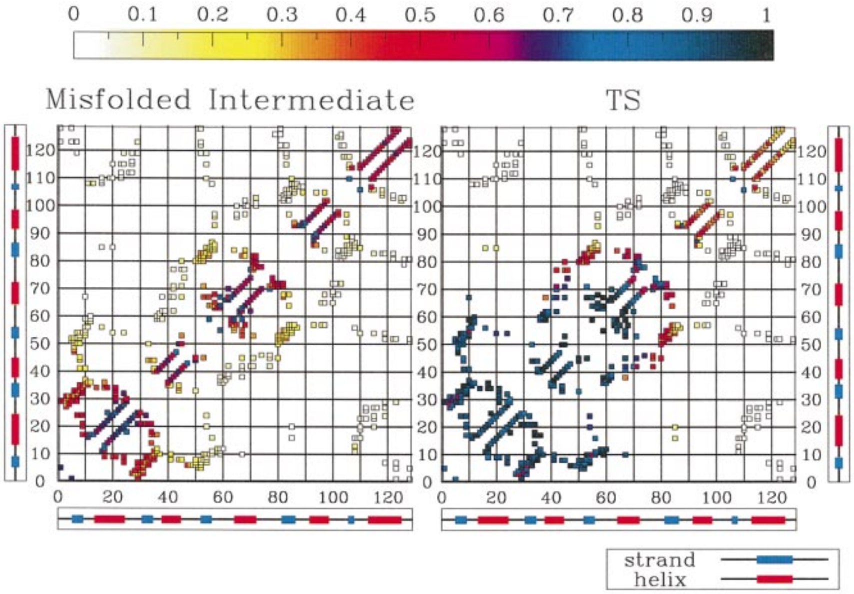

Figure 10. The probability of the native CheY contacts being formed in the "misfolded" intermediate (left panel) and transition state (right panel) for the model protein. Different colors indicate different values from 0 to 1, as quantified by the color scale. In agreement with experimental data, all the helices are mostly formed in the transiently populated "misfolded" structure, while helices 4 and 5 are rather unstructured at the transition state. The two subdomains experimentally detected in the CheY transition state (Lopez-Hernandez & Serrano, 1996; López-Hernández *et al.*, 1997) are evident in the Figure: the first part of the protein (all interactions arising from α-helices 1 and 2 and the β-strands 1-3) is folded, while the second part (interactions among α-helices 4 and 5 and the β-strands 4 and 5) is completely unfolded. Helix 3 is structured but the interactions between helix 3 and the rest of the protein are not completely formed.

Acknowledgments

We thank Viara Grantcharova and David Baker for information about the SH3 structure. We also thank Vladimir Sobolev for the CSU software. We are indebted to Angel García, Peter Wolynes, Steve Plotkin, Jorge Chahine, Joan Shea, Margaret Cheung, Charlie Brooks, Amos Maritan and Jayanth Banavar for helpful discussions. C.C. expresses her gratitude to Giovanni Fossati for his suggestions and for carefully reading the manuscript, and to the Center for Astrophysics and Space Sciences of UCSD for the usage of graphics facilities and computer time. This work has been supported by the NSF (grant 96-03839), by the La Jolla Interfaces in Science program (sponsored by the Burroughs Wellcome Fund) and by the Molecular Biophysics training grant program (NIH T32 GN08326).

References

Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, **96**, 11305-11310.

Betancourt, M. R. & Onuchic, J. N. (1995). Kinetics of proteinlike models: the energy landscape factors that determine folding. *J. Chem. Phys.* **103**, 773-787.

Boczko, E. M. & Brooks, C. L., III (1995). First-principles calculation of the folding free energy of a three-helix bundle protein. *Science*, **269**, 393-396.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways and the energy landscape of protein folding. *Proteins: Struct. Funct. Genet.* **21**, 167-195.

Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 7524-7528.

Bryngelson, J. D. & Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **93**, 6902-6915.

Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. (1997). The energy landscape of a fast-folding protein mapped by ala-gly substitutions. *Nature Struct. Biol.* **4**, 305-310.

Chamberlain, A. K., Handel, T. M. & Marqusee, S. (1996). Detection of rare partially folded molecules in equilibrium with the native conformation of RNase H. *Nature Struct. Biol.* **3**, 782-787.

Chan, H. S. (1998). Matching speed and locality. *Nature*, **392**, 761-763.

Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005-1009.

Clementi, C., Jennings, P. A. & Onuchic, J. N. (2000). How native state topology affects the folding of dihydrofolate reductase and and interleukin-1β. *Proc. Natl Acad. Sci. USA*, in the press.

Dabora, J. M. & Marqusee, S. (1994). Equilibrium unfolding of escherichia coli ribonuclease h: characterization of a partially folded state. *Protein Sci.* **3**, 1401-1408.

Dabora, J. M., Pelton, J. G. & Marqusee, S. (1996). Structure of the acid state of *Escherichia coli* ribonuclease hi. *Biochemistry*, **35**, 11951-11958.

Debe, D. A., Carlson, M. J. & Goddard, W. A. (1999). The topomer-sampling model of protein folding. *Proc. Natl Acad. Sci. USA*, **96**, 2596-2601.

Dill, K. A. & Chan, H. S. (1997). From levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10-19.

Fersht, A. R. (1994). Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79-84.

Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. i. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771-782.

Galzitskaya, O. V. & Finkelstein, A. V. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 11299-11304.

Goldstein, R., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992). Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl Acad. Sci. USA*, **89**, 9029-9033.

Grantcharova, V., Riddle, D., Santiago, J. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714-720.

Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.

Jackson, S. E. & Fersht, A. R. (1991a). Folding of chymotrypsin inhibitor 2. 2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition state of folding. *Biochemistry*, **30**, 10436-10443.

Jackson, S. E. & Fersht, A. R. (1991b). Folding of chymotrypsin inhibitor 2. 1. evidence for a two-state transition. *Biochemistry*, **30**, 10428-10435.

Jackson, S. E., elMasry, N. & Fersht, A. R. (1993a). Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry*, **32**, 11270-11278.

Jackson, S. E., Moracci, M., elMasry, N., Johnson, C. & Fersht, A. R. (1993b). The effect of cavity creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. *Biochemistry*, **32**, 11262-11269.

Kazmirski, S. L., Li, A. & Daggett, V. (1999). Analysis methods for comparison of multiple molecular dynamics trajectories: applications to protein unfolding pathways and denatured ensembles. *J. Mol. Biol.* **290**, 283-304.

Killick, T. R., Freund, S. M. V. & Fersht, A. R. (1999). Real-time NMR studies on a transient folding intermediate of barstar. *Protein Sci.* **8**, 1286-1291.

Kim, D., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA*, **95**, 4982-4986.

Kim, D., Yi, Q., Gladwin, S., Goldberg, J. & Baker, D. (1998). The single helix in protein l is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* **284**, 807-815.

Klimov, D. K. & Thirumalai, D. (1996). Factors governing the foldability of proteins. *Proteins: Struct. Funct. Genet.* **26**, 411-441.

Klimov, D. K. & Thirumalai, D. (1998). Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.* **282**, 471-492.

Lazaridis, T. & Karplus, M. (1997). New view of protein folding reconciled with the old through multiple unfolding simulations. *Science,* **278**, 1928-1931.

Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: kinetic pathways through compact conformational space. *Proc. Natl Acad. Sci. USA,* **89**, 8721-8725.

Li, A. & Daggett, V. (1996). Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* **257**, 412-429.

Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* **1**, 43-55.

López-Hernández, E., Cronet, P., Serrano, L. & Muñoz, V. (1997). Folding kinetics of CheY mutants with enhanced natice α-helix propensities. *J. Mol. Biol.* **266**, 610-620.

Martinez, J. C. & Serrano, L. (1999). The folding transition state between sh3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010-1016.

Martinez, J., Pisabarro, M. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721-729.

Mateu, M. G., Del, Pino M. S. & Fersht, A. R. (1999). Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53. *Nature Struct. Biol.* **6**, 191-198.

Matouschek, A., Otzen, D. K., Itzhaki, L. S., Jackson, S. E. & Fersht, A. R. (1995). Movement of the position of the transition state in protein folding. *Biochemistry,* **34**, 13656-13662.

Micheletti, C., Banavar, J., Maritan, A. & Seno, F. (1999). Protein structures and optimal folding emerging from a geometrical variational principle. *Phys. Rev. Letters,* **82**, 3372-3375.

Mirny, L. A., Abkevich, V. & Shakhnovich, E. I. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Fold. Des.* **1**, 103-116.

Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **96**, 11311-11316.

Nelson, E. D. & Onuchic, J. N. (1998). Proposed mechanism for stability of proteins to evolutionary mutations. *Proc. Natl Acad. Sci. USA,* **95**, 10682-10686.

Nelson, E. D., Eyck, L. T. & Onuchic, J. N. (1997). Symmetry and kinetic optimization of proteinlike heteropolymers. *Phys. Rev. Letters,* **79**, 3534-3537.

Nymeyer, H., García, A. E. & Onuchic, J. N. (1998). Folding funnels and frustration in off-lattice minimalist models. *Proc. Natl Acad. Sci. USA,* **95**, 5921-5928.

Nymeyer, H., Socci, N. D. & Onuchic, J. N. (2000). Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration. *Proc. Natl Acad. Sci. USA,* **97**, 634-639.

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

Onuchic, J. N., Nymeyer, H., García, A. E., Chahine, J. & Socci, N. D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advan. Protein Chem.* **53**, 87-152.

Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. A. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441-450.

Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. A. & Socci, N. D. (1995). Towards an outline of the topography of a realistic protein folding funnel. *Proc. Natl Acad. Sci. USA,* **92**, 3626-3630.

Otzen, D. & Fersht, A. (1995). Side-chain determinants of β-sheet stability. *Biochemistry,* **34**, 5718-5724.

Pande, V. & Rokhsar, D. (1999). Folding pathway of a lattice model for proteins. *Proc. Natl Acad. Sci. USA,* **96**, 1273-1278.

Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.

Plotkin, S. S. & Onuchic, J. N. (2000). Investigation of routes and funnels in protein folding by free energy functional methods. *Proc. Natl Acad. Sci. USA,* In the press.

Raschke, T. M. & Marqusee, S. (1997). The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions. *Nature Struct. Biol.* **4**, 298-304.

Raschke, T. M., Kho, J. & Marqusee, S. (1999). Confirmation of the hierarchical folding of RNase H: a protein engineering study. *Nature Struct. Biol.* **6**, 825-831.

Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in sh3 folding. *Nature Struct. Biol.* **6**, 1016-1024.

Scalley, M. L. & Baker, D. (1997). Protein folding kinetics exhibit an arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc. Natl Acad. Sci. USA,* **44**, 10636-10640.

Scheraga, H. A. (1992). Contribution of physical chemistry to an understanding of protein structure and function. *Protein Sci.* **1**, 691.

Shea, J. E., Nochomovitz, Y. D., Guo, Z. Y. & Brooks, C. L., III (1998). Exploring the space of protein folding hamiltonians: the balance of forces in a minimalist beta-barrel model. *J. Chem. Phys.* **109**, 2895-2903.

Shea, J. E., Onuchic, J. N. & Brooks, C. L., III (1999). Exploring the origins of topological frustration: design of a minimally frustrated model of fragment b of protein a. *Proc. Natl Acad. Sci. USA,* **96**, 12512-12517.

Sheinerman, F. B. & Brooks, C. L., III (1998a). Calculations on folding of segment b1 of streptococcal protein g. *J. Mol. Biol.* **278**, 439-455.

Sheinerman, F. B. & Brooks, C. L., III (1998b). Molecular picture of folding of a small alpha/beta protein. *Proc. Natl Acad. Sci. USA,* **95**, 1562-1567.

Shoemaker, B. A. & Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. *J. Mol. Biol.* **287**, 657-674.

Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble. *J. Mol. Biol.* **287**, 675-694.

Socci, N. D. & Onuchic, J. N. (1995). Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. *J. Chem. Phys.* **103**, 4732-4744.

Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860-5868.

Socci, N. D., Nymeyer, H. & Onuchic, J. N. (1997). Exploring the protein folding landscape. *Physica D,* **107**, 366-382.

Swendsen, R. H. (1993). Modern methods of analyzing monte carlo computer simulations. *Physica A,* **194**, 53-62.

Ueda, Y., Taketomi, H. & Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Res.* **7**, 445-459.

Villegas, V., Martinez, J., Aviles, F. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.

Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc. Natl Acad. Sci. USA,* **93**, 14249-14255.

Wolynes, P. G., Schulten, Z. L. & Onuchic, J. N. (1996). Fast-folding experiments and the topography of protein folding energy landscapes. *Chem. Biol.* **3**, 425-432.

Yamasaki, K., Ogasahara, K., Yutani, K., Oobatake, M. & Kanaya, S. (1995). Folding pathway of *Escherichia coli* ribonuclease HI: a circular dichroism, fluorescence, and NMR study. *Biochemistry,* **34**, 16552-16562.

Appendix

Model and Method

In order to investigate how the native state topology affects the folding of a given protein we follow the dynamics of the protein by using a Gō-like Hamiltonian (Ueda *et al.*, 1975) to describe the energy of the protein in a given configuration. A Gō-like Hamiltonian takes into account only native interactions, and each of these interactions enters into the energy balance with the same weighting. It means that the system gains energy as much as any amino acid residue pair involved in a native contact is close to its native configuration, no matter how strong the actual interaction is in the real protein. Residues in a given protein are represented as single beads centered in their C^α positions. Adjacent beads are strung together into a polymer chain by means of bond and angle interactions, while the geometry of the native state is encoded in the dihedral angle potential and a non-local potential. The energy of a configuration Γ of a protein having the configuration Γ_0 as its native state is thus given by the expression:

$$\begin{aligned} E(\Gamma,\ \Gamma_0) = {} & \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 \\ & + \sum_{\text{dihedral}} K_\phi^{(n)}[1+\cos(n\times(\phi-\phi_0))] \\ & + \sum_{i<j-3}\left\{\varepsilon(i,j)\left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right]\right. \\ & \left. + \varepsilon_2(i,j)\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}\right\} \end{aligned} \tag{A1}$$

In equation (A1) r and r_0 represent the distances between two subsequent residues at, respectively, the configuration Γ and the native state Γ_0. Analogously, θ (θ_0) and ϕ (ϕ_0) represent the angles formed by three subsequent residues and the dihedral angle defined by four subsequent residues along the chain at the configuration Γ (Γ_0). The dihedral potential consists of a sum of two terms for every four adjacent C^α atoms, one with period $n = 1$ and one with $n = 3$. The last term in equation (A1) contains the non-local native interactions and a short-range repulsive term for non-native pairs (i.e. $\varepsilon(i,j) = \text{constant} > 0$ and $\varepsilon_2(i,j) = 0$ if $i - j$ is a native pair while $\varepsilon(i,j) = 0$ and $\varepsilon_2(i,j) = \text{constant} > 0$ if $i - j$ is a non-native pair). The parameter σ_{ij} is taken equal to the $i - j$ distance at the native state for native interactions, while $\sigma_{ij} = 4$ Å for non-native (i.e. repulsive) interactions. Parameters K_r, K_θ, K_ϕ, ε weight the relative strength of each kind of interaction entering in the energy and they are taken to be $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\phi^{(1)} = \varepsilon$ and $K_\phi^{(3)} = 0.5\varepsilon$. With this choice of parameters we found that the stabilizing energy residing in the tertiary contacts is approximately twice the stabilizing energy residing in the torsional degrees of freedom. This balance among the energy terms is optimal to study the folding of our Gō-like protein models. The native contact map of a protein is derived with the CSU software based upon the approach developed by Sobolev *et al.* (1996). Native contacts between pairs of residues (i,j) with $j \leqslant i + 3$ are discarded from the native map as any three or four subsequent residues are already interacting in the angle and dihedral terms. A contact between two residues (i,j) is considered formed if the distance between the C^α atoms is shorter than γ times their native distance σ_{ij}. It has been shown (Onuchic *et al.*, 2000) that the results are not strongly dependent on the choice made for the cutoff distance, γ. Here we used $\gamma = 1.2$. We have used molecular dynamics (entailing the numerical integration of Newton's laws of motion) for simulating the kinetics of the protein models. We employed the simulation package AMBER (version 4.1) at constant temperature, i.e. using the Berendsen algorithm for coupling the system to an external bath (Berendsen *et al.*, 1984). Both temperature and energy are measured in units of the folding temperature T_f in the simulations.

For each protein model, several constant temperature simulations were made and combined using the WHAM algorithm (Ferrenberg & Swendsen, 1988; Ferrenberg & Swendsen, 1989; Swendsen, 1993) to generate a specific heat profile *versus* temperature and a free energy $F(Q)$ as a function of the folding reaction coordinate Q. This algorithm is based on the fact that the logarithm of probability distribution $P(Q)$ of the values taken by a certain variable Q (e.g. the order parameter) at fixed temperature T may serve as an estimate for the free energy profile $F(Q)$ at that temperature. In fact, the probability to have a certain value Q_1 for the variable Q, at temperature $T = 1/k_B\beta$, in the canonical ensemble is given by:

$$P_\beta(Q_1) = \frac{W(Q_1)e^{-\beta E(Q_1)}}{Z_\beta} \qquad (A2)$$

where $W(Q)$ is the density of configurations at a point Q in the configurational space, Z_β is the canonical partition function at temperature $T = 1/k_B\beta$ and $E(Q)$ is the energy of the system at the value Q of the reaction coordinate†. Since the free energy F is:

$$F(Q) = E(Q) - TS(Q) \qquad (A3)$$

and the entropy $S(Q)$ is related to the configurational density $W(Q)$:

$$W(Q) \sim e^{S(Q)/k_B} \qquad (A4)$$

where k_B is the Boltzmann constant, it follows that:

$$\frac{P_\beta(Q_1)}{P_\beta(Q_2)} = \frac{e^{-\beta F(Q_1)}}{e^{-\beta F(Q_2)}} \qquad (A5)$$

and free energy differences can be computed by:

$$-\beta(F(Q_1) - F(Q_2)) = \log\frac{P_\beta(Q_1)}{P_\beta(Q_2)} \qquad (A6)$$

By using the procedure described by Ferrenberg & Swendsen (1988, 1989) and Swendsen (1993), data from a finite set of simulations can be used to obtain complete thermodynamic information over a large parameter region.

Probability distributions are obtained by sampling the configurational space during molecular dynamics simulations.

For the smaller proteins (CI2 and SH3) we determined the errors on the estimates of the transition temperature and contact probabilities (or Φ-values). This has been accomplished by computing these quantities from several (more than ten) uncorrelated sets of simulations. We found that the standard deviation for each single contact probability is 0.06 for CI2 and 0.05 for SH3, while the transition temperature is determined in both cases with an uncertainty smaller than 0.5%. These errors are obtained using about 200 uncorrelated conformations in the transition state ensemble. Since barnase, RNase H and CheY have twice to three times the number of tertiary contacts of SH3 and CI2, in order to have appropriate statistics, we sampled about 500 uncorrelated conformations (thermally weighted) for every transition state ensemble or intermediate.

References

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690.

Ferrenberg, A. M. & Swendsen, R. H. (1988). New Monte Carlo technique for studying phase transitions. *Phys. Rev. Letters*, **61**, 2635-2638.

Ferrenberg, A. M. & Swendsen, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Letters*, **63**, 1195-1198.

Onuchic, J. N., Nymeyer, H., García, A. E., Chahine, J. & Socci, N. D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advan. Protein Chem.* **53**, 87-152

Sobolev, V., Wade, R., Vriend, G. & Edelman, M. (1996). Molecular docking using surface complementarity. *Proteins: Struct. Funct. Genet.* **25**, 120-129.

Swendsen, R. H. (1993). Modern methods of analyzing Monte Carlo computer simulations. *Physica A*, **194**, 53-62.

Ueda, Y., Taketomi, H. & Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific interunit interactions. *Int. J. Pept. Res.* **7**, 445-459.

Edited by F. E. Cohen

(Received 30 August 1999; received in revised form 13 March 2000; accepted 13 March 2000)

---

† Since our model is almost energetically unfrustrated, the energy fluctuations for a set of configurations with fixed Q are strongly reduced such that the energy in a given configuration could be considered as a function of Q.