doi:10.1016/j.jmb.2007.08.032 J. Mol. Biol. (2007) 373, 1348–1360

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Parallel Folding Pathways in the SH3 Domain Protein

A. R. Lam[1]*, J. M. Borreguero[2], F. Ding[3], N. V. Dokholyan[3], S. V. Buldyrev[4], H. E. Stanley[1] and E. Shakhnovich[5]

[1]*Center for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215, USA*

[2]*Center for the Study of Systems Biology, Department of Biology, Georgia Institute of Technology, Atlanta, GA 30318, USA*

[3]*Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

[4]*Department of Physics, Yeshiva University, New York, NY 10033, USA*

[5]*Department of Chemistry, Harvard University, Cambridge, MA 02138, USA*

Received 25 December 2006; received in revised form 6 August 2007; accepted 14 August 2007
Available online 22 August 2007

Edited by K. Kuwajima

The transition-state ensemble (TSE) is the set of protein conformations with an equal probability to fold or unfold. Its characterization is crucial for an understanding of the folding process. We determined the TSE of the src-SH3 domain protein by using extensive molecular dynamics simulations of the Gō model and computing the folding probability of a generated set of TSE candidate conformations. We found that the TSE possesses a well-defined hydrophobic core with variable enveloping structures resulting from the superposition of three parallel folding pathways. The most preferred pathway agrees with the experimentally determined TSE, while the two least preferred pathways differ significantly. The knowledge of the different pathways allows us to design the interactions between amino acids that guide the protein to fold through the least preferred pathway. This particular design is akin to a circular permutation of the protein. The finding motivates the hypothesis that the different experimentally observed TSEs in homologous proteins and circular permutants may represent potentially available pathways to the wild-type protein.

Keywords: transition-state ensemble; protein folding; src-SH3 domain; discrete molecular dynamics; parallel folding pathways

Introduction

The folding process of many small globular proteins can be thermodynamically characterized as a two-state process.[1,2] A subset of the conformations not belonging to these two states forms the transition-state ensemble (TSE),[3–6] which can be defined as either (i) the set of conformations with the same folding and unfolding probabilities or (ii) the set of conformations corresponding to the minimum saddle point in a high-dimensional free-energy landscape between the folded and unfolded states. Determination of the TSE structure allows one to modify the folding rate of a protein by rational design[7] and describe folding intermediates not directly observed in experiments.

The overall TSE structures can be resolved to low resolution using the Φ-value analysis,[8] which determines changes in the free-energy profile after a point mutation[9–11] if one assumes the preservation of TSE structure after a mutation. The Φ-value analysis is suited for proteins folding through a well-defined pathway, but it may result in an average TSE structure if more than one folding pathway is kinetically available. For instance, the existence of different

*Corresponding author. E-mail address: arlam@buphy.bu.edu.

Abbreviations used: TSE, transition-state ensemble; DH, distal hairpin; RT, RT loop; DT, diverging turn; NT, N terminus; DCO, differential contact order.

folding pathways in the 27th immunoglobulin domain of titin was directly observed[12] as the authors varied the denaturant concentration, and other experimental[13] and computational[14] studies report proteins that switch their preferred folding pathway after changes in the environmental conditions. Also, different TSEs were observed in homologous proteins[15–17] and circular permutants.[18,19] These observations point to the existence of different folding pathways. From an evolutionary standpoint, TSE conservation may not be necessary if mutations preserve the fold. However, evolutionary pressure affecting folding rates may effectively lead to conservation of TSEs.[20] From the multidimensional free-energy landscape viewpoint,[21] certain mutations may alter the height of the different energy barriers separating the native basin from the unfolded conformations. This rearrangement of the landscape may result in the emergence of a new preferred folding pathway and TSE, along with the same native state.

The computational methods of finding the TSE of small proteins have given promising results in recent years.[22–33] However, it is still impossible to trace the folding of an atomic explicit-solvent model of a protein. Thus, the attempts to find the TSE are made by using coarse-grained models of a protein. Several computational works have implemented different approaches to determine conformations that belong to the TSE by introducing either the existence of a transmission coefficient as a single transition coordinate[34] or no assumption of a reaction coordinate whatsoever.[35] Many groups employed the concept of reaction coordinate,[24–27,29,30,32,36–42] which aims to record conformations associated to the maximum value of the free energy. Near the folding temperature T_F, the Gō model of small proteins such as SH3 domain allows production of hundreds of folding and unfolding transitions during computationally accessible simulation times. Thus, it is possible to perform accurate analysis, P_{Fold}, for a large number of conformations obtained during these successful folding events. P_{Fold} analysis of the Gō model conformations provides a high-resolution TSE of the protein under study, allowing for the construction of detailed amino acid contact maps.

Using the Gō model, Borreguero *et al.*[43] suggested that the TSE ensemble of the src-SH3 domain consists of partially formed β-hairpins and a long-range contact between the distal hairpin (DH) and the RT loop (RT). The hydrogen bonds L18–M48 and E16–M48 specifically are part of the folding nucleus. In contrast, the contacts between the termini (although more abundant in the unsuccessful unfolding events than in the unsuccessful folding events) were found in the TSE with small probability, from which it was concluded that they are not representing the main folding pathway of the src-SH3. Similar conclusions were drawn by another group[36] where the Φ-value analysis of src-SH3 domain was performed with 58% correlation with experimental values. Borreguero *et al.*[15] suggested that there are two folding pathways of src-SH3 domain: fast and slow. The fast pathway involves formation of the RT–DH contacts, while in the slow pathway the contacts between the termini and the diverging turn (DT) are first formed. This slow pathway gives rise to an intermediate state at low temperatures.

One can argue that although the Gō model gives detailed predictions of the TSE, these predictions can be wrong, because the Gō model does not use the correct energetics of the amino acid interactions. The importance sampling technique of Ding *et al.*,[44] based on the use of the potential energy and radius of gyration as reaction coordinates, determined the putative TSE of the src-SH3 domain. This putative TSE does not contain the RT–DH folding nucleus determined by Borreguero *et al.*,[43] but contains the long-range termini contacts corresponding to the slow folding pathway.[15] Thus, realistic energetics may result in an alternative pathway, which the Gō model tends to avoid. However, both termini in the src-SH3 domain have small experimental Φ values, while DH has high Φ values, suggesting that the kinetic predictions of the Gō model are more accurate than the thermodynamic predictions of importance sampling, which may completely miss the real TSE. Thus, these studies show that the kinetic Gō model has a large potential to determine the TSE. Indeed, the Gō model excessively samples all topologically possible folding trajectories and thus must contain the true TSE. The true amino acid energetics may alter the probability of the possible pathways. Accordingly, the most preferred pathway of the Gō model may not be the correct most preferred pathway. The differences in the experimental TSEs of Sso7d, α-spectrin, and the src-SH3 found by Guerois and Serrano,[45] of which all three proteins have similar folding, prove this assertion. Nevertheless the Gō model dramatically narrows down the putative TSE and thus generates very specific experimentally verifiable predictions.

Several computational studies have investigated the properties of circular permutants of the SH3 domain, S6, and chymotrypsin inhibitor 2.[22,31,46] These studies find that differences between the folding mechanisms of the permuted and wild-type proteins are strongly related to the change in their sequences. Using Φ values and modifying the energetics of certain set of contacts in a protein, we can characterize the possible pathways through which the protein folds, and determine both which contacts are preserved and which contacts could be used to modify the preference of a possible pathway in the TSE structure.

In this paper, we further develop the refinement of the Gō model predictions by dissecting its putative TSE into three clusters, each cluster representing a different structural rearrangement by which the protein may cross the free-energy barrier en route to its native fold. Our goal is to determine not only the preferred folding pathway of src-SH3, but also any additional pathways compatible with the native state. Moreover, we show that we can make one of those pathways to be dominant by altering the Gō model energetics. Pande *et al.*[47] reported the

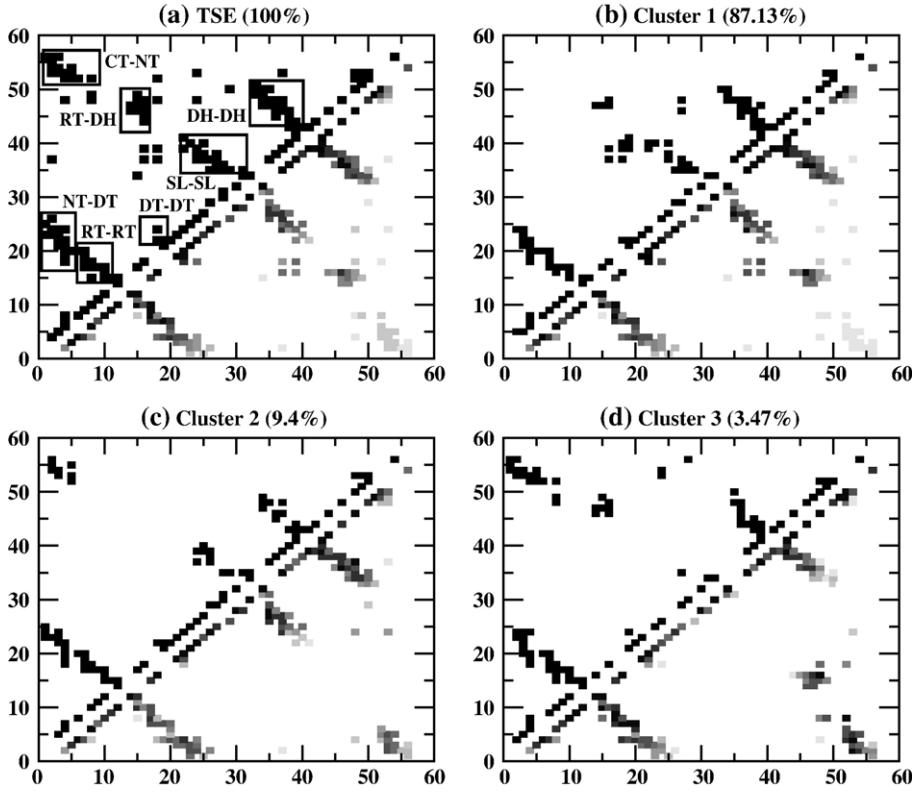

Fig. 1. Contact maps for TSE and clusters. (a) Upper triangle, native contact map. Lower triangle, average contact map of the TSE. (b) through (d) The partition of TSE into three clusters. Upper triangle, contact map of the cluster representative. Lower triangle, average contact map of the cluster. Percentage value on top of the graphs indicates percentage of total TSE.

existence of transition-state classes by using the P_{Fold} method on different lengths of polymers and they found that fast-folding proteins contain a single transition-state class that can go from one state to another through different pathways.

Results

Determination of the transition-state ensemble

We generate 114 folding events at equilibrium conditions and extract 5200 TSE candidates within the [−90, 80] energy window. We employ the P_{Fold} value test to filter out those conformations belonging to either the folded or unfolded states that are in the aforementioned potential energy window for a short period of time (see Methods). We deem a conformation as a TSE member if its P_{Fold} value is located within the [0.4, 0.6] probability range. Out of the 5200 candidates, 1525 conformations turn out to be TSE members after the P_{Fold} value test. We calculate the probability of every native and non-native contact to be present in the set of 1525 conformations and plot all contact probabilities in a grayscale contact map (Fig. 1a, lower triangle). The division of the sequence into the native loops and hairpins allows us to coarse-grain the contact map into a series of medium- and long-range clusters of contacts (Fig. 1a, upper triangle). We group contacts (i) within the RT loop (RT–RT), (ii) within the diverging turn (DT–DT), (iii) within the n-Src loop (SL–SL), (iv) within the distal hairpin (DH–DH), (v) between the N terminus and the diverging turn (NT–DT), (vi) between the RT loop and the distal hairpin (RT–DH), and (vii) between termini (NT–CT). The contact probability for each of these groups of contacts is the average of the contact probabilities for all the contacts included in the group, and we present these probabilities in Table 1. The TSE has the highest probabilities for DT–DT, DH–DH, and RT–RT, while NT–CT contacts have the lowest

Table 1. Average percentile probability of segments of the protein (defined in Fig. 1a) for TSE and clusters

	NT–DT	RT–RT	CT–NT	RT–DH	SL–SL	DH–DH	DT–DT
TSE	51.35	72.34	10.61	51.07	63.82	67.19	66.77
C1	50.35	72.42	4.03	55.98	65.94	67.37	70.22
C2	58.41	66.62	48.34	0.03	63.70	71.94	42.24
C3	57.36	85.91	73.70	65.01	10.87	50.04	46.42

probabilities. The rest of the contacts have moderate probabilities.

Are contacts with moderate or low probability values mere accessory contacts, uncorrelated to the presence/absence of other contacts? To address this question, we partitioned the TSE into clusters (see Methods) and obtain three clusters with differentiated secondary and tertiary structural characteristics. In particular, no two clusters share the same pattern of contacts between different loops and hairpins. When compared to the native state, cluster 1 (C1, Fig. 1b) lacks NT–CT, cluster 2 (C2, Fig. 1c) lacks RT–DH, and cluster 3 (C3, Fig. 1d) lacks SL–SL contacts. We show in Table 1 the actual contact probabilities between the different loops and hairpins for the three clusters. We also considered a finer partition of the TSE containing four clusters but it did not provide any additional pathways. The populations of the clusters vary considerably. C1 comprised 87.13% of the TSE, C2 9.4%, and C3 3.47%. C1, with the highest population, represents the preferred folding pathway. In all three clusters, the contact map of the representative conformation (Fig. 1b–d, upper triangle) is similar to the average contact map (Fig. 1b–d, lower triangle), indicating that each cluster is cohesive. Within each cluster, we observe regions having significantly higher contact probability than other regions (see Table 1).

Φ-Value analysis

We calculate Φ values (see Methods) for the TSE and for each one of the four clusters (Fig. 2). We compare predicted and experimental Φ values[48,49] by the computation of the correlation coefficient, r, and with the average deviation $\sigma^2 = (1/N)\sum_{i=1}^{N}(\Phi^i_{th} - \Phi^i_{exp})^2$ and corresponding p value, assuming a null hypothesis in which the predicted Φ values are independent of each other and randomly distributed within the [−0.22, 1] range, where −0.22 and 1 correspond to the minimum and maximum Φ values recorded in protein-engineering experiments of the src-SH3 domain[49,50] (Table 2). A low σ value indicates agreement between predicted and observed Φ values, in which case the p value is also small. Predicted Φ values for the TSE have $r = 0.56$, $\sigma = 0.34$, and $p = 0.22$ to experimental values, indicating a semiquantitative agreement. C1 has the best fit among all three clusters ($r = 0.39$), followed by C2 ($r = 0.08$) (Table 2). C3 is the least populated cluster and its fit parameters r ($r = -0.39$), σ, and p show disagreement with experimental Φ values. In particular, the p value indicates a fit no better than randomly expected. An important result is that the computed TSE has a higher correlation with the set of experimental Φ values than with any of the three clusters. Thus, it is the combination of all three folding pathways that characterize the TSE. Because

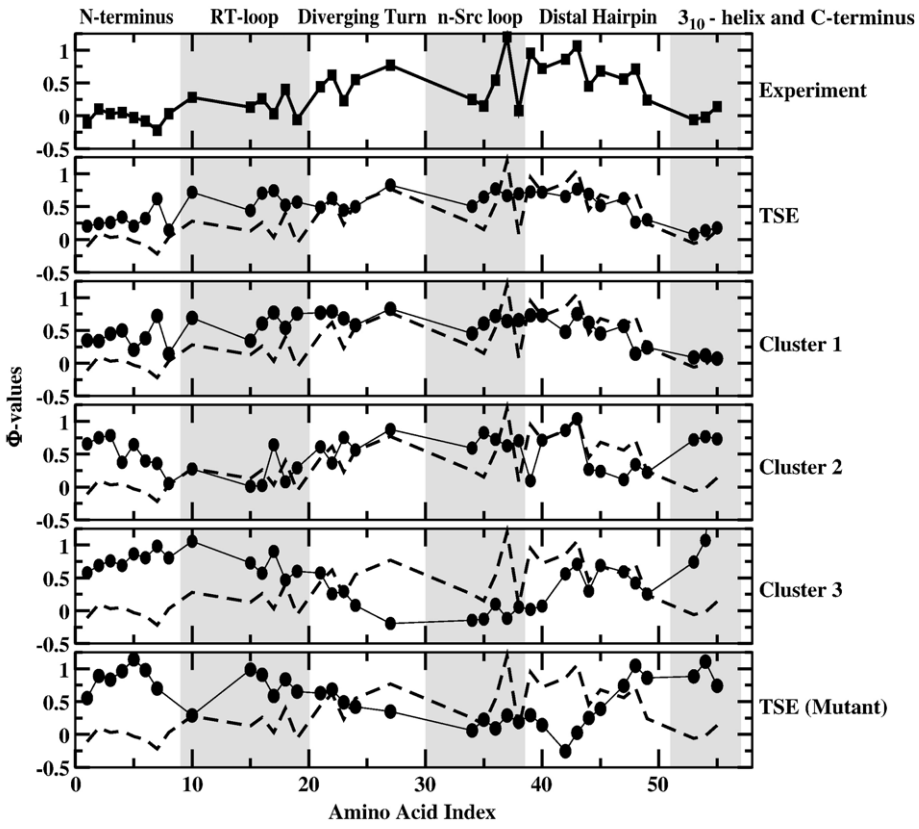

Fig. 2. Experimental (upper panel) and computed Φ values for the TSE, the three clusters, and the TSE for the mutant. Alternating white and gray boxes differentiate between six consecutive elements of secondary structure labeled at the upper panel. For comparison, we plot the experimental Φ values in the dashed line.

Table 2. Correlation coefficient and p value between calculated and experimental Φ values

	r	ALL	NT	RT	DT	n-Src	DH	CT	DCO	DCO′
TSE	0.56	0.22 (0.34)	0.20 (0.41)	0.34 (0.45)	0.11 (0.11)	0.33 (0.45)	0.17 (0.23)	0.10 (0.12)		
C1	0.39	0.26 (0.39)	0.28 (0.51)	0.37 (0.47)	0.19 (0.26)	0.30 (0.43)	0.20 (0.28)	0.10 (0.13)	1.00	1.00
C2	0.08	0.35 (0.47)	0.42 (0.62)	0.20 (0.31)	0.20 (0.28)	0.40 (0.52)	0.29 (0.38)	0.67 (0.73)	1.61	0.51
C3	−0.39	0.89 (0.81)	0.75 (0.83)	0.66 (0.64)	0.65 (0.51)	0.63 (0.66)	0.35 (0.423)	1.00 (1.91)	2.61	0.48
Mutant	−0.54	0.78 (0.73)	0.87 (0.92)	0.97 (0.88)	0.18 (0.25)	0.34 (0.47)	0.75 (0.64)	0.93 (0.98)		

The numbers between parentheses correspond to σ. ALL, whole sequence; NT, N terminus; RT, RT loop; DT, diverging turn; n-Src, n-Src loop; DH, distal hairpin; CT, C terminus. Last two column indicate differential contact order before (DCO) and after (DCO′) circular permutation with cleavage at the n-Src loop.

different fragments are more structured in some particular cluster, we compared computed and experimental Φ values when restricted to different regions of the protein (Fig. 2). C1 had low p values except for the first 20 amino acids including some residues in the RT loop for which the computed Φ values are significantly larger than the experimental ones, indicating that they may be overstructured in the Gō model. This is not surprising, since the Gō model does not take into account the absence of strong hydrogen bonds in the RT loop and treats them in the same way as those in the more stable β-hairpins. This raises the question of whether the DH–RT contacts predicted by Borreguero *et al.*[43] to be the folding nucleus do indeed have this role. Cluster C2 has good agreement with the experiment in the RT loop (RT–DH contacts are not formed) but has no agreement with the experiment for the termini, which are overstructured and form contacts in the Gō model. Cluster C3 has no agreement with the experiment for any structural elements, and correspondingly has a negative correlation coefficient. C3 has both RT–DH and NT–CT contacts, which seem to be serving as folding nuclei in the two different folding pathways of the Gō model, represented by C1 and C2, respectively. The Gō model presents a clear picture of folding: the secondary structural elements are partially formed in the unfolded state; still, its conformational space is too large and the potential energy landscape is flat. The protein spends a certain amount of time searching along this flat landscape until it finds the funnel to the native state, at which point one of these two long-range groups of contacts is formed. This point corresponds to the overcoming of the entropic barrier. Whether this picture is correct remains an open question. In the TSE found by the importance sampling of the all-atom model of Ding *et al.*,[44] the RT–DH contacts are not present at all and the NT–CT contacts have relatively small probability, which is in disagreement with the Gō model TSE. On the other hand, β-sheets are well formed, which is in agreement with the Gō predictions. It is possible that the formation of RT–DH contacts in the TSE is an artifact of the Gō model, but in this case the folding nucleus of the SH3 domain is still unknown. Only careful experiments specifically targeting these long-range contacts may resolve it.

Weak termini linkage and pseudocircular mutation

In order to change the preferred folding pathway, we implement a multiple point mutation (see Methods) that changes the energies of amino acid interactions from ϵ_{ij} to $\epsilon'_{ij}=\epsilon_{ij}(1+\Delta P_{ij}/\alpha)$, where ΔP_{ij} is the probability that residue i will contact residue j for C3 conformations, minus the probability that i will contact j for C1 conformations. α is a tunable parameter indicating the strength of mutation. This mutation reinforces native contacts that are more probable in C3 than in C1 ($\Delta P_{ij}>0$). We empirically determine $\alpha=4$. This value ensures that the new contact energies preserve the native structure and the two-state character of the folding process. Lower values of α parameter resulted in kinetic trapping. According to C1 and C3 conformations (Fig. 1b and d), the mutation amounts to a strengthening of contacts between termini and a weakening of contacts within the n-Src loop. In the extreme case ($\alpha=0$), this mutation would produce contact energies strong enough to permanently bind the termini and weak enough to prevent the formation of any secondary structure in the n-Src loop, thus greatly enhancing the local entropy of this region. Other than the actual cleavage of the sequence, this mutation is akin to a circular permutation with cleavage in the n-Src loop.

After mutation, the energy of the native state is −162.66 (a 1.7% negative increase in energy). The contact map of the mutated TSE is similar to that of the wild-type TSE (Fig. 3). There are differences in NT–CT, SL–SL, and DH–DH contacts. The partition of the mutated TSE contains four clusters, three of them ($\bar{C}1$, $\bar{C}2$, and $\bar{C}3$) having corresponding analogous clusters in the wild-type TSE (Fig. 3b–d). There is a new cluster, which we term cluster 4 ($\bar{C}4$), with no analogous cluster partition in the wild-type TSE (Fig. 3e). $\bar{C}4$ shows no DH–DH contacts, while this set of contacts is present in all clusters of the wild-type TSE and is part of the core in the experimentally observed TSE. The cluster populations are markedly different from those of the wild-type TSE. $\bar{C}1$ comprises 18% of the whole TSE (a 3.2-fold reduction), $\bar{C}2$ comprises 16.2% (a 2.9-fold increase), $\bar{C}3$ comprises 38.3%, and $\bar{C}4$ comprises 27.5%. Thus, $\bar{C}3$ (an 11.0-fold increase) becomes the preferred folding pathway after mutation. Φ values of the mutated TSE (see Fig. 2) are strongly anticorrelated to experimental Φ values ($r=-0.54$, Table

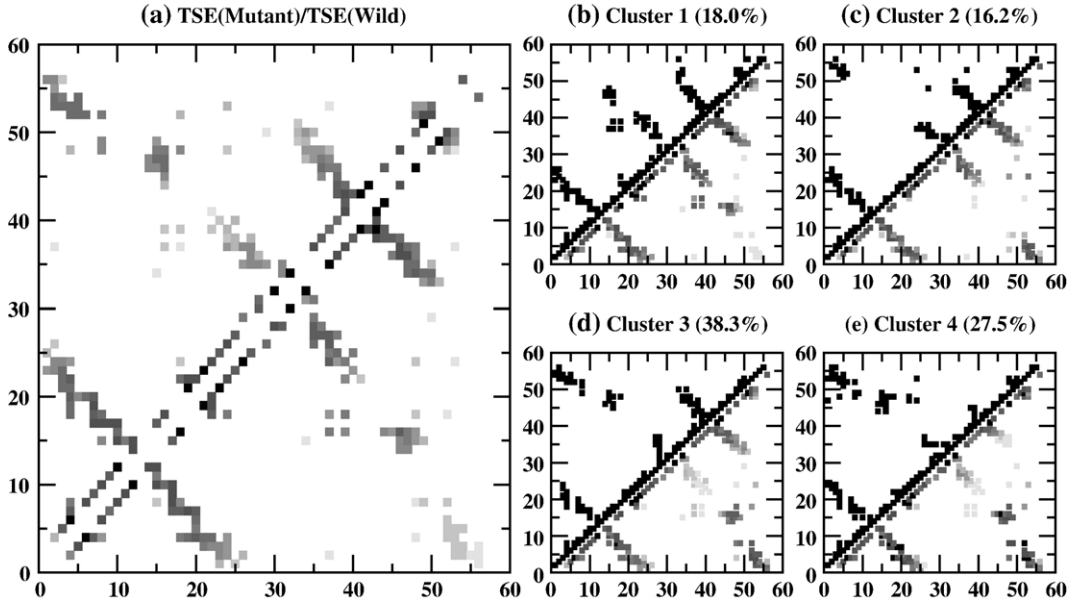

Fig. 3. Contact maps for TSE and clusters after mutation. (a) Lower triangle, average contact map of the TSE before mutation. Upper triangle, average contact map of the TSE after mutation. (b)–(e) Upper triangles, representative of the clusters. Lower triangles, average contact maps of the clusters.

2), indicating the effectiveness of the mutation in selecting a TSE unrelated to the experimental TSE, while preserving the native state and the two-state character of the folding process.

Discussion

The preferred folding pathway of src-SH3 domain was recently investigated.[36] This study validated our simplified protein model, as it reproduces the experimentally observed two-state folding and the main structural features of the TSE. These results encouraged us to obtain a detailed description of the TSE.[44] However, the elusive nature of the TSE conformations (being a maximum of the free energy) makes its determination a time-consuming process and prone to errors, hence the need of simplified models to reproduce enough number of folding events to achieve statistical significance in the results. The use of experimental Φ values in simulations[23,28,33] allows one to alleviate the sampling problem by effectively restricting the conformational search in the neighborhood of the preferred folding pathway. However, we did not use this technique since it precludes observation of less visited pathways that have no correlation to the experimental Φ values, like cluster C3. An interesting variation employed by Hubner *et al.*[51] is the use of all-atom models using only one subset of the Φ value set as restraints. In their study, the authors found that the structure of the TSE was not uniquely defined by the selected subset. Klimov and Thirumalai[35,52] suggested the heterogeneity of the TSE in a protein lattice model. They interpreted that the conformations are grouped in distinctive TSEs that lead to a diversity of pathways. Later, Klimov and Thirumalai[53] used coarse-grained models to represent a set of α/β proteins and determined that the particular arrangement of the secondary structure elements is correlated with the diversity of folding pathways. They point out that a dominant folding pathway through a unique specific nucleus is expected to appear for proteins with different arrangement of helices and strands in the termini, as in the SH3 domain. We observe a dominant folding pathway with dominant nucleus in the TSE (the biggest cluster makes for 87% of all TSE conformations) in our folding simulations, and also in accordance is our observation of a depolarization of the TSE (the biggest cluster now makes for 38% of all TSE conformations) after the multiple point mutation. This mutation is similar to a circular permutant with cleavage in the Src loop, and this permutant has a more similar arrangement of secondary structure in the new termini. According to the observations of Klimov and Thirumalai, the permutant should have a TSE less polarized than the wild-type protein.

Ding *et al.*[44] employed both a full-atom and a two-bead protein model with a P_{Fold} analysis to obtain TSE conformations. They explored three P_{Fold} ranges ([0, 0.2],[0.4, 0.6],[0.8, 1.0]) in order to observe the transition state before and after the conformations reached the top of the free-energy barrier. The clustering algorithm that they used generated a much larger number of clusters since they imposed an rmsd cutoff of 3 Å between conformations of the same cluster. They observe that a conformation belongs to the TSE if it has the following set of contacts: NT–CT, NT–DT, and central β-sheet formed by DT–SL–DH. Also, they put the restriction that the presence of only one of these contacts in a conformation does not guarantee that it belongs to TSE. Our results concentrated on the [0.4, 0.6]

probability range and we made the clustering with no rmsd cutoff between cluster members (see Methods) and with no limit in the cluster population. We obtained three clusters and even with the limitations of our approach, these clusters present the set of contacts that Ding *et al.*[44] observed. C1 has NT–DT contacts and a central β-sheet well formed; C2 has NT–CT and NT–DT contacts and C3 has NT–CT and NT–DT contacts and a light β-sheet. The refinement of the clustering algorithm and the criteria to define the TSE members indicate that this protein has local structural variability and the diversity of these structures lead to the multiple folding pathway scheme. With changes in the strength of specific contacts (point mutation), we obtained a small number of folding pathways for the mutated TSE of which two of these folding pathways are present in the wild TSE and a new one emerges in the conformations. It indicates that the mutated TSE presents some level of homogeneity with respect to the wild-type TSE.

Goldbeck *et al.*[54] used circular dichroism measurements on a globular protein, cytochrome *c*, and they determined heterogeneity in the kinetics of early folding events that is consistent with the dynamics for an ensemble of pathways on the energy landscape. Other studies by Bieri and Kiefhaber[55] observed the folding of lysozyme. They determined that after a rapid collapse, the kinetics of the protein takes two folding pathways, where the fast pathway leads directly to the native structure, whereas the slow pathway goes through a partially folded intermediate with nativelike secondary structure in the α domain. Recent experiments with proline-free Staphylococcal Nuclease made by Kamagata *et al.*[56] concluded that the protein has multiple parallel folding pathways, although each one includes one metastable intermediate. They suggest that even proteins with no apparent folding intermediates may fold though parallel pathways, a hypothesis observed in our studies of the src-SH3 domain.

The overall structure of the computed TSE shows good agreement with experimental results, with a stable core formed by the distal hairpin and diverging turn, concomitant to unstructured termini. As in the experiment, we observed intermediate Φ values that could be the result of an average over two or more parallel folding pathways, where some contacts are absent in one pathway but present in others, a scenario proposed by the authors.[57,58] Still, we find that most of the residues have intermediate Φ values in all three clusters. In addition, since clusters C2 and C3 have a relatively low population when compared to C1, most of their Φ values bear little influence on the resulting Φ values for the whole TSE. It is only for residues 20–24 that one can discern Φ values with an excellent match to experimental values due to average over different pathways. In this case, calculated Φ values for C1 are higher than experimental values, while Φ values for C3 are lower. The net Φ values are closer to experimental values than in any of the two pathways. This adjustment gives rise to the higher correlation of the TSE to experimental Φ values. DH is the most structured region of the TSE in accordance with the nucleating role described by experiment.[35,49,59,60] DH–DH contacts have high probability in all three clusters, which may be a consequence of a previous observation stating that DH–DH contacts are present in pre-TSE conformations.[42] Thus, these contacts may be a common requirement for all subsequent TSE conformations, and hence for all clusters. The presence of NT–CT termini contacts correlate with cluster population. Conformations belonging to C1, the most populated cluster, tend to lack NT–CT contacts. On the other hand, NT–CT contacts are present in conformations of C2, and specially C3, which are clusters of much lower population than C1. A plausible reason for the observed correlation may be the entropic cost of bringing the two termini together. In fact, the differential contact order (DCO; see Methods) and cluster population are anticorrelated, with C1 having the smallest DCO and C3 having the highest value ($DCO_{\{C1,C2,C3\}-C1} = 1.00, 1.61, 2.61$). Given the experimentally determined correlation between contact order and folding rate,[61] it was reassuring to find that most of the simulations folded through C1.

The observed anticorrelation between cluster population and DCO suggests that mutations with very different DCOs, a circular mutation for instance, would alter the relative population of the folding pathways. An ideal circular mutation changes the configurational entropy of the protein while maintaining the enthalpic contribution to the free energy. If this type of mutation would conserve the partition of the TSE into clusters, then a cleavage of the protein at the Src loop and subsequent fuse of the termini would result in C3 having the lowest DCO ($DCO_{C3-C1} = 0.50$). C3 might then become the preferred folding pathway. We attempt to reproduce the effect of a circular mutation not with a cleavage of the chain, but through correlated perturbations in the interaction energies. The implemented perturbations reinforce termini contacts and weaken the contacts stabilizing the n-Src loop. After this "mutation," C3 still has the highest DCO, but the entropic barrier is now overcome by energetically favorable termini contacts. The change can be rationalized as an alteration in the height of the free-energy barriers separating native state and unfolded conformations. The barrier corresponding to C3 becomes the lowest in free energy and C3 becomes the preferred folding pathway. After mutation, the populations of the clusters are more similar to each other, indicating a depolarization of the TSE. The presence of a new cluster ($\bar{C}4$) with a differentiated structure with respect to the others cluster could mean that proper changes in the configurational entropy of the protein *via* circular mutations would lead to the emergence of new pathways. This result suggests that experimentally observed TSE in circular permutants[18,19] and homologous proteins[16,17,62] may be potential folding pathways of the wild-type protein.

Viguera *et al.*[63] performed protein-engineering experiments with all possible permutants that

disrupt the covalent linkage between two β strands forming a β-hairpin in the α-spectrin SH3 domain. They found that those mutants had similar stabilities but different folding kinetics. They concluded that both the order of secondary structure elements and the preservation of any of the β-hairpins present in this domain are key players in determining the folding pathway. They also performed[64] two circular mutations in the α-spectrin SH3 domain to produce mutants with cleavage in RT and the distal loop, respectively. They reported that both cases folded into the same three-dimensional structure as the wild-type SH3. However, the structure of the TSE of each protein as inferred from the Φ values differed significantly from each other.

Several computational studies[22,42,46] have addressed the implications of mutations in the TSE structure. Hubner *et al.*[65] studied the TSE topology of ribosomal α/β S6 and stated that some contacts are preserved in the whole TSE after a mutation, concomitant to changes in other parts of the TSE structure. We observe the preservation of RT–DH contacts as well as local to medium-range contacts in RT and DH, in agreement with the Hubner *et al.* findings. Also, Lindberg *et al.*[66] observed that changes in Φ values could be quantifiably linked to changes in sequence separation. They performed circular permutations in mutants of S6 protein and determined that drastic changes in Φ values occur near the new termini. The changes in the population of the clusters that we determined before and after mutations also result in Φ-value changes that are more pronounced in and close to the termini.

Several reports address the validity of the Gō model and its implementation as a tool to capture the essential elements of potential energy landscapes, TSE, and folding trajectories. Paci *et al.*[67] made a systematic comparison with an atomic model where they used a molecular mechanics energy function with implicit solvation and they found that the energy contacts defined in the Gō models described in very good approximation the folding/unfolding kinetics of the studied proteins such as acylphosphatase (ACP, composed of a β-sheet structure with two short α-helical segments), chymotrypsin inhibitor 2 (CI2, 1 α-helix with a β-sheet structure), α-lactalbumin (α-LA, α plus β), and the third fibronectin type III repeat from tenascin (1TEN, a β sandwich). Rhee and Pande[68] made comparative studies of all-atom and simplified models on a mutated ribonucleoprotein catalyst (BBA5). They concluded that for this protein, the folding kinetics of explicit-solvent atomic models are different from those of implicit-solvent atomic models, and even more different from those of simplified models. They quantified differences between models in terms of correlations between corresponding P_{Fold} values that were calculated for a set of conformations initially at equilibrium in explicit-solvent simulations. BBA5 is a miniprotein (23 residues) with a very low contact order. As a result, BBA4 has a fast and weak cooperative folding transition.[69] Under these folding-favoring conditions, Gō models may fail to reproduce two-state folding behavior and instead produce a "downhill" folding process with no apparent free-energy barriers. Rhee and Pande[68] observed that while all-atom models were able to capture the small free-energy barrier and produce two-state folding, Gō models did not reproduce it. Rhee and Pande concluded that Gō models may be more suited to describe the folding process of bigger proteins where topology is a main factor in determining the free-energy barrier. This is the case of SH3, for which the Gō model produces a TSE in semiquantitative agreement with experimental results, while all-atom models are currently unable to drive the protein from the unfolded to the folded state. Bigger proteins were also studied by Paci,[70] who assessed the validity of Gō models with eight proteins in the 50- to 200-residue range and spanning a range of different folds (two α, three β, and three α–β proteins), including the SH3 fold. The authors found a significant correlation between single-residue energies of Gō and all-atom implicit-solvent models when applied to conformations belonging to the TSE of these proteins. This finding, according to the authors, provides a justification that Gō models are as well suited to describe the TSE of the studied proteins as do more complex models.

The application of the Gō model to the folding of the SH3 domain has provided insights into the role of specific residues as well as whole secondary structure elements. Studies with a Gō model made by Ding *et al.*[36] with C-Src-SH3 (a β protein) provided a comprehensive picture of the folding mechanism with a semiquantitative agreement with experimental data. Borreguero *et al.*[43] studied the thermodynamic properties and folding kinetics of the SH3 and analyzed the contribution of each secondary structure element of the SH3 in the TSE reproducing some of the features observed in experiments. Studies made by Dokholyan *et al.*[5] employed a combination of Gō potential and experimental Φ values and they reported a correlation between the topological properties of proteins with their folding mechanisms. Clementi *et al.*[46] simulated circular permutations of SH3 and CI2 and addressed the question of how these permutations affected the folding mechanism. Their results determined which sequence fragments were more relevant in the folding process and how they could be disrupted in order to change the folding pathways.

Conclusions

In this study, we record an ample set of protein conformations that make up the TSE of src-SH3 domain and group them into three clusters. The least populated cluster represents a folding pathway with Φ values largely deviating from experimental reports, although one needs to combine all three clusters in order to obtain the TSE with the highest correlation to the experimental set of Φ values. Our results show that by tuning the energetics of the Gō model it is possible to change the most preferred

TSE cluster, thus simulating actual TSEs of different representative proteins of the same fold. For example, we have shown that cluster $\bar{C}1$, containing the RT–DH contacts present in the TSE of the src-SH3 domain, becomes less abundant after altering the Gō interactions, while the clusters $\bar{C}2$, $\bar{C}3$, and $\bar{C}4$, containing the CT–NT and DT–NT contacts present in the TSE of the structural homology Sso7d, become more abundant in accord with experimental Φ values of Guerois and Serrano[45]. Moreover, our changes in the Gō interactions lead to changes in Φ values similar to those observed in the circular permutations of Viguera *et al.*,[64] which decreased the role of V44 in the DH and increased the role of V53 in the CT. The implication of this result and our studies suggests the existence of multiple parallel folding pathways each with their own well-defined TSE. The mutations made by experiments that either weaken the RT–DH contacts or the CT–NT contacts will determine the preference of a protein to follow a folding pathway. Our results indicate that these parallel folding pathways are always present, and the change in the TSE structure because of mutations preserving the native fold would result in the prevalence of some pathways over the others.

Methods

Protein model and simulation technique

The src-SH3 domain is an all-β-domain protein made up of two semiperpendicular β-sheets.[50,59] The various loops and hairpins of the native state provide a natural framework to describe the structural differences arising when the protein folds through the different folding pathways. Accordingly, we divide the amino acid sequence into the following fragments (Fig. 4a): N terminus (NT residues 1–7), RT loop (RT 8–20), diverging turn (DT 21–30), n-Src loop (SL 30–38), distal hairpin (DH 39–50), and 3_{10}-helix and C terminus (CT 51–57). Due to the computationally demanding nature of this study, we employed a C^{α}–C^{β} representation per amino acid[36,71] (Fig. 4b). C^{β} beads interact *via* a modified version of the Gō potential,[72] whereby we assign a binding energy of $\epsilon=-1$ to each native and $\epsilon=+1$ to each nonnative contact, and a 7.5-Å interaction range. The resulting set of 162 native contacts reveals the secondary and tertiary structure when plotted on the contact map (Fig. 4c).

We implemented discrete molecular dynamics,[37,71,73,74] which is event-driven, with implicit solvation, and faster than traditional molecular dynamics due to simplified interparticle potentials that are reduced to one or more square wells.[75] A pair of particles moves with constant velocities along straight lines until a distance is reached at which the interparticle potential is discontinuous, and an ensuing collision occurs. At that moment, particle velocities and directions of motion are recalculated such that the total energy, momentum, and angular momentum are conserved.

Determination of the transition-state ensemble

Our strategy to determine the TSE is twofold: we collect a subset of all sampled conformations to build a putative TSE. The collected set will contain conformations with degrees of freedom corresponding to the minimum saddle point in a high-dimensional free-energy landscape between the folded and unfolded states. A free-energy profile versus one or more folding reaction coordinates allows

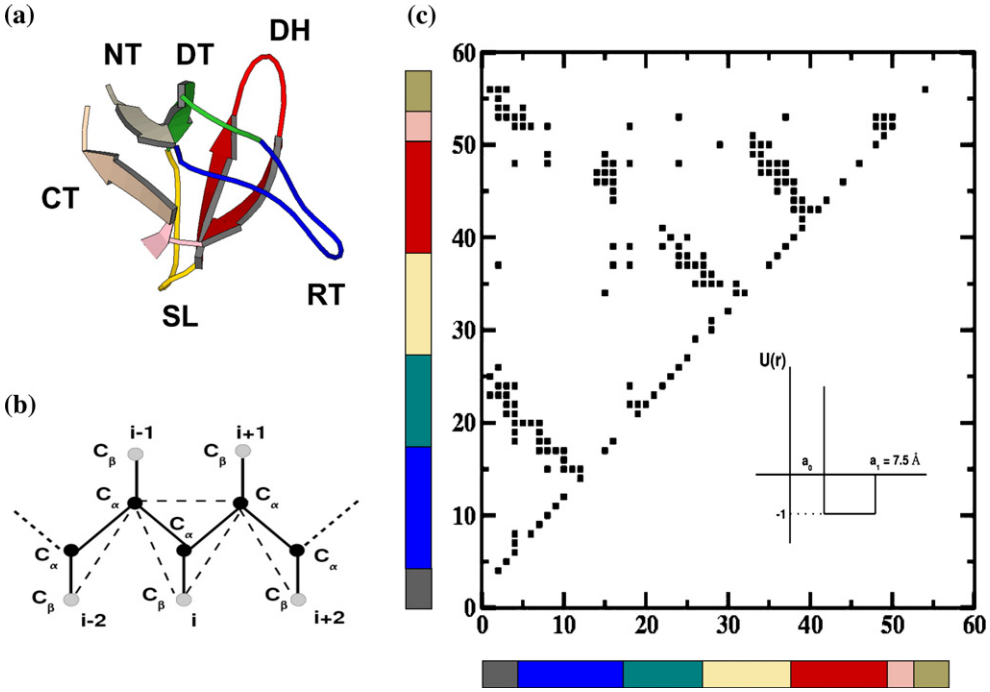

Fig. 4. The protein model. (a) Cartoon representation of the SH3 protein. (b) Two-bead representation of the src-SH3 domain. C^{α} beads represent the backbone and C^{β}s the side chains. Lines represent different types of bonds to reproduce the correct geometry and conformational freedom. (c) Native contact map with gray bands mapping the different segments of secondary structures. We show as an inset the potential energy function of a native contact versus interparticle distance ($a_0=3.3$ Å and $a_1=7.5$ Å).

one to record conformations with reaction coordinates values corresponding to the maximum in the free-energy profile.[24–27,29,30,32,36–42]

Unfortunately, a precise reaction coordinate is not possible for a complex process like the folding of a protein. Therefore, we use the potential energy (U) to play the role of the reaction coordinate to construct a free-energy profile and also take into account the trajectories that follow along the route from the unfolding state toward the folded state. These criteria are not sufficient to consider the putative conformations to be full TSE members, since there are conformations obtained from the simulations either in the folded state or in the unfolded state that stay for a short period of time within the energy window. We call these conformations false positives.

We use the discrete molecular dynamics algorithm to generate 114 folding events at equilibrium conditions at the folding transition temperature $T_F = 0.91$ (Fig. 5a).[36] A folding event is defined when the protein has in its potential energy a bimodal structure or two states, folded and unfolded. We only record conformations with U in the range $[-90, -80]$, which is the region enclosing the minimum of the distribution of U values (Fig. 5b), $P(U)$, or analogously, the maximum of the free-energy profile:

$$F(U) = -k_B T \ln P(U). \quad (1)$$

After recording $N_C = 5200$ candidate conformations, we apply the P_{Fold} test based on the kinetics definition for a conformation belonging to the TSE, which states that the conformation should have the same probability to either fold or unfold.[34,36,37] Thus, we calculated P_{Fold} of a conformation by performing $N = 32$ simulations each starting from the conformation in question and counting the number N_{Fold} of times that the protein folds:

$$P_{Fold} = \frac{N_{Fold}}{N} \pm \frac{1}{2\sqrt{N}}. \quad (2)$$

All N simulations start with the same structure but different particle velocities, generated from a Maxwell–Boltzmann distribution. We consider a simulation to be folded when the potential energy entered in the $U < -120$ region, the typical U's of the folded state, and remained in the region for a time $t \gg T_F \approx 10^3$, the typical folding time (for a comparison of time scales, the typical time of a protein to stay in the folded state at $T = T_F$ in equilibrium condition is about 10^5). Finally, we assign a conformation to the TSE if its P_{Fold} value ranges within [0.4, 0.6] (Fig. 5c). In order to represent the overall TSE structure, we report the contact map averaged over the TSE conformations.

Clustering of the transition-state ensemble

We sort the TSE members into different clusters, with conformations belonging to the same cluster having a similar structure (low rmsd). Another possible criterion like common native contacts between conformations has been employed.[52] A particular arrangement of the TSE into a discrete number of clusters constitutes a *partition*. The clustering procedure starts with every conformation considered as a cluster. Then the algorithm (i) calculates the rmsd between all pairs of clusters and determines the pair with the lowest rmsd value and (ii) merges the pair into one cluster and repeats previous step. We define the rmsd between two clusters A and B as the average rmsd value resulting from all possible pair combinations of one conformation from cluster A and one from cluster B. Every iteration diminishes by one the number of clusters until there is but one cluster encompassing the whole TSE. To determine the partition with an optimal number of clusters, we construct the average contact map of each one of the clusters contained in each partition. If the map of two clusters within the same partition convey the same information in terms of formation of secondary and tertiary structure, then we do not need as many clusters to describe the TSE. On the other hand, a quantitative criterion is implemented. We determined the average rmsd of the largest cluster. From the last stage, the TSE itself, we go up along the tree until the stage where the largest cluster has a significant difference in the average rmsd with respect to the largest cluster of the next stage (see Fig. 6). Thus, the optimal partition will

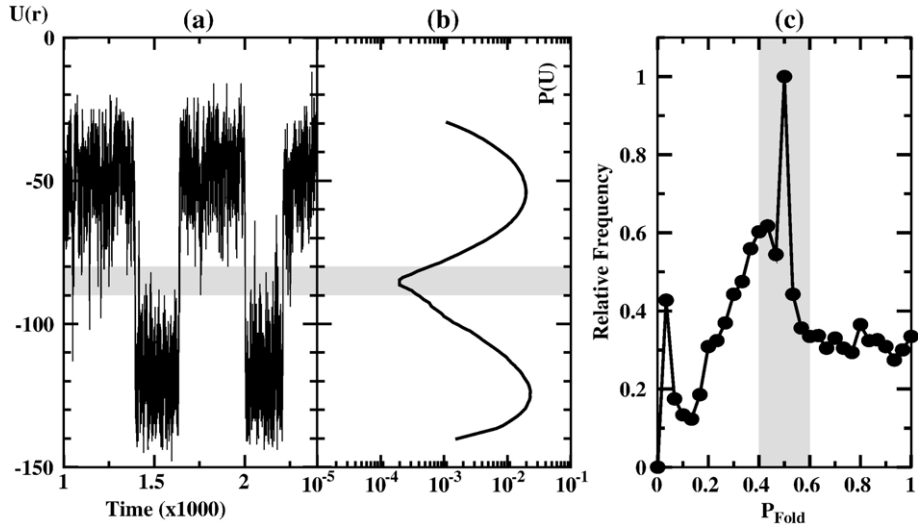

Fig. 5. Two-state folding of src-SH3 model. (a) Histogram of P_{Fold} for the TSE candidates. The area in gray indicates number of candidates belonging to the TSE. (b) Folding events are marked by a sudden decrease of potential energy (U). (c) Normalized distribution of U values. The shadowed region denotes the U range for recruitment of candidate TSE conformations.

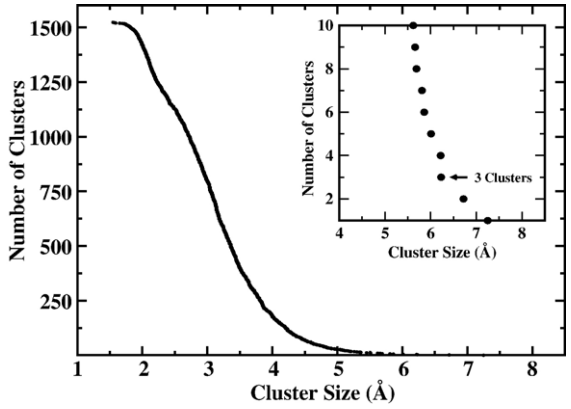

Fig. 6. Number of clusters *versus* size of the largest cluster. The inset shows abrupt increase in the cluster size for less than three clusters.

contain the maximum number of clusters with different average contact maps. We assign to each cluster a representative, which is defined as the member of the cluster having the lowest average rmsd value relative to the other members.

The virtual point mutation method

The virtual mutation allows us to calculate the Φ value for a particular amino acid by removing all nonbonded interactions between C^{β} of the mutated residue and the rest of C^{β}.[76] From a computational point of view, the method has the advantage that it allows one to compute Φ values without the need to run additional molecular dynamics simulations of the mutated protein. In addition, we assume that the virtual mutation preserves the folded, TSE, and unfolded states, in the sense that any conformation belonging to either of these states will still belong to the same state after mutation. The resulting change in free energy is:

$$\Delta G_x^{(i)} = -k_B T \ln\left\langle \exp\left(-\frac{\Delta U^{(i)}}{k_B T}\right)\right\rangle_x, \quad (3)$$

where x indicates folded (F), unfolded (U), or TSE, and $\Delta U^{(i)}$ is the number of native contacts minus the nonnative contacts that amino acid i makes with the other amino acids when in a particular protein conformation. More generally, it is the total nonbonded interaction energy between amino acid i and the rest of the protein. $\langle . \rangle$ indicates the average with the Boltzmann weight. We calculated the Φ_i value as:[8]

$$\Phi_i = \frac{(\Delta G_{TSE}^{(i)} - \Delta G_U^{(i)})}{(\Delta G_F^{(i)} - \Delta G_U^{(i)})}. \quad (4)$$

To compute $\Delta G_{TSE}^{(i)}$, we first obtain the TSE conformations. For ΔG_F and ΔG_U, we record folded and unfolded conformations from simulations in equilibrium at T_F.

The multiple point mutation method

We propose a simple and novel method to introduce controlled changes in the free-energy landscape in order to shift the preferred folding pathway. Let A and B be two ensembles of protein conformations representing the TSE along the preferred folding pathway and the TSE along other less visited folding pathways, respectively. We perturb the set of native binding energies, ϵ_{ij}, *via* the linear transformation

$$\epsilon'_{ij} = \epsilon_{ij}\left(1 + \frac{\Delta P_{ij}}{\alpha}\right), \quad (5)$$

with $-1 < \Delta P_{ij} < 1$. $\Delta P_{ij} = P_{ij}(B) - P_{ij}(A)$ is the probability that amino acids i and j make a contact in conformations of ensemble A minus analogous probability for conformations in ensemble B. $1/\alpha$ is the perturbation strength parameter. In our simulations, we estimate a lower bound to α by trial and error, while ensuring that the energy perturbations preserve the native structure and guarantee the absence of intermediates in the folding process.

Differential contact order

The contact order of a conformation is defined as[61]

$$CO \equiv \frac{1}{LM}\sum^{M} \Delta L_{ij}, \quad (6)$$

where L is the length of the protein, M is the number of contacts, and ΔL_{ij} is the sequence separation between amino acids i and j forming a contact. While CO is a valid measure of global topological complexity, it is not as useful when we compare two conformations differing by a few contacts, because CO will be largely determined by the contacts that are common to both conformations. For this type of comparison, we define the differential contact order as the contact order of contacts in A but not present in B, divided by the contact order of contacts in B but not present in A, that is, $DCO_{A-B} = G(A, B)/G(B, A)$ and:

$$G(A, B) = \frac{\sum_{ij}[1 - \theta(D - |r_i^B - r_j^B|)]\theta(D - |r_i^A - r_j^A|)|i - j|}{\sum_{ij}[1 - \theta(D - |r_i^B - r_j^B|)]\theta(D - |r_i^A - r_j^A|)}, \quad (7)$$

where $D = 7.5$ Å and $\theta(x)$ is the Heaviside step function.

Acknowledgements

Support was provided by a fellowship of the Bechtel foundation (to J.M.B.), NIH, NSF, and the Zenith Award of the Alzheimer Association, and the Petroleum Research Fund (to H.E.S.), Muscular Dystrophy Association grant MDA3720 and the March of Dimes Research grant no. 5-FY03-155 (to N.V.D.), and NIH RO1 GM52126 (to E.I.S.).

References

1. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9.
2. Jackson, S. E. (1998). How do small single-domain proteins fold? *Folding Des.* **3**, R81–R91.
3. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, **33**, 10026–10036.

4. Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.
5. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc. Natl. Acad. Sci. USA*, **99**, 8637–8641.
6. Dokholyan, N. V., Borreguero, J. M., Buldyrev, S. V., Ding, F., Stanley, H. E. & Shakhnovich, E. I. (2003). Identifying the importance of amino acids for protein folding from crystal structures. *Methods Enzymol.* **374**, 618–640.
7. Khare, S. D., Caplow, M. & Dokholyan, N. V. (2004). The rate and equilibrium constants for a multistep reaction sequence for the aggregation of superoxide dismutase in amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. USA*, **101**, 15094–15099.
8. Fersht, A. R. (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.
9. Ervin, J. & Gruebele, M. (2001). Quantifying protein folding transition states with $\Phi(T)$. *J. Biol. Phys.* **28**, 115–128.
10. Northey, J. G. B., Di Nardo, A. A. & Davidson, A. R. (2002). Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol.* **9**, 126–130.
11. Sosnick, T. R., Dothager, R. S. & Krantz, B. A. (2004). Differences in the folding transition state of ubiquitin indicated by Φ and Ψ analyses. *Proc. Natl. Acad. Sci. USA*, **101**, 17377–17382.
12. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways revealed and mapped. *Nat. Struct. Biol.* **10**, 658–662.
13. Kammerer, R. A., Kostrewa, D., Zurdo, J., Detken, A., Garcia-Echeverria, C., Green, J. D. *et al.* (2004). Exploring amyloid formation by a *de novo* design. *Proc. Natl. Acad. Sci. USA*, **101**, 4435–4440.
14. Ding, F., LaRocque, J. J. & Dokholyan, N. V. (2005). Direct observation of protein folding, aggregation, and a prion-like conformational conversion. *J. Biol. Chem.* **280**, 40235–40240.
15. Borreguero, J. M., Ding, F., Buldyrev, S. V., Stanley, H. E. & Dokholyan, N. V. (2004). Multiple folding pathways of the SH3 domain. *Biophys. J.* **87**, 521–533.
16. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003). Structural analysis of the rate-limiting transition states in the folding of lm7 and lm9: similarities and differences in the folding of homologous proteins. *J. Mol. Biol.* **326**, 293–305.
17. Teilum, K., Thormann, T., Caterer, N. R., Poulsen, H. I. & Jensen, P. H. (2005). Different secondary structure elements as scaffolds for protein folding transition states of two homologous four-helix bundles. *Proteins: Struct. Funct. Bioinf.* **59**, 80–90.
18. Martinez, J. C., Viguera, A. R., Berisio, R., Wilmanns, M., Mateo, P. L., Filimonov, V. V. & Serrano, L. (1999). Thermodynamic analysis of alpha-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. *Biochemistry*, **38**, 549–559.
19. Lindberg, M., Tangrot, J. & Oliveberg, M. (2002). Complete change of the protein folding transition state upon circular permutation. *Nat. Struct. Biol.* **9**, 818–822.
20. Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **1**, 177–196.
21. Onuchic, J. N., Nymeyer, H., García, A. E., Chahine, J. & Socci, J. N. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Adv. Protein Chem.* **53**, 152–870.
22. Li, L. & Shakhnovich, E. I. (2001). Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA*, **98**, 13014–13018.
23. Gsponer, J. & Caflisch, A. (2002). Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. USA*, **99**, 6719–6724.
24. Shea, J. E., Onuchic, J. N. & Brooks III, C. L. (2002). Probing the folding free-energy landscape of the src-SH3 protein domain. *Proc. Natl. Acad. Sci. USA*, **99**, 16064–16068.
25. Ozkan, S. B., Dill, K. A. & Bahar, I. (2003). Computing the transition state populations in simple protein models. *Biopolymers*, **68**, 35–46.
26. Kussell, E., Shimada, J. & Shakhnovich, E. I. (2003). Side-chain dynamics and protein folding. *Proteins: Struct. Funct. Genet.* **52**, 303–321.
27. Garbuzynskiy, S. O., Finkelstein, A. V. & Galzitskaya, O. V. (2004). Outlining folding nuclei in globular proteins. *J. Mol. Biol.* **336**, 509–525.
28. Hubner, I. A., Oliveberg, M. & Shakhnovich, E. I. (2004). Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. *Proc. Natl. Acad. Sci. USA*, **101**, 8354–8359.
29. Singhal, N., Snow, C. D. & Pande, V. S. (2004). Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* **121**, 415–425.
30. Henry, E. R. & Eaton, W. A. (2004). Combinatorial modeling of protein folding kinetics: free energy profiles and rates. *Chem. Phys.* **307**, 163–185.
31. Weikl, T. R. & Dill, K. (2003). Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* **332**, 953–963.
32. Chang, I., Cieplak, M., Banavar, J. R. & Maritan, A. (2004). What can one learn from experiments about the elusive transition state? *Protein Sci.* **13**, 2446–2457.
33. Sato, S., Religa, T. L., Daggett, V. & Fersht, A. R. (2004). Testing protein-folding simulations by experiment: B domain of protein A. *Proc. Natl. Acad. Sci. USA*, **101**, 6952–6956.
34. Du, R., Pande, V. P., Grosberg, Y., Tanaka, T. & Shakhnovich, E. S. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350.
35. Klimov, D. K. & Thirumalai, D. (2002). Stiffness of the distal loop restricts the structural heterogeneity of the transition-state ensemble in SH3 domains. *J. Mol. Biol.* **317**, 721–737.
36. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2002). Direct molecular dynamics observation of protein folding transition-state ensemble. *Biophys. J.* **83**, 3525–3532.
37. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183–1188.
38. Kameda, T. (2003). Importance of sequence specificity for predicting protein folding pathways: perturbed Gaussian chain model. *Proteins: Struct. Funct. Genet.* **53**, 616–628.
39. Karanicolas, J. & Brooks III, C. L. (2003). The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: lessons for protein design? *Proc. Natl. Acad. Sci. USA*, **100**, 3954–3959.

40. Lee, S. Y., Fujitsuka, Y., Kim, D. H. & Takada, S. (2004). Roles of physical interactions in determining protein folding mechanisms: molecular simulation of protein G and α-spectrin SH3. *Proteins: Struct. Funct. Bioinf.* **55**, 128–138.
41. Rao, F. & Caflisch, A. (2004). The protein folding network. *J. Mol. Biol.* **342**, 299–306.
42. Weikl, T. R., Palassini, M. & Dill, K. A. (2004). Cooperativity in two-state protein folding kinetics. *Protein Sci.* **13**, 822–829.
43. Borreguero, J. M., Dokholyan, N. V., Buldyrev, S. V., Shakhnovich, E. I. & Stanley, H. E. (2002). Thermodynamics and folding kinetics analysis of the SH3 domain from discrete molecular dynamics. *J. Mol. Biol.* **318**, 863–876.
44. Ding, F., Guo, W., Dokholyan, N. V., Shakhnovich, E. I. & Shea, J. E. (2005). Re-construction of the src-SH3 protein domain transition-state ensemble using multi-scale molecular dynamics simulations. *J. Mol. Biol.* **350**, 1035–1050.
45. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
46. Clementi, C., Jennings, P. & Onuchic, J. N. (2001). Prediction of folding mechanism for circular-permuted proteins. *J. Mol. Biol.* **311**, 879–8909.
47. Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **1**, 68–79.
48. Grantcharova, V. P., Riddle, D. S., Santiago, J. N. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src-SH3 domain. *Nat. Struct. Biol.* **8**, 714–720.
49. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Ruczinski, A. E. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* **6**, 1016–1024.
50. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src-Sh3 domain. *Biochemistry*, **36**, 15685–15692.
51. Hubner, I. A., Edmonds, K. A. & Shakhnovich, E. I. (2005). Nucleation and the transition state of the SH3 domain. *J. Mol. Biol.* **349**, 424–434.
52. Klimov, D. K. & Thirumalai, D. (2001). Multiple protein folding nuclei and the transition-state ensemble in two-state proteins. *Proteins: Struct. Funct. Genet.* **43**, 465–475.
53. Klimov, D. K. & Thirumalai, D. (2005). Symmetric connectivity of secondary structure elements enhances the diversity of folding pathways. *J. Mol. Biol.* **353**, 1171–1186.
54. Goldbeck, R. A., Thomas, Y. G., Chen, E., Esquerra, R. M. & Kliger, D. S. (1999). Multiple pathways on a protein-folding energy landscape: kinetic evidence. *Proc. Natl. Acad. Sci. USA*, **96**, 2782–2787.
55. Bieri, O. & Kiefhaber, T. (2001). Origin of apparent fast and non-exponential kinetics of lysozyme folding measured in pulsed hydrogen exchange experiments. *J. Mol. Biol.* **310**, 919–935.
56. Kamagata, K., Sawano, Y., Tanokura, M. & Kuwajima, K. (2003). Multiple parallel-pathway folding of proline-free staphylococcal nuclease. *J. Mol. Biol.* **332**, 1143–1153.
57. Fersht, A. R. (2000). A kinetically significant intermediate in the folding of barnase. *Proc. Natl. Acad. Sci. USA*, **97**, 14121–14126.
58. Fersht, A. R., Itzhaki, L. S., Elmasry, N. F., Matthews, J. M. & Otzen, D. E. (1994). Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl. Acad. Sci. USA*, **91**, 10426–10429.
59. Viguera, A. R., Jimenez, M. A., Rico, M. & Serrano, L. (1996). Conformational analysis of peptides corresponding to β-hairpins and a β-sheet that represent the entire sequence of the α-spectrin SH3 domain. *J. Mol. Biol.* **255**, 507–521.
60. Gnanakaran, S. & García, A. E. (2003). Folding of a highly conserved diverging turn motif from the SH3 domain. *Biophys. J.* **84**, 1548–1562.
61. 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.
62. Kiang, C. H. (2001). Single particle study of protein assembly. *Phys. Rev. E*, **64**, 041911-1–041911-3.
63. Viguera, A. R., Blanco, F. J. & Serrano, L. (1995). The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.* **247**, 670–681.
64. Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nat. Struct. Biol.* **10**, 874–880.
65. Hubner, I. A., Lindberg, M., Haglund, E., Oliveberg, M. & Shakhnovich, E. I. (2006). Common motifs and topological effects in the protein folding transition state. *J. Mol. Biol.* **359**, 1075–1085.
66. Lindberg, M., Haglund, E., Hubner, I. A., Shakhnovich, E. I. & Oliveberg, M. (2006). Identification of the minimal protein-folding nucleus through loop-entropy perturbations. *Proc. Natl. Acad. Sci. USA*, **103**, 4083–4088.
67. Paci, E., Vendruscolo, M. & Karplus, M. (2002). Native and non-native interactions along protein folding and unfolding pathways. *Proteins: Struct. Funct. Genet.* **47**, 379–392.
68. Rhee, Y. M. & Pande, V. S. (2006). On the role of chemical detail in simulating protein folding kinetics. *Chem. Phys.* **323**, 66–77.
69. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, **420**, 102–106.
70. Paci, E. (2002). High pressure simulations of biomolecules. *Biochim. Biophys. Acta.* **1595**, 185–200.
71. Zhou, Y., Karplus, M., Wichert, J. M. & Hall, C. K. (1997). Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte Carlo simulations of systems with square-well interactions. *J. Chem. Phys.* **107**, 10691–10708.
72. Gō, N. & Abe, H. (1981). Non-interacting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers*, **20**, 991–1011.
73. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Molecular dynamics studies of folding of a protein-like model. *Folding Des.* **3**, 577–587.
74. Zhou, Y. Q. & Karplus, M. (1999). Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. *J. Mol. Biol.* **293**, 917–951.
75. Rapaport, D. C. (1997). *The Art of Molecular Dynamics Simulation,* Cambridge University Press, Cambridge, UK.
76. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition-state ensemble and "enroute" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **5**, 937–953.