Topological determinants of protein folding

Nikolay V. Dokholyan*†, Lewyn Li*, Feng Ding§, and Eugene I. Shakhnovich*

*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138; and §Center for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215

Edited by Alan Fersht, University of Cambridge, Cambridge, United Kingdom, and approved April 18, 2002 (received for review February 7, 2002)

The folding of many small proteins is kinetically a two-state process that represents overcoming the major free-energy barrier. A kinetic characteristic of a conformation, its probability to descend to the native state domain in the amount of time that represents a small fraction of total folding time, has been introduced to determine to which side of the free-energy barrier a conformation belongs. However, which features make a protein conformation on the folding pathway become committed to rapidly descending to the native state has been a mystery. Using two small, well characterized proteins, CI2 and C-Src SH3, we show how topological properties of protein conformations determine their kinetic ability to fold. We use a macroscopic measure of the protein contact network topology, the average *graph connectivity*, by constructing graphs that are based on the geometry of protein conformations. We find that the average connectivity is higher for conformations with a high folding probability than for those with a high probability to unfold. Other macroscopic measures of protein structural and energetic properties such as radius of gyration, rms distance, solvent-accessible surface area, contact order, and potential energy fail to serve as predictors of the probability of a given conformation to fold.

The concept of the protein transition state ensemble (TSE) (1, 2) is the foundation of modern views on protein folding. Conformations of proteins belonging to the TSE are unstable and by definition have a 50% probability to fold to the protein native state and a 50% probability to unfold or misfold. The TSE conformations belong to the free-energy barrier separating native and unfolded or misfolded domains for two-state proteins. To understand the structure of the TSE conformations we must determine the difference between "pretransition" states (conformations that are *en route* to the native domain from the unfolded state but the transition barrier has not been crossed) and "posttransition" states (conformations that are *en route* to the unfolded domain from the native state but the transition barrier has not been crossed). The distinguishing kinetic feature between pre- and posttransition conformations is their probability to reach the native state domain, p_{FOLD} (3). Because in the posttransition conformations the nucleus (4) is not disrupted, these conformations are more probable to fold than pretransition conformations in which the nucleus is not formed. If both pre- and posttransition states are structurally and energetically close to the TSE, the question then is what global properties distinguish these states from each other?

To answer this question, we selected pre- and posttransition states (see *Methods*) of two proteins: chymotrypsin inhibitor 2 (CI2) and C-Src SH3 domain, by using two different approaches (ref. 5 and F.D., N.V.D., S. Buldyrev, H. E. Stanley, and E.I.S., unpublished data) and two different simulation techniques (6, 7). Both the CI2 and C-Src SH3 domain proteins have been extensively studied experimentally (8–10), and both are known to be two-state proteins. We verify that the p_{FOLD} of the selected pre- and posttransition conformations is $\approx$0 and 1 correspondingly for both proteins (Table 1).

We find that such structural properties of protein conformations as radius of gyration (R_G), rms displacement (RMSD) from the native state, solvent-accessible surface area, and contact order (11) cannot distinguish the pre- and posttransition conformations (Table 1). Correspondingly, the entropy of the pre- and posttransition conformations cannot account for the difference between these conformations. We also find that the potential energies (E) and the total number of contacts between amino acids are within error bars from each other in the pre- and posttransition conformations. If the pre- and posttransition conformations are similar to each other structurally, we hypothesize that there may be a difference in the topology of the network of amino acid interactions in these conformations.

To study the topology of pre- and posttransition conformations, we construct graphs corresponding to these conformations in which nodes represent amino acids and edges represent those pairs of amino acids that are geometrically located within interaction distance from each other. Vendruscolo *et al.* (12) have shown recently that the "small-world" feature (13–15) of proteins can be used to identify the key residues that stabilize the structure of the transition state. Our hypothesis is that the network of amino acid interactions in posttransition conformations is more small world-like (13–15) than that in pretransition conformations. The small-world graphs are a special class of random graphs that are connected as strongly as regular graphs (the clusters have a similar structure to regular graphs), but the average path that spans two nodes via a minimal set of graph edges is as low as that for random graphs (refs. 12 and 16 and figure 2 of ref. 13). The difference between regular, small-world, and random graphs is the "wiring" of these graphs: regular graphs are connected strongly locally, with no long-range edges; random graphs are disconnected locally but have many long-range edges; and small-world graphs are the blend of the high local connectivity with a number of the long-range contacts. Small-world graphs are characterized by small separation of nodes from each other, which for proteins means a higher degree of *interaction cooperativity*. Thus, we hypothesize that the wiring of the posttransitional conformation graphs is "tighter" than that of the pretransition conformation graphs, resulting in a cooperative folding to the native state domain.

To measure the wiring properties of pre- and posttransition conformation graphs, we compute the average minimal distance L between any pair of nodes of a graph by counting the minimal set of edges that connect these nodes (13). We find that the L values for posttransition conformation graphs are distinctly smaller than those for the pretransition conformation graphs, thus fully supporting our hypothesis (Table 1). We also observe that the posttransition conformation graphs have more edges that are of intermediate and long range than pretransition ones (Fig. 1), which shortens the minimal path for each node k, $L(k)$ (Fig. 2), thus creating a more cooperative network for the former graphs. A similar mechanism was observed by Watts and Strogatz (13), who, by rewiring circular graphs by removing local edges and creating a few long-range edges, were changing the graph properties from the regular to the small world. Interestingly, some CI2 pretransition conformations have N and C termini in contact, in contrast to posttransition conformations (Figs. 1 and 2). Although the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: TSE, transition state ensemble; RMSD, rms displacement.

†To whom reprint requests should be addressed. E-mail: dokh@wild.harvard.edu.

Table 1. The structural [R_G, RMSD, solvent-accessible surface area (SASA), contact order, and number of contacts], energetic (*E*), and topological properties (*L*) of pre- and posttransition states of CI2 and C-Src SH3 domain proteins

Protein	Relation to TSE	Number of conf.	p_{FOLD}	R_G, Å	RMSD, Å	SASA, $\times 10^3$ Å^2	Contact order, %	No. of contacts	*E*	*L*
CI2	post-	20	0.89 ± 0.07	13.0 ± 0.5	5.2 ± 0.9	6.5 ± 0.2	19 ± 1	183 ± 5	−102 ± 13	3.5 ± 0.1
	pre-	6	0.02 ± 0.04	13.0 ± 0.2	5.9 ± 0.3	7.1 ± 0.3	19 ± 2	171 ± 4	−130 ± 19	4.4 ± 0.4
C-Src	post-	10	0.96 ± 0.01	11.2 ± 0.3	4.9 ± 0.3	4.5 ± 0.1	22 ± 3	110 ± 4	−85 ± 2	2.73 ± 0.03
SH3	pre-	10	0.26 ± 0.08	11.8 ± 0.3	4.7 ± 0.1	4.4 ± 0.1	16 ± 2	102 ± 7	−84 ± 3	3.31 ± 0.06

The values of p_{FOLD} correlate only with *L* values; the posttransition states are characterized by $p_{FOLD} \approx 1$, and their *L* values are smaller than that for the pretransition states, which are characterized by $p_{FOLD} \approx 0$. conf., conformations.

contact between the N and C termini is of the longest range, the lack of intermediate-range contacts nevertheless makes pretransition conformation networks less "cooperative" than posttransition ones. The difference between the numbers of long-range contacts in pre- and posttransition conformations is not statistically significant, thus the average contact orders for both conformation ensembles cannot discriminate between pre- and posttransition ensembles.

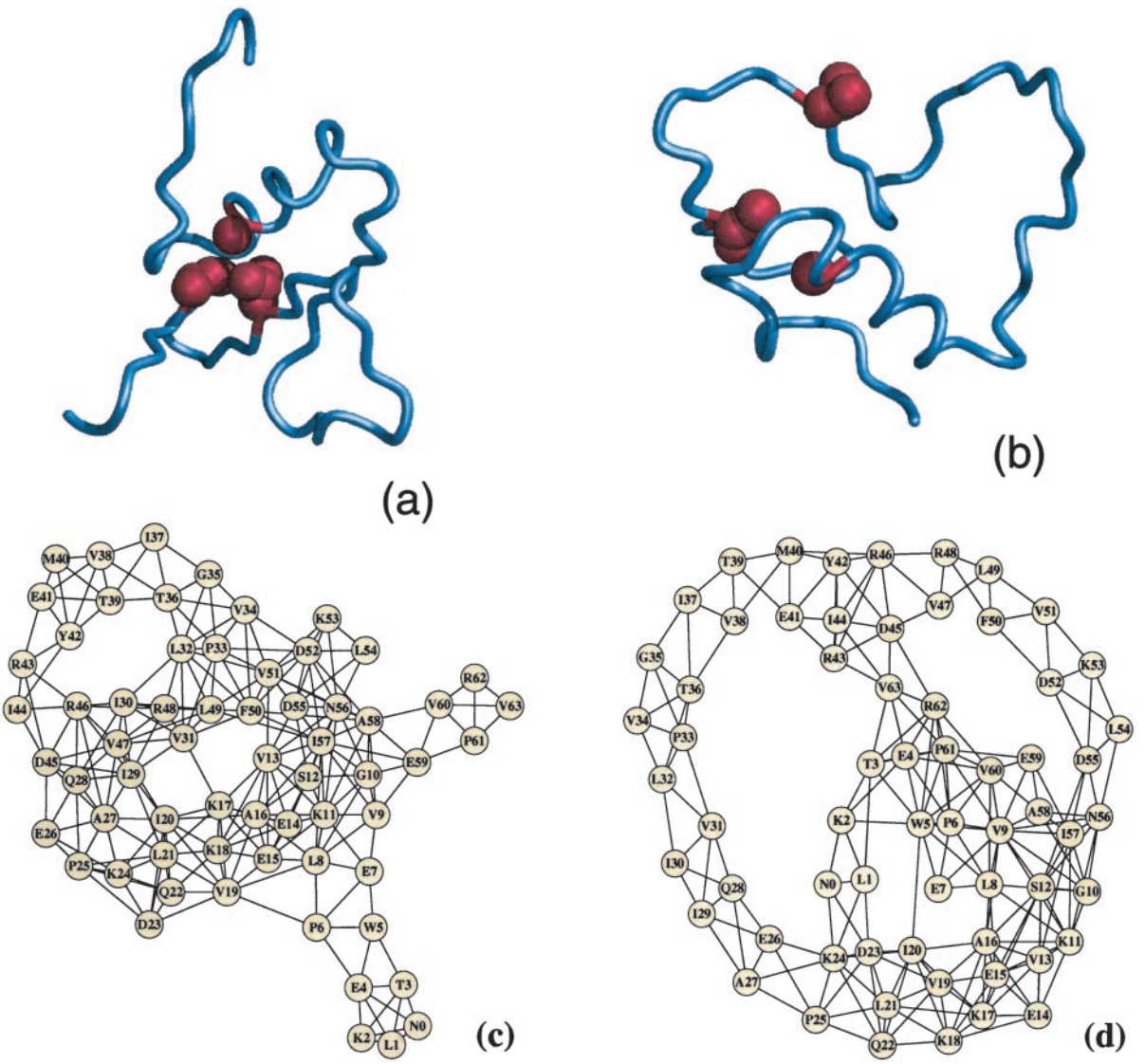

Fig. 1. The three-dimensional structure of the CI2 protein in post- (*a*) and pretransition (*b*) states. The protein graphs are constructed based on the structure of post- (*c*) and pretransition (*d*) states. Each node of protein graphs corresponds to an amino acid, whereas each edge between a pair of nodes corresponds to that pair of amino acids that are geometrically in contact with each other. For both CI2 and C-Src SH3 domain proteins' graph constructions, the contact between two amino acids is considered to be present if the distance between corresponding C_α atoms is less than 8.5 Å. In *a* and *b*, residues A16, L49, and I57 belonging to the specific nucleus of CI2 (8) are denoted by red spheres. A16, L49, and I57 form a triad of contacts in posttransition conformations (*a*), whereas such contacts are missing in the pretransition conformations. In both pre- and posttransition states the number of edges (contacts) are approximately the same.

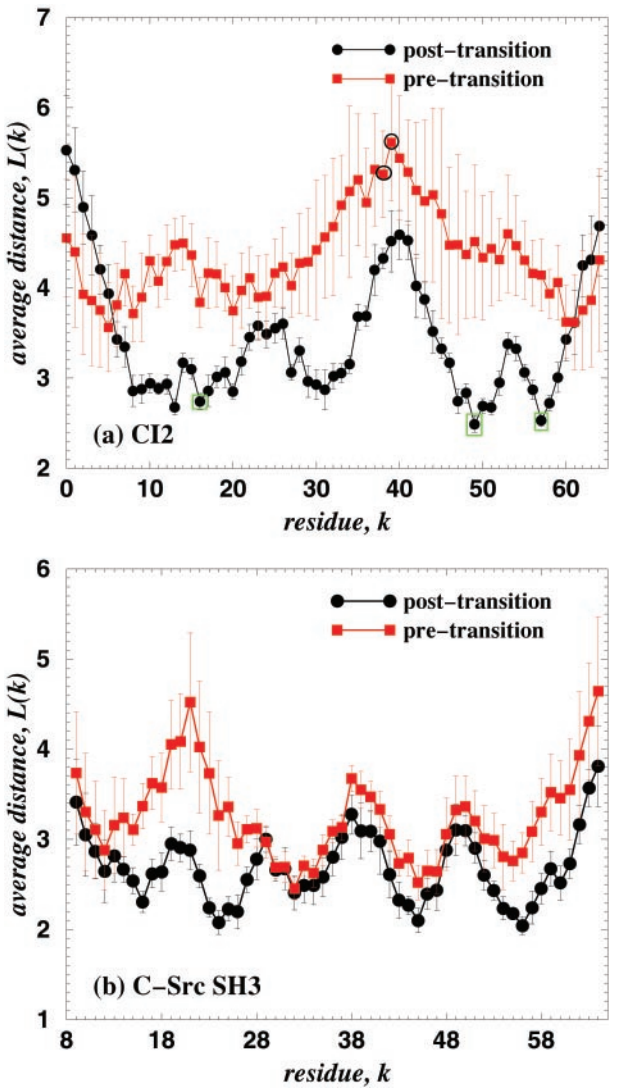

Fig. 2. The dependence of the average minimal distance $L(k)$ between a node k and the rest of the nodes on CI2 (*a*) and C-Src SH3 domain (*b*) proteins' graphs for post- (●) and pretransition (■) states. The error bars represent the standard deviation from the average values of $L(k)$ over all post- and pretransition states. In *a*, by the open circles (○) we denote amino acids M40 and E41 that do not affect the protein three-dimensional structure after cleavage of the 40–41 bond (19), and by the open boxes (□) we denote the folding nucleus of CI2 (8), A16, L49, and I57.

An important property of the L values of protein conformations is that they can serve as a structurally reliable determinant of the pre- ($p_{\mathrm{FOLD}} \approx 0$) and posttransition ($p_{\mathrm{FOLD}} \approx 1$) states. The principal difficulty in selecting TSE conformations, the basis of the protein-engineering experiments, is the identification of the reaction coordinate for protein folding. The reaction coordinate for folding is not well defined (3, 17, 18) and has yet to be identified. The fact that average graph connectivity distinguishes the protein pre- and posttransition states, which can be close along the reaction coordinate to the TSE, tells us that any future constructions of the reaction coordinate should strongly depend on the structure of protein interaction networks.

Interestingly, in experimental studies of CI2, the cleavage between amino acids M40 and E41 is the only one that does not destroy the protein's three-dimensional structure (19). Neira *et al.* (19) cut CI2 at M40–E41 (without circular permutation) to separate fragments 1–40 and 41–64 and found that these fragments reassociate into CI2. We find that amino acids M40 and E41 have the largest values of $L(k)$ in the pretransition states and among the largest values in the posttransition states (Fig. 2*a*), indicating that these amino acids are the most separated on interaction network from the rest of the amino acids. Weak participation of amino acids in the protein interaction network in pre- and posttransition states means that these amino acids have weak impact on protein folding kinetics and on the final native state of the protein (because the folding pathway is not altered). Thus, our findings are in agreement with ref. 19.

A crucial factor that distinguishes pre- and posttransition states is the protein folding nucleus, the formation of which in the TSE results in the rapid folding transition to the native state, and the disruption of which results in the global unfolding (4). Pretransition states lack the folding nucleus, whereas posttransition states have it intact (Fig. 1 *a* and *b*). Thus, the difference of $L(k)$ between the pre- and posttransition states, $\Delta L(k)$, is most pronounced for those amino acids that are part of the protein folding nucleus. We find that for CI2 (Fig. 2*a*), the experimentally identified folding nucleus (8), A16, L49, and I57, has one of the largest $\Delta L(k)$ values.

We also find that for C-Src SH3 domain (Fig. 2*b*) $\Delta L(k)$ is most pronounced for two fragments, RT–loop (16–26) and β4 (54–61), suggesting a crucial role of the connectivity between these fragments in the TSE. This observation is in agreement with the finding of F.D., N.V.D., S. Buldyrev, H. E. Stanley, and E.I.S., (unpublished data), in which the nucleus of C-Src SH3 domain is identified on the RT–loop and β4.

The evolution of the protein graphs from pre- to posttransition states may be the key to better understanding the protein folding dynamics. It is possible that the formation of a specific nucleus (17) is the consequence of "specific rewiring" of protein graphs when a protein crosses the free-energy barrier en route to its folded state. Further studies are necessary to shed light on the relation of protein graph properties to the formation of a specific nucleus. The evolution of various networks has been extensively studied recently. Barabasi and Albert (14) recently proposed a model to explain scale-free networks, in which the distribution of the number of edges per node (node degree) scales as a power law. Their model is based on the idea of "preferential attachments": the most connected nodes are more probable to acquire new edges with other graphs' nodes in the course of graph evolution. Although because of the finite size of protein graphs there is no evidence that our graphs are scale-free, we test whether evolution of the protein graphs during protein folding follows the preferential attachment scenario. We find (Fig. 3) that most connected residues actually decrease the number of edges rather than increase them (e.g., A16). In addition, we find that some residues that were less connected in the pretransition states become more connected (e.g., refs. 47–51). The correlation between change in the node degrees between post- and pretransition states and posttransition states is 0.61. We believe that when protein crosses its folding transition barrier, the network topology changes toward a specific one. Such rewiring herds the topology into a less random conformation. Thus, we observe specific rewiring rather than a preferential attachment.

In this work we presented a new structure-based topological criterion that seems to be a good predictor of kinetic ability to fold for a given conformation. The fact that this criterion performed equally well for two different proteins, simulated within different models by using different techniques, suggests its generality. Moreover, our recent all-atom Monte Carlo analysis of TSE of protein G (J. Shimada and E.I.S., unpublished data) also shows consistency with the proposed criterion. Further theoretical understanding of the deep connection between topological properties of protein conformations and their kinetic ability to fold is a challenging task for future studies.

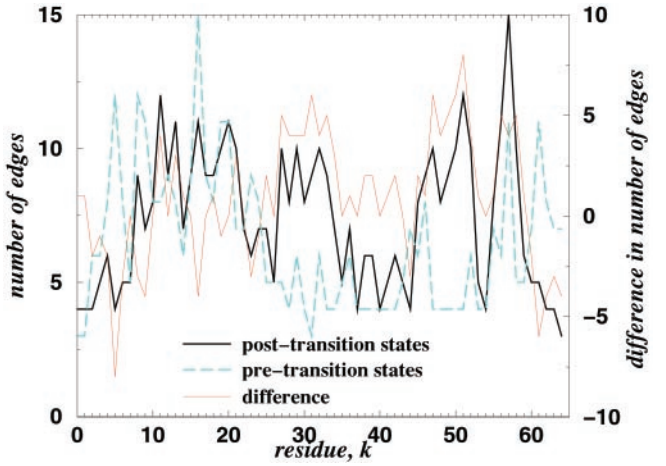

Fig. 3. Plot of the degrees of each node versus the residue number for post- (thick solid line) and pretransition (thick broken line) states for CI2. The thin line represents the difference in node degrees between pre- and posttransition states.

Appendix: Methods

CI2. We perform Monte Carlo simulations of the all-atom model of CI2. The all-atom Monte Carlo has been described in detail elsewhere (7). We identify our putative TSE by using the method of Vendruscolo *et al.* (5). In our simulations, we approximate the experimentally determined ϕ values of amino acids ψm by the ratio of the number of contacts each amino acid makes in the TSE, $N^{\ddagger}$, to that in the native state of CI2, $N^{\rm NS}$,

$$\phi_{\rm sim} \equiv N^{\ddagger}/N^{\rm NS}. \qquad [1]$$

To generate the putative TSE, we perform unfolding simulations from the native structure of CI2 at $T = 2.3$ by using the form of the potential energy of amino acid interactions

$$E = E_{G\bar{o}} + \Lambda \cdot \sum_{k=1}^{N} [\phi_{\rm sim}(k) - \phi_{\rm exp}(k)]^2, \qquad [2]$$

where $E_{G\bar{o}}$ is the Gō potential energy (20–22), where the attractive potential ($E_{G\bar{o}} = -1$) between residues is assigned to the pairs that are in contact in the native state, and the repulsive potential ($E_{G\bar{o}} = +1$) is assigned to the pairs that are not in contact in the native state. The parameter $\Lambda = 10^3$ is set to a large value compared with the Gō potential to enforce the environment of each amino acid, k, in our simulations to be close to that observed experimentally. The $\phi_{\rm exp}$ are the experimental ϕ values at 0 M GdHCl from Itzhaki *et al.* (8) for the following 39 mutations: K2M, T3A, P6A, E7A, L8A, S12A, E14N, E15Q, A16G, K17A, K18G, I20V, L21A, Q22G, K24G, P25A, E26A, I29A, I30A, L32A, V34T, T36V, V38A, T39A, E41A, Y42G, R43G, D45A, V47A, L49A, F50A, V51A, D52A, N56D, I57A, A58G, V60G, P61A, and V63A. ψm for I57A has been set to 0.5, because this residue is part of the nucleus despite its low ϕ value (23). The summation is over the $n = 39$ positions stated above.

For each putative TSE conformation (40 in total) we compute the probability to fold to its native state by performing 20 independent simulations at $T = 1.2$ for 5×10^7 Monte Carlo steps, which is less than 5% of the folding time from a random coil (data not shown; ref. 24). We consider the system folded when the RMSD for the backbone of the protein is less than 1Å. The average probability of folding $p_{\rm FOLD}$ is 0.59, which confirms that the selected conformations do belong to the TSE. In Table 1 we present the data for the 20 conformations with $p_{\rm FOLD} \geq 0.8$, which represent posttransition states.

We construct the pretransition conformations in a similar way to the putative TSE conformations except that the $\phi_{\rm exp}$ values are set artificially and are not taken from experiments (24). We choose $\phi_{\rm exp}$ to create structures with approximately the same energy and RMSD as the TSE conformations (Table 1). For the construction of the pretransition conformations we choose the $\phi_{\rm exp}$ to be 0.5 for K2, E4, E7, L8, K11, E15, K18, V19, and D23 and 0.4 for V60, P61, R62, and V63. To create pretransition conformations that differ from each other we alternate the $\phi_{\rm exp}$ values between these conformations. In Table 1 we present the data for the six conformations with $p_{\rm FOLD} \approx 0$, which represent pretransition states.

C-Src SH3 Domain. We model the C-Src SH3 domain by beads representing C_α and C_β atoms. To mimic the flexibility of real proteins, we apply additional constraints (F.D., N.V.D., S. Buldyrev, H. E. Stanley, and E.I.S., unpublished data): (*i*) "covalent" bonds between $C_{\alpha i}$ and $C_{\beta i}$, (*ii*) "peptide" bonds between $C_{\alpha i}$ and $C_{\alpha(i\pm1)}$, (*iii*) effective bonds between $C_{\beta i}$ and $C_{\alpha(i\pm1)}$, (*iv*) effective bonds between $C_{\alpha i}$ and $C_{\alpha(i\pm2)}$, where the subscript i denotes the amino acid sequence number. We use the Gō potential (20–22) to model interactions between C_β atoms of the C-Src SH3 domain. It has been shown (F.D., N.V.D., S. Buldyrev, H. E. Stanley, and E.I.S., unpublished data) that our model of the C-Src SH3 domain can reproduce faithfully the thermodynamic and kinetic properties observed in experiments (9, 10).

To identify the TSE and then the pre- and posttransition states, we follow the method developed in ref. 32 (F.D., N.V.D., S. Buldyrev, H. E. Stanley, and E.I.S., unpublished data). We perform the discrete molecular dynamics simulation of the C-Src SH3 domain at the folding transition temperature, $T_{\rm f} \approx 0.92$. We select conformations that belong to the putative TSE from those with the potential energy in the range $\{E_{\rm TS}\}$, corresponding to the minimum of the probability of the potential energy histogram at $T = T_{\rm f}$ (figure 1a in ref. 4). We distinguish four types of fluctuations during the simulation of the C-Src SH3 domain that pass through the unstable states within energies in the range $\{E_{\rm TS}\}$: (*i*) FF, when the folded protein unfolds to $\{E_{\rm TS}\}$ and then refolds rapidly to its native state, (*ii*) UU, when the unfolded protein partly folds into $\{E_{\rm TS}\}$ and then unfolds rapidly, (*iii*) FU, when the folded protein unfolds to $\{E_{\rm TS}\}$, and then proceeds unfolding further, and (*iv*) UF, when the unfolded protein traverses the energy range $\{E_{\rm TS}\}$ on its way to folded conformations.

Next, we compute $p_{\rm FOLD}$ for each selected conformation by performing 10^2 independent simulations for 2×10^3 time units at $T_{\rm f}$. The average time the C-Src SH3 domain spends in the unfolded state is 10^5 time units, whereas the average and maximum times of C-Src SH3 domain folding from an unfolded state at which C-Src SH3 domain is already committed to the folding transition is 10^2 and 10^3 time units correspondingly.

We find that the $p_{\rm FOLD}$ of the most representative UF and FU conformations are $\approx$0.5, which suggests that these conformations belong to the TSE. For UU and FF conformations, we find that the $p_{\rm FOLD}$ values are $\approx$0 and $\approx$1, respectively. Thus, we select the pre- and posttransition conformations from the UU and FF ensembles of conformations correspondingly.

The Protein Graphs. The protein graphs are constructed on the basis of the C_α representation of proteins. Each graph node represents an amino acid. Each graph edge connects pairs of nodes that correspond to pairs of amino acids that are located geometrically within an interaction threshold radius, which we set to $R_{\rm c} = 8.5$ Å. We test graph connectivity properties for

various definitions of contacts and find that these properties are qualitatively invariant under contact definitions.

The average minimal distance $L(k)$, also known as *the chemical distance* (25), between a node k and the rest of the graph nodes is defined as

$$L(k) \equiv \left\langle \frac{1}{N-1} \sum_{j=1}^{N} \ell_{kj} \right\rangle, \qquad [3]$$

where N is the number of nodes in the graph or number of amino acids in proteins, ℓ_{kj} is the minimal number of edges one must transverse to reach a node j from a node $k \neq j$. The averaging is done over all pre- or posttransition states. Analogously, we define the average minimal distance L between all pairs of nodes of the protein graph

$$L \equiv \left\langle \frac{1}{N(N-1)} \sum_{k \neq j = 1}^{N} \ell_{kj} \right\rangle. \qquad [4]$$

For example, the minimal distance between nodes N1 and V20 in the graph of Fig. 1*c* is 3, the path corresponding to the minimal distance passes through nodes E5 and P7: N1 → E5 → P7 → V20.

We thank J. Shimada for helpful discussion of the folding kinetics of protein G. We acknowledge S. V. Buldyrev for help with simulations of C-Src SH3 and M. Karplus and M. Vendruscolo for critical reading of the manuscript. This work is supported by National Institutes of Health Grant GM52126 (to E.I.S.) and National Institutes of Health National Research Service Award Fellowship GM20251 (to N.V.D.).

1. Fersht, A. R. (1995) *Curr. Opin. Struct. Biol.* **5,** 79–84.
2. Mirny, L. A. & Shakhnovich, E. I. (2001) *Annu. Rev. Biophys. Biomol. Struct.* **30,** 361–396.
3. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. I. (1998) *J. Chem. Phys.* **108,** 334–350.
4. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000) *J. Mol. Biol.* **296,** 1183–1188.
5. Vendruscolo, M., Paci, E., Dobson, C. & Karplus, M. (2001) *Nature* (*London*) **409,** 641–645.
6. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998) *Folding Des.* **3,** 577–587.
7. Shimada, J., Kussell, E. L. & Shakhnovich, E. I. (2001) *J. Mol. Biol.* **308,** 79–95.
8. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 260–288.
9. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998) *Nat. Struct. Biol.* **5,** 714–720.
10. Riddle, D.S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999) *Nat. Struct. Biol.* **6,** 1016–1024.
11. Plaxco, K. W., Simons, K. T. & Baker, D. (1998) *J. Mol. Biol.* **277,** 985–994.
12. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002) *Phys. Rev. E*, in press.
13. Watts, D. J. & Strogatz, S. H. (1998) *Nature* (*London*) **393,** 440–442.
14. Barabasi, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
15. Jeong, H., Tombor, B., Albert, R., Oltval, Z. N. & Barabasi, A. L. (2000) *Nature* (*London*) **407,** 651–654.
16. Bollobás, B. (1985) *Random Graphs* (Academic, London).
17. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *Biochemistry* **33,** 10026–10036.
18. Klimov, D. K. & Thirumalai, D. (2001) *Proteins Struct. Funct. Genet.* **43,** 465–475.
19. Neira, J. L., Davis, B., Ladorner, A. G., Buckle, A. M., Gay, G. D. & Fersht, A. R. (1996) *Folding Des.* **1,** 189–208.
20. Gō, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12,** 183–210.
21. Zhou, Y. & Karplus, M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 14429–14432.
22. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000) *J. Mol. Biol.* **278,** 937–953.
23. Ladurner, A. G., Itzhaki, L. S. & Fersht, A. R. (1997) *Folding Des.* **2,** 363–368.
24. Li, L. & Shakhnovich, E. I. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13014–13018.
25. Havlin, S. & Ben-Avraham, D. (1987) *Adv. Phys.* **36,** 695–798.