doi:10.1016/j.jmb.2005.05.017 J. Mol. Biol. (2005) 350, 1035–1050

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Reconstruction of the src-SH3 Protein Domain Transition State Ensemble using Multiscale Molecular Dynamics Simulations

Feng Ding[1]†, Weihua Guo[2]†, Nikolay V. Dokholyan[1] Eugene I. Shakhnovich[3] and Joan-Emma Shea[2]*

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

[2]*Department of Chemistry and Biochemistry, University of California, Santa Barbara CA 93106, USA*

[3]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA*

**Corresponding author*

We use an integrated computational approach to reconstruct accurately the transition state ensemble (TSE) for folding of the src-SH3 protein domain. We first identify putative TSE conformations from free energy surfaces generated by importance sampling molecular dynamics for a fully atomic, solvated model of the src-SH3 protein domain. These putative TSE conformations are then subjected to a folding analysis using a coarse-grained representation of the protein and rapid discrete molecular dynamics simulations. Those conformations that fold to the native conformation with a probability (P_{fold}) of approximately 0.5, constitute the true transition state. Approximately 20% of the putative TSE structures were found to have a P_{fold} near 0.5, indicating that, although correct TSE conformations are populated at the free energy barrier, there is a critical need to refine this ensemble. Our simulations indicate that the true TSE conformations are compact, with a well-defined central β sheet, in good agreement with previous experimental and theoretical studies. A structured central β sheet was found to be present in a number of pre-TSE conformations, however, indicating that this element, although required in the transition state, does not define it uniquely. An additional tight cluster of contacts between highly conserved residues belonging to the diverging turn and second β-sheet of the protein emerged as being critical elements of the folding nucleus. A number of commonly used order parameters to identify the transition state for folding were investigated, with the number of native C^{β} contacts displaying the most satisfactory correlation with P_{fold} values.

Keywords: protein folding; transition state ensemble; folding nucleus; molecular dynamics simulations; src-SH3 protein domain

Introduction

Most small, single-domain proteins fold in a two-state manner, populating either the unfolded or the folded state, but not any detectable, partly structured intermediate state.[1,2] Much of the effort aimed at understanding the folding mechanism of such proteins has focused on characterizing the transition state for folding. Given "the" reaction coordinate for folding, this transition state presents itself as the maximum in the free energy barrier separating the native and denatured state. This barrier arises from the incomplete cancellation of the entropic and enthalpic contributions to folding.[3–6] Folding can proceed through a multiplicity of routes, and the transition state consists of an ensemble of structures rather than of a single conformation. Despite decades of effort, the identification of this transition state ensemble (TSE) still poses serious challenges, both experimentally and computationally.[7–9]

The main experimental methodology for characterizing the folding transition state is the φ-value analysis developed by Fersht and co-workers.[1,10,11] The φ-value probes the extent of destabilization

† F.D. & W.G. contributed equally to this work.

Abbreviations used: TS, transition state; TSE, transition state ensemble; DMD, discrete molecular dynamics.

E-mail address of the corresponding author: shea@chem.ucsb.edu

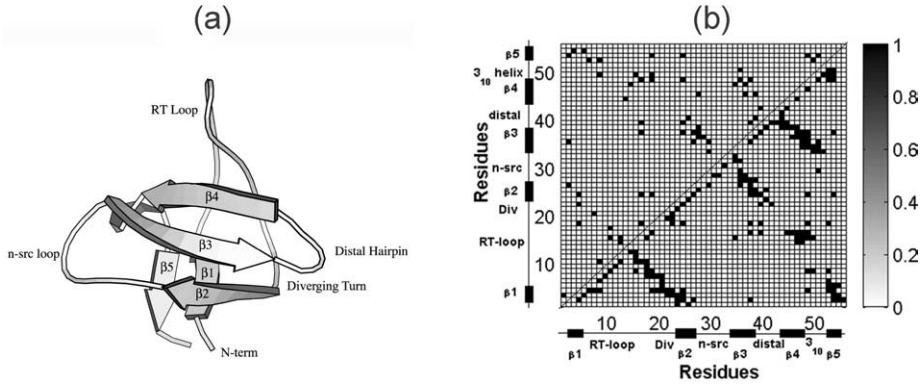

Figure 1. A cartoon of (a) the src-SH3 protein domain, ((b) upper quadrant) native side-chain contact map and ((b) lower quadrant) native C^{β} contact map. A native contact is defined between two residues if the center of geometry of the side-chains is within 6.5 Å. A native C^{β} contact is defined if the C^{β} atoms of the two residues (C^{α} is used for Gly) are within 7.5 Å. A total of 57 native side-chain contacts and 162 C^{β} contacts are obtained in this manner.

of the transition state upon mutation of a given side-chain, $\phi = (\Delta\Delta G_{T-D}/\Delta\Delta G_{N-D})$, where $\Delta\Delta G_{T-N}$ is the free energy change between the transition and the native state and $\Delta\Delta G_{D-N}$ is the free energy change between the denatured and the native state upon mutation. A ϕ-value of 0 indicates that the conformation is as unstructured in the transition state as it is in the denatured state, while a ϕ-value of 1 indicates as much structure as in the native state at the site of the mutation.[1] Intermediate ϕ-values are notoriously difficult to interpret, as they indicate either partially formed structure or the presence of different transition state structures belonging to different folding pathways. The meaning of an abnormal (>1 or negative) ϕ-value is also a contentious issue.[12] The validity of experimental ϕ-values for which $\Delta\Delta G_{D-N} < 1.7$ kcal/mol has recently been brought into question by Kiefhaber,[13] with a rebuttal by Fersht[11] stating that reliable ϕ-values can be obtained in the range $0.6 < \Delta\Delta G_{D-N} < 1.7$ kcal/mol.

Computer simulations are critical complements to the ϕ-values analysis, as they can provide additional atomistic detail of the transition state.[14–23] Significant success has been achieved in locating the transition state for simplified protein models; however, the identification of the TSE when both the protein and solvent are described explicitly poses a formidable computational challenge. While a full kinetic study of folding involving several hundred simulations from a random coil to the folded state is impractical, the thermodynamics of folding can be characterized using special sampling techniques, such as a combination of high-temperature unfolding and umbrella sampling,[24] or replica exchange molecular dynamics methods.[25–28] In principle, the TSE can be obtained from the maximum in the free energy surface projected onto the reaction coordinate for folding. In practice, we do not know the "correct" reaction coordinate and can hence only infer a putative TSE by projecting the free energy surface on aptly chosen order parameters characteristic of the folding reaction. This choice of order parameters which simplifies the problem from $3N-6$ dimensions to a few dimensions causes a loss of information about the true TSE. The structures residing at the top of the free energy barrier in simulations include structures that may not belong to the true TSE.

The purpose of the present research is to identify a putative TSE from free energy surfaces projected onto chosen order parameters for folding and then further refine this putative TSE to extract the true transition state structures. This refinement can be performed using the probability of folding (P_{fold}) analysis method.[29] From a kinetics point of view, structures belonging to the transition state should have equal probability (0.5) of proceeding either to the folded or unfolded basins. Hence, an appropriate test for whether a structure belongs to the TSE would involve launching a series of simulations to see whether this $P_{fold} = 0.5$ criteria is fulfilled. It is still a challenging task to compute the P_{fold} value of a given conformation by multiple all-atom molecular dynamics simulations. While traversing the transition state is expected to occur 100 to 1000 times faster than the actual folding time for the protein,[30] this event is still prohibitively long to simulate for proteins that fold on time-scales of milliseconds or longer. In recent simulations, Pande and co-workers implemented a 5 ns cutoff for P_{fold} calculations of the microsecond folding miniprotein BBA5, which would translate to simulations of the order of microseconds for a millisecond folder.[31] In addition, a reliable characterization of the transition state using the P_{fold} criterion would require multiple folding runs from each of the multiple putative transition state structures. Hence, rather than using fully-atomic, solvated simulations to determine P_{fold}, we employ the discrete molecular dynamics simulations,[32,33] using a Go interaction model to calculate the P_{fold} values for the selected putative TSE conformations.[20] This exercise would be impractical if all conformations occurring during a folding trajectory were considered. By narrowing the range of test structures

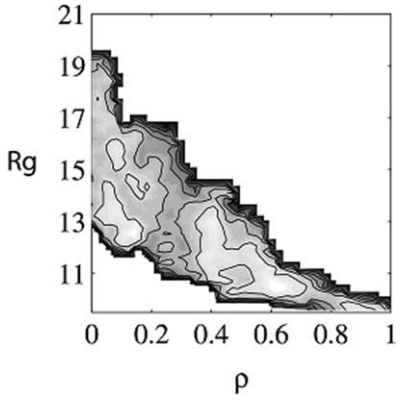

Figure 2. Free energy surface at 343 K as a function of the fraction of native contacts ρ and the radius of gyration R_g. Contour lines are drawn every 1 kcal/mol.

by using only those belonging to the putative TSE, the P_{fold} method becomes a viable means to identify transition state structures.

The subject of our study is the src-SH3 protein domain (PDB code 1SRL). The protein domain has a 56 residue β-barrel structure (Figure 1(a)), consisting of two hydrophobic sheets, packed orthogonally to form the hydrophobic core of the protein. The first sheet consists of the three central strands of the protein (β2-β3-β4) and the second sheet of the two terminal strands (β1 and β5) and a portion of the RT loop. There is also a small 3_{10} helix between β4 and β5. Due to its small size and multiple homologues, it has been the target of extensive experimental and theoretical studies.[20,34–39] Experimentally, this protein folds with kinetic and thermodynamic signatures of a two-state folder, without any detectable intermediates.[34] The φ-value studies have revealed an unusually polarized transition state for src-SH3, in which only the first hydrophobic sheet (β2-β3-β4) is highly structured (high φ-values), while the rest of the protein appears mostly unstructured (intermediate to low φ-values).[40] Theoretical work using both coarse-grained and atomically detailed models were found to be in good agreement with experimental findings.[20,36,40,41]

In earlier work, we located a putative TSE for a fully atomic model of the src-SH3 protein domain in explicit solvent from free energy surfaces generated by importance sampling molecular dynamics.[39] Conformations belonging to this putative TSE were in good agreement with both experimental and other theoretical studies. The structural characteristics of the putative TSE do not vary significantly with temperature, with the folding temperature for this protein being around 343 K.[42] In the present work, we subject the structures belonging to the putative TSE to P_{fold} analysis using a coarse-grained model of src-SH3. This analysis enables the extraction of the true transition state conformations and the identification of the key contacts distinguishing pre- and post- from true structures. A detailed discussion of the nature of the true TSE and the reliability of commonly used order parameters to identify TSE conformations is presented in Results and Discussion. We refer the reader to Methods for details of the protein model and simulations.

Results and Discussion

Structural features of the putative TSE from the free energy landscape

The folding free energy landscape of the src-SH3 domain generated by importance sampling

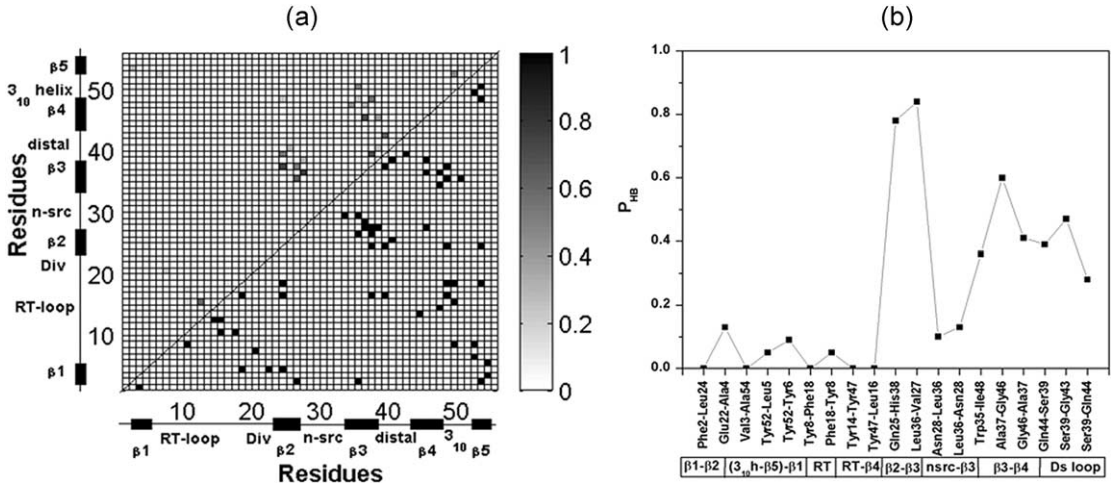

Figure 3. (a) Probability map of native side-chain contacts obtained from the structures residing at the free energy barrier of ρ = 0.3 (left-hand quadrant). The native contact map is shown in the right-hand quadrant. (b) Probability of forming native hydrogen bonds obtained from the structures residing at the free energy barrier of ρ = 0.3. Both (a) and (b) show that structure is present in the first hydrophobic sheet of the protein consisting of the central β2-β3-β4 strands, whereas the rest of the protein is less structured.

molecular dynamics is plotted in Figure 2 as a function of the fraction of native contacts ρ and the radius of gyration R_g. The free energy surface presents two barriers, a major one of 2.5 kcal/mol ($3.5k_BT$) located at ρ=0.3 and a minor one of 1 kcal/mol ($1.4k_BT$) located around ρ=0.8. The major barrier is attributed to the putative transition state,[39] while the second barrier corresponds to the desolvation of the hydrophobic core occurring in the late stage of folding.[39,42] This second barrier is more pronounced at low temperatures than at high temperatures,[42] consistent with recent findings by Chan and co-workers that elementary hydrophobic desolvation barriers tend to decrease at high temperatures.[43] This barrier may be responsible for the alternative folding pathways in simulations reported recently.[44]

Our initial putative transition state consisted of structures residing at the top of the major barrier (ρ=0.3) with 17–18 native side-chain contacts. The probabilities of forming native side-chain contacts for these structures are plotted in Figure 3. The contact probability map shows clearly that most high-probability contacts lie in the central three-stranded β-sheet, in good agreement with experimental φ-value. Additional details about the transition state obtained from the free energy surface can be found in the publications by Guo *et al*.[39,42]

P_{fold} analysis

Since the accuracy of the location of the TSE from the free energy surfaces is not known *a priori*, we chose to probe an expanded region around ρ=0.3. P_{fold} analysis was performed for structures ρ=0.25 to 0.5 (with 14–30 native side-chain contacts). This range allows us to probe structural details right before (pre-TS), at (TS) and after (post-TS) the transition. A total of 1661 structures were selected randomly out of all (about 40,000) structures in this region, with most structures coming from the middle of our chosen ρ range, and subject to P_{fold} calculations. The distribution of number of side-chain contacts and C^β contacts for the selected structures are shown in Figure 4. In all, 348 structures were found to have P_{fold} values between 0.4 and 0.6 and hence correspond to TS structures. 375 and 444 structures have $P_{fold}<0.2$ and >0.8, respectively, corresponding to pre-TS and post-TS structures.

The contact probabilities of the TS, pre-TS and post-TS structures are illustrated by contact maps in Figure 5 for both side-chain and β-carbon (C^β) native contacts, and given in Tables 1 and 2. C^β native contacts were used in addition to the side-chain native contacts, as they provide additional structural probes. Before the transition state, both side-chain and C^β contact maps show that the most structured element is the central three-stranded β sheet, with the highest-probability contacts lying between β2-β3 and β3-β4. Most side-chain contacts between β2-β3-β4 have intermediate probabilities: between β2-β3, Leu24-Ala37 (0.71), Leu24-Ser39 (0.58), Ile26-Trp35 (0.64), Ile26-Ala37 (0.58), Val27-Leu36 (0.74), Val27-Ala37 (0.19) and Val27-His38 (0.41); between β3 and β4 and the distal loop connecting them: Trp34-Tyr47 (0.46), Trp35-Ile48 (0.34), Leu36-Thr45 (0.63), Leu36-Tyr47 (0.24), Ala37-Ile48 (0.48), His38-Thr45 (0.37) and Ser39-Thr42 (0.53). Contacts in the n-src loop, which connects β2-β3, have lower probability: Asn29-Asp33 (0.01) and Asn29-Trp35 (0.06), indicating that this region is still quite open. Contacts in the 3_{10} helix region also have intermediate probability, with Pro49-Tyr52 (0.43) and Ser50-Val53 (0.30). The rest of the protein is mostly unstructured, except for a few contacts with intermediate probabilities, Thr12-Asp15 (0.42) located in the tip of the RT loop and Phe2-Val53 (0.31) between β1 and β5. In summary, pre-TS, the central β sheet β2-β3-β4 is more structured than other regions of the protein, yet remains incompletely formed (as most contacts have only intermediate probabilities of formation). The plot of C^β contact probabilities (Figure 5(c), upper quadrant) shows a similar distribution of structured units.

Comparing the side-chain native contact maps for pre-TS (Figure 5(a), upper quadrant) and post-TS (Figure 5(a), lower quadrant), the most

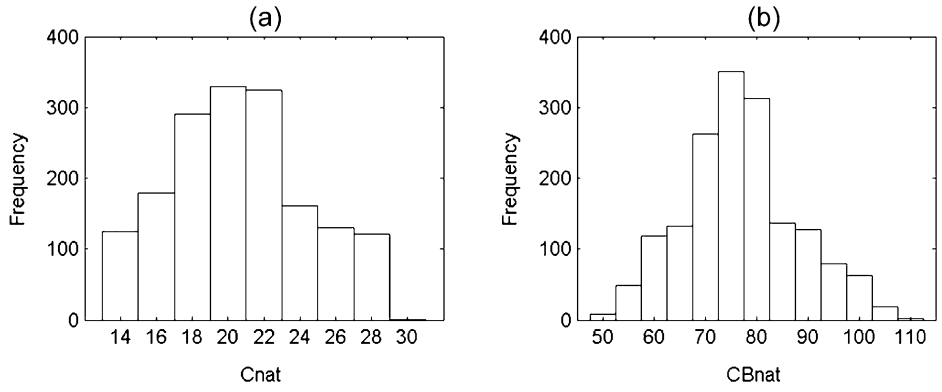

Figure 4. The distribution of structures selected around the putative TSE by number of (a) side-chain contacts and (b) C^β contacts. More structures were selected close to the center than further out.

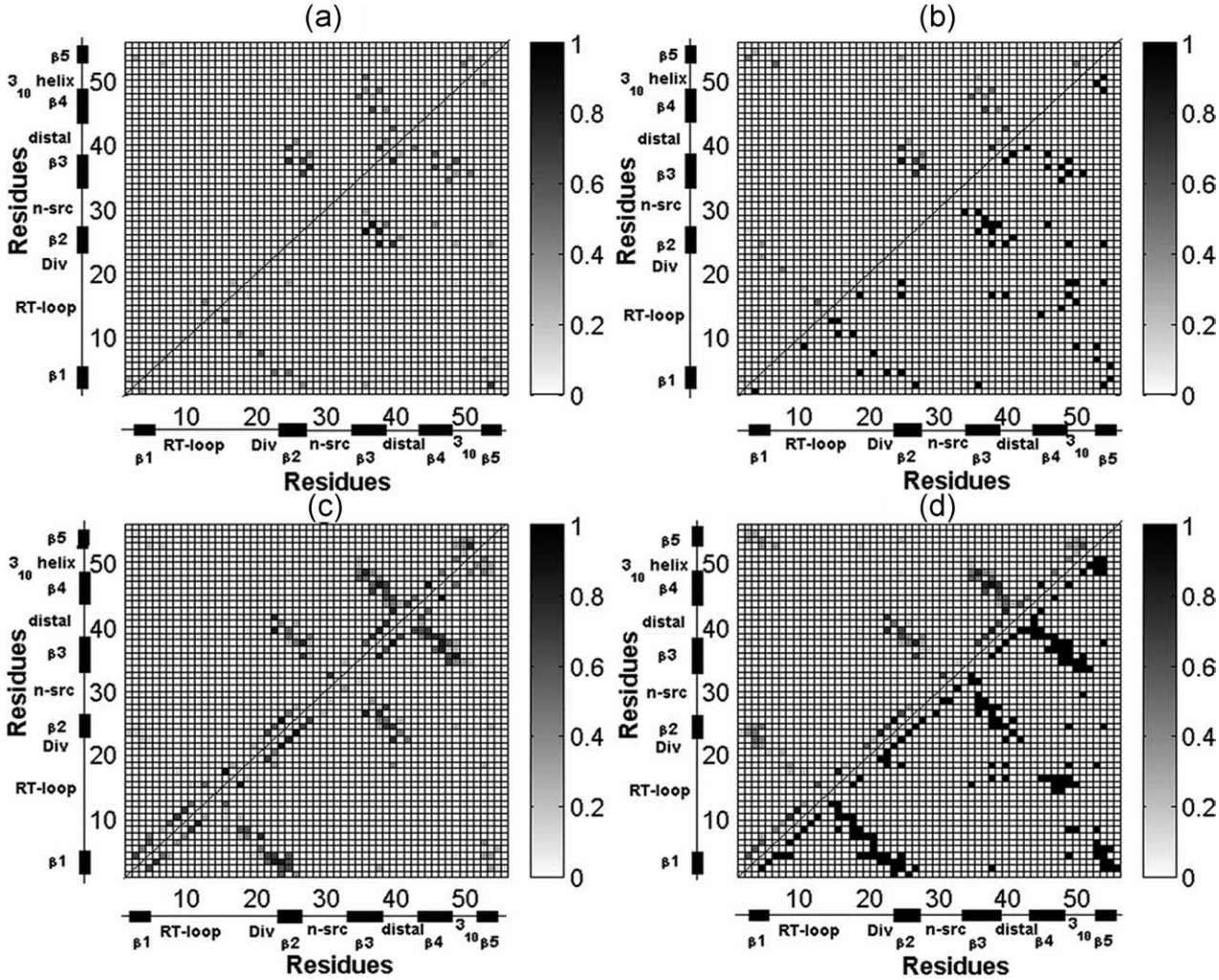

Figure 5. Probabilities of forming (a) and (b) native side-chain contacts and (c) and (d) native C^{β} contacts for pre-TS ((a) and (c), upper quadrant), post-TS ((a) and (c), lower quadrant) and TS ((b) and (d), upper quadrant) determined from the P_{fold} test. The (b) and (d) lower quadrant shows the native side-chain and C^{β} contact maps for the folded state, respectively. Both side-chain contacts and C^{β} contacts show that right before the transition ((a) and (c), upper quadrant), the most structured region is the central three-stranded β-sheet β2-β3-β4, with very little structure in other portions of the sequence; right after the transition ((a) and (c), lower quadrant), the central three-stranded β-sheet is refined, contacts between the β1 sheet and diverging turn-β2 region are significantly more structured, and the β1-β5 interactions slightly more structured.

significant increase in side-chain contact probabilities post transition occurs between β1 and β2, and between the lower ends of the RT loop (due to the sequence connectivity). Side-chain contacts Phe2-Ile26, Ala4-Glu22, Ala-4-Leu24, Asp7-Lys20 and Ser10-Ser17 all have increased from around 0.1 to 0.4. This is even more obvious in the C^{β} contact map, which shows that the probabilities of most C^{β} contacts in this region have increased from around 0.1 to 0.6. Also increased are the contacts between β1 and β5, although not as systematically as in the β1-β2 region, with only the side-chain contact Phe2-Val53 increased from 0.31 to 0.70, while other contacts retained low probabilities of formation. The C^{β} contact map (Figure 5(c)) shows a more systematic increase in this region, albeit only from 0.1 to 0.3. Despite the presence of contacts between β1-β5 (such as Phe2-Val53), the secondary structure is not fully formed. The side-chain contacts in the central β-sheet have been refined to higher probabilities: especially, Leu24-Ala37 (0.84), Leu24-Ser39 (0.70), Ile26-Trp35 (0.88), Ile26-Ala37 (0.75) and Val27-Leu36 (0.84). The contacts between the RT loop and β3-β4, which connect the two sheets and close the hydrophobic core in the native state, remain unchanged, with low probabilities of contact formation both before and after the transition.

As expected, most (side-chain and C^{β}) native contact probabilities of the TS structures (Figure 5(b) and (d), upper quadrant) lie between the pre-TS and post-TS contact probabilities. The transition state determined from the P_{fold} analysis appears to possess structured central β-sheet (β2-β3-β4), in good accord with experimental φ-value. The β1-β2 contacts and the RT loop, although not as well

Table 1. Probabilities of forming native side-chain contacts for pre-TS, TS and post-TS structures as determined by the P_{fold} test

Contacts	Pre	TS	Post	Contacts	Pre	TS	Post
Thr1-Val3	0.27	0.22	0.32	Phe18-Val53	0.00	0.05	0.16
Phe2-Ile26	0.10	0.16	0.36	Leu24-Ala37	0.71	0.83	0.84
Phe2-Trp35	0.09	0.12	0.29	Leu24-Ser39	0.58	0.67	0.70
Phe2-Val53	0.31	0.45	0.70	Leu24-Ile48	0.19	0.20	0.26
Val3-Ala54	0.05	0.25	0.11	Leu24-Val53	0.08	0.29	0.32
Ala4-Phe18	0.01	0.01	0.09	Gln25-Leu40	0.37	0.30	0.33
Ala4-Glu22	0.13	0.33	0.47	Ile26-Trp35	0.64	0.68	0.88
Ala4-Leu24	0.11	0.35	0.49	Ile26-Ala37	0.58	0.61	0.75
Leu5-Ala54	0.03	0.14	0.19	Val27-Leu36	0.74	0.69	0.84
Tyr6-Tyr52	0.22	0.51	0.29	Val27-Ala37	0.19	0.21	0.13
Asp7-Lys20	0.11	0.43	0.56	Val27-His38	0.41	0.40	0.43
Tyr8-Ser10	0.04	0.02	0.03	Val27-Thr45	0.12	0.19	0.35
Tyr8-Pro49	0.06	0.06	0.05	Asn28-Leu36	0.05	0.15	0.18
Tyr8-Tyr52	0.05	0.09	0.10	Asn29-Asp33	0.01	0.01	0.15
Ser10-Asp15	0.20	0.11	0.15	Asn29-Trp35	0.06	0.15	0.30
Ser10-Ser17	0.10	0.23	0.39	Trp34-Tyr47	0.46	0.53	0.48
Thr12-Thr14	0.14	0.13	0.14	Trp35-Ile48	0.34	0.37	0.32
Thr12-Asp15	0.42	0.50	0.47	Trp35-Ser50	0.49	0.28	0.45
Glu13-Gln44	0.00	0.00	0.00	Leu36-Thr45	0.63	0.61	0.56
Thr14-Tyr47	0.00	0.00	0.00	Leu36-Tyr47	0.24	0.32	0.41
Asp15-Pro49	0.00	0.00	0.00	Ala37-Ser39	0.62	0.72	0.73
Leu16-Phe18	0.27	0.20	0.16	Ala37-Ile48	0.48	0.51	0.59
Leu16-Leu24	0.00	0.00	0.00	His38-Leu40	0.31	0.27	0.27
Leu16-Ala37	0.00	0.00	0.00	His38-Thr45	0.37	0.41	0.50
Leu16-Ser39	0.00	0.00	0.00	Ser39-Thr42	0.53	0.62	0.45
Leu16-Ile48	0.00	0.00	0.03	Ile48-Val53	0.11	0.13	0.21
Phe18-Leu24	0.02	0.04	0.31	Pro49-Tyr52	0.43	0.44	0.38
Phe18-Ile48	0.03	0.01	0.13	Ser50-Val53	0.30	0.24	0.19
Phe18-Pro49	0.02	0.03	0.04				

Table 2. Probabilities of forming native C^{β} contacts for pre-TS, TS and post-TS structures as determined by the P_{fold} test

Contacts	Pre	TS	Post	Contacts	Pre	TS	Post
Thr1-Arg23	0.23	0.19	0.54	Asp7-Phe18	0.05	0.18	0.43
Thr1-Gln25	0.10	0.04	0.38	Asp7-Lys19	0.01	0.00	0.00
Thr1-Ser56	0.11	0.32	0.38	Asp7-Lys20	0.07	0.26	0.75
Phe2-Ala4	0.65	0.57	0.62	Tyr8-Ser10	0.70	0.71	0.70
Phe2-Arg23	0.15	0.39	0.57	Tyr8-Asp15	0.10	0.10	0.23
Phe2-Leu24	0.14	0.28	0.62	Tyr8-Ser17	0.15	0.26	0.55
Phe2-Ile26	0.01	0.02	0.05	Tyr8-Phe18	0.08	0.34	0.57
Phe2-Ala37	0.04	0.17	0.30	Tyr8-Ile48	0.00	0.01	0.01
Phe2-Val53	0.15	0.34	0.30	Tyr8-Pro49	0.02	0.00	0.00
Phe2-Ala54	0.13	0.28	0.13	Tyr8-Tyr52	0.00	0.00	0.02
Phe2-Pro55	0.04	0.19	0.15	Glu9-Arg11	0.68	0.84	0.81
Phe2-Ser56	0.06	0.14	0.16	Glu9-Ser17	0.00	0.00	0.03
Val3-Leu5	0.41	0.55	0.72	Ser10-Thr12	0.49	0.33	0.29
Val3-Gly21	0.14	0.38	0.50	Ser10-Asp15	0.20	0.09	0.15
Val3-Glu22	0.19	0.49	0.80	Ser10-Leu16	0.02	0.05	0.01
Val3-Arg23	0.15	0.45	0.82	Ser10-Ser17	0.09	0.18	0.23
Val3-Leu24	0.12	0.33	0.61	Arg11-Asp15	0.17	0.05	0.19
Val3-Val53	0.07	0.22	0.12	Thr12-Thr14	0.05	0.03	0.03
Val3-Ala54	0.06	0.30	0.12	Thr12-Asp15	0.47	0.57	0.53
Val3-Ser56	0.06	0.22	0.12	Thr14-Gly46	0.00	0.00	0.00
Ala4-Tyr6	0.56	0.54	0.48	Thr14-Tyr47	0.00	0.00	0.00
Ala4-Asp7	0.45	0.29	0.27	Asp15-Ser17	0.89	0.84	0.86
Ala4-Tyr8	0.01	0.02	0.00	Asp15-Trp34	0.00	0.00	0.00
Ala4-Phe18	0.01	0.01	0.12	Asp15-Gly46	0.00	0.00	0.00
Ala4-Lys19	0.02	0.03	0.14	Asp15-Tyr47	0.00	0.00	0.00
Ala4-Lys20	0.01	0.07	0.23	Asp15-Ile48	0.00	0.00	0.00
Ala4-Gly21	0.09	0.31	0.64	Asp15-Pro49	0.00	0.00	0.00
Ala4-Glu22	0.15	0.42	0.65	Leu16-Phe18	0.28	0.22	0.20
Ala4-Leu24	0.13	0.45	0.57	Leu16-Ala37	0.00	0.00	0.00
Ala4-Ile48	0.07	0.14	0.11	Leu16-Ser39	0.00	0.00	0.02
Ala4-Tyr52	0.23	0.35	0.33	Leu16-Gln44	0.03	0.02	0.07
Ala4-Val53	0.20	0.40	0.33	Leu16-Thr45	0.00	0.00	0.00
Leu5-Lys20	0.07	0.20	0.44	Leu16-Gly46	0.00	0.00	0.00
Leu5-Tyr52	0.03	0.11	0.10	Leu16-Tyr47	0.00	0.00	0.00
Leu5-Val53	0.06	0.15	0.12	Leu16-Ile48	0.00	0.00	0.01
Leu5-Ala54	0.05	0.16	0.29	Phe18-Glu22	0.05	0.06	0.30

Table 2 (*continued*)

Contacts	Pre	TS	Post	Contacts	Pre	TS	Post
Tyr6-Tyr8	0.23	0.42	0.35	Phe18-Leu24	0.00	0.00	0.01
Tyr6-Lys20	0.00	0.00	0.00	Phe18-Ala37	0.03	0.00	0.02
Tyr6-Tyr52	0.18	0.33	0.39	Phe18-Ser39	0.11	0.14	0.39
Asp7-Glu9	0.53	0.33	0.36	Phe18-Ile48	0.02	0.01	0.00
Asp7-Ser17	0.04	0.03	0.22	Phe18-Tyr52	0.02	0.21	0.26
Lys19-Gly21	0.25	0.51	0.66	Trp34-Ile48	0.43	0.48	0.42
Lys19-Glu22	0.05	0.09	0.18	Trp34-Pro49	0.26	0.22	0.29
Lys20-Glu22	0.16	0.25	0.43	Trp34-Ser50	0.30	0.22	0.23
Gly21-Arg23	0.62	0.77	0.87	Trp35-Ala37	0.50	0.65	0.81
Glu22-Leu24	0.54	0.59	0.79	Trp35-Tyr47	0.47	0.43	0.57
Glu22-Ser39	0.82	0.74	0.72	Trp35-Ile48	0.74	0.74	0.83
Glu22-Thr41	0.75	0.70	0.48	Trp35-Ser50	0.16	0.07	0.08
Arg23-Gln25	0.89	0.91	0.89	Leu36-His38	0.70	0.61	0.58
Arg23-Leu40	0.38	0.47	0.44	Leu36-Thr45	0.66	0.59	0.53
Leu24-Ile26	0.47	0.61	0.61	Leu36-Gly46	0.53	0.78	0.76
Leu24-Ala37	0.23	0.29	0.27	Leu36-Tyr47	0.18	0.15	0.20
Leu24-His38	0.48	0.59	0.67	Leu36-Ile48	0.38	0.40	0.26
Leu24-Ser39	0.49	0.63	0.65	Ala37-Ser39	0.77	0.81	0.87
Leu24-Ile48	0.17	0.18	0.26	Ala37-Gly46	0.81	0.89	0.81
Leu24-Val53	0.06	0.24	0.22	Ala37-Tyr47	0.42	0.51	0.51
Gln25-Val27	0.28	0.33	0.62	Ala37-Ile48	0.17	0.29	0.30
Gln25-Ala37	0.45	0.42	0.41	Ala37-Val53	0.02	0.07	0.11
Gln25-His38	0.21	0.39	0.47	His38-Leu40	0.71	0.60	0.61
Gln25-Leu40	0.31	0.22	0.25	His38-Gly43	0.34	0.27	0.41
Ile26-Asn29	0.03	0.10	0.21	His38-Gln44	0.55	0.46	0.66
Ile26-Trp35	0.66	0.63	0.81	His38-Thr45	0.57	0.64	0.72
Ile26-Ala37	0.74	0.75	0.73	His38-Gly46	0.67	0.70	0.65
Val27-Trp35	0.22	0.20	0.09	Ser39-Thr41	0.02	0.03	0.01
Val27-Leu36	0.17	0.17	0.11	Ser39-Thr42	0.75	0.76	0.63
Val27-Ala37	0.22	0.26	0.21	Ser39-Gly43	0.45	0.49	0.57
Val27-His38	0.41	0.38	0.66	Ser39-Gln44	0.57	0.53	0.66
Asn28-Thr30	0.04	0.12	0.21	Leu40-Gly43	0.26	0.25	0.30
Asn28-Glu31	0.05	0.01	0.03	Thr41-Gly43	0.65	0.55	0.55
Asn28-Trp35	0.03	0.09	0.05	Thr42-Gln44	0.47	0.44	0.50
Asn28-Leu36	0.06	0.19	0.34	Gln44-Gly46	0.89	0.94	0.91
Asn29-Trp35	0.00	0.00	0.01	Gly46-Ile48	0.47	0.47	0.49
Asn29-Ser50	0.00	0.00	0.00	Ile48-Ser50	0.46	0.41	0.42
Thr30-Gly32	0.71	0.64	0.30	Ile48-Tyr52	0.34	0.39	0.43
Glu31-Trp34	0.03	0.05	0.20	Ile48-Val53	0.25	0.21	0.25
Glu31-Trp35	0.07	0.02	0.11	Pro49-Asn51	0.25	0.28	0.38
Gly32-Trp34	0.26	0.20	0.14	Pro49-Tyr52	0.37	0.31	0.25
Asp33-Pro49	0.04	0.01	0.06	Pro49-Val53	0.39	0.34	0.30
Asp33-Ser50	0.02	0.01	0.06	Ser50-Tyr52	0.74	0.58	0.43
Asp33-Asn51	0.01	0.00	0.00	Ser50-Val53	0.38	0.30	0.18
Trp34-Tyr47	0.52	0.52	0.64	Ala54-Ser56	0.51	0.71	0.68

formed as the central β-sheet, appear to be in the middle stages of formation and better formed than predicted by experimental ϕ-value analysis. Few contacts have formed between the RT loop and the central β-sheet in our TS structures. Interestingly, earlier work in which the TSE for SH3 was extracted directly from discrete molecular dynamics simulations showed more structure in this region, in particular between residues Leu16 and Gly46,[20] the latter being a kinetically important residue for folding.[45] The method of identification of the transition state in this previous work was based on the P_{fold} analysis presented here. Hence, structures with the Leu16-Gly46 contact formed and P_{fold} values of 0.5 are true TSE conformations (based on the P_{fold} definition used here). The fact that structures with this contact are not found in the present study implies that they were not present in the putative TSE obtained from the structures residing at the top of the free energy barrier. Clearly, the P_{fold} refinement method can identify true TSE structures only if they belong to the putative TSE in the first place, indicating that the TSE obtained here is a subset of the complete TSE. The absence of structures with the Leu16-Gly46 contact in the putative TSE can be attributed to insufficient sampling in the importance sampling protocol used to generate the free energy surfaces, as well as to the choice of reaction coordinates onto which these surfaces were projected. The fact that true TSE structures may or may not present the Leu16-Gly46 contact highlights the possible existence of multiple parallel pathways for folding, each with their own TSE. A comparison of the contact probabilities in the TSE determined from discrete molecular dynamics (DMD) simulations,[20] and that found here is offered in Figure 6. The transition state obtained from DMD simulations is seen to be more structured, with additional contacts between the RT loop and the central hydrophobic sheet. The robust features of the TSE, which are found in both the DMD and the present study are discussed below.

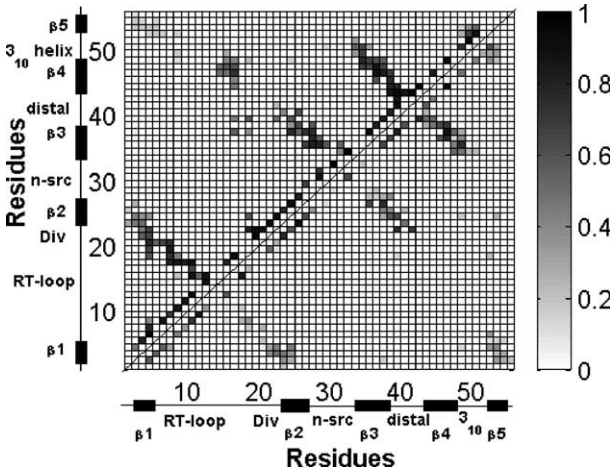

Figure 6. Contact probability maps of the TSE obtained from DMD simulations by Ding *et al.* (upper quadrant)[20] and in the present study (lower quadrant).

The contacts that exhibit the most dramatic changes between pre- and true TSE structures are of key interest. Such contacts are comprised of the nucleus residues, whose contacts both define and guarantee that the TSE is reached. The differential native side-chain center of geometry (C_{nat}) and C^{β} contact probability maps are represented in Figure 7 (a) and (b). The mean difference in C_{nat} contact probability (over the 57 total native side-chain contacts) is very small (0.04), with a probability difference for 44 of the side-chain contacts between -0.10 and 0.10. The most significant increases (≥ 0.2) in contact probability occur near the diverging turn as well as in the second β-sheet. In particular, contacts Val3-Ala54 between β1-β5 and Tyr6-Tyr52 between β1-3_{10}-helix increase by 0.20 and 0.29, respectively, and contacts Ala4-Glu22 and Ala4-Leu24 between β1-(diverging turn-β2) increase by 0.20 and 0.24, respectively. Furthermore, due to chain connectivity, contacts Asp7-Lys20 at

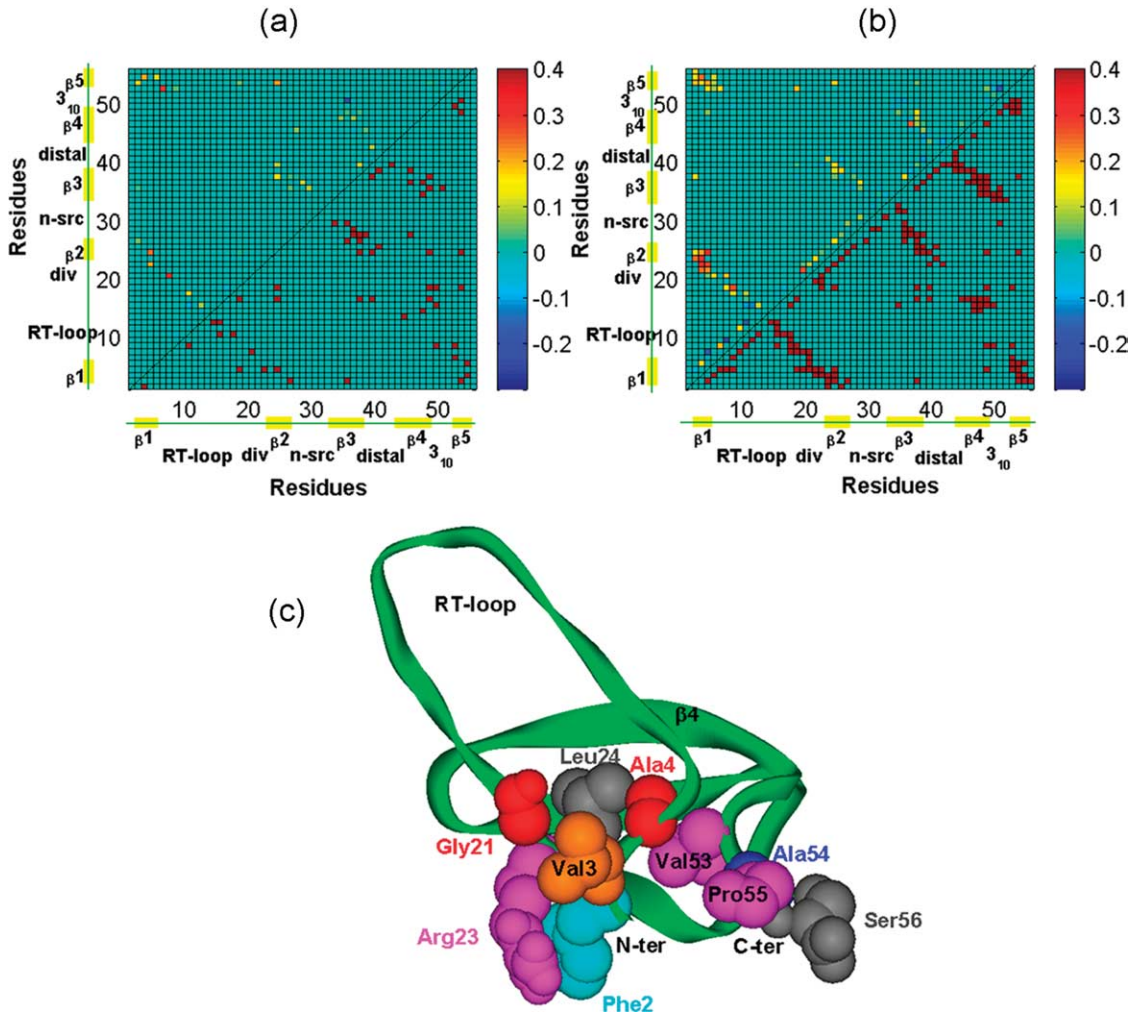

Figure 7. TS–preTS differential probability contact maps for native side-chain center of geometry C_{nat} contacts ((a), upper quadrant) and C^{β} contacts ((b), upper quadrant). C_{nat} and C^{β} contacts for the folded configuration are plotted in the lower quadrants of (a) and (b), respectively. Significant changes in contact probabilities occur in the second β-sheet, while the central β-sheet experiences only slight rearrangements, with positive/negative fluctuations in contact probabilities. (c) A cartoon of a sample TS structure determined by P_{fold} analysis. Residues with an average C^{β} contact probability change of >0.1 (as determined from Figure 6(b)) are shown in spacefill scheme. The structure is presented with Accelrys ViewLite.

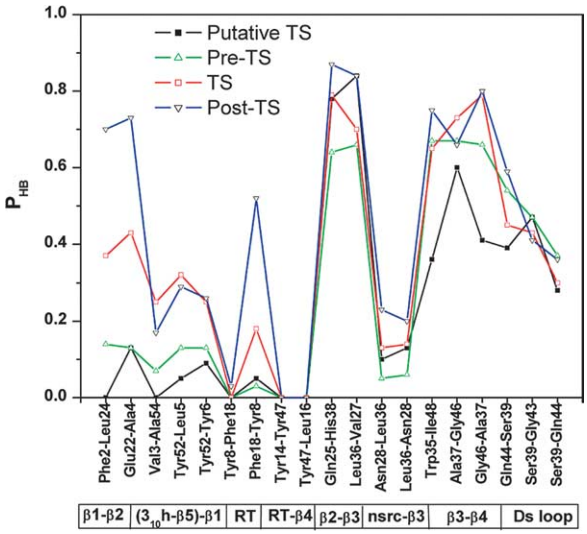

Figure 8. Probabilities of forming native hydrogen bonds for transition state determined from free energy surface (black curve), P_{fold}-determined pre-TS (green curve), P_{fold}-determined TS (red curve) and P_{fold}-determined post-TS (blue curve). Before the transition (pre-TS), the hydrogen bonds are already quite well formed in the central β-sheet, but not between β1-β2 and β5-β1. After the transition (post-TS), hydrogen bonds between β1-β2 are significantly enhanced, and those between β5 and β1 slightly increased. The TS determined from the free energy surface is closer to the pre-TS determined from the P_{fold} analysis.

the base of the RT loop increase by 0.32. Leu24-Val53 between β2-3_{10}-helix increases by 0.21, while Trp35-Ser50 between β2-3_{10}-helix decreases by 0.21. The latter may contribute to the disruption of 3_{10}-helix (discussed in Clustering analysis). None of the 22 side-chain contact probabilities in the central β2-β3-β4 sheet increased by more than 0.1, except Leu24-Ala37 (by a marginal 0.12). Similar trends are observed in the differential contact map for C^{β} contacts (Figure 7(b), upper quadrant). Out of 84 non-local C^{β} contacts (of more than three residues apart), 12 outside the central β-sheet increase by more than 0.2, compared to only one out of 34 in the central β-sheet. In addition, only residues outside the central β-sheet display average C^{β} contact probability changes of >0.1; namely, Phe2, Val3 and Ala4 from β1, Gly21, Arg23 and Leu24 from the diverging turn-β2 and Val53, Ala54, Pro55, Ser56 from β5. These residues are key elements of the folding nucleus. While they are spread throughout the sequence of the protein, they form a localized nucleus in space. Figure 7(c) shows a sample TS structure determined by P_{fold} analysis, with the above key residues shown in spacefill scheme. In sum, the major difference between pre- and true TSE conformations is a cluster of additional contacts between the β1 strand and diverging turn-β2 region as well as between the terminal β-strands. The importance of the diverging turn in the transition state of the src-SH3 domain has been demonstrated experimentally through ϕ-value analysis.[46] The central β-sheet does not differ significantly between pre- and true TSE conformations, and hence emerges as a necessary, but not sufficient element of the true TSE.

The hydrogen bond formation probabilities for the pre-TS, TS and post-TS structures are shown in Figure 8 and given in Table 3. In good agreement with the results drawn from the contact maps, all pre-TS have quite well-structured hydrogen bonds in the central β-sheet β2-β3-β4 hydrogen bonds, which are enhanced slightly during the transition. The most significant increase in hydrogen bond probability occurs between β1 and β2, from around 0.1 to 0.7 for both hydrogen bonds Phe2-Leu24 and Glu22-Ala4, with the first residue denoting donor (backbone O atom) and the second acceptor (backbone H atom). The hydrogen bond probabilities between β1 and β5 have slightly increased during the transition, from around 0.1 to 0.3 for all the three hydrogen bonds Val3-Ala54, Tyr52-Leu5 and Tyr52-Tyr6. Although the β2-β3 are quite well formed, the n-src loop that connects them remains quite unstructured even after the transition, despite the fact that the two hydrogen bonds between Asn28 and Leu36 are also slightly enhanced during the transition from 0.05 to 0.2. The hydrogen bonds between the RT loop and the central β-sheet are not formed at the transition state, in agreement with our previous results, which indicate that this step occurs only in the final stage of folding.[39] Interestingly, the hydrogen bond O(Phe18)–H(Tyr8) enhances significantly from 0.05 to 0.55 during the

Table 3. Probabilities of forming native side-chain contacts for pre-TS, TS and post-TS structures as determined by the P_{fold} test

H-bond	Pre	TS	Post	H-bond	Pre	TS	Post
Phe2-Leu24	0.14	0.37	0.70	Leu36-Asn28	0.06	0.14	0.20
Val3-Ala54	0.07	0.25	0.17	Ala37-Gly46	0.67	0.73	0.66
Tyr8-Phe18	0.00	0.00	0.03	Ser39-Gly43	0.47	0.43	0.41
Thr14-Tyr47	0.00	0.00	0.00	Ser39-Gln44	0.37	0.30	0.36
Phe18-Tyr8	0.03	0.18	0.52	Gln44-Ser39	0.54	0.45	0.59
Glu22-Ala4	0.13	0.43	0.73	Gly46-Ala37	0.66	0.79	0.80
Gln25-His38	0.64	0.79	0.87	Tyr47-Leu16	0.00	0.00	0.00
Asn28-Leu36	0.05	0.13	0.23	Tyr52-Leu5	0.13	0.32	0.29
Trp35-Ile48	0.67	0.65	0.75	Tyr52-Tyr6	0.13	0.25	0.26
Leu36-Val27	0.66	0.70	0.84				

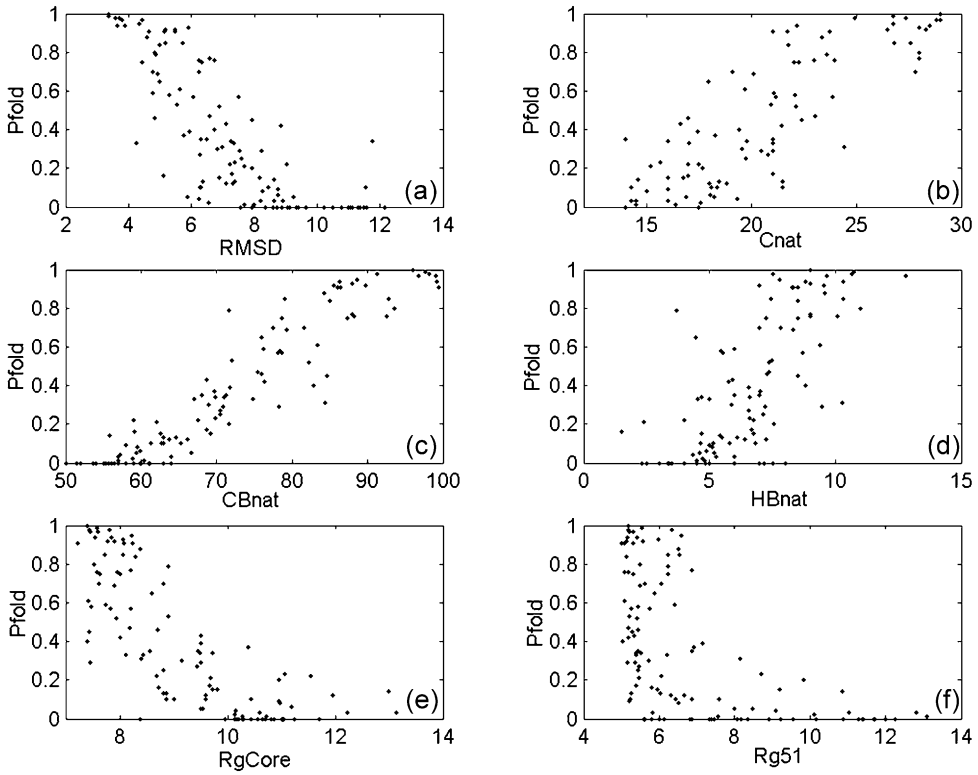

Figure 9. Correlation of average P_{fold} after clustering with (a) average backbone RMSD, (b) number of native side-chain contacts, (c) C^{β} contacts, (d) hydrogen bonds, (e) radius of gyration of residues forming the hydrophobic core (five β-strands), and (f) radius of gyration of the terminal β-strands. Although none of the order parameters (a) through (d) shows a clear-cut correlation with the P_{fold}, number of C^{β} contacts, (c) seems to behave better than the others. True transition state structures cannot have (e) a very open core, or (f) a large distance between β1 and β5.

transition, while its complement O(Tyr8)–H(Phe18) remains essentially unformed. This can be attributed to the sequence connectivity of the protein: with the β1-β2 formed earlier, the structured unit zips from the two ends of the RT loop up to its hinge region (Glu13), leading to the bottom of the RT loop being better structured than the hinge. This can be seen in the C^{β} contact map of the transition state (Figure 5(d), upper quadrant): starting from the β1-β2 contacts region, the probabilities decrease as the contact approaches the diagonal.

Comparing the hydrogen bond probabilities for the transition state structures determined by P_{fold} and from the free energy surface, we find that the putative transition state obtained from the free energy surface is slightly off towards the pre-transition state side, with less structure formed in the β1-β2 region than determined from P_{fold} test.

Clustering analysis

In order to further probe the nature and possible diversity of the pre-TS, TS and post-TS structures, we clustered all the analyzed conformations based on mutual heavy-atom RMSD after a least-squares fit. The maximum RMSD between each structure in a cluster, and the cluster center is constrained to be no greater than 3 Å. The 1661 structures were divided into 113 clusters after clustering, and the average P_{fold} for structures in each cluster calculated. The clustering resulted in 521 (pre-TS) clusters with average $P_{fold} < 0.2$, 12 (TS) clusters with average P_{fold} between 0.4 and 0.6 and 20 (post-TS) clusters with average $P_{fold} > 0.8$. Out of the 52 pre-TS clusters, 31 have less than five members, revealing, not surprisingly, that the pre-TS ensemble has much wider structural variety than TS and post-TS ensembles. In addition, among the 21 larger pre-TS clusters, 11 showed a structured helix in the 3_{10} helix region (Figure 10). All 12 TS clusters have at least five structures each and a total of 438 structures. Only three out of 12 TS cluster centers show helical structure in the 3_{10} helix region. Out of 20 post-TS clusters, 14 have more than five members, and none of them shows a helix in the 3_{10} helix region. Therefore, prior to the transition, the protein has more helix formed in the 3_{10} region, with the helix component decreasing significantly during the transition. The 3_{10} helix region has a short sequence Ser50-Asn51-Tyr52-Val53, and is expected to adopt a turn structure based on the Chou–Fasman secondary structure prediction when in isolation.[47] The helix may be induced by adjacent components of the protein and form early in the folding process, as it involves only local contacts. However, to form contacts between the terminal strands, it may be necessary to extend this region, thus disrupting the rigid helix. Contacts

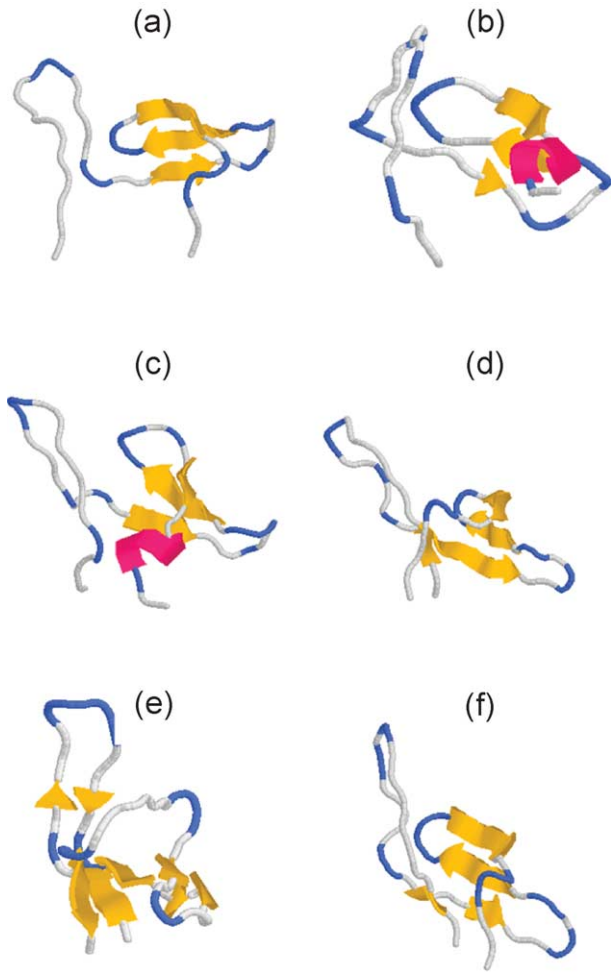

Figure 10. Some cluster centers with average $P_{fold} < 0.2$ ((a) and (b), pre-TS), $0.4 < P_{fold} < 0.6$ ((c) and (d), TS) and $P_{fold} > 0.8$ ((e) and (f), post-TS). (a) A pre-TS with well-structured central β-sheet, but an open core without contacts between β1-β2 or β1-β5. (b) A pre-TS structure with structured central β-sheet in intermediate stage, helix formed in the 3_{10} region, but unstructured β1-β2. (c) and (d) Two true-TS structures with well-structured central β-sheets and more contacts in the β1-β5 and β1-β2 regions than in the pre-TS structures. The (e) and (f) post-TS structures have more structured β1-β2 and β5-β1 and RT loop than TS, and more compact core, but less helix in the 3_{10} region. Structures were prepared with Rasmol.

between terminal strands can provide a better protection to the hydrophobic core of the protein and promote folding. It has been shown for the Fyn SH3 domain that folding can be accelerated dramatically by stabilizing the transition state through a non-native backbone conformation.[45] Early formation of helical structure followed by a transition to a sheet has been observed in the folding mechanism of β-lactoglobulin.[48,49]

Correlation between order parameters and P_{fold}

In order to evaluate how well a given order parameter can discriminate between pre-TS, TS and post-TS structures, we plotted the correlation between P_{fold} and a number of order parameters used commonly in theoretical studies of protein folding (Figure 9). While none of the order parameters in Figure 9 shows a striking difference between pre-TS, TS and post-TS, a number of C^{β} contacts appear to perform significantly better than the others, with less "outliers" from the correlation line. The C^{β} contacts total 162 in the native state, compared with 57 native side-chain contacts and 19 native hydrogen bonds and hence offer more "probes" throughout the protein, yielding more information than the other two types of probes and improving the correlation with the P_{fold} value. The backbone RMSD appears not to be as good an order parameter as the C^{β} contacts. This is likely due to the fact that the SH3 domain has a quite polarized transition state, which can tolerate small RMSD in the central β-sheet but large RMSD in the other sections. The C^{β} contacts, on the other hand, are more concentrated in specific secondary structural regions, hence yielding a better correlation with P_{fold}. Figure 9(e) and (f) indicate that TS structures cannot have a very exposed core ($R_{gCore} > 10$ Å), or a large distance ($R_{g51} > 8$ Å) between the two termini (else the newly formed central β-sheet may not survive under the attack of water molecules and unfold).

Representative cluster centers for pre-TS, TS and post-TS structures are given in Figure 10. Requirements for belonging to the transition state (as identified by P_{fold} analysis) involve the following features: a structured central β-sheet, and an appropriately compact core, involving β1-β2 contacts, β1-3_{10} helix-β5 contacts and a structured diverging turn. Any one of these structural elements alone does not guarantee membership in the true TSE. For instance, the structure in Figure 10(a) has a well-structured central β-sheet, but it is a pre-TS because the core is wide open; structure *b* has a less-structured central β-sheet, and a 3_{10} helix, but an open core with no contact between either β1-β2 or β1-β5, therefore it is a pre-TS too. The structures in Figure 10(c) and (d) both have a relatively compact core, with (c) β1-β5 contact or (d) β1-β2 contact and are TS structures. Post-TS structures (Figure 10(e) and (f)) have more compact structures, more contact between β1-β5 and β1-β2 than TS structures, although some secondary structures may not be well formed, for example, the central β-sheet in Figure 10(e) and β1-β5 in Figure 10(f). Recent studies on SH3 homologues showed that structures obtained from a putative TSE have the topology characteristic of an SH3 domain despite local structural variability.[50] In addition, the diversity of these structures agrees well with the multiple pathway scheme for a homologue of this protein, the c-Crk SH3 domain.[44]

Conclusions

Simulations are a particularly attractive means of studying the TSE of proteins, as they provide a

picture of the TSE at a detailed atomistic level not available from mere inspection of experimental φ-values. A number of computational methods have been proposed in the literature to probe the TSE of small proteins, including the use of high-temperature unfolding simulations,[51] and φ-value based methods.[15] The method of Daggett identifies TS conformations on the basis of structural fluctuations occurring during unfolding trajectories. This method involves arbitrary assumptions relating observed clustering of conformations along a very small number of unfolding trajectories. The fact that putative transition states are obtained from a small number of trajectories precludes, in principle, derivation of the TSE from the simulations described by Daggett *et al.*, and raises questions as to the statistical significance of their conclusions. An additional drawback of this method includes the use of unfolding simulations to infer folding events; however, such microscopic reversibility is unlikely to hold under the dramatic perturbations imposed by high temperatures. Moreover, recent simulations have shown that elementary hydrophobic interactions (in particular the height of the desolvation barrier) are highly sensitive to temperature, rendering the extraction of accurate information about folding at low temperature from high-temperature simulations questionable.[42,52] The method of Vendruscolo and co-workers, on the other hand, seeks to increase the statistical weight of transition states by imposing a restraining potential based on experimental φ-values, with the purpose of focusing sampling of conformational space to the TS region. While a powerful approach to identifying the TSE, it requires the availability of experimental φ-values as input. The combined use of importance sampling molecular dynamics simulations and P_{fold} analysis presented here overcomes the limitations of the above methods and provides a powerful means to identify the transition state ensemble for folding of small proteins accurately and cost-effectively. In addition, our study enables us to evaluate the suitability of a set of parameters (such as the number of native contacts and the radius of gyration) commonly used as reaction coordinates for folding. While the approach taken here makes use of two different Hamiltonians (a fully atomic one to generate the free energy surfaces and a simpler one for the P_{fold} calculations), we anticipate that this discrepancy will not significantly affect the validity of our calculations. Indeed, the TSE of small proteins appears to be robust, largely determined by chain topology, rather than by the specific details of the interactions.[53–56] Our recent simulations using both all-atom Go as well as non-Go potentials showed that the transition states for three different SH3 domains (Src, Fyn and α-spectrin) obtained by P_{fold} analysis are all very similar, and are mostly insensitive to the specific potential used.[56] P_{fold} values obtained using Go-models tend to be slightly higher than those obtained using more frustrated potentials, but are in good agreement overall. In a similar vein, recent simulations by Pande on the BBA5 mini-protein indicate that P_{fold} values computed in both implicit and explicit solvent are in qualitative agreement.[31] Use of a fully atomic potential can enrich the putative TSE and, when combined with P_{fold} calculations using a coarse-grained model, can yield atomically detailed insight into at least a subset of the entire TSE.

We found that conformations of the SH3 domain with P_{fold} near 0.5 showed a structured central β-sheet with less structure throughout the rest of the protein, in agreement with both experimental φ-value analysis and theoretical studies. Interestingly, pre-TS structures can possess a well-defined central sheet, indicating that this element is a necessary, but not sufficient criterion for membership in the transition state. In addition to a structured central region, the protein must adopt a relatively compact conformation ($R_{\mathrm{gCore}} < 10$ Å) in the transition state with well defined cluster of contacts formed by residues in the diverging turn and the second β-sheet. Remarkably, these residues appear to be highly universally conserved in SH3 fold structures ("CoC central" database†).[57,58] Certain flexibility is permitted in the rest of the structure, with varying degrees of β1-β2 and β1-β5 contacts formed in the TSE, a possible indication of multiple folding routes. Overly compact structures always belong to the post-transition state. An interesting observation that arises from a comparison of the TSE determined here and the one we found previously,[59] is that there may be multiple parallel pathways for folding, each possessing its own TSE. This observation suggests that, in addition to performing standard φ value analysis, it may be useful to examine complementary probes of the TSE aimed specifically at identifying the presence of multiple TSE for folding. Sosnick and co-workers suggested recently that their ψ-value method may be capable of discerning parallel pathways,[60,61] although we note that a number of concerns have recently been raised regarding the ability of ψ-values to determine TSE heterogeneity unambiguously.[62]

The free energy surfaces for SH3 yielded a putative TSE that, on average, leaned more towards the pre-transition state, with only 20% of structures belonging to the true TSE (as defined by P_{fold} analysis). The small percentage of true TSE conformations obtained may be due to insufficient sampling used to generate the free energy surface and to the fact that the free energy surfaces were obtained using a fully atomic model in explicit solvent, while the P_{fold} analysis was performed on a coarse-grained protein model. In addition, the selection of "reaction coordinates" onto which the free energy is projected has a strong influence on the nature of the resulting TSE. As discussed in the main body of the text, conformations residing on the top of the barrier belong to the "true"

† http://kulibin.mit.edu/coc/

transition state only if the free energy surface is projected onto "the" reaction coordinate for folding. Since such a coordinate is not known for protein folding (and, if it exists, is likely to be protein-specific), we projected the free energy onto a chosen order parameter describing the progress of the folding reaction. In our case, we selected the number of native side-chain contacts C_{nat}. This choice is an approximation for the true reaction coordinate and will hence lead to a free energy surface for which not all the structures lying at the maxima are true transition state. It is apparent from the P_{fold} analysis that finding a simple geometric reaction coordinate to describe the transition state is a non-trivial task. None of the standard order parameters considered (such as the number of native side-chain contacts and radius of gyration) yielded a perfect correlation with the P_{fold} values. The number of native β-carbon (C^{β}) contacts did correlate significantly better than other parameters, indicating that for SH3, the C^{β} contacts are a more reliable indicator of folding progress.

Methods

Free energy surface

Details of the importance sampling methodology are as described,[24,39] and are summarized below. The protein is described in atomic detail using the CHARMM19 force-field with a TIP3P water model.[63] All molecular dynamics simulations were performed using the CHARMM software,[64] using the computational cluster at Argonne National Laboratory. Covalent bonds between hydrogen atoms and the heavy atoms were fixed using the SHAKE algorithm, allowing for 2 fs time-steps in the Verlet leap-frog integration. All long-range forces were treated using the particle mesh Ewald method. Free energy surfaces (potentials of mean force) as a function of a number of reaction coordinates were generated using extensive importance sampling molecular dynamics simulations. In a first step, the native state of the protein was characterized through two 2 ns molecular dynamics simulations at 298 K, from which two descriptors of the native state were defined, the number of native side-chain contacts and native hydrogen bonds. A native side-chain contact is formed if the distance is less than 6.5 Å between the centers of geometry of side-chains of two non-adjacent residues. A native hydrogen bond is formed if the backbone hydrogen and oxygen atoms of two residues are less than 2.5 Å apart. A total of 57 native contacts (plotted in the contact map in Figure 1(b), upper quadrant) and 19 native hydrogen bonds were identified in this manner. In the next step, an ensemble of structures spanning the unfolded to the folded state was generated using three 2 ns high-temperature (400 K to 450 K) unfolding simulations. In all, 76 cluster centers were found by clustering these structures with the number of native contacts, the number of native hydrogen bonds and the protein solvation energy in the dissimilarity function. These cluster centers were then resolvated and equilibrated by 100 ps of molecular dynamics run at 343 K at constant pressure, and then used as the starting points for biased sampling at 343 K (close to the folding temperature).[42] Biased sampling was then performed at constant volume for 400–800 ps on each cluster center, using a harmonic restraint in the fraction of native contacts (ρ) with a force constant between 500 kcal/mol and 1000 kcal/mol. In a final step, the sampling data were combined using a constant-temperature weighted histogram analysis method.[65] The density of states as a function of the fraction of native contacts and radius of gyration were obtained and used to generate the free energy surface at 343 K.

The fraction of native side-chain contacts (ρ) used in the umbrella sampling was defined as described.[66] First, the state of each contact $x(i)$ is defined using a continuous function:

$$x(i) = \frac{1}{\left(1 + \exp\left(N_{bins}\left(d(i) - \left(K + \frac{tol}{2}\right)\right)\right)\right)}$$

which gives essentially 1 if the distance $d(i)$ between the centers of geometry of the side-chains is less than the cutoff K (6.5 Å), and 0 if $d(i)$ is larger than the cutoff plus a given tolerance (*tol*, 0.3 Å). The fraction of native contacts ρ is the sum of states of all contacts, with ρ = 0 and 1 representing the completely unfolded and folded states, respectively. In addition to the fraction of native contacts, we used the number of side-chain β-carbon atom (C^{β}) contacts as an order parameter. A C^{β} contact is formed when the distance between C^{β} atoms (except for Gly, for which C^{α} is used) is within 7.5 Å. A total of 162 C^{β} contacts were found in the native state structure.

P_{fold} calculation

From the free energy surface we found that the putative TSE is located close to ρ = 0.3, with about 18 native side-chain contacts. A total of 1661 structures with 14–29 native side-chain contacts (ρ = 0.25–0.5) (~4% of ~40,000 structures sampled in this region) were selected randomly and subjected to P_{fold} calculations.

We use the discrete molecular dynamics (DMD) algorithm to compute the P_{fold} values. DMD has been applied recently to study protein folding and aggregations.[32,59,67–69] We model the src-SH3 domain by the "bead-on-the-string" model as described,[20] with two beads per amino acid residue corresponding to C^{α} and C^{β} atoms, and constraints between neighboring beads to mimic the real protein flexibility. We use the Go potential to model the interactions between different amino acids.[70] The interaction potentials are assigned between C^{β} atoms (C^{α} for Gly) and a cutoff of 7.5 Å has been used. We perform DMD simulations of src-SH3 at the previously determined folding transition temperature T_F to compute the P_{fold} values.[20] For each putative TSE conformation, we run 100 independent DMD simulations with different initial velocities at T_F. Since all initial conformations may fold into the native state at T_F with the simulation time long enough to overcome the free energy barrier, we limit each DMD simulation within the time that is much smaller than the average barrier-crossing time but longer than the relaxation time.[20] We determine the P_{fold} value as the percentage of the 100 runs that the model protein folds. We count the final state as folded if both the RMSD is less than 4 Å and the potential energy is in the folded basin, as identified from our equilibrium studies at T_F.[20] Out of the 1661 putative TSE structures, 125 had less than 4 Å backbone RMSD from the native state, of which 122 were found to have $P_{fold} > 0.8$ and thus counted as post-TS structures.

Clustering of the structures

To explore the structural diversity of the transition state, we clustered the structures using the kclust module[71] from the MMTSB tool set†.[72] The clustering is based on RMSD of heavy atoms after a least-squares fit. The algorithm optimizes cluster assignment subject to the constraint on cluster radius (set to 3 Å), such that no member of a cluster is more than the specified distance from the cluster center.

Acknowledgements

This work is supported, in part, by Muscular Dystrophy Association grant MDA3720, Research Grant no. 5-FY03-155 from the March of Dimes Birth Defect Foundation, and the UNC/IBM Junior Investigator Award (to N.V.D.). J.E.S. acknowledges support from the NSF (CAREER # MCB 0133504), the David and Lucile Packard Foundation and the A. P. Sloan Foundation. E.S. acknowledges support from NIH (grant RO1 GM52126).

References

1. Fersht, A. R. (1999). *Structure and Mechanism in Protein Science*, W. H. Freeman, New York.
2. Englander, S. W. (2000). Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 213–238.
3. Sali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein-folding—a lattice model study of the requirements for folding to the native-state. *J. Mol. Biol.* **235**, 1614–1636.
4. Onuchic, J. N., LutheySchulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600.
5. Gruebele, M. (2002). Protein folding: the free energy surface. *Curr. Opin. Struct. Biol.* **12**, 161–168.
6. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Free-energy landscape for protein-folding kinetics—intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062.
7. Sosnick, T. R., Mayne, L. & Englander, S. W. (1996). Molecular collapse: the rate-limiting step in two-state cytochrome c folding. *Proteins: Struct. Funct. Genet.* **24**, 413–426.
8. Nymeyer, H., Socci, N. D. & Onuchic, J. N. (2000). Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration. *Proc. Natl Acad. Sci. USA*, **97**, 634–639.
9. Oliveberg, M. (2001). Characterisation of the transition states for protein folding: towards a new level of mechanistic detail in protein engineering analysis. *Curr. Opin. Struct. Biol.* **11**, 94–100.
10. Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. 1. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771–782.
11. Fersht, A. R. & Sato, S. (2004). Phi-value analysis and the nature of protein-folding transition states. *Proc. Natl Acad. Sci. USA*, **101**, 7976–7981.
12. Ozkan, S. B., Bahar, I. & Dill, K. A. (2001). Transition states and the meaning of Phi-values in protein folding kinetics. *Nature Struct. Biol.* **8**, 765–769.
13. Sanchez, I. E. & Kiefhaber, T. (2003). Origin of unusual Phi-values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* **334**, 1077–1085.
14. Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R. (1998). Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc. Natl Acad. Sci. USA*, **95**, 8473–8478.
15. Paci, E., Vendruscolo, M., Dobson, C. M. & Karplus, M. (2002). Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* **324**, 151–163.
16. Hubner, I. A., Shimada, J. & Shakhnovich, E. I. (2004). Commitment and nucleation in the protein G transition state. *J. Mol. Biol.* **336**, 745–761.
17. Onuchic, J. N. & Wolynes, P. G. (2004). Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75.
18. Shea, J. E., Onuchic, J. N. & Brooks, C. L. (1999). Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Natl Acad. Sci. USA*, **96**, 12512–12517.
19. Shea, J. E., Onuchic, J. N. & Brooks, C. L. (2000). Energetic frustration and the nature of the transition state in protein folding. *J. Chem. Phys.* **113**, 7663–7671.
20. 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.
21. 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.
22. Chan, H. S. & Dill, K. A. (1994). Transition states and the folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* **100**, 9238–9257.
23. Settanni, G., Gsponer, J. & Caflisch, A. (2004). Formation of the folding nucleus of an SH3 domain investigated by loosely coupled molecular dynamics simulations. *Biophys. J.* **86**, 1691–1701.
24. Shea, J. E. & Brooks, C. L. (2001). From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.* **52**, 499–535.
25. Gnanakaran, S. & Garcia, A. E. (2003). Folding of a highly conserved diverging turn motif from the SH3 domain. *Biophys. J.* **84**, 1548–1562.
26. Pitera, J. W. & Swope, W. (2003). Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl Acad. Sci. USA*, **100**, 7587–7592.
27. Rao, F. & Caflisch, A. (2003). Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **119**, 4035–4042.
28. Sugita, Y. & Okamoto, Y. (2000). Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Letters*, **329**, 261–270.

† http://mmtsb.scripps.edu/software/mmtsbToolSet.html

29. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350.
30. Pande, V. S. & Rokhsar, D. S. (1999). Molecular dynamics simulations of unfolding and refolding of a -hairpin fragment of protein G. *Proc. Natl Acad. Sci. USA*, **96**, 9062–9067.
31. Rhee, Y. M., Sorin, E. J., Jayachandran, G., Lindahl, E. & Pande, V. S. (2004). Simulations of the role of water in the protein-folding mechanism. *Proc. Natl Acad. Sci. USA*, **101**, 6456–6461.
32. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (1998). Discrete molecular dynamics studies of the folding of a protein-like model. *Fold. Des.* **3**, 577–587.
33. Zhou, Y. & Karplus, M. (1997). Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl Acad. Sci. USA*, **94**, 14429–14432.
34. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry*, **36**, 15685–15692.
35. Yi, Q., Bystroff, C., Rajagopal, P., Klevit, R. E. & Baker, D. (1998). Prediction and structural characterization of an independently folding substructure in the src SH3 domain. *J. Mol. Biol.* **283**, 293–300.
36. Tsai, J., Levitt, M. & Baker, D. (1999). Hierarchy of structure loss in MD simulations of src SH3 domain unfolding. *J. Mol. Biol.* **291**, 215–225.
37. Grantcharova, V. P., Riddle, D. S. & Baker, D. (2000). Long-range order in the src SH3 folding transition state. *Proc. Natl Acad. Sci. USA*, **97**, 7084–7089.
38. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.
39. Guo, W. H., Lampoudi, S. & Shea, J. E. (2003). Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain. *Biophys. J.* **85**, 61–69.
40. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024.
41. Gsponer, J. & Caflisch, A. (2001). Role of native topology investigated by multiple unfolding simulations of four SH3 domains. *J. Mol. Biol.* **309**, 285–298.
42. Guo, W. H., Lampoudi, S. & Shea, J. E. (2004). Temperature dependence of the free energy landscape of the src-SH3 protein domain. *Proteins: Struct. Funct. Genet.* **55**, 395–406.
43. Moghaddam, M. S., Shimizu, S. & Chan, H. S. (2005). Temperature dependence of three-body hydrophobic interactions: potential of mean force, enthalpy, entropy, heat capacity, and nonadditivity. *J. Am. Chem. Soc.* **127**, 303–316.
44. 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.
45. Di Nardo, A. A., Korzhnev, D. M., Stogios, P. J., Zarrine-Afsar, A., Kay, L. E. & Davidson, A. R. (2004). Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation. *Proc. Natl Acad. Sci. USA*, **101**, 7954–7959.
46. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714–720.
47. Chou, P. Y. & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, **13**, 222–245.
48. Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C. A. & Goto, Y. (2000). Is folding of beta-lactoglobulin non-hierarchic? Intermediate with native-like beta-sheet and non-native alpha-helix. *J. Mol. Biol.* **296**, 1039–1051.
49. Fernandez, A., Colubri, A. & Berry, R. S. (2000). Topology to geometry in protein folding: beta-lactoglobulin. *Proc. Natl Acad. Sci. USA*, **97**, 14062–14066.
50. Lindorff-Larsen, K., Vendruscolo, M., Paci, E. & Dobson, C. M. (2004). Transition states for protein folding have native topologies despite high structural variability. *Nature Struct. Mol. Biol.* **11**, 443–449.
51. Li, A. J. & Daggett, V. (1994). Characterization of the transition-state of protein unfolding by use of molecular-dynamics—chymotrypsin inhibitor-2. *Proc. Natl Acad. Sci. USA*, **91**, 10430–10434.
52. Kaya, H. & Chan, H. S. (2003). Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? *J. Mol. Biol.* **326**, 911–931.
53. Clarke, J., Cota, E., Fowler, S. B. & Hamill, S. J. (1999). Folding studies of immunoglobulin-like -sandwich proteins suggest that they share a common folding path. *Structure*, **7**, 1145–1153.
54. Chiti, F., Taddei, N., White, P., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein foldi. *Nature Struct. Biol.* **6**, 1005–1009.
55. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.
56. Hubner, I. A., Edmonds, K. A. & Shakhnovich, E. (2005). Nucleation and the transition state of the SH3 domain. *J. Mol. Biol.* **349**, 424–434.
57. Mirny, L. & Shakhnovich, E. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding knetics and function. *J. Mol. Biol.* **291**, 177–196.
58. Li, L., Mirny, L. & Shakhnovich, E. (2000). Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature Struct. Biol.* **7**, 336–342.
59. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2002). Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J. Mol. Biol.* **324**, 851–857.
60. Krantz, B. A. & Sosnick, T. R. (2001). Engineered metal binding sites map the heterogeneous folding landscape of a coiled coil. *Nature Struct. Biol.* **8**, 1042–1047.
61. 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.
62. Fersht, A. R. (2004). ϕ Values *versus* ψ analysis. *Proc. Natl Acad. Sci. USA*, **101**, 17327–17328.
63. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
64. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

65. Boczko, E. M. & Brooks, C. L., III (1993). Constant-temperature free energy surfaces for physical and chemical processes. *J. Phys. Chem.* **97**, 4509–4513.
66. Sheinerman, F. B. & Brooks, C. L. (1998). Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.* **278**, 439–456.
67. Dokholyan, N. V., Borreguero, J. M., Buldyrev, S. V., Ding, F., Stanley, H. E. & Shakhnovich, E. I. (2003). Identifying importance of amino acids for protein folding from crystal structures. In *Macromolecular Crystallography, part D*, vol. 374, Academic Press, San Diego pp. 616–640.
68. Smith, A. V. & Hall, C. K. (2001). Protein refolding *versus* aggregation: computer simulations on an intermediate-resolution protein model. *J. Mol. Biol.* **312**, 187–202.
69. Zhou, Y. & Karplus, M. (1999). Interpreting the folding kinetics of helical proteins. *Nature*, **401**, 400–403.
70. Go, N. & Abe, H. (1981). Non-interacting local-structure model of folding and unfolding transition in globular-proteins. 1. Formulation. *Biopolymers*, **20**, 991–1011.
71. Karpen, M. E., Tobias, D. T. & Brooks, C. L., III (1993). Statistical clustering techniques for analysis of long molecular dynamics trajectories. I: analysis of 2.2 ns trajectories of YPGDV. *Biochemistry*, **32**, 412–420.
72. Feig, M., Karanicolas, J. & Brooks, C. L., III (2004). MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **22**, 377–395.

Edited by C. R. Matthews

(Received 9 February 2005; received in revised form 24 April 2005; accepted 10 May 2005)