Critical nucleation size in the folding of small apparently two-state proteins

YAWEN BAI,[1] HONGYI ZHOU,[2] AND YAOQI ZHOU[2]

[1]Laboratory of Biochemistry, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA
[2]Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, University at Buffalo, State University of New York, Buffalo, New York 14214, USA

(RECEIVED December 18, 2003; FINAL REVISION February 5, 2004; ACCEPTED February 5, 2004)

Abstract

For apparently two-state proteins, we found that the size (number of folded residues) of a transition state is mostly encoded by the topology, defined by total contact distance (TCD) of the native state, and correlates with its folding rate. This is demonstrated by using a simple procedure to reduce the native structures of the 41 two-state proteins with native TCD as a constraint, and is further supported by analyzing the results of eight proteins from protein engineering studies. These results support the hypothesis that the major rate-limiting process in the folding of small apparently two-state proteins is the search for a critical number of residues with the topology close to that of the native state.

Keywords: topology; total contact distance; folding rate; nucleation size

To uncover the principles that govern the protein folding process, recent studies have focused on small single-domain proteins. So far, the folding behavior of more than three dozen small proteins has been investigated. They very generally fold in an apparent two-state manner in the absence of detectable early folding intermediates (Jackson 1998; Krantz et al. 2002). Some of the small proteins have been studied in further detail, including the characterization of the rate-limiting transition states by a protein engineering procedure (Goldenberg 1999) and the detection of hidden intermediates that exist after the rate-limiting step by a native-state hydrogen exchange method (Bai et al. 1995; Englander 2000; Chu et al. 2002). Interestingly, the folding rates of these proteins, despite the fact that they span over six orders of magnitude from microseconds to seconds (Munoz et al. 1997; Chiti et al. 1999), correlate with various simple parameters describing the structural properties of the native state including contact order (CO; Plaxco et al. 1998), long-range order (LRO; Gromiha and Selvaraj 2001), fraction of local contacts (Mirny and Shakhnovich 2001), total contact distance (TCD; Zhou and Zhou 2002), and secondary structure contents (Gong et al. 2003).

The earlier observation of the empirical correlations has spurred the development of several quantitative kinetic theories (Plaxco et al. 1998; Alm and Baker 1999; Debe et al. 1999; Munoz and Eaton 1999; Makarov et al. 2002; Kaya and Chan 2003; Makarov and Plaxco 2003; Weikl and Dill 2003) to predict folding rates. Among the theoretical models, the topomer search model (Makarov et al. 2002) also provides a theoretical basis for the empirical correlation between folding rate and the number of native contacts separated by a long distance in sequence (LRO; Gromiha and Selvaraj 2001). The rate-limiting step in the topomer search model is the search for a topomer with full native-like backbone structure in the unfolded state before the formation of the transition state (Makarov and Plaxco 2003). This requirement, however, makes it difficult to explain the native-state hydrogen exchange results that indicate partially unfolded intermediates exist after the rate-limiting step (Bai et al. 1995; Englander 2000; Chu et al. 2002). Moreover, the model leaves little role for transition states in determining folding rates.

Reprint requests to: Yawen Bai, Laboratory of Biochemistry, National Cancer Institute, NIH, Building 37, Room 6114E, Bethesda, MD 20892, USA; e-mail: yawen@helix.nih.gov; fax: (301) 402-3095; or Yaoqi Zhou, Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, University at Buffalo, State University of New York, 124 Sherman Hall, Buffalo, NY 14214, USA; e-mail: yqzhou@buffalo.edu; fax: (716) 829-2344.

Article published ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.03587604.

We found that a more unified model could be obtained if the rate-limiting step of folding is the search for the topomers of a critical nucleus with a native-like topology rather than the full native backbone structure. The hypothesis that the rate-limiting step is the search for a critical nucleus with a native-like topology was initially proposed by Sosnick et al. (1996; see Fig. 1) and formulated quantitatively later by Plaxco et al. (1998) using a simple theoretical model. However, a quantitative relationship between the properties of nuclei and folding rates so far has not been found for real proteins. To seek the evidence for these hypotheses in real proteins, we analyzed the topology, described by a total contact distance (Zhou and Zhou 2002), of the hidden intermediates and the size (number of folded residues in the native conformation) of the transition states of small apparently two-state proteins. We discovered a quantitative correlation between folding rate and topologically determined size of the transition state.

Results

Topology of hidden intermediates

If the rate-limiting transition state has a native-like topology, we reasoned that any state between the rate-limiting transition state and the native state should also have native-like topology. To test this, we first investigated the topology of the hidden intermediates of cytochrome *c* (cyt. *c*; Bai et al. 1995) and Rd-apocyt b_{562} (Chu et al. 2002) using TCD because the folded regions of these hidden intermediates are well defined. TCD (Zhou and Zhou 2002) is very similar to the contact order proposed earlier by Plaxco et al. (1998) except that the sum of the sequence separations of the contacts is normalized by using the total number of residues rather than using the total number of contacts as in the contact order (Zhou and Zhou 2002):

$$\mathrm{tcd} = \frac{1}{\mathrm{n_r^2}} \sum_{k=1}^{n_c} |i-j| \qquad (1)$$

where n_r is the number of residues of a protein, and n_c is the number of nonlocal residue–residue contacts. A nonlocal contact is defined as two residues having heavy atoms within a cutoff distance R_{cut} and separated by at least a sequence cutoff value $\mathrm{l_{cut}}$. In this paper, we used $R_{\mathrm{cut}} = 6$ Å and $\mathrm{l_{cut}} = 2$. *i* and *j* are the residue numbers in the sequence. This renormalization not only has good physical basis (Zhou and Zhou 2002) but also allows very small proteins and short peptides to be included in the correlation between topology and folding rates, whereas the contact order has failed in such cases (see Discussion).

An immediate question related to the calculation of TCD value of a partially unfolded structure is how to model the effect of unfolded loops on the folding rate. Previously, Munoz and Eaton (1999) simply ignored the loop effect in their model for calculating folding rates. Alm and Baker (1999) and Fersht (2000), however, have modeled the loop effect on folding rates using a polymer theory, which correctly predicted the small effect (< fourfold) of short loops (~10 amino acids) on the folding rates (Ladurner and Fersht 1997; Viguera and Serrano 1997). More recent experimental study by Scalley-Kim et al. (2003) has shown that the effect of a very long loop (~100 amino acids) on the folding rate is still small (< fivefold), suggesting that unfolded loops generally have little effect on folding rates. Because the folding rates of small apparently two-state proteins span six orders of magnitude, the loop effects on the folding rate therefore are relatively small. Therefore, we also ignored the loop effect on the folding rates in our analysis. The unfolded regions were substituted with putative linkers without sequence separation (see Fig. 2 for illustration). The TCD values calculated under this assumption for the hidden intermediates of cyt. *c* (Bai et al. 1995) and Rd-apocyt b_{562} (Chu et al. 2002) are shown in Table 1. Indeed, the TCD values of the hidden intermediates are very close to the value of native state (81%–102%; see Table 1) despite the fact that they involve the deletion of significant numbers of residues (up to 67 in the case of cyt. *c*) from the native protein.

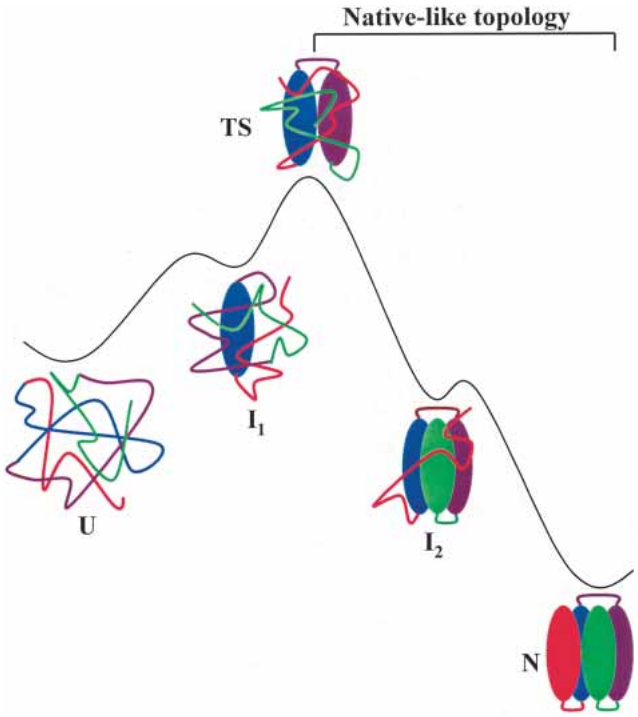

Figure 1. A schematic diagram of a putative folding pathway for an apparent two-state protein and the change of topology along the folding coordinate. In this illustration, the formation of the middle region is the rate-determining step. The two hidden intermediates are not observable in the conventional kinetic folding experiments. I_2 can be identified by native-state hydrogen exchange method (Chu et al. 2002).

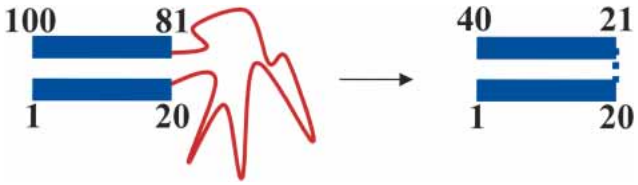

Figure 2. Illustration for calculation of TCD of a partially unfolded structure. The unfolded region was substituted with a putative bond without sequence separation. In this illustration, the C-terminal region following the unfolded loop was renumbered after the loop was substituted with a putative bind. The TCD value was calculated using the structured region with renumbered sequence to obtain $|i - j|$ values in equation 1.

Correlation between the size of the transition state and the folding rate

The finding that the TCD values of the hidden intermediates are close to those of native states in cyt. *c* and Rd-apocyt b_{562} led us to use TCD values as constraints to reduce native structure in the search for the size of the transition state assuming that the transition state is the smallest substructure with a TCD value close to the native state. This is done using a computer program. The program deletes a stretch of six contiguous residues that have the least effect on the TCD value in each step. This process may be viewed as unfolding proteins block by block on a pathway that causes the least change of TCD values in each step. However, it should be noted that this process does not necessarily reflect the actual unfolding process. Figure 3A illustrates the results from four small proteins: cyt. *c*, Rd-apocyt b_{562}, mAcP (Chiti et al. 1999), and TNfn3 (Hamill et al. 2000). These proteins have similar sizes (90–106 residues) but very different native TCD values (0.60–1.26). During the reduction (or unfolding) process, the TCD values of these proteins are unchanged initially, and start to decrease more rapidly and almost monotonically after a significant number of residues were unfolded. This monotonic feature provides a basis for defining and obtaining the size of transition state (see below).

We found that the native structures with smaller native TCD values (cyt. *c* and Rd-apocyt b_{562}) can be reduced to very small sizes without significantly affecting the TCD values. In contrast, more "complex" topologies (higher TCD values, mAcP, and TNfn3) are more sensitive to the size reduction. In general, the smaller the TCD value of a native state, the slower the decreasing rate of the TCD value. It is independent of cutting size (1–20) in each step. Thus, these results suggest that there is an intrinsic relationship between the topology of the native state and the size of

Table 1. *TCDs values of hidden partially unfolded forms (PUF)*

Proteins	Unfolded regions	Tcd
Cyt. c (Bai et al. 1995)[a]		**0.84**
PUF1[a]	[70–85]	0.80
PUF2[a]	[70–85, 36–60]	0.86
PUF3[a]	[19–85]	0.81
Rd-apocyt b_{562} (Chu et al. 2002)		**0.81**
PUF1	[1–22]	0.73
PUF2	[1–22, 95–106]	0.66

[a] In these calculations, the effect of heme is not considered.

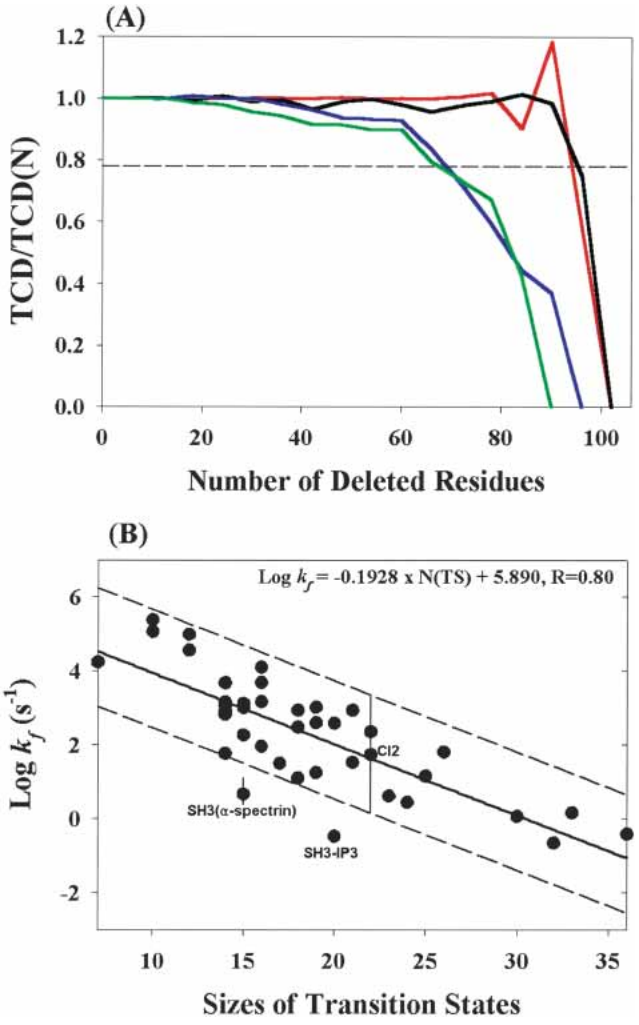

Figure 3. Reduction of the native structures. (*A*) Plots of TCD/TCD(N) vs. the number of deleted residues for cyt. *c* (red, TCD(N) = 0.84), Rd-apocyt b_{562} (blue, TCD(N) = 0.79), mAcP (pink, TCD(N) = 1.52), and TNfn3 (green, TCD(N) = 1.24). TCD is the corresponding topology of the structure after deletion of a six-residue segment in each step (see Materials and Methods) and TCD(N) is the topology for the native state structure. The dashed line represents 78% of the TCD(N) value. (*B*) Correlation between folding rates and the sizes of the "transition states" generated using the reduction procedures. The solid line is the result of a linear fitting. The dotted line indicates a ~50-fold deviation from the solid line determined by the maximum range of the folding rates from single mutants of CI2 (Jackson 1998). Note: The molecule (SH3–PI3) in the experiment had additional four residues (WNSS) at the C terminus (Guijarro et al. 1998) of the structure in the PDB file (2PNI) that is used to calculate the TCD value.

Table 2. *Structural and kinetic parameters for the 41 small apparent two-state proteins*

Proteins	PDB code	Size of NS	Size of TS	Log k_f	TCD(N)
Villin 14T	2VIK[a]	126	18	2.95	0.97
Ribosomal protein L9	1DIV[b]	56	14	2.86	0.88
FKBP12	1FKB[a]	107	23	0.63	1.30
Ubiquitin	1UBO[a]	76	16	3.18	1.07
Procarboxipeptidase A2	1PBA[a]	81	21	2.95	1.08
McrP	2HQI[b]	72	30	0.08	1.48
U1A	1URN[a]	96	18	2.49	1.20
HPr	1HDN[a]	85	25	1.17	1.35
Tendamistat	2AIT[a]	75	26	1.82	1.42
CspA	1MJC[a]	69	15	2.28	1.14
CspB	1CSP[a]	67	19	3.03	1.10
Twitchin	1WIT[b]	93	33	0.18	1.48
Titin	1TIT[b]	89	17	1.51	1.26
FN_III	1FNF_10[a]	94	22	2.37	1.14
FN_III	1FNF_9[a]	90	36	–0.40	1.30
α-spectrin SH3 domain	1SHG[c]	57	15	0.61	1.35
Src SH3 domain	1FMK	57	21	1.54	1.28
Sso SH3 domain	1BF4	64	14	3.02	0.88
PI3 SH3 domain	2PNI	86	20	–0.46	1.21
Fyn-SH3 domain	1NYF[a]	58	16	1.97	1.22
E2P	2PDD[b]	43	7	4.25	0.75
Im9	1IMQ[b]	86	14	3.17	0.93
ACBP	2ABD[a]	86	14	2.84	1.04
Monomeric λ-repressor	1LMB[a]	79	14	3.69	0.75
CD2	1HNG[b]	98	19	1.26	1.25
N-Ribosomal/Protein L9	1CQU	56	14	2.95	0.91
Cyt. c	1HRC[a]	104	14	3.08	0.84
Rd-apocyt b_{562}	1M6T	106	16	3.70	0.81
CI2	1COA[a]	64	22	1.75	1.14
Protein L	2PTL[a]	62	14	1.78	1.13
Protein G	3GB1	56	20	2.60	1.10
mAcP	1APS[a]	98	32	–0.64	1.52
TNfn3	1TEN[a]	90	24	0.46	1.24
WW domain	1PIN	34	16	4.11	0.95
Protein A B-domain	1BDC	60	10	5.08	0.71
Engrailed homeodomain	1ENH	54	12	4.57	0.83
Villin head piece	1VII	36	12	4.34	0.67
Barnase	1A2P	108	18	1.11	0.86
Ribosomal protein S6	1RIS	97	19	2.61	1.33
NTL9 (1–39)	1COU_1_39	39	15	3.13	1.51
β-hairpin (41–56)	3GB1	16	10	5.39	0.89

[a] From Jackson (1998).
[b] From H. Zhou and Zhou (2002).
[c] References: β-hairpin (Munoz et al. 1997); 1RIS (Miller et al. 2002); T9 to D65 for src-SH3 domain (Riddle et al. 1999); E15 to F76 for protein I (Kim et al. 2000), 1 to 39 for NTL9 (1–39) (Horng et al. 2003); G1326 to T1415 for 1FNF_9 and Vall416 to T1509 for 1FNF_10 (Plaxco et al. 1997); SH3-PI3 (Guijarro et al. 1998); 1A2P (Chu and Bai 2002); 1CQU (Luisi and Raleigh 2000); 1M6T (Chu et al. 2002); 1PIN (Jager et al. 2001); 1BDC (Myers and Oas 2001); IVII (Wang et al. 2003); 1ENH (Mayor et al. 2003); 3GB1 (McCallister et al. 2000); 1BF4 (Guerois and Serrano 2000); 1SHG (Martinez et al. 1999).

the substructures. If one reverses the reduction process, that is, in the folding direction, these curves suggest that the same fraction of the topology, relative to the native value, is reached by different sizes (number of residues undeleted) of the substructures for different proteins, independent of the original sizes (total number of amino acids) of the native proteins. These results suggest that there may be a quantitative correlation between the folding rates and the sizes of the substructures with the same fraction of the topology of the native states.

To determine the sizes of these "transition states," we use the smallest substructures with TCD values above 78% of those of the native states by choosing contiguous six residues as the cutting size (see Materials and Methods). Indeed, there is a significant correlation between the folding rates and the sizes of the reduced structures for the 41 apparently two-state proteins (see Table 2, Fig. 3B). The correlation coefficient is 0.80 (0.85 if excluding SH3–α-spectrin and SH3–IP3; for SH3–IP3, the protein used in the experiment has four more residues [WNSS] at the C terminus than the pdb structure; Guijarro et al. 1998). This correlation coefficient is, in fact, slightly higher than that of the correlation between TCD values of the native state and the folding rates ($r = 0.76$) for this data set. The regression line indicates that the size of the transition state can account for about five orders of magnitude differences in folding rates.

Sizes of the transition states from Φ values

To test the correlation between the sizes of the "transition states" from the reduction procedure and the folding rates independently, we examined the sizes of the transition states characterized by the ϕ-value analysis from protein engineering studies (Fersht et al. 1992). We modeled transition-state structures of eight extensively studied proteins in a simple binary manner with "folded" and "unfolded" regions by setting a threshold ϕ value of 0.35. A residue with ϕ value larger than 0.35 and its nearest neighbors are considered folded. Otherwise, a residue is considered unfolded (see details in methods). We found that the sizes of the transition states (or number of folded residues in the transition state) indeed correlate with the folding rates ($r = 0.88$ and $p = 0.004$, see Fig. 4), in agreement with the result from the reduction procedure. We also examined other ways in defining transition structures (see Materials and Methods) and found that this correlation is very robust. This result provides the independent experimental support for the dependence of the folding rate on the size of the transition states obtained from the reduction procedure.

The fact that the number of residues with $\phi > 0.35$ in the transition state correlates with the folding rate but not the m_f values (Sosnick et al. 1996), the derivative of the logarithm of the folding rate constants with respect to denaturant concentrations, or transition state placement (Plaxco 1998) suggests that only the residues that make strong interactions in the transition states are important for determining the fold-

ing rate. Thus, the size presented in this paper may be considered as an "effective size." The concept of an "effective size" further rationalizes the above treatment of the residues with missing ϕ values. Because these residues are on the surface or in the loop regions they unlikely interact strongly with other residues in the transition state. Therefore, they do not contribute to the "effective size" of the transition state.

Discussion

Size and topology of transition states and folding rates

The importance of topology and size of transition state in determining the folding rate of protein folding has been clearly suggested in the earlier studies of protein folding. Sosnick et al. (1996) proposed that the intrinsic rate-limiting step in the folding of a protein is the search for a large nucleus with a native-like topology that could support later downhill folding to the native state (Krantz et al. 2004). Plaxco et al. (1998) used contact order to quantitatively represent the topology of protein structure and found that the contact order of the native structure correlates with folding rate. In the same paper, these authors also proposed a simple model based on the earlier work of Zwanzig (1995) to show that a correlation between a transition state placement and folding rate should exist. This model, however, also implied a correlation between the folding rate and the size of the transition state, although the relationship between transition placement and the folding rate was the focus (Plaxco et al. 1998). However, these predictions so far have not been verified for real proteins. Now, it appears that our discovery of the correlation between the size of the transition state and the folding rate has provided strong evidence for these theoretical hypotheses.

Topomer-search of transition state

In the previous topomer search model (Debe et al. 1999; Makarov et al. 2002), the rate-limiting process involved the search of all contacts in the native structure. Using a Gaussian chain model, Makarov et al. (2002) derived a rate equation:

$$\ln k = \text{constant} + \ln N - N\Delta F/k_B T \qquad (2)$$

where k is the folding rate, N is number of contacts, ΔF is the mean free-energy gained for contact formation, k_B is the Boltzmann constant, and T is the temperature. It was found that only the number of contacts for those residues with sequence separation larger than 12 were important for determining the folding rates rather than the total number of contacts in the native state. We found that the number of residues in the transition state, N^{TS}, satisfies equation 2 for both the transition states determined by the reduction procedure ($r = 0.73$ for 41 proteins) and the experimental ϕ values ($r = 0.87$ for eight proteins). Thus, the topomer search model, if used for searching a transition state with native-like topology, provides a theoretical explanation for the correlation between folding rates and the number of folded residues in the transition state.

Implications of the size of the transition states in the folding mechanism

The discovery of the correlation between the sizes of the transition states (or number of folded residues in the transition states) and folding rates has a number of implications in understanding protein folding. First, it provides a quantitative evidence for a topologically controlled nucleation model for real proteins. The size dependence of folding rates suggests that conformational entropy plays a dominant role in determining the kinetic barrier of folding. On average, the more residues folded in the transition state, the greater the reduction of conformational entropy and the higher the kinetic barrier. Thus, the rate-limiting step appears to involve the search for a group of folded residues that is large enough to have a similar topology as the native state so that it can support further folding in a downhill manner to the native state (see Fig. 1; Sosnick et al. 1996). Second, the determination of transition-state size by the native TCD value (Fig. 3A) provides a rationale for the observation of hidden intermediates in two-state proteins. A large protein with a simple topology (small TCD value) will have a small transition state, and therefore, leaves a large number of residues to be folded after the formation of the transition state. This creates the condition for a population of multiple hidden intermediates after the rate-limiting transition states. Both cyt. c and Rd-apocyt b_{562} have large native sizes but small TCD values of the native state. Finally, the size dependence might provide a simple explanation to the recent finding that the folding rates of three-state proteins correlate with the size of proteins (Galzitskaya et al. 2003) because these proteins have late transition states whose sizes would be close to those of the native states.

It needs to be pointed out that the above results provide no information on how a protein reaches its nucleation size. The prenucleation event can be dominated by any one of the following processes: diffusion-collision (Karplus and Weaver 1976), a hierarchical process with high energetic folding intermediates (Baldwin and Rose 1999), a hydrophobic collapse-reorganization (Dill 1999), or a combination of them, depending on the energetic interactions in the native structure (Zhou and Karplus 1999).

Energetic interactions in transition states and folding rates

Although the size and topology of the transition state can explain the major difference in the rate of folding for small

apparently two-state proteins, the deviation of the folding rate from the regression line in Figure 3B can be as large as ±50-fold, suggesting that other factors can modulate folding rates. An obvious factor that can affect the folding rate is the energetic interactions among the residues in the transition states. The range of folding rates of the single nondisruptive mutants of CI2 (Jackson 1998) illustrated in Figure 3B suggests that this factor essentially can fully account for the ±50-fold deviation.

In addition, we found that the reduction procedure did not accurately predict the structure of transition states as specified by the ϕ values (<0.35) for most of the eight proteins that have been extensively characterized. Figure 5 shows the comparison between the experimentally determined structure using the ϕ values and the calculated structure from the reduction procedure for the eight proteins. These results suggest that there might be many native-like partially unfolded structures with similar sizes and topology. The reduction procedure only selected one of them. Indeed, a complete enumeration study shows that there are ~10^6 of very different substructures with 25 residues that have TCD values between 85% and 90% of the native TCD value of CI2. We speculate that the exact structure of a transition state is also determined by the detailed interactions at the atomic level. For example, the protein G B-domain and its variant NuG1 have similar structures in the native state; however, the structures of their transition states are completely different (Nauli et al. 2001). Although the transition state of the protein G B domain mainly involves the formation of the C-terminal β-hairpin, the transition state of NuG1 involves the formation of the N-terminal β-hairpin. Both the size and the topology of the N- and C-terminal β-hairpins are similar.

It is very important to point out that the failure to predict the exact structures of transition states by the reduction procedure does not mean the correlation between the size of transition state derived from the reduction procedure and folding rate is not physically real. The same correlation was also found by the analysis of experimental ϕ values of transition states as illustrated in Figure 4. It is likely that the size and topology are more robust properties than the exact structure of transition state. For instance, earlier theoretical models for predicting folding rates and the reduction procedure yield similar results. Munoz and Eaton's model is able to correlate the energy barrier to the folding rates ($r = 0.87$) but unable to accurately predict the ϕ values or the structures of transition states (Munoz and Eaton 1999; if the 0.35 threshold value is used to define the folded structure). Similarly, a significant correlation between the calculated energy of transition state and the folding rate has been obtained ($r = 0.67$) from a more sophisticated model of Alm et al. (2002). However, among the 19 proteins studied by Alm et al. (2002), only half of them yield correlation coefficients that are larger than 0.5 between the experimental and predicted ϕ values. It has negative correlation coefficients for four of the proteins. Because it is still not possible to predict reliably how a single mutation would affect the thermodynamic stability of the native state based on the high-resolution structure, we reasoned that a reliable prediction of ϕ values that are very much associated with the detailed energetic interactions in both transition states and native states from theoretical model alone would be even more difficult. In particular, it has been shown that nonnative hydrophobic interactions might occur generally in partially unfolded structures including the transition states (Feng et al. 2003). These nonnative interactions so far have been completely ignored in all of the above models, including the reduction procedure.

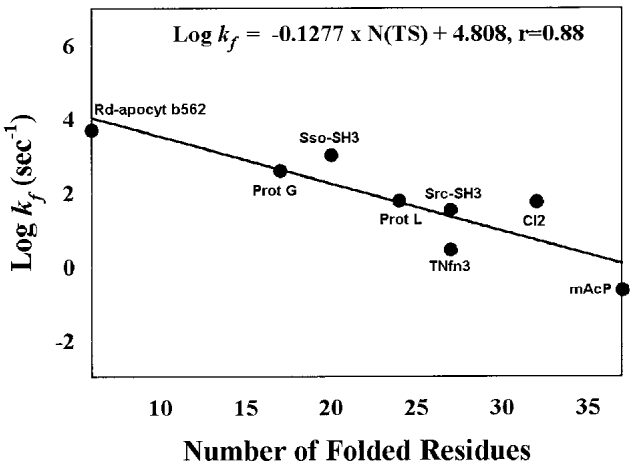

Figure 4. The correlation between the folding rate and the size of the transition state derived based on the experimental ϕ values (see Materials and Methods).

Other empirical correlation parameters

In addition to TCD values, there are several other empirical parameters that previously were found to correlate with folding rates. For this set of 41 proteins, we find that the folding rates correlate with long-range order ($R = 0.85$) but not contact order for structures of native protein ($R = 0.35$). Mainly contact order failed to account for very small proteins and short peptides such as the WW domain and the β-hairpin. Using lro as a parameter to represent the topology in the reduction procedure, the correlation between the size of the reduced substructure and folding rate is basically reproduced ($R = 0.71$). This result suggests the robustness of the dependence of folding rates on transition-state sizes.

Recently, Gong et al. (2003) also found a correlation between folding rates and a parameter that combines the secondary structure contents and the inverse size of native structures. For the 41 proteins discussed here, the correlation coefficient is 0.79. Using the reduced substructures generated from the reduction procedure, we also found a

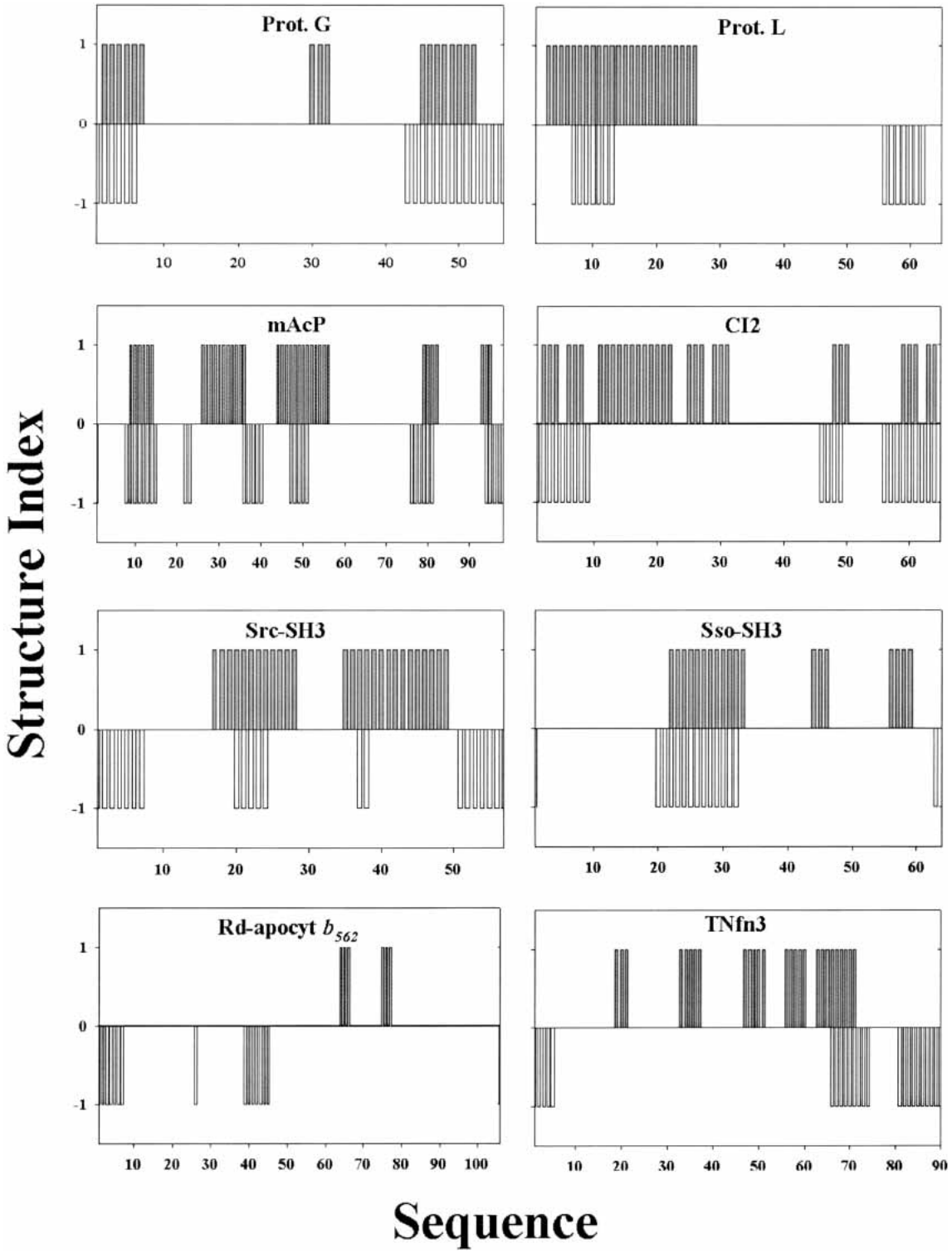

Figure 5. Comparison of the structures of transition states determined from experimental ϕ values and from the reduction procedure using structure indexes for the eight proteins. A residue or its nearest neighbor with $\phi > 0.35$ is considered structured and given a structure index 1. Otherwise, it is given a structural index 0 to represent that the residue is unfolded in the transition state. Similarly, a residue that is folded in the transition state based on the reduction procedure is given a structure index of −1 (the minus sign is for the purpose of comparison). It is given a structural index of 0 if the residue is unfolded in the transition state.

significant correlation between this parameter and the folding rate ($R = 0.75$). However, in contrast to the native state, the correlation between this term and folding rate for the transition state is mostly contributed by the size dependence ($R_{size} = 0.70$).

Conclusion

In summary, the results presented here provide a quantitative evidence for the hypothesis that the search for a critical nucleus with the topology close to that of the native protein is the rate-limiting step for small apparently two-state proteins. The size of the transition states, encoded by the topology of the corresponding native state, is the dominant determinant for the five orders of magnitude difference in the folding rates. The detailed energetic interactions at the atomic level appear to be responsible for modulating the folding rate in the range of ±50-fold and for determining the exact structure of the transition state. The results also help to understand why hidden intermediates are observed in some small proteins and the folding rates of larger proteins are correlated with their sizes.

Materials and methods

The size of the "transition state" derived from a reduction procedure

We derived the size of a "transition state" using a reduction procedure with the native TCD value as a constraint. In each step, contiguous six residues in the structure that have the minimum effect on TCD values are deleted. The deletion is continued until the smallest number of folded residues with the closest value above 78% of the native TCD value is found. The 78% cutoff value is chosen because it gives the best correlation between the size of the "transition state" and folding rate (see Results) from the survey of the parameter space for the sizes of deletion from three to nine and the cutoff value of TCD from 50% to 98% of the native TCD value (with a grid size of 1% and 2%, respectively). A block size of five or six produces stronger correlations over a wide space of TCD values than other block sizes. The physical basis for this observation is not entirely clear. It may be connected to the fact that nucleation of an α-helix involves five residues and the formation of the transition state of a β-hairpin involves six residues.

The sizes of transition states derived from experimental ϕ values

The ϕ value of a residue is the ratio of the free energy change in the transition state over that of the native state upon a nondisruptive mutation at the corresponding site (Fersht et al. 1992). The experimental ϕ values of the eight proteins were obtained from published results: mAcP (Chiti et al. 1999; Vendruscolo et al. 2001), SH3-src (Riddle et al. 1999), Protein G (Nauli et al. 2001), Protein L (Kim et al. 2000), CI2 (Itzhaki et al. 1995), TNfn3 (Hamill et al. 2000), SH3-sso (Guerois and Serrano 2000), Rd-apocyt b_{562} (Chu et al. 2002). For mAcP, theoretical values generated after using limited experimental values as constraints (Vendruscolo et al. 2001) were used because the experimental ϕ values were not measured for some hydrophobic residues that are significantly buried. The highest ϕ value was taken when there are multiple ϕ values at the same position. The transition state structures are modeled with the following two rules: (1) missing ϕ values are treated either as zero or linearly interpolated. The former is based on the fact that most missing ϕ values are from more flexible surface residues that are unlikely to form strong interactions with other residues in transition states. The latter assume that the neighboring residues are located in a similar native environment. (2) A residue was considered to be folded if it or its nearest neighbor has a ϕ value (or an interpolated ϕ value) greater than 0.35. All other residues are considered unfolded. The 0.35 threshold value is chosen to make the sizes of the transition states close to those from the reduction procedure. For this threshold value, the correlation coefficient between the sizes of transition states and folding rates is 0.88 (zero-value approximation) and 0.83 (interpolation approximation), respectively.

Acknowledgments

We thank Drs. S. Walter Englander, Martin Karplus, and Tobin R. Sosnick for comments on the manuscript. The work at Buffalo was supported by NIH (R01 GM 966049 and R01 GM 068530), a grant from HHMI to SUNY Buffalo, and by the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Y.Z. was also supported in part by a two-base fund from the National Science Foundation of China.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

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

Alm, E., Morozov, A.V., Kortemme, T., and Baker, D. 2002. Simple physical models connect theory and experiment in protein folding kinetics. *J. Mol. Biol.* **322:** 463–476.

Bai, Y., Sosnick, T.R., Mayne, L., and Englander, S.W. 1995. Protein folding intermediates: Native-state hydrogen exchange. *Science* **269:** 192–197.

Baldwin, R.L. and Rose, G.D. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24:** 26–33.

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

Chu, R.A., Pei, W.H., Takei, J., and Bai, Y. 2002. Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein. *Biochemistry* **41:** 7998–8003.

Debe, D.A., Carlson, M.J., and Goddard III, W.A. 1999. The topomer-sampling model of protein folding. *Proc. Natl. Acad. Sci.* **96:** 2596–2601.

Dill, K.A. 1999. Polymer principles and protein folding. *Protein Sci.* **8:** 1166–1180.

Englander, S.W. 2000. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 213–238.

Feng, H., Takei, J., Lipsitz, R., Tjandra, N., and Bai, Y. 2003. Specific non-native hydrophobic interactions in a hidden intermediate: Implications for protien folding. *Biochemistry* **42:** 12461–12465.

Fersht, A.R. 2000. Transition-state structure as a unifying basis in protein folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci.* **97:** 1525–1529.

Fersht, A.R., Matouschek, A., and Serrano, L. 1992. The folding of an enzyme.

I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224:** 771–782.
Galzitskaya, O.V., Garbuzynskiy, S.O., Ivankov, D.N., and Finkelstein, A. 2003. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* **51:** 162–166.
Goldenberg, D.P. 1999. Finding the right fold. *Nat. Struct. Biol.* **6:** 987–990.
Gong, H., Isom, D.G., Srinivasan, R., and Rose, G.D. 2003. Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* **327:** 1149–1154.
Gromiha, M.M. and Selvaraj, S. 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol.* **310:** 27–32.
Guerois, R. and Serrano, L. 2000. The SH3-fold family: Experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304:** 967–982.
Guijarro, J.I., Morton, C.J., Plaxco, K.W., Campbell, I.D., and Dobson, C.M. 1998. Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* **276:** 657–667.
Hamill, S.J., Steward, A., and Clarke, J. 2000. The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297:** 165–178.
Horng, J.C., Moroz, V., and Raleigh, D.P. 2003. Rapid cooperative two-state folding of a miniature a-b protein and design of a thermostable variant. *J. Mol. Biol.* **326:** 1261–1270.
Itzhaki, L.S., Otzen, D.E., and Fersht, A.R. 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254:** 260–288.
Jackson, S.E. 1998. How do small single-domain proteins fold? *Fold. Des.* **3:** R81–R91.
Jager, J., Nguyen, H., Crane, J.C., Kelly, J.W., and Gruebele, M. 2001. The folding mechanism of a β-sheet: The WW domain. *J. Mol. Biol.* **311:** 373–393.
Karplus, M. and Weaver, D.L. 1976. Protein-folding dynamics. *Nature* **260:** 404–406.
Kaya, H. and Chan, H.S. 2003. Contact order dependent protein folding rates: Kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. *Protein Sci.* **52:** 524–533.
Kim, D.E., Fisher, C., and Baker, D. 2000. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298:** 971–984.
Krantz, B.A., Mayne, L., Rumbley, J., Englander, S.W., and Sosnick, T.R. 2002. Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324:** 1–13.
Krantz, B.A., Dothager, R.S., and Sosnick, T.R. 2004. Discerning the structure and energy of multiple transition states in protein folding using ψ-analysis. *J. Mol. Biol.* (in press)
Ladurner, A.G. and Fersht, A.R. 1997. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* **273:** 330–337.
Luisi, D.L. and Raleigh, D.P. 2000. pH-dependent interactions and the stability and folding kinetics of the N-terminal domain of L9. Electrostatic interactions are only weakly formed in the transition state for folding. *J. Mol. Biol.* **299:** 1091–1100.
Makarov, D. and Plaxco, K.W. 2003. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **12:** 17–26.
Makarov, D., Keller, D.A., Plaxco, K.W., and Metiu, H. 2002. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci.* **99:** 3535–3539.
Martinez, J.C., Viguera, A.R., Berisio, R., Wilmanns, M., Mateo, P.L., Filimonov, V.V., and Serrano, L. 1999. Thermodynamic analysis of α-spectrin SH3 and two of its circular permutants with different loop lengths: Discerning the reasons for rapid folding in proteins. *Biochemistry* **38:** 549–559.
Mayor, U., Guydosh, N.R., Johnson, C.M., Grossmann, J.G., Sato, S., Jas, G.S., Freund, S.M.V., Alonso, D.O.V., Daggett, V., and Fersht, A.R. 2003. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature* **421:** 863–867.
McCallister, E.L., Alm, E., and Baker, D. 2000. Critical role of β-hairpin formation in protein G folding. *Nat. Struct. Biol.* **7:** 669–673.
Miller, E.J., Fisher, K.F., and Marqusee, S. 2002. Experimental evaluation of topological parameters determining protein-folding rates. *Proc. Natl. Acad. Sci.* **99:** 10359–10363.
Mirny, L. and Shakhnovich, E. 2001. Protein folding theory: From lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* **30:** 361–396.
Munoz, V. and Eaton, W.A. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci.* **96:** 11311–11316.
Munoz, V., Thompson, P.A., Hofrichter, J., and Eaton, W.A. 1997. Folding dynamic and mechanism of β-hairpin formation. *Nature* **390:** 196–199.
Myers, J.K. and Oas, T.G. 2001. Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol.* **8:** 552–558.
Nauli, S., Kuhlman, B., and Baker, D. 2001. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **8:** 602–605.
Plaxco, K.W., Spitzfaden, C., Campbell, I.D., and Dobson, C.M. 1997. A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* **270:** 763–770.
Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277:** 985–994.
Riddle, D.S., Grantcharova, V.P., Santiago, J.V., Alm, E., Ruczinski, I., and Baker, D. 1999. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* **6:** 1016–1024.
Scalley-Kim, M., Minard, P., and Baker, D. 2003. Low free energy cost of very long loop insertions in proteins. *Protein Sci.* **12:** 197–206.
Sosnick, T.R., Mayne, L., and Englander, S.W. 1996. Molecular collapse: The rate-limiting step in two-state cytochrome *c* folding. *Proteins* **24:** 413–426.
Vendruscolo, M., Paci, E., Dobson, C.M., and Karplus, M. 2001. Three key residues from a critical contact network in a protein folding transition state. *Nature* **409:** 641–645.
Viguera, A.R. and Serrano, L. 1997. Loop length, intramolecular diffusion and protein folding. *Nat. Struct. Biol.* **4:** 939–946.
Wang, M., Tang, Y., Sato, S., Vugmeyster, L., McKnight, C.J., and Raleigh, D.P. 2003. Dynamic NMR line-shape analysis demonstrates that the Villin headpiece subdomain folds on the microsecond time scale. *J. Am. Chem. Soc.* **125:** 6032–6033.
Weikl, T. and Dill, K. A. 2003. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* **329:** 585–598.
Zhou, Y. and Karplus, M. 1999. Interpreting the folding kinetics of helical proteins. *Nature* **401:** 400–403.
Zhou, H. and Zhou, Y. 2002. Folding rate prediction using total contact distance. *Biophys. J.* **82:** 458–463.
Zwanzig, R. 1995. Simple model of protein folding kinetics. *Proc. Natl. Acad. Sci.* **92:** 9801–9804.