JMB

Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins

Kevin W. Plaxco, Kim T. Simons and David Baker*

Department of Biochemistry, Box 357350, University of Washington, Seattle, WA 98195 USA

*Corresponding author

Theoretical studies have suggested relationships between the size, stability and topology of a protein fold and the rate and mechanisms by which it is achieved. The recent characterization of the refolding of a number of simple, single domain proteins has provided a means of testing these assertions. Our investigations have revealed statistically significant correlations between the average sequence separation between contacting residues in the native state and the rate and transition state placement of folding for a non-homologous set of simple, single domain proteins. These indicate that proteins featuring primarily sequence-local contacts tend to fold more rapidly and exhibit less compact folding transition states than those characterized by more non-local interactions. No significant relationship is apparent between protein length and folding rates, but a weak correlation is observed between length and the fraction of solvent-exposed surface area buried in the transition state. Anticipated strong relationships between equilibrium folding free energy and folding kinetics, or between chemical denaturant and temperature dependence-derived measures of transition state placement, are not apparent. The observed correlations are consistent with a model of protein folding in which the size and stability of the polypeptide segments organized in the transition state are largely independent of protein length, but are related to the topological complexity of the native state. The correlation between topological complexity and folding rates may reflect chain entropy contributions to the folding barrier.

Keywords: nucleation-condensation mechanism; two-state kinetics

Introduction

Numerous theoretical studies have suggested that the size (Wolynes, 1997; Finkelstein & Badredtinov, 1997; Klimov & Thirumalai, 1997; Gutin *et al.*, 1996; Thirumalai, 1995), stability (Finkelstein, 1991; Sali *et al.*, 1994; Bryngelson *et al.*, 1995; Onuchic *et al.*, 1995; Pande *et al.*, 1997) and topology (Doyle *et al.*, 1997; Gross, 1996; Unger & Moult, 1996; Wolynes, 1996; Abkevich *et al.*, 1995; Fersht, 1995a,b; Govindarajan & Goldstein, 1995; Karplus & Weaver, 1994; Orengo *et al.*, 1994; Dill *et al.*, 1993) of a protein influence the rate and mechanisms by which it folds. Unfortunately, attempts to demonstrate such relationships (e.g. see Munoz & Serrano, 1996; Scalley *et al.*, 1997) have been hindered by the difficulties associated with analyzing complex, multiphasic folding kinetics and by the limited amount of experimental evidence available. The recent characterization of the refolding of a number of single domain proteins lacking *cis* proline residues or disulfide bonds, however, motivated us to re-investigate these relationships. Here we report potentially significant correlations between the folding kinetics and the native, equilibrium properties of a set of kinetically simple, single domain proteins.

Comparisons of the refolding of proteins under differing experimental conditions, of mutant proteins (e.g. Fersht, 1995b; Burton *et al.*, 1996) and of homologous proteins (Kragelund *et al.*, 1996; Mines *et al.*, 1996; Plaxco *et al.*, 1997, 1998) indicate that minor changes in solvent or sequence can dramatically alter the kinetics of folding. The large range of kinetic behaviors exhibited by a single protein under differing solvent conditions, or by multiple

This paper is dedicated to W. A. Goddard III, on the occasion of his 60th birthday.

Abbreviations used: ACBP, acyl carrier binding protein; CO, relative contact order; cspB, *Bacillus* cold shock protein; FN3, fibronectin type III domain; ln(k), the natural logarithm of the intrinsic folding rate; SH3, src-homology 3 domain.

proteins adopting nearly identical folds, suggests that the resolution of sequence and experiment specific effects from the potentially more subtle effects of size, stability and topology might prove very difficult. Our solution to this problem is to search for relationships in a large, diverse data set so that the kinetic consequences of equilibrium properties may be assessed despite this noise. Causal relationships should thus appear as statistically significant, albeit imperfect, correlations between kinetic parameters and equilibrium properties.

Results

We have investigated the influence of three general equilibrium properties, the size, stability and topological complexity of the native state, on the folding kinetics of a non-homologous set of simple single domain proteins. The size (length) and stability (ΔG_u) of the native state are easily quantified and were taken directly from the literature. Topological complexity is somewhat more difficult to specify numerically. We have used relative contact order, (*CO*), which reflects the relative importance of local and non-local contacts to a protein's native structure, as a measure of this property. Relative contact order is the average sequence distance between all pairs of contacting residues normalized by the total sequence length:

$$CO = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j} \qquad (1)$$

where N is the total number of contacts, $\Delta S_{i,j}$ is the sequence separation, in residues, between contacting residues i and j, and L is the total number of residues in the protein.

There is a statistically significant relationship between protein folding kinetics and native state topological complexity. The logarithm of the intrinsic refolding rate (the extrapolated folding rate in the absence of denaturant: ln(k)), which is proportional to the height of the transition state barrier, is well correlated with relative contact order (with a correlation coefficient, r, of 0.81; Figure 1(A)). The p-value associated with this correlation, $p = 0.001$, is extremely low, suggesting that the observed correlation is highly unlikely to have arisen by chance in the 12 member test set. That this strong correlation occurs despite the wide variation in experimental conditions employed in generating the data set (discussed more fully in Materials and Methods) suggests that topology is a very significant determinant of protein folding rates. The inclusion of eight homologous proteins and two circular permutants into the test set (listed in Materials and Methods) does not significantly alter this correlation but does improve its statistical significance (p-value; data not shown).

A statistically significant relationship is also observed between estimates of folding transition state placement, θ_m, and relative contact order ($r = 0.68$; $p = 0.01$; Figure 1(B)). θ_m is computed from the ratio of the denaturant dependencies (m-values) of the relative free energy of folding of the native (m_{eq}) and folding transition states ($m_f^{\ddagger}$) and is thought to reflect the fraction of solvent-accessible surface buried in the native state that is also buried in the transition state. Thus, θ_m, which can theoretically vary from near 0% for a highly unstructured transition state to near 100% for a transition state of native-like compactness, monitors the placement of the transition state on a reaction coordinate that describes the degree of burial of hydrophobic surface area. The correlation

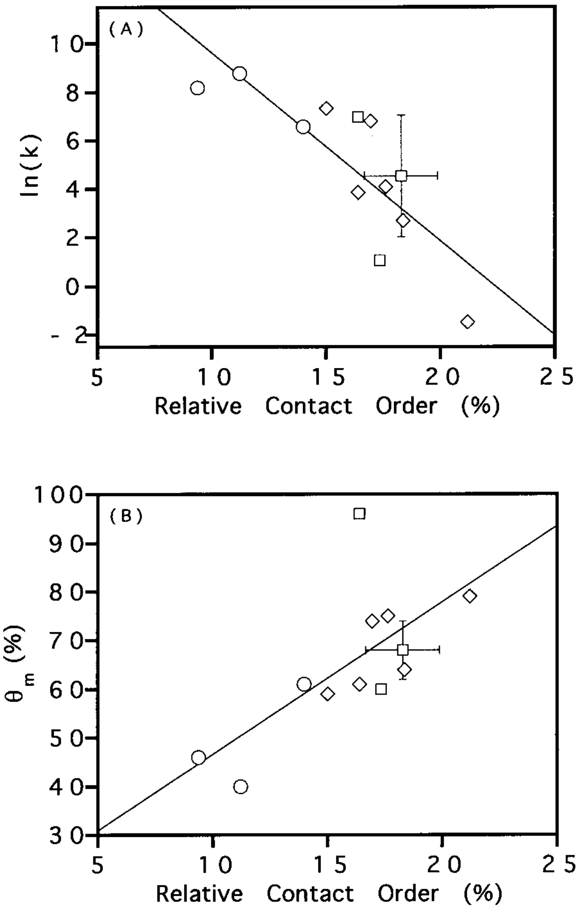

Figure 1. The relationships between topology and folding kinetics. The correlation between the relative contact order of the native state and (A) the natural logarithm of the intrinsic folding rate (ln(k)) or (B) transition state placement (θ_m) for the 12 protein test set. The lines represent linear fits with correlation coefficients of 0.81 and 0.68, respectively. The error bars on the point corresponding to the FynSH3 domain represent the scatter (standard deviation) of values from the four SH3 domains for which the appropriate data have been reported (listed in Materials and Methods). These homologues cover a broad range of sequence identities (from 28 to 70%) but share very similar native state topologies. They therefore provide an indication of the magnitude of sequence specific effects. Circles denote helical proteins, diamonds mixed sheet-helix proteins and squares proteins comprised predominantly of sheet structures.

between contact order and θ_m is perhaps somewhat less surprising than the correlation between CO and ln(*k*) as, unlike ln(*k*), θ_m is largely insensitive to changes in temperature (Scalley & Baker, 1997; Schindler & Schmid, 1996) or solvent conditions (Viguera *et al.*, 1994; V. Grantcharova, personal communication) and thus should be less subject to scatter arising from differing experimental conditions. The inclusion of eight homologous proteins and two circular permutants into the test set (listed in Materials and Methods), while not significantly changing this correlation, improves somewhat its statistical significance (data not shown).

Contact order is related to secondary structural content but the latter is a significantly poorer predictor of folding kinetics. Because helices are characterized by numerous *i*, $i+3$ contacts the contact orders of helical proteins tend to be low and thus there is a significant correlation between contact order and helical content ($r = 0.72$; $p = 0.009$). However, the correlations between helical content and both ln(*k*) ($r = 0.48$; $p = 0.12$) and θ_m ($r = 0.58$, $p = 0.05$) are much less significant than the corresponding correlations with contact order. This suggests that contact order is the more important determinant of protein folding kinetics.

Relationships between the size or stability of members of the test set and their refolding kinetics are weak or non-existent. No statistically significant correlation ($r = 0.13$; $p = 0.68$) is observed between native state stability and folding rates for the proteins of the test set (Figure 2(A)). Also not apparent in the data set is any significant correlation ($r = 0.20$; $p = 0.53$) between the length of single domain proteins and the rates at which they fold (Figure 2(B)). Casual inspection of Figure 2(B) suggests that the lack of a correlation could arise from the potentially anomalous folding rate of cytochrome *c*. After removal of this protein from the data set (which cannot be justified on statistical grounds but may be justified on the biochemical grounds that it is the only member of the test set with a covalently attached prosthetic group) a statistically weak correlation ($r = 0.56$; $p = 0.07$) is observed. Multivariable fits of the full data set to contact order and length dependencies also suggest that length may, after topology, be a secondary determinant of folding rates (K.W.P & I. Ruczinski, unpublished data) although with only a 12 member data set this, too, remains difficult to justify statistically. Thus, while there exist hints in the data set of a length dependence to folding rates, size is at best a significantly less important determinant of kinetics than topology. A statistically weak correlation ($r = 0.51$; $p = 0.09$) hints that length may be a determinant of θ_m (Figure 2(C)). The inclusion of eight homologous proteins and two circular permutants into the test set (listed in Materials and Methods) increases the significance of this correlation while reducing further the statistical significance of the correlations between ln(*k*) and both stability and length (data not shown).

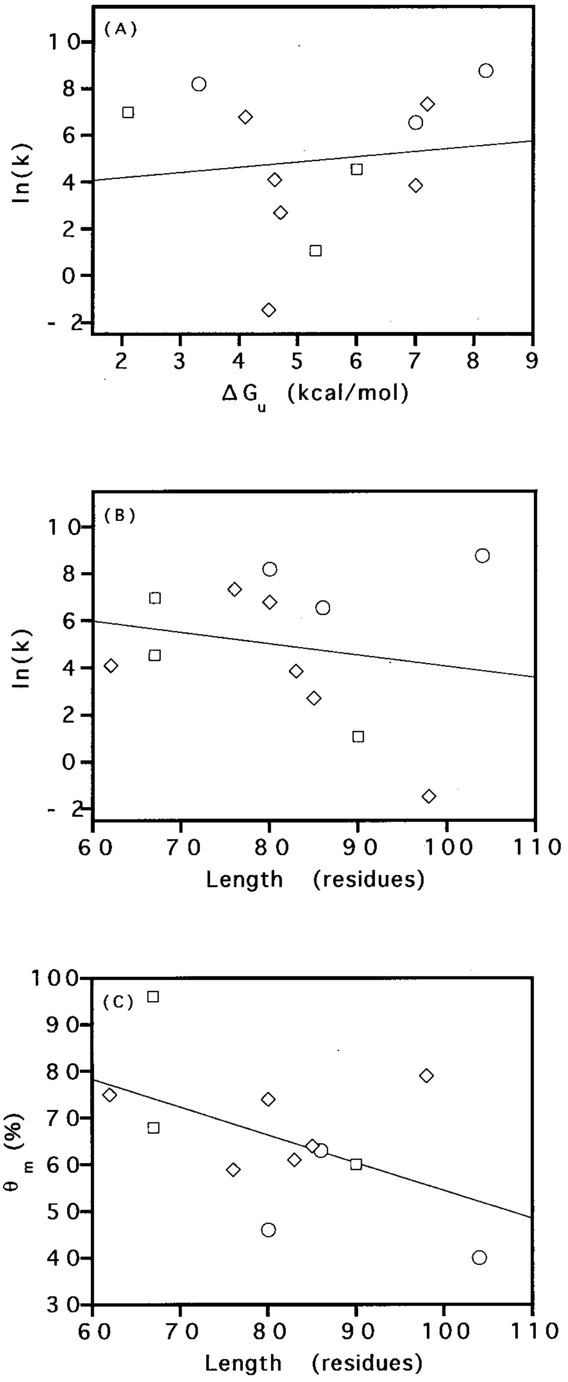

Figure 2. The relationships between stability or length and folding kinetics. No significant correlation is observed between either the (A) equilibrium stability ($r = 0.13$; $p = 0.68$) or (B) length of single domain proteins ($r = 0.20$; $p = 0.53$) and the rates at which they fold. (C) A potential, albiet weak correlation ($r = 0.51$; $p = 0.09$) is, however, observed between length and θ_m. It should also be noted that length may play a significant role in defining the folding rates of larger proteins which often contain multiple independently folding domains and exhibit complex, multiphasic folding kinetics (Roder & Colón, 1997). Symbols are as for Figure 1.

Analysis of the data set also suggests a potential systematic error in one or both of two commonly used descriptors of transition state placement. The excellent correlation between equilibrium measurements of the change in heat capacity upon folding (ΔC_p) and the denaturant dependencies of protein stability (*m*-value) of characterized proteins (Myers *et al.*, 1995) suggests that, by analogy, θ_m should correlate well with the ratio of the relative heat capacities of the transition and native states (θ_{C_p}). The correlation of these values for the six simple, single domain proteins for which data are available is indicated in Figure 3. The correlation is poor and appears to be systematically skewed from the proportional relationship suggested by equilibrium studies: for all six single domain proteins for which we have data θ_{C_p} is less than θ_m. While the data are admittedly limited, the probability of such a systematic bias arising by chance is only 1/64 and for several proteins the two quantities deviate dramatically from the expected equality (e.g. for muscle acyl-phosphatase the values are 79% and 40% for θ_m and θ_{C_p}, respectively: F. Chiti, N. A. J. van Nuland, N. Taddei, F. Magherini, M. Stefani, G. Ramponi & C. M. Dobson, personal communication).

Discussion

Recent years have seen a large increase in studies of the refolding of simple, single domain proteins. We have used this rapidly increasing data base to investigate the roles played by general, equilibrium properties such as length, topology or stability in defining the rates and mechanisms by which proteins fold. Due to the relatively small size of the data set presently available the results of these investigations should be considered preliminary. However, several statistically significant relationships are already apparent and relatively weak or non-existent relationships are observed between several parameters that might have been expected to be strongly correlated.

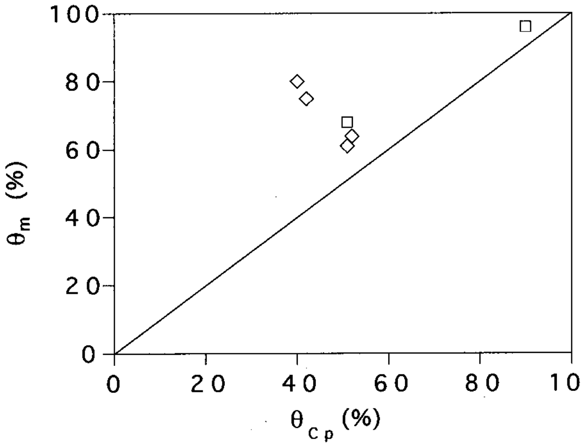

Figure 3. Measures of transition state placement. The correlation between θ_{C_p} and θ_m is poorer than might be expected from equilibrium studies which suggest that both ΔC_p and *m*-values measure the burial of solvent-accessible surface. The continuous line represents this theoretical equality. That θ_{C_p} is always less than θ_m suggests that there may a systematic error in the method used to determine one or both of these values. Symbols are as for Figure 1.

Both equilibrium ΔC_p and *m*-values are well described as proportional to the change in solvent accessible surface area upon folding (Myers *et al.*, 1995). In light of this strong correlation the non-equivalence of θ_m and θ_{C_p} is somewhat surprising. This discrepancy may simply arise due to difficulty in accurately measuring θ_{C_p} (Tan *et al.*, 1996). However, that θ_{C_p} is smaller than θ_m for all six of the simple, single domain proteins studied to date suggests that there may be a systematic error arising from either experiment or its theoretical interpretation. A potentially significant contributor to this discrepancy may be the solvent viscosity dependence of protein folding rates. Several studies have indicated that the "pre-exponential factor" in the Arrhenius equation that describes folding rates is inversly proportional to solvent viscosity (Jacob *et al.*, 1997; K.W.P. & D.B., unpublished results). Since solvent viscosity increases with increasing denaturant concentration and decreases with increasing temperature this would tend to cause θ_m and θ_{C_p} to over and underestimate transition state compactness, respectively. Differing temperature and denaturant-induced movements of the folding transition state, which are observed in an adaptation of Zwanzig's simple model of protein folding (Zwanzig, 1995), may also account for part of this discrepancy (R. Baldwin & D.B., unpublished results).

No statistically significant correlation is observed between the stability of a single domain protein and the rate at which it folds. Studies of the denaturant dependence of protein stability and refolding kinetics have demonstrated that there exists a near-perfect correlation between the equilibrium stability and folding rate of a protein under differing solvent conditions (Chen *et al.*, 1989; Tanford, 1970). Additional experimental evidence suggests that there is often a good correlation between the stability and folding rates of pairs of homologous proteins sharing the same topology (Mines *et al.*, 1996; Plaxco *et al.*, 1997, 1998). Numerous theoretical studies have also suggested that native state thermodynamics may be a significant determinant of protein folding kinetics (Finkelstein, 1991; Bryngelson *et al.*, 1995; Onuchic *et al.*, 1995; Pande *et al.*, 1997), even to the exclusion of other properties such as topology (Sali *et al.*, 1994). The apparent lack of a relationship between the stabilities and folding rates of the topologically diverse proteins in the test set indicates, however, that topology may be a much more critical determinant of relative folding kinetics.

No significant correlation is apparent between folding rates and protein size, suggesting that the relative free energy of the transition state is much less dependent on the total length of the polypeptide chain than on its topology. This would be consistent with a nucleation-condensation model in which the rate-limiting folding step is the formation of a region of native-like structure, a "folding nucleus" (e.g. see Go, 1993; Fersht, 1995a), the size and energetics of which may be independent of protein length. The weak correlation between θ_m and length might also be explained if the sizes of these nuclei are independent of the size of the protein. Then the fraction of the polypeptide chain organized in the transition state, which should roughly parallel θ_m, would tend to be larger for smaller proteins. Such a trend is observed, for example, in the SH3 family of proteins. The three characterized SH3 domains that are of similar lengths exhibit near-identical θ_m values (ranging from 69 to 71%). The PI3 K SH3 domain, which exhibits a similar overall topology but contains an 18 residue insertion, exhibits a θ_m of only 61% (see Plaxco *et al.*, 1998, and references cited therein). However, while this trend is apparent in a set of homologous proteins sharing a common topology, the strong correlation between θ_m and contact order suggests that topology is also a major determinant of transition state structure (see below), which may partially obscure the correlation between length and θ_m for proteins of differing topologies.

There is a very strong correlation between the relative contact order of a protein and the rate at which it folds. While many theoretical studies have predicted relationships between folding kinetics and the relative importance of local and non-local contacts, the exact nature of this relationship has been the subject of much debate in recent years. One theoretical camp has predicted that proteins with predominantly non-local interactions should fold more rapidly because significant local interactions increase the "roughness" of the energy landscape (Govindarajan & Goldstein, 1995), increase the stability of the unfolded state (Fersht, 1995b), or decrease the uniqueness of the native state (Abkevich *et al.*, 1995). The opposing camp (e.g. see Dill *et al.*, 1993; Karplus & Weaver, 1994; Gross, 1996; Unger & Moult, 1996; Doyle *et al.*, 1998) predicts that proteins characterized primarily by local interactions would be the more rapidly folding. Our results provide support for this conclusion. Consistent with this is the experimental observation that mutations that increase the propensity (if not number) of local interactions have been demonstrated to significantly accelerate the folding of a simple, single domain protein (Viguera *et al.*, 1997).

A weak, but statistically significant, correlation is observed between native state CO and folding transition state placement, θ_m. Proteins characterized by large contact orders tend to exhibit more well-ordered transition states, presumably because more of the polypeptide must be ordered to form the requisite number of favorable contacts. This correlation is also consistent with numerous experimental studies demonstrating that topologically similar homologues exhibit very similar θ_m values (Kragelund *et al.*, 1996; Mines *et al.*, 1996; Plaxco *et al.*, 1997, 1998).

Somewhat surprisingly, as both parameters are related to the relative contact order of the native state, the correlation between θ_m and refolding rates is relatively poor ($r = 0.34$, $p = 0.29$). Inspection of the data set (Figure 4), however, suggests that this may be due to the rather anomalous behavior of a single protein, the *Bacillus* cold shock protein (cspB). cspB exhibits the most highly collapsed transition state of any protein characterized to date and yet it folds very rapidly (Schindler *et al.*, 1995; Schindler & Schmid, 1996). The relatively simple, all β-sheet topology of cspB (Gross, 1996) has an intermediate contact order, and thus the protein does not significantly perturb the correlation observed between relative contact order and ln(*k*). However, the protein lies far from the best fit lines for relationships between θ_m and both contact order (a Cook's distance of 2.5 suggests cspB is extremely anomalous) and ln(*k*) (Cook's distance 1.5). Removal of cspB from the data set significantly increases the statistical significance of the relationships between θ_m and ln(*k*) ($r = 0.63$; $p = 0.04$) and between θ_m and contact order ($r = 0.87$, $p = 0.0005$). No explanation for the unusually high θ_m value of cspB is apparent.

Further insights into the relationship between contact order and folding kinetics may be gained from an adaptation of a simple model of folding originally described by Zwanzig (1995). We have used such a model to investigate the contributions of local and non-local interactions to the thermodynamics and kinetics of folding (Doyle *et al.*, 1997). A protein is modeled by *N* residues which can each take on a number of possible orientations, a particular one of which is adopted in the native

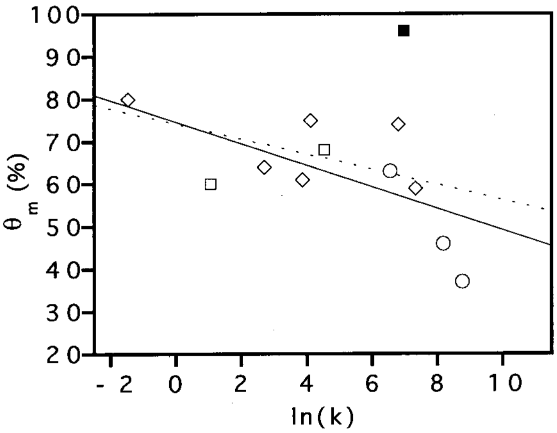

Figure 4. Rate and transition state placement. θ_m and ln(*k*) are poorly correlated (dotted line: $r = 0.34$; $p = 0.29$) unless the anomalously behaving cspB (filled square) is omitted from the data set (continuous line, $r = 0.63$; $p = 0.04$). The reasons cspB folds *via* such a highly compact transition state are unknown (Schindler *et al.*, 1995; Schindler & Schmid, 1996). Symbols are as for Figure 1.

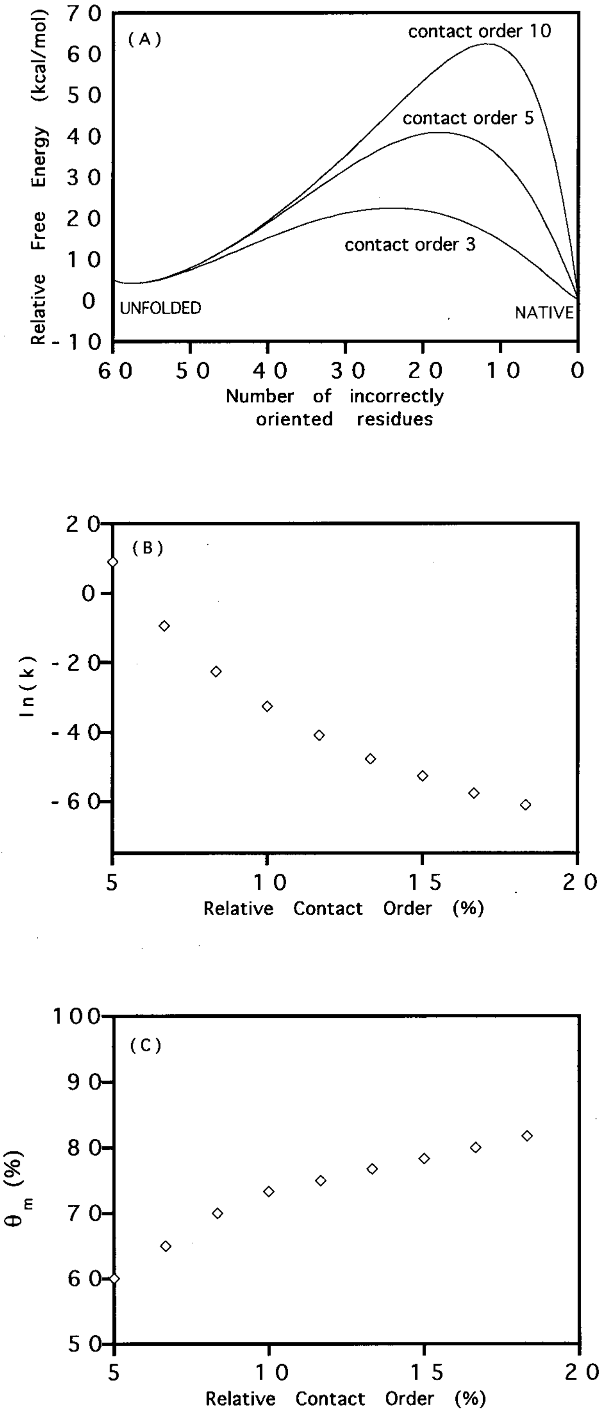

Figure 5. Relationship between contact order, folding rate and transition state placement in a simple model of folding. The calculations employed the adaptation of Zwanzig's simple model of folding (Zwanzig, 1995) described by Doyle *et al.* (1997) with a protein length (*N*) of 60 residues. The kinetics and thermodynamics of folding were computed for different models as described (Doyle *et al.*, 1997). Free parameters were fixed wherever possible to values obtained from the experimental protein folding literature (D.B. & R. Baldwin, unpublished results). The energy was assumed to be directly proportional to the average number of properly formed contacts, a decreasing function of the number of incorrectly oriented angles (*S*) (described in the text). The constant of proportionality was chosen so that the overall free energy of unfolding was between 5 and 7 kcal/mol. (A) Free energy as a function of *S* for different average contact orders. The transition state barrier increases and moves towards the native state as the average contact order increases. (B) Correlation between contact order and the logarithm of the folding rate. (C) Correlation between contact order and transition state placement. Both relationships are qualitatively similar to those shown for real proteins (Figure 1).

protein. Provided that the energy is a function of the number but not the identity of correctly oriented residues, the very high dimensional free energy surface can be collapsed into one dimension where the number of incorrectly oriented residues, *S*, serves as a convenient reaction coordinate. Two residues are considered to form a contact if all the residues between them are properly oriented. The probability that any one residue is properly oriented is *[(N-S)/N]*, and the probability that the average number of residues between contacting residues are simultaneously properly oriented is $[(N\text{-}S)/N]^{CO\cdot N}$, where *CO* is the relative contact order. Taking, for simplicity, all contacts to be equally favorable, the energy as a function of *S* is then proportional to the number of contacts $[(N\text{-}S)/N]^{CO\cdot N}$. The conformational entropy and the overall free energy as a function of *S* and the thermodynamics and kinetics of folding may be readily determined for different contact orders using the methods described by Doyle *et al.* (1998). As indicated in Figure 5(A), the transition state barrier increases and moves towards the native state as the average contact order increases. The dependence of both rate (Figure 5(B)) and transition state placement (Figure 5(C)) on contact order are similar to that of the proteins in our data set (Figure 1).

Theoretical models of folding dating to Anfinsen (1973) have emphasized the importance of the loss of chain entropy upon folding as a rate determining factor. However, while a significant body of experimental evidence supports enthalpic folding barriers even for simple, single domain proteins (Alexander *et al.*, 1992; Schindler & Schmid, 1996; Tan *et al.*, 1996; Scalley & Baker, 1997; Plaxco *et al.*, 1998; van Nuland *et al.*, 1998), chain entropy contributions to the folding barrier have remained difficult to verify experimentally. In part, this is due to a lack of an accurate value for the configurational diffusion constant (Socci *et al.*, 1996; Dill & Chan, 1997), in the absence of which there is no currently available experimental measure of the entropy of a folding transition state. Moreover, even if the diffusion factor were well established it would remain difficult, if not impossible, to distinguish experimentally the relative contributions of chain and solvent entropy. We propose, however, that the correlations reported here represent qualitative experimental evidence for the contribution of chain entropy loss to the free energy barrier of folding. To wit, the available data suggest that proteins that fold *via* poorly packed transition states (i.e. proteins that surmount the rate-limiting step with smaller chain entropy losses) tend to fold more rapidly than those characterized by highly organized transition states. Moreover, because the barrier to folding in the simple model described above is entirely due to loss of configurational entropy, the similarity in the contact order dependencies of the folding rate and transition state placement in Figures 1 and 5 suggests that chain entropy losses also contribute to the free energy barriers of the folding of real proteins.

Conclusions

The recent characterization of the refolding properties of a number of simple, single domain proteins has provided an opportunity to demonstrate that the relative contact order of the native state is a determinant of both the height and placement of the folding transition state barrier. The influences of other factors, such as equilibrium stability and chain length, are either not apparent or only weakly supported by the test set presently available. No doubt the rapidly increasing protein folding data base will soon provide an opportunity to more fully characterize the importance of such factors.

Materials and Methods

We are aware of 22 monomeric, single domain proteins which lack disulfide bonds and *cis* proline residues, which have been suggested to fold *via* two-state kinetics under at least some conditions and for which most of the appropriate structural and kinetic data are available. Multiple members of homologous families were not included in the test set in order to avoid over representation of a single topology or length. Thus, eight of the 22 proteins were excluded because they exhibit significant (>25%) sequence identity with proteins already in the set: three src-homology 3 domains (spectrin SH3, Viguera *et al.*, 1994; PI3K SH3, Guijarro *et al.*, 1998; src-SH3, Grantcharova & Baker, 1997), two fibronectin type III domains (^{9}FN3 and ^{10}FN3, Plaxco *et al.*, 1997), two acyl carrier binding proteins (rat and yeast ACBP, Kragelund *et al.*, 1996), and one cytochrome *c* (yeast cytochrome *c*, Mines *et al.*, 1996). The representative family members included in the test set were either the first published example (bovine ACBP and equine cytochrome *c*) or the homologue for which the most complete data set is available (FynSH3 and TnFN3). Also omitted from the test set were two circular permutants of the spectrin SH3 domain (Viguera *et al.*, 1995, 1996).

The wide variety of experimental conditions under which the folding data were collected generates ''noise'' which may obscure any underlying correlations. The refolding data were collected within the pH range 5.0 to 7.2. Experiments with protein L (Yi *et al.*, 1997), CI-2 (Oliveberg & Fersht, 1996), ^{9}FN3, ^{10}FN3 (Plaxco *et al.*, 1997; K.W.P. & C. M. Dobson, unpublished data) and srcSH3 (V. Grantcharova & D.B., unpublished data) suggest that pH changes over this range are unlikely to alter ln(*k*) by more than 1 ln unit or to measurably affect θ_m. Studies of cspB (Schindler *et al.*, 1995) and protein L (Scalley & Baker, 1997) indicate that θ_m values are also unlikely to change significantly over the 17°C temperature range over which the data were collected. Most of the data were collected within an 8°C temperature range. Experiments with protein L (Scalley & Baker, 1997), fynSH3 (Plaxco *et al.*, 1998), CI-2 (Tan *et al.*, 1996), HPr (van Nuland *et al.*, 1998), cspB (Schindler *et al.*, 1995), ^{10}FN3 (K.W.P. & C. M. Dobson, unpublished data) and AcP (F. Chiti, N. A. J. van Nuland, N. Taddei, F. Magherini, M. Stefani, G. Ramponi & C. M. Dobson, personal communication) suggest that over this temperature range folding rates are unlike to vary by more than 1 ln unit. The data for λ-repressor was collected 12°C above the median experimental conditions. Investigations of a double mutant λ-repressor, which folds somewhat more rapidly than the wild-type sequence (Burton *et al.*, 1996), indicate however that ln(k) is unlikely to vary by more than 1.4 ln units over this temperature range (T. Oas, personal communication), an amount which will not significantly affect the observed correlations.

Intrinsic folding rates, kinetic *m*-values and folding free energies were taken directly as reported (Table 1). Data for folding free energies and kinetics were taken under identical conditions in all cases. In cases for which multiple values of a given parameter are reported the more recently reported data was used. Lengths were taken as the total length of the construct studied. Using instead the total number of well-ordered residues does not significantly affect the reported correlations (data not shown). For a two-state folding protein the equilibrium *m*-value can be obtained from the difference between the kinetic *m*-values of folding and unfolding. In order to reduce the likelihood of any bias arising from systematic errors in the direct determination of m_{eq} the ratio θ_m, which is given by $m_f^\ddagger\ /m_{eq}$, was calculated as $m_f^\ddagger/(m_f^\ddagger - m_u^\ddagger)$ (Jackson & Fersht, 1991). Equilibrium (Myers *et al.*, 1995) and kinetic (Plaxco *et al.*, 1997; Clarke *et al.*, 1997; V. Grantcharova & D.B., unpublished data; T. Schindler, personal communication) studies indicate that θ_m values obtained using urea can be directly compared to those obtained using guanidine hydrochloride and thus we have made no distinction between the two. θ_{C_p} values were taken as reported and typically calculated as the ratio $\Delta C_p^{\ddagger}(f)$ to ΔC_p(eq). The intrinsic folding rates extrapolated from two-state folding conditions were used as a convenient parameter for the comparison of two-state refolding rates. It should be noted, however, that under the conditions employed two proteins in the data set (cytochrome *c* and ubiquitin) exhibit ''roll-over'' (Sosnick *et al.*, 1996; Khorasanizadeh *et al.*, 1993). Thus the extrapolated intrinsic refolding rate may not accurately reflect the actual folding rate in the absence of denaturant.

Contact orders and helical content were determined from structural coordinates recorded in the Brookhaven Data Bank (Bernstein *et al.*, 1977). Relative contact order was calculated using the formula described in Results (equation (1)). Residues were considered contacting if they contained non-hydrogen atoms that are within 6.0 Å. Cutoffs from 3.5 to 8 Å were investigated but do not significantly affect the correlations described in this work (data not shown). Limiting contacts to hydrophobic residues slightly reduces the significance of the correlations reported here and limiting contacts to potential backbone hydrogen bonding pairs increases somewhat the significance of *CO versus* θ_m and decreases somewhat the significance of *CO versus* ln(*k*) (data not shown). Helical content was calculated using the second-

Table 1. The non-homologous set of simple, single domain proteins used in this study

Name	PDB code	Length	ΔG_u (kcal mol^{-1})	ln(k_f)	θ_m (%)	θ_{C_p} (%)	*CO* (%)	Helix (%)	Temperature (°C)
A. *Helical*									
λ-Repressor[a]	1LMB3	80	3.3	8.19	46	–	9.4	73	37
Equine cyt *c*[b]	1HRC	104	8.2	8.76	40	–	11.2	41	23
Bovine ACBP[c]	2ABD	86	7.0	6.55	63	–	14.0	60	20
B. *Mixed*									
Ubiquitin[d]	1UBQ	76	7.2	7.33	59	–	15.1	24	25
CI-2[e]	1CIS	83	7.0	3.87	61	51	16.4	17	25
ADA2h[f]	1PCA	80	4.1	6.80	74	–	17.0	25	25
Protein L[g]	2PTL	63	4.6	4.10	75	42	17.6	19	22
HPr[h]	1HDN	85	4.7	2.70	64	52	18.4	38	20
Muscle AcP[i]	1APS	98	4.5	−1.48	79	40	21.2	18	28
C. *Sheet*									
CspB[j]	1CSP	67	2.1	6.98	96	90	16.4	4	25
TnFN3[k]	1TEN	90	5.3	1.06	60	–	17.4	0	20
FynSH3[l]	1SHFA	67	6.0	4.55	68	51	18.3	5	20

[a] Huang & Oas (1995a,b)

[b] Mines *et al.* (1996), Sosnick *et al.* (1996), extrapolated folding rate: J. Winkler & H. Gray (personal communication).

[c] Kragelund *et al.* (1995, 1996), Data at 20°C: B. Kragelund (personal communication).

[d] Khorasanizadeh *et al.* (1993).

[e] Jackson & Fersht (1991), Tan *et al.* (1996).

[f] Villegas *et al.* (1995), kinetics: human ADA; structure: porcine ADA.

[g] Scalley *et al.* (1997), Scalley & Baker (1997).

[h] van Nuland *et al.* (1998).

[i] F. Chiti, N. A. J. van Nuland, N. Taddei, F. Magherini, M. Stefani, G. Ramponi & C. M. Dobson (personal communication).

[j] Schindler *et al.* (1995), Schindler & Schmid (1996).

[k] Clarke *et al.* (1997).

[l] Plaxco *et al.* (1998).

ary structure assignments of the dictionary of secondary structural preferences (DSSP: Kabsch & Sander, 1983).

Single variable linear relationships were assumed. Multivariable linear models were also investigated and suggest that length may be an important secondary determinant of ln(k) and θ_m (K.W. P. & I. Ruczinski, unpublished data), but the small size of the present data set makes difficult the justification of these more complex models. These linear models may be very poor approximations of the actual relationships between the variables under investigation. Correctly addressing this issue, however, will also require significantly larger data sets. The reported correlation coefficients, p-values and Cook's distances were calculated using S-plus (MathSoft, Inc). Reported p-values are the probability that an observation from a t-distribution with ten degrees of freedom would exceed the ratio of the estimate of the slope to the estimate of the standard deviation. The p-values thus represent the probability that, if the null hypothesis were true (i.e. that there exists no relationship and the true slope is zero), an estimate of the slope would be generated as far or farther from zero than that actually observed. Cook's distance is a measure of the impact of a given data point on a postulated relationship. A Cook's distance of >0.5 suggests that the data point is a significant outlier.

Acknowledgments

The authors gratefully acknowledge a long-standing collaboration with Chris Dobson as the source of much of the data used in this analysis. The authors also thank R. Baldwin, H. S. Chan, F. Chiti, K. Dill, V. Daggett, C. Dobson, K. Fiebig, H. Gray, M. Gross, B. Kragelund, T. Oas, M. Scalley, D. Shortle, D. Teller and D. Thirumalai for helpful reviews of this paper and L. Plaxco and I. Ruczinski for invaluable aid and advice on the statistical analysis. We are also deeply indebted to F. Chiti, J. Clarke, C. Dobson, V. Grantcharova, J. I. Guijarro, B. Kragelund, N. vanNuland, M. Scalley, T. Schindler, T. Oas and J. Winkler for generously providing important unpublished and pre-publication results. K.T.S. was supported by a NIH predoctoral training grant (GM07270). This work was supported by young investigator awards to D.B. from the NSF and the Packard Foundation.

References

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995). Impact of local and non-local interactions on the thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460–471.

Alexander, P., Orban, J. & Bryan, P. (1992). Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G. *Biochemistry*, **31**, 7243–7248.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

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

Burton, R. E., Huang, G. S., Daugherty, M. A., Fullbright, P. W. & Oas, T. G. (1996). Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **263**, 311–322.

Chen, B. L., Baase, W. A. & Schellman, J. A. (1989). Low temperature unfolding of a mutant of phage T4 lysozyme. 2. Kinetic investigations. *Biochemistry*, **28**, 691–699.

Clarke, J., Hamill, S. J. & Johnson, C. M. (1997). Folding and stability of a fibronectin type III domain of human tenascin. *J. Mol. Biol.* **270**, 771–778.

Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struc. Biol.* **4**, 10–19.

Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993). Cooperativity in protein-folding kinetics. *Proc. Natl Acad. Sci. USA*, **90**, 1942–1946.

Doyle, R., Simons, K., Qian, H. & Baker, D. (1997). Local interactions and the optimization of protein folding. *Proteins: Struct. Funct. Genet.* **29**, 282–291.

Fersht, A. R. (1995a). Mapping the structures of transition states and intermediates in folding: delineation of pathways at high resolution. *Phil. Trans. Roy. Soc. London*, **348**, 11–15.

Fersht, A. R. (1995b). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. U.S.A*, **92**, 10869–10873.

Finkelstein, A. V. (1991). Rate of β-structure formation in polypeptides. *Proteins: Struct. Funct. Genet.* **9**, 23–27.

Finkelstein, A. V. & Badretdinov, A. Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Folding Design*, **2**, 115–121.

Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210.

Govindarajan, S. & Goldstein, R. A. (1995). Optimal local propensities for model proteins. *Proteins: Struct. Funct. Genet.* **95**, 413–418.

Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry*, **36**, 15685–15692.

Gross, M. (1996). Linguistic analysis of protein folding. *FEBS Letters*, **390**, 249–252.

Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D. & Dobson, C. M. (1998). Folding kinetics of the SH Domain of PI3 K by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* In the press.

Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1996). Chain length scaling of protein folding time. *Phys. Rev. Letters*, **77**, 5433–5436.

Huang, G. S. & Oas, T. G. (1995a). Submillisecond folding of monomeric λ-repressor. *Proc. Natl Acad. Sci. U.S.A*, **92**, 6878–6882.

Huang, G. S. & Oas, T. G. (1995b). Structure and stability of monomeric λ-repressor: NMR evidence for two-state folding. *Biochemistry*, **34**, 3884–3892.

Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2.1. Evidence for a two-state transition. *Biochemistry*, **30**, 10428–10435.

Jacob, M., Schindler, T., Balbach, J. & Schmid, F. X. (1997). Diffusion control in an elementary protein folding reaction. *Proc. Natl Acad. Sci. USA*, **94**, 5622–5627.

Kabsch, W. & Sander, C. (1983). Dictionary of secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karplus, M. & Weaver, D. L. (1994). Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* **3**, 650–668.

Khorasanizadeh, S., Peters, I. D., Butt, T. R. & Roder, H. (1993). Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry*, **32**, 7054–7063.

Klimov, D. K. & Thirumalai, D. (1997). Factors governing the foldability of proteins. *Proteins: Struct. Funct. Genet.* **26**, 411–441.

Kragelund, B. B., Robinson, C. V., Knudsen, J., Dobson, C. M. & Poulsen, F. M. (1995). Folding of a four-helix bundle: studies of acyl-coenzyme A binding protein. *Biochemistry*, **34**, 7217–7224.

Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J. & Poulsen, F. M. (1996). Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* **256**, 187–200.

Mines, G. A., Pascher, T., Lee, S. C., Winkler, J. R. & Gray, H. B. (1996). Cytochrome *c* folding triggered by electron transfer. *Chem. Biol.* **3**, 491–497.

Munoz, V. & Serrano, L. (1996). Local *versus* non-local interactions in protein folding and stability–an experimentalist's point of view. *Folding Design*, **1**, R71–R77.

Myers, J. K., Pace, C. N. & Scholtz, J. M. (1995). Denaturant *m* values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148.

Oliveberg, M. & Fersht, A. R. (1996). Formation of electrostatic interactions on the protein folding pathway. *Biochemistry*, **35**, 2726–2737.

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

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). On the theory of folding kinetics for short proteins. *Folding Design*, **2**, 109–114.

Plaxco, K. W., Spitzfaden, C., Campbell, I. D. & Dobson, C. M. (1997). Comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* **270**, 763–770.

Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998). The folding kinetics and thermodynamics of the Fyn-SH3 Domain. *Biochemistry*, In the press.

Roder, H. & Colón, W. (1997). Kinetic role of early intermediates in protein folding. *Curr. Opin. Struc. Biol.* **7**, 15–28.

Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature*, **369**, 248–251.

Scalley, M. L. & Baker, D. (1997). Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc. Natl Acad. Sci. USA*, **94**, 10636–10640.

Scalley, M. L., Yi, Q., Gu, H., McCormack, A., Yates, J. R. & Baker, D. (1998). Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry*, **36**, 3373–3382.

Schindler, T. & Schmid, F. X. (1996). Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry*, **35**, 16833–16842.

Schindler, T., Herrler, M., Marahiel, M. A. & Schmid, F. X. (1995). Extremely rapid protein folding in the absence of intermediates. *Nature Struct. Biol.* **2**, 663–673.

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

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.

Tan, Y.-J., Oliveberg, M. & Fersht, A. R. (1996). Titration properties and thermodynamics of the transition state for folding: comparison of two-state and multistate folding pathways. *J. Mol. Biol.* **264**, 377–389.

Tanford, C. (1970). Protein denaturation. Part C. Theoretical models for the mechanism of denaturation. *Advan. Protein Chem.* **24**, 1–95.

Thirumalai, D. (1995). From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys. I. (France),* **5**, 1457–1467.

Unger, R. & Moult, J. (1996). Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **259**, 988–994.

van Nuland, N. A. J., Meijberg, W., Warner, J., Forge, V., Scheek, R. M., Robillard, G. T. & Dobson, C. M. (1998). Slow cooperative folding of a small globular protein HPr. *Biochemistry,* **37**, 622–637.

Viguera, A. R., Martínez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L. (1994). Thermodynamic and kinetic analysis of the SH3 domain of spectrin shows a two-state folding transition. *Biochemistry,* **33**, 2142–2150.

Viguera, A. R., Blanco, F. J. & Serrano, L. (1995). The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.* **247**, 670–681.

Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874–880.

Viguera, A. R., Villegas, V., Aviles, F. X. & Serrano, L. (1997). Favourable native-like helical local interactions can accelerate protein folding. *Folding Design*, **2**, 23–33.

Villegas, V., Azuaga, A., Catasús, L., Reverter, D., Mateo, P. L., Avilés, F. X. & Serrano, L. (1995). Evidence for a two-state transition in the folding process of the activation domain of human procarboxypeptidase A2. *Biochemistry,* **34**, 15105–15110.

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

Wolynes, P. G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl Acad. Sci. USA,* **94**, 6170–6175.

Yi, Q., Scalley, M. L., Simons, K. T., Gladwin, S. T. & Baker, D. (1997). Characterization of the free energy spectrum of peptostreptococcal protein L. *Folding Design*, **2**, 271–280.

Zwanzig, R. (1995). Simple model of protein folding kinetics. *Proc. Natl Acad. Sci. USA,* **92**, 9801–9804.

Edited by P. E. Wright

(Received 16 October 1997; received in revised form 16 January 1998; accepted 16 January 1998)

Note added in proof: Since this paper was accepted for publication, it has come to our attention (B. Kuhlman & D. Raleigh, personal communication) that the 56 residue amino-terminal domain of ribosomal protein L9 exhibits two-state folding kinetics with $\ln(k) = 6.57$ and $\theta_m = 60\%$ (at 25°C). The contact order of this domain, 12.7%, corresponds to $\ln(k) = 7.55$ and $\theta_m = 55\%$ using the correlations noted in Figure 1. The correspondance between these values is fully consistent with the results reported here and improve their statistical significance.