Prediction of Folding Rates of Small Proteins: Empirical Relations Based on Length, Secondary Structure Content, Residue Type, and Stability[†]

N. Prakash Prabhu and Abani K. Bhuyan*

School of Chemistry, University of Hyderabad, Hyderabad 500046, India

Received October 17, 2005; Revised Manuscript Received January 21, 2006

ABSTRACT: Delineating the determinants of folding of small proteins is essential for decoding the folding code. By considering the literature data for the folding of 45 two-state proteins that differ widely in sequence, structure, and function, this work suggests two empirical relations for the prediction of the folding rate. One relation is based on the content of secondary structure elements, and the other uses the number of residues of each of the following types: hydrophobic, positively charged, and negatively charged. Both relations incorporate the chain length as an indirect descriptor. The correlation between experimental values for folding rates and free energy is poor, providing little support to the "stability gap" hypothesis based on the folding of simple lattice and off-lattice polymers. The average rate-determining barrier height for these two-state proteins is much larger than the small barriers for the transition-state ensemble envisaged by theoretical models.

Excitement in the protein folding field over the past 20 years has been fueled by very significant development along several lines of research. The most notable examples include atomic-level characterization of temporally resolved folding intermediates for proteins that apparently fold via multistate kinetics (*1*, *2*), the finding of a substantially large set of two-state or class II proteins (*3*) for which no structural intermediate is populated to a detectable level (*4*−*6*), and the emergence of theoretical models inspired by the folding of lattice polymers (*7*−*10*). The idea that obligates intermediates and the observation of straight two-state folding appear mutually restrictive. While folding models that engage kinetic intermediates sequentially (*11*, *12*) seem to provide clues to Levinthal's conformational search enigma, the volume of experimental data for two-state proteins raises doubts that the accumulation of kinetic intermediates resolves the problem. The theoretical landscape model leaves it to the sequence-based funneled organization of the energy landscape that dominates the folding kinetics (*13*). Folding on a sufficiently smooth funnel is rapid with no traps, the analogue of the phenomenological two-state model. Folding on a rugged funnel would appear to be multistate, since discrete intermediates, labeled "frustrated" or "misfolded", appear trapped at much later stages of folding (*14*, *15*).

A unified mechanism is needed for the conceptual understanding of the folding problem. Theory and simulations suggest that the funnel landscape does provide a unified picture, because the great variety of detailed mechanisms at work in the funnel influences the folding trajectories (*13*, *16*, *17*). However, conflicts between experimental observations and theory-based tenets have been found (*6*, *18*−*20*). Another basic problem with the funnel perspective is that it does not attach much significance to the decoding of the primary structure in the three-dimensional native structure (see ref *18*). In a recent work, Kuwajima and co-workers have attempted to unify the folding mechanisms, in a classical sense, of non-two-state and two-state proteins (*21*).

One approach to understanding the basic mechanism of folding, first taken up in Baker's laboratory (*22*), is to define the determinants of folding kinetics of small two-state proteins by considering the available literature on sequence, structure, and folding. For a data set of 12 proteins, percent contact order was found to correlate strongly with folding rate constants (*22*). A later analysis using 24 proteins showed a stronger correlation between relative contact order and folding rate (*23*). Several related studies, including the prediction of folding rates based on the first principles of protein folding (*24*), calculation of folding rates from three-dimensional structure (*25*), and correlation of folding rates with stability and contact order (*26*), long-range order (*27*, *28*), and total contact distance (*29*), have been reported since then. In the most recent study of this kind, transition-state contact orders for 10 proteins determined by molecular dynamics simulation have been found to correlate with folding rates (*30*). In a somewhat different approach, prediction of folding rates by analysis of local secondary structure content for a set of 24 proteins has been described (*31*).

The scaling of folding times, τ_f, with protein size has also been considered. While theoretical studies have projected the predictive value of the number of residues in the protein, N, through the simple relation $\log(\tau_f) \approx N^{\beta}$, in which the value of β lies in the range of 0.5−0.67 (*32*−*34*), a recent database analysis of various size proteins and peptides with

[†] This work was supported by grants from the Department of Biotechnology (BRB/15/227/2001), the Department of Science and Technology (4/1/2003-SF), and the University Grants Commission (UPE Funding), Government of India. N.P.P. acknowledges a senior research fellowship from the Council of Scientific & Industrial Research. A.K.B. is the recipient of a Swarnajayanti Fellowship from the Department of Science and Technology, Government of India.

* To whom correspondence should be addressed. E-mail: akbsc@uohyd.ernet.in. Fax: 91-40-23012460.

10.1021/bi0521137 CCC: $33.50 © 2006 American Chemical Society
Published on Web 03/03/2006

Table 1: Proteins Used in This Study[a]

label	protein	PDB entry	structure type
1	λ-repressor	1lmb	α
2	ACBP bovine	1hb6	α
3	ACBP rat	models	α
4	ACBP yeast	models	α
5	ACBP *Plasmodium falciparum*	1hbk	α
6	horse ferricytochrome *c*	1hrc	α
7	horse ferrocytochrome *c*	1giw	α
8	yeast ferricytochrome *c*	1yic	α
9	yeast ferrocytochrome *c*	1ycc	α
10	Psb D	2pdd	α
11	Im 9	1imq	α
12	reduced cytochrome b_{562}	1m6t	α
13	Engrail homeo domain	1enh	α
14	villin head piece	1vii	α
15	Tendamistat	2ait	β
16	*Bacillus subtilis* Csp	1csp	β
17	*Bacillus caldolylicus* Csp	1c9o	β
18	*Thermotoga maritima* Csp	1g6p	β
19	Csp A	3mef/1mjc	β
20	α-spexctrin	1aey	β
21	Src SH3 domain	1srl	β
22	SH3-Fyn	1a0n	β
23	9-fibronectin-III	1fnf-3	β
24	10-fibronectin-III	1fnf-4	β
25	Twitchin	1wit	β
26	Tenascin (short)	1ten	β
27	Tenascin (long)	−	β
28	Psa E	1pse	β
29	Ti-I27	1tit	β
30	SH3 pl 3-kinase	1pks	α and β
31	CI2	2ci2	α and β
32	procarboxy peptidase A2	1pba	α and β
33	Arc repressor	1arr	α and β
34	ubiquitin	1ubq	α and β
35	protein L	2ptl	α and β
36	protein U1A	1urn	α and β
37	Hpr	1hdn	α and β
38	FKBP 12	1fkb	α and β
38A	FKBP 12	1fkb	α and β
39	muscle AcP	1aps	α and β
40	villin 14T	1vik	α and β
41	Che Y	3chy	α and β
42	protein G B1 domain	1em7	α and β
43	protein L9 (1−56)	1div	α and β
44	Sso 7D	1bnz	α and β
45	Mer P	1osd	α and β

[a] The labels correspond to those shown in Figure 1.

sizes ranging from 16 to 396 residues has shown that $\beta = 0.5$ (*35*; see also ref *36*). This scaling simply involves the folding rate and the number of residues and does not consider any structural or chemical details of the proteins. We were interested in knowing how the folding rates of simple benchmark proteins are correlated with the details of secondary structure and random coil content, the content of amino acid residues classified according to their chemical nature, and the total number of residues in the proteins. We also sought to find the dependence of the equilibrium surface exposure of residues in global unfolding ($m_{\rm eq}$) and the approximate position of the folding transition state along the reaction coordinate ($\alpha^{\ddagger}_{\rm f}$) on protein length.

Here, we consider the experimental folding parameters in a data set of 45 two-state proteins (Figure 1 and Table 1) and establish empirical relations between chain length and equilibrium *m* value, between chain length with percent secondary structure elements and folding rate, and between chain length with classified residue types and folding rate. The correlations are sufficiently strong to enable prediction of folding rates from the primary sequence data. The data have also been viewed from the energy landscape perspective. Consistent with earlier finding of Plaxco et al. (*23*) and others (refs *5* and *37*, for example), the native-state stability is poorly correlated with the folding rate. Additionally, we find no considerable correspondence between theory and experiments regarding the size of the barrier energy, although the location of the classical transition-state barrier along the folding coordinate derived from experiments agrees fairly well with the theoretical prediction of the region for the transition-state ensemble.

METHODS

Forty-five two-state proteins from different families with different secondary structures were taken for the analysis. Structural properties of the proteins were obtained from Protein Data Bank. All thermodynamic and kinetic properties were collected from the literature. Secondary structure propensities were calculated from the PROSS program developed in the laboratory of G. Rose. Contact orders were obtained from Plaxco's Contact Order Calculator. The amino acid composition and numbers of hydrophobic, positively charged, and negatively charged residues were calculated with SAPS.

RESULTS AND DISCUSSION

Length of Two-State Proteins and Folding Parameters. The number of residues (N) is not directly correlated with any of the folding parameters, namely, equilibrium surface exposure of residues during unfolding ($m_{\rm eq}$), free energy of unfolding ($\Delta G°$),[1] rates of folding and unfolding ($k_{\rm f}$ and $k_{\rm u}$, respectively), and position of the folding transition state ($\alpha^{\ddagger}_{\rm f}$). N may have an indirect contribution to all or some of these parameters. For example, Figure 2 shows a linear analysis of N and $m_{\rm eq}$ for 44 proteins by the empirical relation

$$\frac{1}{N} = a + \frac{b}{Nm_{\rm eq}} \tag{1}$$

where $a = 0.009$ and $b = 0.403$. Both the correlation coefficient ($r = 0.81$) and the statistical significance ($p < 0.0001$) are fairly good, suggesting that N could be an indirect determinant for folding parameter values.

A very strong correlation between N and the calculated change in total surface area in unfolding ($A_{\rm u} - {\rm A}_{\rm n}$, where $A_{\rm u}$ and $A_{\rm n}$ are total solvent accessible surface areas in unfolded and native states, respectively) has been shown for a mixed set of two- and higher-state proteins (*38*). Then, a linear correlation of N with $m_{\rm eq}$ is also expected, since $m_{\rm eq}\alpha(A_{\rm u} - A_{\rm n})$ (*39*). The reason this direct correlation is not observed here to any significant extent ($r = 0.37$) must be due to different levels of unfoldedness and chain configuration for different proteins. Incomplete chain expansion and the persistence of loops or local structural elements in the unfolded state affect the total surface area accessible to water. Equation 1 provides a means of estimating the effective $m_{\rm eq}$ value based on the length of any two-state protein.

[1] Abbreviations: GdnHCl, guanidine hydrochloride; $\Delta G_{\rm D}$, Gibbs energy of denaturation; $\Delta G°$, Gibbs energy of denaturation in the absence of denaturant; cyt *c*, cytochrome *c*; ferricyt *c*, ferricytochrome *c*; ferrocyt *c*, ferrocytochrome *c*; CD, circular dichroism.

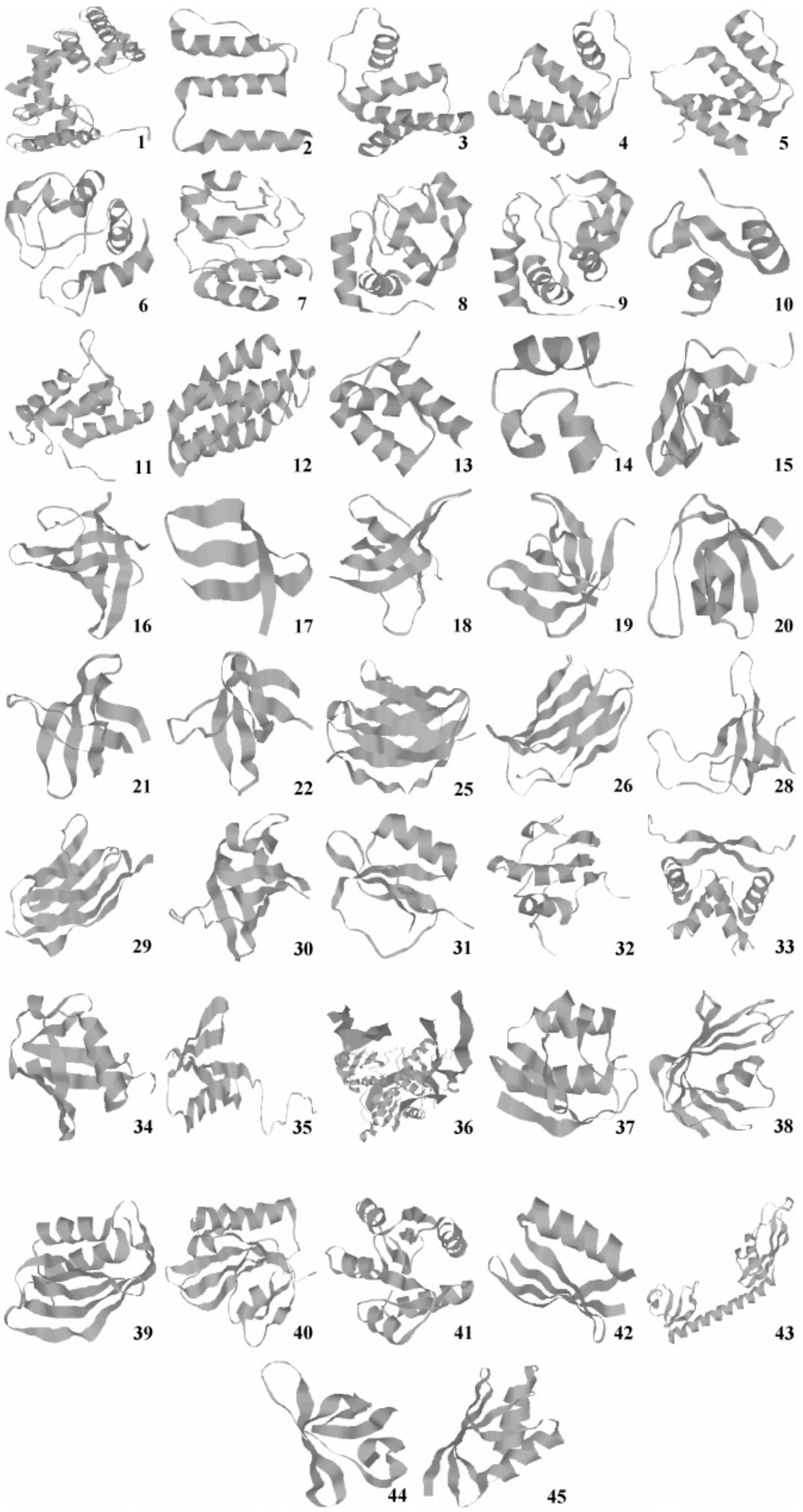

FIGURE 1: Structures of 42 of the 45 proteins used for analysis in this work. The proteins corresponding to the labels are listed in Table 1. Additional information, including length, number of residue types, secondary structure content, and equilibrium and kinetic folding parameters, is given in the Supporting Information.

Although a correlation between the folding rate and the chain length has been reported for Gô model proteins (*40*), we did not find a direct relationship between the length and the folding rate of the naturally occurring two-state proteins, consistent with earlier reports (*23*, *41*). We therefore examined the correlations between secondary structure with random coil content and experimentally observed folding rates for 39 two-state proteins in our test set by setting N as an indirect determinant. Figure 3a shows plots of the results. The best-fit linear relationship between ($\ln k_f$)/N and the

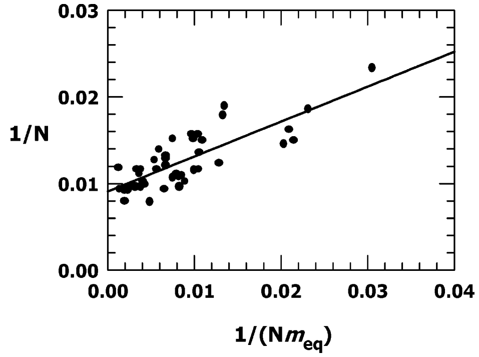

FIGURE 2: Correlation between 1/*N* and 1/(*Nm*$_{eq}$) for 44 proteins given by eq 1. The solid line is the best fit for these data.

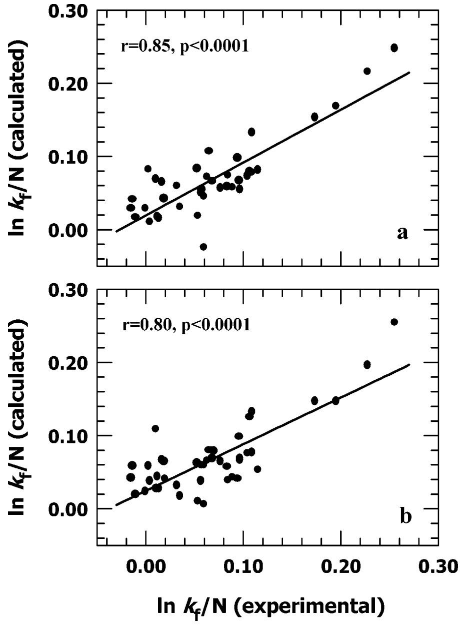

FIGURE 3: Predicted and experimental folding rates. (a) Prediction based on secondary structure and random coil content for 39 proteins, with coefficients from eq 2. (b) Prediction on the basis of the content of hydrophobic, positively charged, and negatively charged residues for 43 proteins. The coefficients are from eq 3.

percentages of residues forming different secondary structural elements is

$$\frac{\ln k_f}{N} = -0.136(\pm 0.0802) + 0.0013(\pm 0.0008)a + 0.0007b - 0.001c + 0.0016(\pm 0.0014)d + 0.001e + \frac{10.7374(\pm 1.5193)}{N} \quad (2)$$

where *a*−*e* are percentages of residues forming α-helix, *β*-turn, *β*-strand, polyproline II, and random coil, respectively. The numbers in parentheses represent standard errors associated with the coefficients. The correlation is strong ($r =$ 0.85) with a good level of statistical significance ($p <$ 0.0001). A similar analysis using a set of 24 two-state proteins has been reported recently, where % helix, % hairpin, and % turn are used in conjunction with chain length (*31*). This work considers % polyproline II and % random coil, in addition. Since each type of secondary structure content enters as a discrete term in the summation (eq 2), a zero value is registered for a term when the corresponding structural element is absent.

In the search for a residue-type trend, amino acids were classified as hydrophobic, positively charged, and negatively charged. The linear correlation between the sum of numbers of classified residue types and the folding rate was poor (r = 0.43). We therefore incorporated the chain length as a descriptor. The best-fit relationship between (ln k_f)/*N* and the number of different residue types for 43 two-state proteins was found to be

$$\frac{\ln k_f}{N} = -0.2094(\pm 0.0712) + 0.0001l + 0.0043(\pm 0.0015)m + 0.0023(\pm 0.0018)n + \frac{15.3449 \pm 2.62}{N} \quad (3)$$

where *l*−*n* are the number of hydrophobic, positively charged, and negatively charged residues, respectively. A fairly strong correlation ($r = 0.80$, $p < 0.0001$) is evident.

These empirical relations provide bases for predicting folding rates from sequence alone, without a detailed knowledge of the three-dimensional molecular structure (*25*). While the number of hydrophobic and charged residues can be obtained by inspection of the primary structure, the fraction of the sequence in each secondary structural element can be calculated by any available secondary structure identification programs, PROSS, for example. We note that eqs 2 and 3 provided here for the two-state proteins ($r \geq$ 0.8) use the protein length as an indirect descriptor and are clearly distinct from the general relation $\log(\tau_f) = N^{\beta}$ that scales the folding time simply with the number of amino acids (*33*, *34*). By using a β of 0.5 as prescribed (*35*), one finds $r = 0.6$ for 36 of the two-state proteins considered here. Clearly, the folding rate is predicted well when *N* is used in conjunction with secondary structure content or residue type details.

Relations predictive of two-state folding rates based on contact order in both native (*22*) and transition states (*30*), relative contact order (*23*), long-range order (*28*), total contact distance (*29*), and percent secondary structural elements (*31*) have been advanced by several studies in which smaller-size data sets were analyzed. In this study, we repeated some of those analyses using a relatively larger data set (Table 2). To analyze the contact order−rate relationship, we took 29 proteins, 16 from the data set of 24 proteins reported previously (*23*) and 13 from our collection. The analysis showed poor linear relationships between ln k_f and contact order and between ln k_f and relative contact order. The prediction of folding rate based on local secondary structure content was also poor (Table 2). The lower records of the strength of the linear relationships could be due to a gross error of measurement or an error in recording the data for some of the proteins in our data set, producing many outliers in the final correlation plots. The data for various proteins comprising the data set used in this work were acquired under various pH, temperature, denaturant, and solvent conditions. To exclude the effect of different experimental conditions for different proteins, we carried out the correlation analyses using data from the recently reported kinetic data set of 30 two-state proteins that were studied under a “standard” set of experimental conditions (*42*). Of these 30, only those that also figured in our data set were chosen. The contact order versus ln k_f and the relative contact order versus ln k_f relationships were further analyzed using 19 proteins, all from

Table 2: Predicted and Experimentally Observed Folding Rates

		r value from this work (Table 1)				literature		
independent varible(s)	dependent variable	all proteins	all α	all *β*	α + *β*	no. of proteins	*r* value	ref
secondary structure and random coil content and *N*	(ln k_f)/*N*	0.85	0.97	0.83	0.77			
no. of hydrophobic, positively charged, and negatively charged residues and *N*	(ln k_f)/*N*	0.80	0.95	0.68	0.80			
secondary structure content and *N*	ln k_f	0.58	0.60	0.79	0.34	24	0.91	*31*
contact order	ln k_f	0.62	0.68	0.64	0.44	12	0.81	*22*
relative contact order	ln k_f	0.63	0.68	0.64	0.44	24	0.92	*23*
Δ*G*°	ln k_f	0.31	0.18	0.30	0.52	24	0.40	*23*

Table 3: Analyses Using the "Standard" Two-State Data Set Reported in Ref *42*

independent variable(s)	dependent variable	*r* value	no. of proteins
secondary structure and random coil content and *N*	(ln k_f)/*N*	0.81	14[a]
no. of hydrophobic, positively charged, and negatively charged residues and *N*	(ln k_f)/*N*	0.50	14[a]
secondary structure content and *N*	ln k_f	0.78	14[a]
contact order	ln k_f	0.79	11[a]
relative contact order	ln k_f	0.69	11[a]
Δ*G*°	ln k_f	0.26	13[a]
contact order	ln k_f	0.33	19[b]
relative contact order	ln k_f	0.49	19[b]

[a] Number of proteins common to our surveyed data set and the experimental data set of Maxell et al. (*42*). [b] Number of proteins taken from the study of Maxell et al. (*42*).

the standard data set of Maxell et al. (*42*). Table 3 lists the correlation coefficients that were found. The coefficients for these two determinants are poor again. The lower correlation coefficients may then arise from those proteins for which chain topology is not a critical determinant of folding kinetics. These proteins may be the outliers in the final correlation plots. Indeed, mutations that do not significantly alter the topology, and hence the contact order, may affect the folding rate (*23*). The analyses presented indicate better predictive values of secondary structure content. The prediction becomes even better when % polyproline II and % random coil are considered.

Experimental Results and Theoretical Predictions of Two-State Folding. Numerous theoretical and computer simulation studies have endeavored to provide a general framework, popularly called funnel perspective, for the understanding of the protein folding problem. Just as the fast folding of lattice- and simplified off-lattice polymer models (*7*−*10*) results from a smooth energy landscape, one can imagine a smooth funnel for rapid folding of a perfect two-state folder. As folding progresses, the energy of the folding ensemble decreases in a smooth and monotonic manner all the way to the bottom of the funnel. The large data set compiled here for two-state proteins provides an opportunity to check the extent to which the predictions of funnel-guided folding properties are met.

Funnel Depth and Folding Speed. Figure 4a shows the dependence of ln k_f on Δ*G*° for 45 two-state proteins. The correlation is very poor ($r = 0.32$), indicating that the folding speed is hardly dependent on the protein stability. The observed rate constant for unfolding in water is considerably correlated with Δ*G*° (Figure 4b), apparently suggesting that higher equilibrium stability retards the unfolding rate. An extreme case of this relationship is provided by the folding−unfolding properties of four test tube variants of ferrocytochrome *c* (ferrocyt *c*). Figure 4c shows the dependences of ln k_f and ln k_u on Δ*G*° for ferrocyt *c*, alkaline ferrocyt *c*, carbonmonoxyferrocyt *c* (cyt-CO), and nitrosylferrocyt *c* (cyt-NO). They share identical primary and secondary structures, and very similar native-state tertiary structure, but are far apart in terms of equilibrium stability (*6*, *20*). It would then appear that the natural sequence rather than the stability controls the folding speed. On the other hand, it is the native-state stability that dictates the unfolding speed.

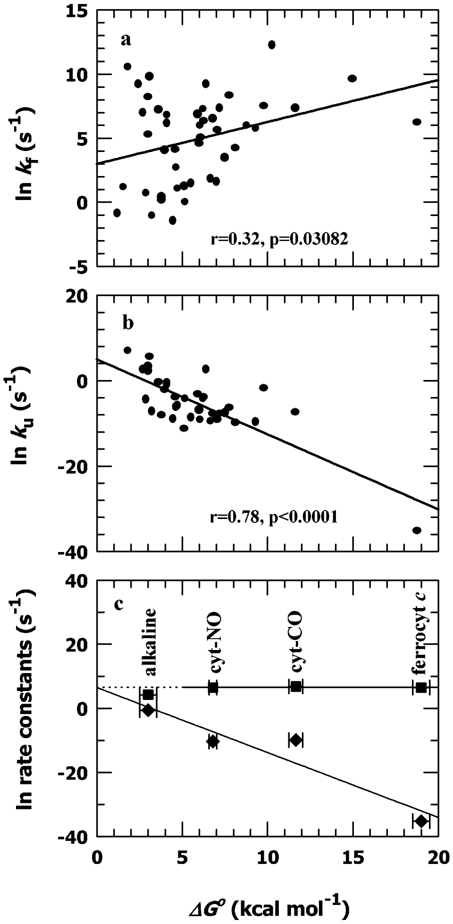

FIGURE 4: Test of the stability gap hypothesis with data for 45 proteins. (a) Ln k_f is very poorly correlated with the free energy of folding. (b) The correlation between ln k_u and Δ*G*° is good. (c) Correlation of natural logarithm of folding and unfolding rates with Δ*G*° for four structurally very similar test tube variants of horse ferrocytochrome *c*: (■) folding rate and (◆) unfolding rate.

The energy landscape approach associates the fast folding speed with the "energy gap" between the native or nativelike states and the set of states with little structural similarity to the native state (*43*−*45*). Translation of this contextual energy gap, called the "stability gap" by Bryngelson et al.

(*45*), to ΔG° ($G_U - G_N$) would associate higher equilibrium stability with larger free energy gradient or energetic drive leading to the native state, especially when no glassy states exist, and hence rapid folding. The energy gap hypothesis, in a subtly different form, has also been advanced by others (*46*, *47*), but the effective predictions are the same. The data survey presented here indicates that the folding speed is not correlated with the free energy gradient or the funnel depth. Studies show that proteins from the same family can produce a good correlation between the folding rate and free energy (*23*, *48*). When we consider all the proteins together, the relation is not significant. Thus, even though the rapid folding of simple lattice polymers may be related to the size of the stability gap, the folding speed of natural two-state proteins is effectually independent of the equilibrium stability. It should be mentioned that a better measure of foldability, for lattice models of proteins at least, is the *Z*-score parameter that estimates the energy of the native conformation relative to the average energy of compact nonnative conformations (*49*). Folding rates of small natural proteins may show good correlation with *Z*-scores. The test of the hypothesis would require experimental determination of the "relative energy of the native conformation".

The energy gap rather appears to be a determinant of the unfolding rate of these small proteins (Figure 4b,c). The considerable correlation between unfolding rates and protein stability may reflect the bearing of the native-state stability on the unfolding activation barrier. It seems the differences in the rate-limiting unfolding barrier heights are similar to the differences in the equilibrium stabilities for these small proteins. This proportionality requires that the unfolding transition state be nativelike structurally, and presumably thermodynamically, so that the energetic stabilization of the transition state is controlled by the native-state stability. More experimental data will be needed to test this conjecture.

Barrier Heights. By assuming that the apparent activation energies for folding and unfolding in water are given by an expression of the type $E_a = -RT \ln(k/A_o)$, where k is the rate coefficient (k_f and k_u) and A_o is the front factor, the heights of activation barriers for folding and unfolding can be estimated, although the appropriateness of the use of this thermally activated rate law in its general form is not quite clear (*50*, *51*). We take A_o as the time needed for two residues of the unfolded chain to come together to form a contact or geminate pair. The most recent ultrafast folding experiment from this laboratory has shown that for ferrocytochrome *c* the diffusion time constant for formation of a contact between two regions separated by 46−60 residues is ~400 ns (*52*). The diffusion times for the first steps in formation of short turns, loops, and helices fall in the 5−40 ns range (*53*; see also ref *54*). From these considerations, the value of A_o can be set in the range of 10^7-10^8 s^{-1}. Using an A_o of 1×10^8 s^{-1}, E_a values associated with refolding and unfolding reactions for 42 and 35 proteins, respectively, were estimated (Figure 5a). The barrier heights average to 7.92 ± 2.1 and 13.3 ± 2.77 kcal/mol, respectively. These averages decrease by 1 ± 0.5 kcal/mol when an A_o of 1×10^7 s^{-1} is used to estimate the barrier sizes.

The landscape perspective is not strict about such barriers. Folding in an extremely smooth funnel may not even encounter any significant barrier (*55*); the retardation of folding here arises from the integral over a number of changing dynamic processes, including motional diffusion, as folding progresses (*13*, *45*, *56*). According to simulation studies, general downhill folding may encounter, at later stages of folding, tiny barriers ($\sim 3k_BT \approx 1.68$ kcal/mol at 10 °C; *14*, *45*) that are significantly smaller in magnitude than the average value of 7.92 kcal/mol shown by the experimental data (Figure 5a).

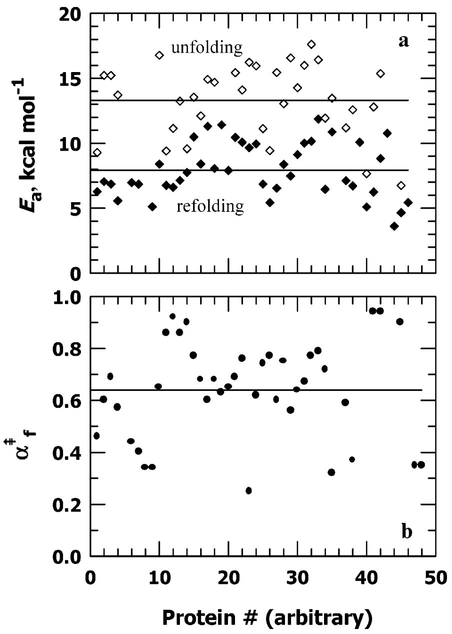

FIGURE 5: (a) Activation energies of folding and unfolding for 42 and 35 proteins, respectively. (b) Barrier locations shown for 41 proteins. The solid lines in both panels show the mean values.

Barrier Location. Classically, an approximation of the extent of structure formation in the transition-state ensemble (TSE) along the folding coordinate is made from the relation $\alpha = m_{f(u)}^‡/m_g$, where $m_f^‡$ and $m_u^‡$ are kinetic *m* values ($=2.3RT\,\partial \log k_{f(u)}/\partial$[GdnHCl]) associated with refolding and unfolding, respectively (*57*). The average of α values, estimated from folding data for 41 two-state proteins, is 0.63 (Figure 5b). This average must be taken at a coarse level, since curvature in the chevron limbs for some two-state proteins introduces large errors in the estimation of $m_f^‡$. Nevertheless, it does provide the general idea that the folding transition states for small proteins structurally resemble the native states to an extent of ~60%, suggesting that roughly two-thirds of the surface area that is buried in the native state becomes buried in the transition state. The α value per se does not indicate the amount of nativelike structure present in the transition state. However, several studies (reviewed in ref *58*), including ϕ value analysis (*59*) and restrained molecular dynamics simulations (*30*), show the presence of substantial nativelike structures in the folding transition states of many proteins, suggesting that the transition barrier is placed closer to the native state along the folding coordinate.

Mapping of the behavior of lattice simulations to real proteins shows that the TSE of small proteins has 60% of the native contacts (*8*, *14*, *60*). The same result has been obtained from an all-atom simulation study (*61*); the TSE has ~80−85% of its surface area buried (*15*, *60*). Thus, a good correspondence between experimental data and theoretical predictions exists for the barrier location.

Limitations. Understandably, the strength of the equations provided for the prediction of folding rates is subject to the quality of data in the data set that has been used to obtain the equations. Proteins in the database have been studied at various temperatures ranging from 5 to 40 °C (see the Supporting Information). Inherent errors in the experiments vary with different techniques. Also, the minimum denaturant concentration used for different proteins in the set is not the same, and the effects of urea and guanidinium hydrochloride have not been considered separately.

SUMMARY AND CONCLUSION

Although the number of residues in the sequence does not directly determine the folding rate, it can be used as an indirect descriptor for prediction. For this purpose, we have provided two empirical relations: one requires the number of residues separately forming α-helix, *β*-turn, *β*-strand, polyproline II, and random coil and the other the number of residues classified as hydrophobic, positively charged, and negatively charged.

The size of the classical rate-determining folding barrier is substantially larger than the tiny TSE barriers on the order of a few k_BT encountered by the funnel model, although there is a fair correspondence between experiments and theory regarding their location along the folding coordinate.

SUPPORTING INFORMATION AVAILABLE

A detailed list of the proteins, experimental conditions, experimental and calculated parameters, and references. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Udgaonkar, J. B., and Baldwin, R. L. (1988) NMR evidence for an early intermediate in the folding pathway of ribonuclease A, *Nature 335*, 694−699.
2. Roder, H., Elove, G. A., and Englander, S. W. (1988) Structural characterization of folding intermediates in cytochrome *c* by H-exchange labeling and proton NMR, *Nature 335*, 700−704.
3. Baldwin, R. L., and Rose, G. D. (1999) Is protein folding hierarchic? II. Folding intermediates and transition states, *Trends Biochem. Sci. 24*, 77−83.
4. Jackson, S. E., and Fersht, A. R. (1991) Folding of chymotrypsin inhibitor-2. 1. Evidence for a two-state transition, *Biochemistry 30*, 10428−10435.
5. Jackson, S. E. (1998) How do small single-domain proteins fold? *Folding Des. 3*, 81−91.
6. Prabhu, N. P., Kumar, R., and Bhuyan, A. K. (2004) Folding barrier in horse cytochrome *c*: Support for a classical folding pathway, *J. Mol. Biol. 337*, 195−208.
7. Bryngelson, J. D., and Wolynes, P. G. (1987) Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. U.S.A. 84*, 7524−7528.
8. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., and Socci, N. D. (1995) Towards an outline of the topography of a realistic protein folding funnel, *Proc. Natl. Acad. Sci. U.S.A. 92*, 3626−3630.
9. Thirumalai, D., Ashwin, V., and Bhattacharjee, J. K. (1996) Dynamics of random hydrophobic−hydrophilic copolymers, *Phys. Rev. Lett. 77*, 5385−5388.
10. Nymeyer, H., Garcia, A. E., and Onuchic, J. N. (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes, *Proc. Natl. Acad. Sci. U.S.A. 95*, 5921−5928.
11. Matthews, C. R. (1993) Pathways of protein folding, *Annu. Rev. Biochem. 62*, 653−683.
12. Ptitsyn, O. B. (1995) Molten globule and protein folding, *Adv. Protein Chem. 47*, 83−229.
13. Onuchic, J. N., and Wolynes, P. G. (2004) Theory of protein folding, *Curr. Opin. Struct. Biol. 14*, 70−75.
14. Wolynes, P. G., Onuchic, J. N., and Thirumalai, D. (1995) Navigating the folding routes, *Science 267*, 1619−1620.
15. Socci, N. D., Onuchic, J. N., and Wolynes, P. G. (1998) Protein folding mechanisms and the multidimensional folding funnel, *Proteins 32*, 136−158.
16. Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J., and Socci, N. D. (2000) The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios, *Adv. Protein Chem. 53*, 87−152.
17. Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M., and Karplus, M. (2000) Understanding protein folding via free energy surfaces from theory and experiment, *Trends Biochem. Sci. 25*, 331−339.
18. Rumbley, J., Hoang, L., Mayne, L., and Englander, S. W. (2001) An amino acid code for protein folding, *Proc. Natl. Acad. Sci. U.S.A. 98*, 105−112.
19. Gillespie, B., and Plaxco, K. W. (2004) Using protein folding rates to test protein folding theories, *Annu. Rev. Biochem. 73*, 837−859.
20. Bhuyan, A. K., Rao, D. K., and Prabhu, N. P. (2005) Protein folding in classical perspective: Folding of horse cytochrome *c*, *Biochemistry 44*, 3034−3040.
21. Kamagata, K., Arai, M., and Kuwajima, K. (2004) Unification of the folding mechanisms of non-two-state and two-state proteins, *J. Mol. Biol. 339*, 951−965.
22. Plaxco, K. W., Simons, K. T., and Baker, D. (1998) Contact order, transition state placement and the refolding rate of single domain proteins, *J. Mol. Biol. 277*, 985−994.
23. Plaxco, K. W., Simons, K. T., Ruczinski, I., and Baker, D. (2000) Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics, *Biochemistry 39*, 1177−11183.
24. Debe, D. A., and Goddard, W. A. (1999) First principle prediction of protein folding rates, *J. Mol. Biol. 294*, 619−625.
25. 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. U.S.A. 96*, 11311−11316.
26. Dinner, A., and Karplus, M. (2001) The roles of stability and contact order in determining protein folding rates, *Nat. Struct. Biol. 8*, 21−22.
27. Ghaemmaghami, S., Word, J. M., Burton, R. E., Richardson, J. S., and Oas, T. G. (1998) Folding kinetics of a fluorescent variant of monomeric *λ* repressor, *Biochemistry 37*, 9179−9185.
28. 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.
29. Zhou, H., and Zhou, Y. (2002) Folding rate prediction using total contact distance, *Biophys. J. 82*, 458−463.
30. Paci, E., Lindorff-Larsen, K., Dobson, C. M., Karplus, M., and Vendruscolo, M. (2005) Transition state contact orders correlate with protein folding rates, *J. Mol. Biol. 352*, 495−500.
31. 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.
32. Wolynes, P. G. (1997) Folding funnels and energy landscapes of larger proteins within the capillarity approximation, *Proc. Natl. Acad. Sci. U.S.A. 94*, 6170−6175.
33. Gutin, A. M., Abkevich, V. I., and Shakhnovich, E. I. (1996) Chain length scaling of protein folding time, *Phys. Rev. Lett. 77*, 5433−5436.
34. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) Contact order revisited: Influence of protein size on the folding rate, *Protein Sci. 12*, 2057−2062.
35. Naganathan, A. N., and Munoz, V. (2005) Scaling of folding times with protein size, *J. Am. Chem. Soc. 127*, 480−481.
36. Li, M. S., Klimov, D. K., and Thirumalai, D. (2004) Thermal denaturation and folding rates of single domain proteins: size matters, *Polymer 45*, 573−579.
37. Scott, K. A., Batey, S., Hooton, K. A., and Clark, J. (2004) The folding of spectrin domain I: Wild-type domains have the same stability but very different kinetic properties, *J. Mol. Biol. 344*, 195−205.
38. Myers, J. K., Pace, C. N., and 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.
39. Schellman, J. A. (1978) Solvent denaturation, *Biopolymers 17*, 1305−1322.

40. Koga, N., and Takada, S. (2001) Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gô-like model, *J. Mol. Biol. 313*, 171−180.
41. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N., and Finkelstein, A. V. (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics, *Protein Sci. 51*, 162−166.
42. Maxell, K. A., Wildes, D., Zarrine-Afsar, A., de Los Rios, M. A., Brown, A. G., et al. (2005) Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins, *Protein Sci. 14*, 602−616.
43. Shakhnovich, E. I., and Gutin, A. M. (1993) Engineering of stable and fast-folding sequences of model proteins, *Proc. Natl. Acad. Sci. U.S.A. 90*, 7195−7199.
44. Socci, N. D., and Onuchic, J. N. (1994) Folding kinetics of protein-like heteropolymer, *J. Chem. Phys. 101*, 1519−1528.
45. Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995) Funnels, pathways and the energy landscape of protein folding: A synthesis, *Proteins 21*, 167−195.
46. Sali, A., Shakhnovich, E., and Karplus, M. (1994) How does a protein fold? *Nature 369*, 248−251.
47. Sali, A., Shakhnovich, E., and Karplus, M. (1994) Kinetics of protein folding. A lattice model study of the requirements for folding to the native state, *J. Mol. Biol. 235*, 1614−1636.
48. Clarke, J., Cota, E., Fowler, S. B., and Hamill, S. J. (1999) Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway, *Struct. Folding Des. 7*, 1145−1153.
49. Gutin, A. M., Abkevich, V. I., and Shakhnovich, E. I. (1995) Evolution-like selection of fast-folding model proteins, *Proc. Natl. Acad. Sci. U.S.A. 92*, 1282−1286.
50. Portman, J. J., Takada, S., and Wolynes, P. G. (2001) Microscopic theory of protein folding rates. II. Local reaction coordinates and chain dynamics, *J. Chem. Phys. 114*, 5082−5096.
51. Kaya, H., and Chan, H. S. (2002) Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: How applicable is the transition state picture to folding and unfolding? *J. Mol. Biol. 315*, 899−909.
52. Kumar, R., Prabhu, N. P., and Bhuyan, A. K. (2005) Ultrafast events in the folding of ferrocytochrome *c*, *Biochemistry 44*, 9359−9367.
53. Krieger, F., Fierz, B., Bieri, O., Drewello, M., and Kiefhaber, T. (2003) Dynamics of unfolded polypeptide chains as model for the earliest steps in protein folding, *J. Mol. Biol. 332*, 265−274.
54. Ivankov, D. N., and Finkelstein, A. V. (2001) Theoretical study of a landscape of protein folding-unfolding pathways. Folding rates at midtransition, *Biochemistry 40*, 9957−9961.
55. Schonbrun, J., and Dill, K. A. (2003) Fast protein folding kinetics, *Proc. Natl. Acad. Sci. U.S.A. 100*, 12678−12682.
56. Dill, K. A., and Chan, H. S. (1997) From Levinthal to pathways to funnels, *Nat. Struct. Biol. 4*, 10−19.
57. Tanford, C. (1968) Protein denaturation, *Adv. Protein Chem. 23*, 121−282.
58. Matthews, C. R. (1993) Pathways of protein folding, *Annu. Rev. Biochem. 62*, 653−683.
59. Fersht, A. R. (1997) Nucleation mechanism in protein folding, *Curr. Opin. Struct. Biol. 7*, 3−9.
60. Wolynes, P. G. (2004) Latest folding game results: Protein A barely frustrates computationalists, *Proc. Natl. Acad. Sci. U.S.A. 101*, 6837−6838.
61. Alonso, D. O., and Daggett, V. (2000) Staphylococcal protein A: Unfolding pathways, unfolded states and differences between the B and E domains, *Proc. Natl. Acad. Sci. U.S.A. 97*, 133−138.

BI0521137