Biochemistry

© Copyright 2000 by the American Chemical Society Volume 39, Number 37 September 19, 2000

Current Topics

Topology, Stability, Sequence, and Length: Defining the Determinants of Two-State Protein Folding Kinetics

Kevin W. Plaxco,*,[‡] Kim T. Simons,[§,||] Ingo Ruczinski,[⊥] and David Baker*,[§]

Department of Chemistry and Biochemistry and Interdepartmental Program in Biochemistry and Molecular Biology, University of California, Santa Barbara, Santa Barbara, California 93106, and Departments of Biochemistry and Statistics, University of Washington, Seattle, Washington 98195

Received February 1, 2000; Revised Manuscript Received May 19, 2000

ABSTRACT: The fastest simple, single domain proteins fold a million times more rapidly than the slowest. Ultimately this broad kinetic spectrum is determined by the amino acid sequences that define these proteins, suggesting that the mechanisms that underlie folding may be almost as complex as the sequences that encode them. Here, however, we summarize recent experimental results which suggest that (1) despite a vast diversity of structures and functions, there are fundamental similarities in the folding mechanisms of single domain proteins and (2) rather than being highly sensitive to the finest details of sequence, their folding kinetics are determined primarily by the large-scale, redundant features of sequence that determine a protein's gross structural properties. That folding kinetics can be predicted using simple, empirical, structure-based rules suggests that the fundamental physics underlying folding may be quite straightforward and that a general and quantitative theory of protein folding rates and mechanisms (as opposed to unfolding rates and thus protein stability) may be near on the horizon.

In the previous decade, more than 3 dozen small, single domain proteins have been reported to fold via two-state kinetic mechanisms (reviewed in ref *1*). The simplicity of such two-state folding suggests that these proteins might provide a clear, compelling picture of the means by which the "protein folding problem" is solved. This same simplicity, however, also poses a significant experimental challenge: how can we characterize the mechanisms that underlie two-state folding when a myriad of time-resolved biophysical probes all report identical kinetics? Biophysics does not provide us the luxury of directly observing intermediates in an ostensibly two-state process, forcing experimentalists to adopt indirect methods in order to characterize the mechanisms that define two-state folding.

Cytochrome b_{562} folds 1 million times more rapidly than muscle acylphosphatase (*2*, *3*). This simple statement illustrates one such indirect method of studying two-state folding; a quantitative accounting of the factors that define this 6 orders of magnitude range could provide invaluable constraints on theories of protein folding. While much progress has been reported regarding theoretical models of the process (reviewed, for example, in refs *4*−*7*), here we focus on recent *experimental* efforts aimed at elucidating these determinants that are beginning to provide important and apparently general insights into the nature of folding.

* To whom correspondence should be addressed. K.W.P.: e-mail, kwp@chem.ucsb.edu. D.B.: e-mail, dabaker@u.washington.edu.

[‡] University of California, Santa Barbara.

[§] Department of Biochemistry, University of Washington.

[||] Current address: Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Ave., Cambridge, MA 02138.

[⊥] Department of Statistics, University of Washington.

10.1021/bi000200n CCC: $19.00 © 2000 American Chemical Society
Published on Web 08/25/2000

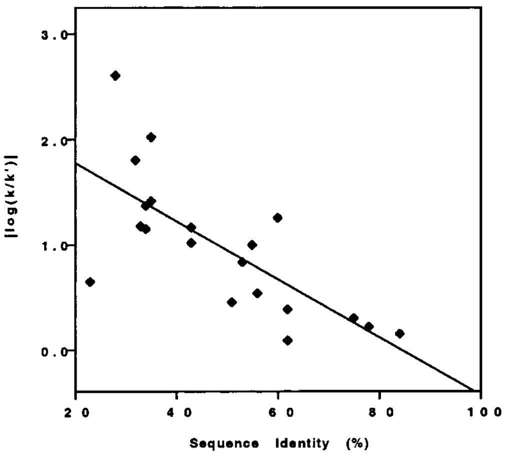

FIGURE 1: The determining role that sequence plays in folding kinetics is readily apparent in the significant correlation ($r = 0.73$; $p < 10^{-3}$) between the relative sequence identities and relative folding rates of pairs of homologous proteins. However, the relative stability, size, and topology of homologues are also most similar the more closely related the proteins are. Thus the question remains, are folding kinetics related to these more global parameters? Shown are all homologous pairings of the appropriately characterized two-state proteins (*21*−*27* and references therein).

Determinants of Two-State Folding Kinetics. Anfinsen demonstrated that protein folding is a spontaneous, first-order process, suggesting that a protein's primary sequence defines both its structure and the rate with which that structure is formed (*8*). Thus, in a very real sense, sequence is the *only* determinant of the rates and mechanisms of folding. The interplay between sequence and kinetics is readily observed; for example, there is a statistically significant correlation ($r = 0.73$; $p < 10^{-3}$) between the pairwise sequence identities and relative folding kinetics across homologous, single domain proteins (Figure 1).

A protein's sequence defines its size, stability, structure, and folding kinetics. The question remains, however, does sequence directly define folding kinetics? Or does it define folding rates primarily by defining other, more global equilibrium properties of a protein? These are separable issues. For example, there are many pairs of proteins of a given length that share little or no sequence identity. Similarly, there are many examples of proteins with the same stability or sharing a common topology but lacking significant evidence of homology. Thus, more simply put, do two proteins with the same length fold with similar rates? Do proteins with the same stability or topology? Or are folding kinetics so sensitively encoded in sequence that point mutations lead to millionfold changes in rate? This is a critical issue—if folding kinetics are very sensitive to fine sequence details, the development of quantitative theoretical models of the process will prove exceptionally difficult. In contrast, if simple stability-, length-, or topology-based rules can be used to predict folding rates, then it can be assumed that fundamental polymer physics dominates the folding problem and that a common mechanism underlies the diverse kinetic properties of these proteins. This, in turn, would suggest that a quantitative theoretical treatment of folding will be relatively easier to achieve.

Folding Rates Are Insensitive to Fine Sequence Details. Several lines of evidence suggest that folding rates are relatively insensitive to even large-scale sequence changes as long as the size, stability, and topology of the native state are maintained. Of the >200 characterized, two-state point mutants that actually fold, for example, few fold even 10 times more slowly than wild type, and we are aware of none that alter folding rates by more than a factor of 50 (*9*−*18*). Compared to the millionfold range of characterized two-state rates, these small perturbations suggest that two-state folding is relatively insensitive to minor changes in a protein's sequence.

Compared to naturally occurring sequence variation, most of the characterized in vitro mutations reflect relatively trivial sequence changes. Recent evidence suggests, however, that even large-scale, in vitro sequence alterations do not significantly perturb folding rates. This question has been addressed by the use of phage display techniques to obtain large collections of divergent SH3 and protein L sequences that adopt their correct native folds (*19*, *20*). If folding rates are sensitive to fine details of sequence, the folding rates of these artificially generated sequences (some of which share less than 50% sequence identity with the wild-type sequence) would be expected to be considerably altered from those of their naturally occurring counterparts. Moreover, if evolution optimizes folding kinetics at the level of sequence details, these variants might be expected to fold more slowly than the corresponding wild-type sequences. In contrast to these expectations, none of the characterized variants exhibit folding rates altered by more than a factor of 10, and half fold more rapidly than the wild-type sequences from which they were derived. These results provide further support for the suggestion that, as long as native structure and stability are maintained, folding kinetics are relatively insensitive to even rather large-scale sequence changes.

Further insight into the sequence dependence of folding kinetics has come from studies of naturally occurring sets of homologous proteins. The folding rates of six homologous sets of two-state proteins have been reported (*21*−*27*). Three of the six, with pairwise identities in the range 43−84%, exhibit only modest (<20-fold) rate dispersions (*21*−*24*). The remaining three homologous families (*25*−*27*), ranging from 23% to 78% pairwise identity, exhibit rather larger rate variations: for example, the fyn SH3 domain folds 2 orders of magnitude more rapidly than the drk SH3 domain with which it shares 35% sequence identity (*27*) (Figure 1). Does a 2 order of magnitude spread between the folding rates of homologous proteins indicate that sequence changes that do not change length, stability, or topology can vastly alter folding rates? Only if such studies properly control for the contributions of these more global properties. Recent evidence suggests that they do not.

Stability as a Determinant of Folding Rates. Homology studies control relatively well for the kinetic effects of length and topology because protein size and structure are generally conserved. These studies do not, however, control for potentially differing native state stabilities. What might be the origin of a relationship between stability and rates? For a two-state reaction, stability and folding rate are clearly not independent quantities: $\Delta G_u = RT \ln(k_f/k_u)$. Changes in ΔG_u need not be correlated, however, with changes in folding

[1] Abbreviations: ADAh2, activation domain of procarboxypeptidase A2; CI-2, chymotrypsin inhibitor 2; CO, relative contact order; FNIII, fibronectin type III domain; SH3, src homology region 3.

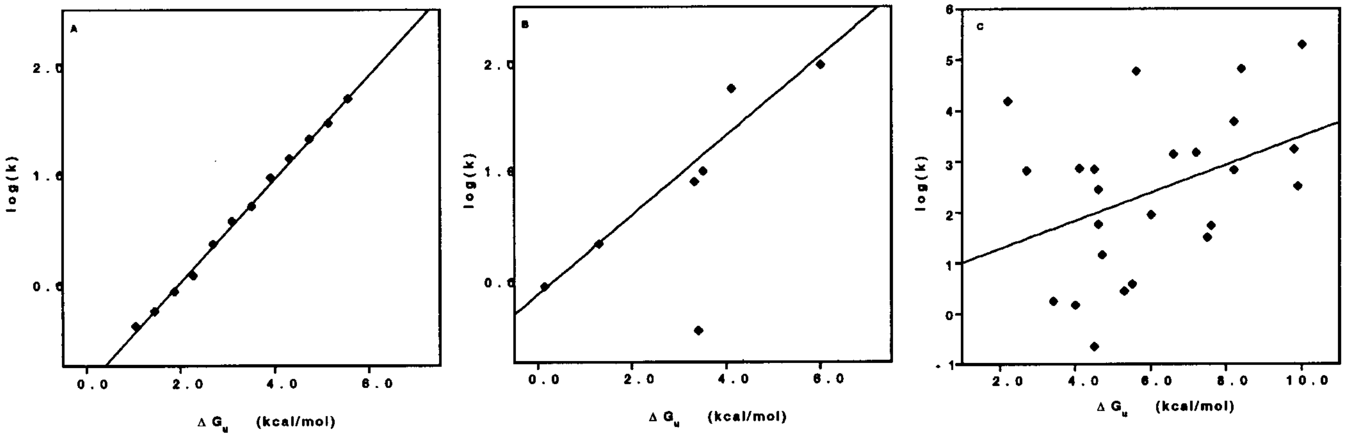

FIGURE 2: Native state stability as a determinant of folding kinetics. (A) The correlation between relative free energy of folding (ΔG_u) and folding rates of proteins under differing solvent conditions is typically extremely strong. This includes, as shown here, the folding of the fyn SH3 domain ($r = 0.999$; $p < 10^{-5}$) (*27*; M. de los Rios, personal communication). (B) A similar correlation is observed across a set of homologous proteins; there is a strong correlation ($r = 0.96$; $p = 0.002$) between the native state stabilities and folding rates of six of the seven characterized SH3 domains (*27*, *29*, *64*−*65*; Northey and Davidson, personal communication; Camarero, Sato, Raleigh, and Muir, personal communication). The obvious outlier, the pI3K SH3 domain (*29*), contains an 18-residue insertion and thus differs significantly from the other SH3 domains in both length and topology (it falls well within the scatter in Figure 4B). (C) A statistically marginal correlation ($r = 0.40$; $p = 0.05$) is observed across a large set of nonhomologous two-state proteins; the large degree of scatter in this plot clearly indicates that other factors must play an important role in defining folding kinetics. (References as for Table 1.)

rate; for example, in the limit of a "golf course-shaped" energy landscape, changes in sequence which change the depth of the energy well change k_u (and thus ΔG_u) but not k_f. More generally, changes in sequence that alter native state stability but do not affect the free energies of conformations in the transition state ensemble will also not affect the folding rate. In the limit of a symmetric, funnel-shaped folding landscape, on the other hand, all interactions are partially formed at the rate-limiting step in folding. In this extreme, the effects of sequence changes on stability will be perfectly correlated with their effects on the folding rate. Thus, the extent of correlation between stability and folding rate might provide insights into the nature of the folding transition state.

The highly linear arms of folding "chevron plots" (semilog plots of folding rate versus cosolvent concentration) (e.g., refs *2*, *3*, *14*, *27*−*29*) demonstrate that for a given protein there is a nearly perfect correlation between stability and folding rates across a broad range of solvent conditions (Figure 2A). Presumably this correlation arises because solvent alterations that affect the stability of interactions in the native state also affect the stability of interactions formed during the rate-limiting step of folding. Unlike changing solvent conditions, mutations affect only a small subset of the interactions that are formed during the rate-limiting step. Despite this limitation, statistically significant correlations have been reported between thermodynamics and kinetics for large sets of point mutations in CI-2 (*30*), the spectrin SH3 domain (*13*), ADAh2 (*14*), FKBP12 (*17*), and acylphosphatase (*18*; F. Chiti and C. M. Dobson, personal communication). More recently, several groups have reported that mutations and cosolvents increasing nativelike secondary structural preferences, and thus protein stability, also increase the folding rates of two-state proteins (*11*, *31*, *32*).

Consistent with the role that equilibrium stability plays in defining relative folding rates across differing solvent conditions and sets of point mutations, stability also predicts relative folding rates across some sets of homologous proteins. No set of homologous proteins has been reported that exhibits folding rates differing more than an order of magnitude when their differing stabilities are taken into account (e.g., refs *21*−*27*), and there is a statistically significant correlation between folding kinetics and stability for the six characterized SH3 domains of similar length and topology (Figure 2B; $r = 0.96$; $p = 0.002$). Recently, Clarke and co-workers have extended these observations to topologically similar proteins lacking significant sequence identity (*33*). These data strongly suggest that, at least across sets of structurally similar proteins, stability-specific effects can account for up to 2 orders of magnitude in the range of characterized folding rates.

While native state stability is often correlated with the relative folding rates of homologous proteins, a perhaps more important question is: does the correlation hold across unrelated proteins? Examination of a large, nonhomologous data set of simple, single domain proteins (Table 1) is, at best, consistent with this hypothesis (Figure 2C; $r = 0.40$; $p = 0.05$). Even a cursory inspection of the scatter associated with this correlation clearly indicates that other factors play a significant role in defining folding rates.

Length as a Determinant of Folding Rates. Length is usually (although not always—e.g., ref *34*) a gross determinant of two-state versus no-two-state behaviors; proteins less than ∼110 amino acids usually exhibit two-state kinetics (*1*). There is, however, effectively no correlation between length and the relative folding rates of nonhomologous, two-state proteins (Figure 3; $r = 0.16$; $p = 0.53$). Clearly, other determinants are responsible for the diversity of two-state folding rates.

Topology as a Determinant of Folding Rates. Topology might be the missing determinant, but quantitatively testing this hypothesis is not straightforward and requires the creation of a single value descriptor of topological complexity. The measure we have used previously (*35*), contact order, is defined as the average sequence separation of contacting residue pairs. Thus, protein structures featuring predominantly long-range interactions have high contact order, while those built of predominantly local structures are of low contact order. To make this measure of topology independent

Table 1

protein[a]	log (k_f)[b]	CO[c] (%)	ΔG_u (kcal/mol)	length[d] (residues)	temp (°C)	ref
cyt b_{562}	5.30	7.47	10.0	106	20	*2*
myoglobin	4.83[e]	8.50	8.4	154	25	*54*
λ-repressor	4.78	9.37	5.6	80	20	*55*
PSBD	4.20	11.20	2.2	41	41	*37*
cyt *c*	3.80[f]	11.22	8.2	104	23	*25*
Im9	3.16	12.07	6.6	85	10	*34*
ACBP	2.85	13.99	8.2	86	25[g]	*21*
villin 14T	3.25	12.31	9.8	126	25	*56*
N-terminal L9	2.87	12.74	4.5	56	25	*36*
ubiquitin	3.19	15.11	7.2	76	25	*44*
CI-2	1.75	16.40	7.6	64	25	*9*
U1A	2.53	16.91	9.9	102	25	*57*
ADAh2	2.88	16.96	4.1	79	25	*14*
protein G[h]	2.46	17.30	4.6	56	25	*58*
protein L	1.78	17.62	4.6	62	22	*59*
FKBP	0.60	17.70	5.5	107	25	*60*
HPr	1.17	18.35	4.7	85	20	*61*
MerP	0.26[i]	18.90	3.4	72	25	*62*
mAcP	−0.64	21.20	4.5	98	28	*3*
CspB	2.84	16.40	2.7	67	25	*23*
TnFNIII	0.46	17.35	5.3	92	20	*33*
TiI27	1.51	17.82	7.5	89	25	*33*
fyn SH3	1.97	18.28	6.0	59	20	*27*
twitchin	0.18	19.70	4.0	93	20	*33*

[a] A nonhomologous set of simple, single domain, non-disulfide-bonded proteins that have been reported to fold via two-state kinetics under at least some conditions. Reported data and representative members of homologous families were selected as previously described (*35*). [b] Extrapolated folding rates in water. The rates may differ from the true folding rate in water (e.g., cyt *c*, protein G, ubiquitin, and others) due to "roll over" at low denaturant concentrations. [c] Calculated as previously described (*35*). [d] Length of the protein in residues from the first structured residue to the last. The length may differ from the number of residues in the construct characterized. No significant correlation exists between folding rates and length of the construct (*35*; data not shown). [e] Extrapolated folding rate of deoxymyoglobin in water (P. Wittung-Stafshede, personal communication). [f] Extrapolated folding rate of reduced cytochrome *c* in water (J. Winkler and H. Gray, personal communication). [g] Folding rate at 25 °C (B. Kraglund, personal communication). [h] Ala-53 mutant (two-state fit parameters not available for the wild-type protein). [i] Extrapolated folding rate of MerP in water (G. Aronsson, personal communication).

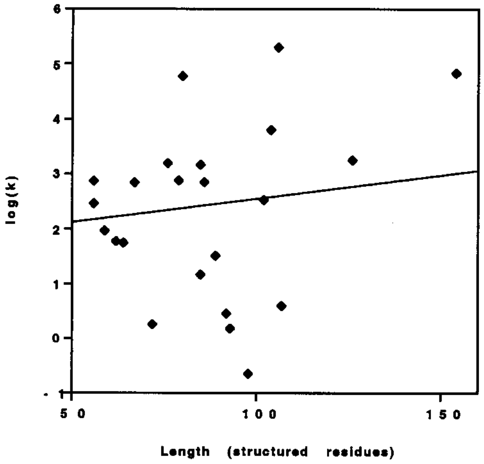

FIGURE 3: Length is not significantly correlated with the folding kinetics of nonhomologous, two-state proteins ($r = 0.16$; $p = 0.53$). (References as for Table 1.)

of length, absolute contact order is normalized by chain length to generate relative contact order (CO). The CO of kinetically characterized, single domain proteins ranges from 7% to 21% (Table 1).

Topology, as described by CO, is highly correlated with the relative folding rates of simple, single domain proteins (Figure 4A; $r = 0.92$; $p < 10^{-9}$) (*35*). Out of the millionfold range of characterized folding rates, only 1 protein in 6 falls off the best fit line by more than a factor of 10, and no protein in the data set deviates by more than a factor of 20. With a median deviation of only 4-fold, CO is of significant predictive value and has been used for the successful (median error a factor of 3; maximum error a factor of 8), *blind* prediction of the folding rates of almost a dozen proteins (*2*, *36*, *37*; P. Wittung-Stafshede, H. Gray, S. Jackson, B. Khulman, D. Raleigh, G. Aronsson, J. Fernandez, and B. Gillespie, personal communication).

Due to variations in the length of secondary structural elements, proteins sharing a common fold can exhibit significantly differing CO. Nevertheless, CO appears to reflect a fundamental kinetic determinant. For example, the five characterized, nonhomologous proteins that share the $\alpha-\beta$ pleat fold (AcP, ADA2h, U1A, HPr, and MerP) span a 3300-fold range of rates, suggesting that topology is not, per se, a strong determinant of rates. Yet Chiti et al. have demonstrated that the folding rates of these proteins are highly correlated with their CO ($r = 0.96$; $p = 0.01$) (*18*), and none fall more than an order of magnitude off of the CO−rate line (Figure 4A).

While the predictive value of CO suggests that topology is a critical determinant of kinetics, mutations that do not significantly alter CO still affect folding rates; clearly, other factors contribute to the folding barrier. To characterize the magnitude of these additional effects, it is informative to investigate the correlation between CO and folding kinetics when all reported homologous and mutant proteins are included (Figure 4B). Across this larger data set the correlation between CO and log(k_f) remains extremely significant ($r = 0.89$, $p \sim 10^{-18}$). Again, the indication is that CO reflects the single most important characterized determinant of the rate with which a protein folds.

An independent test of CO as a kinetic determinant, provided by recent work in the laboratories of Fersht and Serrano (*38, 39*), sheds further light on the origins of the scatter inherent in CO−rate plots. These groups characterized the folding kinetics of sets of variant proteins differing in the number of residues inserted into solvent-exposed loops. In addition to directly altering chain length, these changes also alter CO (by increasing the sequence separation of contacting residues on opposite sides of the loop) *without significantly perturbing core sequence or stability*. In the absence of these potentially confounding effects, the correlation coefficient of the topology−kinetics relationship improves significantly (Figure 4C; $r = \geq 0.97$; $p < 0.04$) (*40*).

Stability effects are presumably a major source of the scatter surrounding the CO−rate relationship. Unfortunately, however, a nontrivial correlation ($r = -0.44$; $p = 0.03$) between CO and stability—arising because CO is highly correlated with k_f and uncorrelated with k_u—confounds efforts at deconvoluting the individual contributions of each to the folding kinetics of nonhomologous proteins. Nonetheless, there is some evidence for the putative role of native state stability in defining folding kinetics. For example, of the >250 sequences illustrated (Figure 4B), only 10 fall more than a factor of 20 off of the best fit line. Six of these are

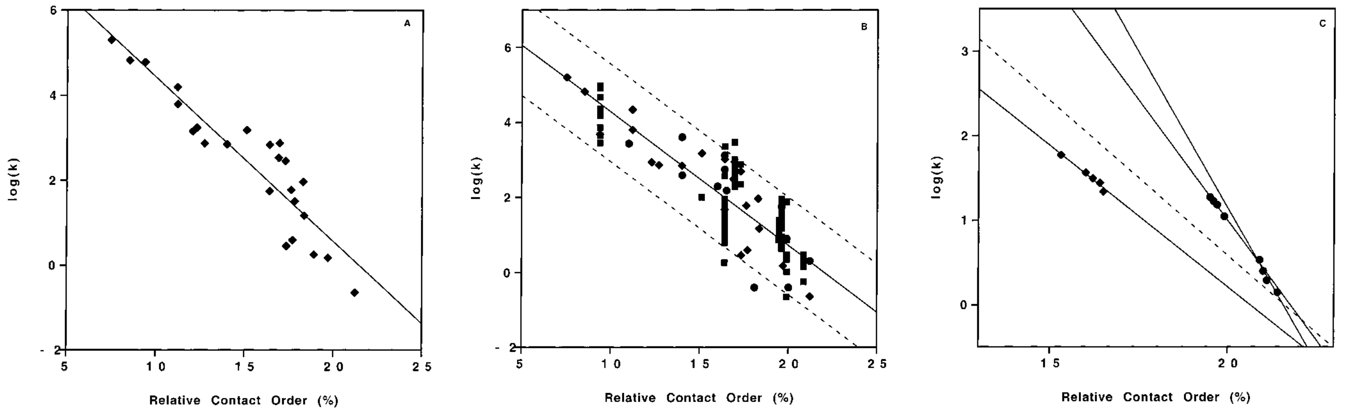

FIGURE 4: A simple metric of topological complexity, CO, predicts relative folding rates. (A) The relative refolding kinetics of a large set of nonhomologous, simple, single domain proteins can be accurately predicted ($r = 0.92$; $p < 10^{-9}$) solely on the basis of a knowledge of their CO. (B) The correlation remains robust ($r = 0.89$; $p < 10^{-18}$) even when the >250 characterized homologous or mutant proteins are included (CO calculated assuming wild-type structures unless independent structural information is available). Dotted lines denote ±20-fold rate variations. (References as for Table 1; also refs *9−27* and *62−65*.) (C) Fersht, Serrano, and their co-workers have generated three sets of variant proteins differing in the number of residues inserted into solvent-exposed, unstructured loops (*38*, *39*). Within each set of variants, CO is systematically altered (the loop-length changes increase the average sequence separation of contacting residues on either sides of the loop) without significantly altering the structure or stability of the protein's core. In the absence of these potentially confounding alterations, the correlation between CO and rate is significantly improved ($r = 0.97$; $p < 0.04$) (*40*). Of note, a set of variants constructed using the "average" amino acid glutamine (diamonds) features a slope within error of that of the all-protein line (dotted line from panel A). Those constructed with the more flexible glycine residue (circles) exhibit significantly greater slopes. This provides perhaps the first direct, quantitative experimental demonstration of the contribution of chain entropy to the folding barrier.

slow-folding, relatively unstable mutants of muscle acylphosphatase (*18*). Three (of the remaining four) fold more slowly than predicted by their topologies: they are the two least stable of all characterized CI-2 (*9*) and SH3 (*16*) point mutants and the wild-type ^{9}FNIII sequence (*26*). The latter is both the most significant outlier (a factor of 46 off the fit) and, by a large margin, the least stable wild-type proteins for which complete kinetic characterization has been reported. The one sequence that falls significantly above the line (23-fold) is, to within experimental error, the most stable mutant of CI-2 reported to date and is substantially more stable that the wild-type protein (*9*). Thus stability may reflect an important *secondary* determinant of folding kinetics.

Determinants of Transition State Structure. The correlation between CO and folding rates argues strongly that topology dominates the thermodynamics of the rate-limiting step in folding and thus defines the rates with which proteins fold. Several, rather weaker lines of evidence suggest that topology also plays a role in defining the structure of the folding transition state.

A crude measure of the relative solvent accessibility of the transition state provides evidence for the role of topology in defining transition state structure. This solvent accessibility is typically conserved between topologically similar homologues (*21−27*)—but not between topologically dissimilar homologues (*27*)— and, with a notable exception (*11*), is rarely affected by mutations (*9*, *10*, *12−20*). Consistent with these observations, transition state solvent accessibility also correlates with CO, although a number of outliers reduce the statistical significance of this relationship (Figure 5; $r = 0.57$; $p < 0.01$) (*35*).

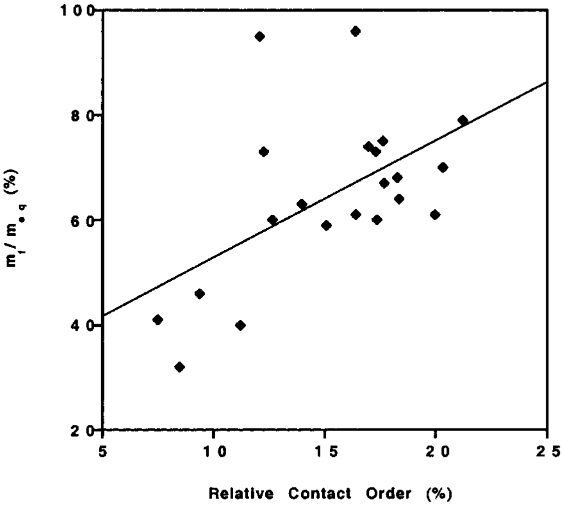

FIGURE 5: CO is modestly correlated ($r = 0.57$; $p < 10^{-2}$) with the relative reduction of solvent-accessible surface that occurs before the rate-limiting step in folding (given by the ratio $m_{\mathrm{f}}/m_{\mathrm{eq},}$ which relates the denaturant sensitivity of the transition state to that of the native state). This is consistent with several lines of evidence suggesting that topology is an important determinant of the structure of the folding transition state (*13*, *18*, *33*, *35*, *40*, *41*). (Data taken from references in Table 1.)

More recently, several groups have characterized folding transition state structures across sets of structurally related, two-state proteins. Serrano and co-workers have characterized the folding transition states of a series of circularly permuted SH3 domains and found these topological alterations significantly perturb transition state structure without significantly modifying the protein's native core packing (*41*). Coupled mutagenesis−kinetics studies of homologous proteins demonstrate that, while the sequence identities of kinetically critical residues are not conserved (*42*), their positioning along the peptide chain is maintained (*13*, *42*). Of course, such similarity across homologous proteins could reflect conservation of transition state structure rather than a more general relationship between transition state structure and protein topology. In contrast, Dobson and co-workers have demonstrated that the transition state structures of topologically similar proteins lacking any significant sequence identity are also closely related (*18*), and Clarke and co-workers have reported that the folding pathways of several pairs of two-state "structural homologues" are similarly "conserved" (*33*).

A Common Folding Mechanism? The correlation between CO and folding rates is extremely strong; across the 6 orders of magnitude range of characterized two-state folding rates, few sequences fall more than a factor of 10 off of the best fit line and none by more than a factor of 50. That this is observed despite 1−2 orders of magnitude scatter arising from sequence- and stability-specific effects (plus that due to differing experimental conditions) suggests that CO predicts folding rates as accurately as possible with *any* single determinant model. The empirical observation that all two-state proteins fall within reasonable spread around the CO−rate line implies, in turn, that despite a vast diversity of structures and functions, there are fundamental similarities in the folding mechanisms of single domain proteins. Although the precise nature of this similarity remains unclear (e.g., refs *7*, *40*, *43*), the empirical correlation between CO and folding rates might reflect a balance between the gain of attractive native interactions and the loss of chain configurational entropy as a protein folds. The major determinant of this balance is the chain topology, as this determines how much chain entropy is lost as native interactions are progressively formed.

Non-Two-State Folding. Do non-two-state proteins fold via the same mechanism? Many proteins, including the small, simple, single domain proteins cytochrome *c* (*25*), ubiquitin (*44*), and Im7 [a homologue of the two-state protein Im9 (*34*)], exhibit kinetic "roll over" (depressed folding rates under low denaturant conditions) that implies non-two-state folding. Extrapolation of a two-state model to no-denaturant conditions thus overestimates the true folding rates of these proteins. These extrapolated rates, however, fall well within the scatter of the CO−rate plot, suggesting these potentially non-two-state proteins share common mechanistic features with their more clearly two-state brethren. The majority of larger (>110 residues) proteins also exhibit roll over. The few of these we have analyzed—RNase H (*45*), barnase (*46*), and the two-domain T4 lysozyme (*47*)—all fold (both extrapolated and observed rates) significantly more slowly than would be predicted by the topology−kinetics relationship (CO = 13%, 11%, and 7% respectively). Thus, while the correlations described above appear to accurately predict the folding kinetics of most simple, single domain proteins, additional factors (perhaps escape from more significant kinetic traps) appear to complicate the folding of larger proteins.

Conclusions. Emerging from studies of two-state proteins is a picture of folding as a simple, mechanistically uniform process not dependent on fine details of sequence or highly optimized by selective pressures but determined rather by the global, equilibrium properties of proteins. These studies indicate that, for small, kinetically simple, single domain proteins, (1) length is not significantly correlated with folding rates, (2) folding rates are relatively insensitive to sequence changes that do not significantly alter the structure or stability of the native state, (3) native state stability is correlated with folding kinetics across differing solvent conditions and often across sets of topologically similar proteins, and (4) the relative folding kinetics of nonhomologous proteins are quantitatively related to the gross topologies of their native states.

Prospects for a Simple Model of Protein Folding Kinetics. A few years ago, in a comprehensive review of the so-called "New View" of protein folding, Chan and Dill put forth a "wish list" of experimental characterizations of which theorists were desirous (*4*). Here we have turned the tables and presented a list of empirical observations that we feel any successful theoretical model of folding must address, if not quantitatively predict. The results of recent theoretical studies (*48*−*51*) leave us optimistic that such a quantitative model can be achieved.

A problem confronting quantitative modeling of protein stability (and consequently protein structure) is its dependence on the intricate nonbonding interactions in densely packed native states. Due to this high packing density, small changes in the sequence [such as the addition of a small number of methyl groups in the core (e.g., ref *52*)] can destabilize proteins significantly. These large destabilizations are almost always reflected in large increases in the rate of unfolding (e.g., refs *9*−*18* and *52*), since such mutations generally increase the free energy of the native state considerably more than the more loosely packed transition state ensemble.

Protein folding rates, in contrast, are relatively insensitive to such changes, presumably because the interactions determining the folding process are considerably more coarse grained that those that define the stability (structure) of the native state. [The conformations populated during the rate-limiting step are likely to be much less well packed than the native state (*53*) and thus much less sensitive to small changes in core volume.] This insensitivity to fine detail, coupled with the important role of native state topology suggested by the experiments described above, bodes well for the prospects of developing a simple, quantitative model of protein folding kinetics. A theory that, by emphasizing these coarse-grained parameters, accurately describes the loss in configurational entropy and gain in attractive interactions that transpire during folding may succeed in quantitatively predicting all of the critical features of the folding reaction.

ACKNOWLEDGMENT

The conclusions presented in this review could not have been reached without the generous assistance and insights provided by our many collaborators. These include (but are not limited to) Goran Aronsson, Fabrizio Chiti, Jane Clarke, Alan Davidson, Chris Dobson, Julio Fernandez, Alan Fersht, Blake Gillespie, Harry Gray, Iñaki Guijarro, Sophie Jackson, Brian Khulman, Birthe Kragelund, Dan Raleigh, Luis Serrano, and Pernilla Wittung-Stafshede, all of whom kindly provided unpublished or prepublication results critical to this work. Additional commentary by Michael Gross, Dan Raleigh, and members of our research groups is also gratefully acknowledged.

REFERENCES

1. Jackson, S. E. (1998) *Folding Des. 3*, R81−R91.
2. Wittung-Stafshede, P., Lee, J. C., Winkler, J. R., and Gray, H. B. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 6587−6590.
3. van Nuland, N. A. J., Chiti, F., Taddei, N., Raugei, G., Ramponi, G., and Dobson, C. M. (1998) *J. Mol. Biol*. *283*, 883−891.

4. Chan, H. S., and Dill, K. A. (1997) *Nat. Struct. Biol. 4*, 10–19.
5. Pande, V. S., Grosberg, A. Y., Tanaka, T., and Rokhsar, D. S. (1998) *Curr. Opin. Struct. Biol. 8*, 68–79.
6. Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem. 48*, 545–600.
7. Alm, E., and Baker, D. (1999) *Curr. Opin. Struct. Biol. 9*, 189–196.
8. Anfinsen, C. B. (1973) *Science 181*, 223–230.
9. Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) *J. Mol. Biol. 254*, 260–288.
10. Viguera, A. R., Serrano, L., and Wilmanns, M. (1996) *Nat. Struct. Biol. 3*, 874–880.
11. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L., and Oas, T. G. (1997) *Nat. Struct. Biol. 4*, 305–310.
12. Gu, H., Kim, D., and Baker, D. (1997) *J. Mol. Biol. 274*, 588–596.
13. Martinez, J. C., and Serrano, L. (1999) *Nat. Struct. Biol. 6*, 1010–1016.
14. Villegas, V., Martinez, J. C., Aviles, F. X., and Serrano, L. (1998) *J. Mol. Biol. 6*, 1027–1036.
15. Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J., and Poulsen, F. M. (1999) *Nat. Struct. Biol. 6*, 594–601.
16. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., and Baker, D. (1999) *Nat. Struct. Biol. 6*, 1016–1024.
17. Fulton, K. F., Main, E. R. G., Daggett, V., and Jackson, S. E. (1999) *J. Mol. Biol. 291*, 445–461.
18. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M., and Dobson, C. M. (1999) *Nat. Struct. Biol. 6*, 1005–1009.
19. Riddle, D. S., Santiago, J. V., BrayHall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q., and Baker, D. (1997) *Nat. Struct. Biol. 4*, 805–809.
20. Kim, D. E., Gu, H. D., and Baker, D. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 4982–4986.
21. Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J., and Poulsen, F. M. (1996) *J. Mol. Biol. 256*, 187–200.
22. Reid, K. L., Rodriguez, H. M., Hillier, B. J., and Gregoret, L. M. (1998) *Protein Sci. 7*, 470–479.
23. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R., and Schmid, F. X. (1998) *Nat. Struct. Biol. 5*, 229–235.
24. Taddei, N., Chiti, F., Paoli, P., Fiaschi, T., Bucciantini, M., Stefani, M., Dobson, C. M., and Ramponi, G. (1999) *Biochemistry 38*, 2135–2142.
25. Mines, G. A., Pascher, T., Lee, S. C., Winkler, J. R., and Gray, H. B. (1996) *Chem. Biol. 3*, 491–497.
26. Plaxco, K. W., Spitzfaden, C., Campbell, I. D., and Dobson, C. M. (1997) *J. Mol. Biol. 270*, 763–770.
27. Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D., and Dobson, C. M. (1998) *Biochemistry 37*, 2529–2537.
28. Jackson, S. E., and Fersht, A. R. (1991) *Biochemistry 30*, 10428–10435.
29. Guijarro, J. I., Morton, C. J., Plaxco, K. W., Campbell, I. D., and Dobson, C. M. (1998) *J. Mol. Biol. 276*, 657–667.
30. Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M., and Otzen, D. E. (1994) *Proc. Natl. Acad. Sci. U.S.A. 91*, 10426–10429.
31. Viguera, A. R., Virtudes, V., Aviles, F. X., and Serrano, L. (1996) *Folding Des. 2*, 23–33.
32. Chiti, F., Taddei, N., Webster, P., Hamada, D., Fiaschi, T., Ramponi, G., and Dobson, C. M. (1999) *Nat. Struct. Biol. 6*, 380–387.
33. Clarke, J., Cota, E., Fowler, S. B., and Hamill, S. J. (1999) *Structure 7*, 1145–1153.
34. Ferguson, N., Capaldi, A. P., James, R., Kleanthous, C., and Radford, S. E. (1999) *J. Mol. Biol. 286*, 1597–1608.
35. Plaxco, K. W., Simons, K. T., and Baker, D. (1998) *J. Mol. Biol. 277*, 985–994.
36. Kuhlman, B., Luisi, D. L., Evans, P. A., and Raleigh, D. P. (1998) *J. Mol. Biol. 284*, 1661–1670.
37. Spector, S., and Raleigh, D. P. (1999) *J. Mol. Biol. 293*, 763–768.
38. Ladurner, A. G., and Fersht, A. R. (1997) *J. Mol. Biol. 273*, 330–337.
39. Viguera, A. R., and Serrano, L. (1997) *Nat. Struct. Biol. 4*, 939–946.
40. Fersht, A. R. (2000) *Proc. Natl. Acad. Sci. U.S.A. 97*, 1525–1529.
41. Viguera, A. R., Serrano, L., and Wilmanns, M. (1996) *Nat. Struct. Biol. 3*, 874–880.
42. Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R., and Baker, D. (1999) *J. Mol. Biol. 298*, 303–312.
43. Chan, H. S. (1998) *Nature 392*, 761–763.
44. Khorasanizadeh, S., Peters, I. D., Butt, T. R., and Roder, H. (1993) *Biochemistry 32*, 7054–7063.
45. Parker, M. J., and Marqusee, S. (1999) *J. Mol. Biol. 293*, 1195–1210.
46. Dalby, P. A., Oliveberg, M., and Fersht, A. R. (1998) *J. Mol. Biol. 276*, 625–646.
47. Llinás, M., Gillespie, B., Dahlquist, F. W., and Marqusee, S. (1999) *Nat. Struct. Biol. 6*, 1072–1078.
48. Alm, E., and Baker, D. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 11305–11310.
49. Muñoz, V., and Eaton, W. A. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 11311–11316.
50. Galzitskaya, O. V., and Finkelstein, A. V. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 11299–11304.
51. Debe, D. A., and Goddard, W. A., III (1999) *J. Mol. Biol. 294*, 619–625.
52. O'Brien, R., Wynn, R., Driscoll, P. C., Davis, B., Plaxco, K. W., Sturtevant, J. M., and Ladbury, J. E. (1997) *Protein Sci. 6*, 1325–1332.
53. Plaxco, K. W., and Baker, D. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 13591–13596.
54. Wittung-Stafshede, P., Malmstrom, B. G., Winkler, J. R., and Gray, H. B. (1998) *J. Phys. Chem. A 102*, 5599–5601.
55. Ghaemmaghami, S., Word, J. M., Burton, R. E., Richardson, J. S., and Oas, T. G. (1998) *Biochemistry 37*, 9179–9185.
56. Choe, S. E., Matsudaira, P. T., Osterhout, J., Wagner, G., and Shakhnovich, E. I. (1998) *Biochemistry 37*, 14508–14518.
57. Silow, M., and Oliveberg, M. (1997) *Biochemistry 36*, 7633–7637.
58. Smith, C. K., Bu, Z. M., Anderson, K. S., Sturtevant, J. M., Engelman, D. M., and Regan, L. (1996) *Protein Sci. 5*, 2009–2019.
59. Scalley, M. L., Yi, Q., Gu, H., McCormack, A., Yates, J. R., and Baker, D. (1997) *Biochemistry 36*, 3373–3382.
60. Main, E. R. G., Fulton, K. F., and Jackson, S. E. (1999) *J. Mol. Biol. 291*, 429–444.
61. van Nuland, N. A. J., Meijberg, W., Warner, J., Forge, V., Scheek, R. M., Bovillard, G. T., and Dobson, C. M. (1998) *Biochemistry 37*, 622–637.
62. Aronsson, G., Brorsson, A. C., Sahlman, L., and Jonsson, B. H. (1997) *FEBS Lett. 411*, 359–364.
63. Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L., and Serrano, L. (1994) *Biochemistry 33*, 2142–2150.
64. Farrow, N. A., Zhand, O., Forman-Kay, J. D., and Kay, L. E. (1995) *Biochemistry 34*, 868–878.
65. Grantcharova, V. P., and Baker, D. (1997) *Biochemistry 36*, 15685–15692.

BI000200N