doi:10.1016/S0022-2836(02)00706-4 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 322, 463–476

JMB

Simple Physical Models Connect Theory and Experiment in Protein Folding Kinetics

Eric Alm[1], Alexandre V. Morozov[2], Tanja Kortemme[3] and David Baker[3]*

[1]*Lawrence Berkeley National Lab, Physical Biosciences Division, Berkeley, CA 94720 USA*

[2]*Department of Physics University of Washington Box 351560, Seattle, WA 98195 USA*

[3]*Department of Biochemistry Howard Hughes Medical Institute, University of Washington, J-567 Health Sciences, Box 357350, Seattle WA 98195-7350, USA*

**Corresponding author*

Our understanding of the principles underlying the protein-folding problem can be tested by developing and characterizing simple models that make predictions which can be compared to experimental data. Here we extend our earlier model of folding free energy landscapes, in which each residue is considered to be either folded as in the native state or completely disordered, by investigating the role of additional factors representing hydrogen bonding and backbone torsion strain, and by using a hybrid between the master equation approach and the simple transition state theory to evaluate kinetics near the free energy barrier in greater detail. Model calculations of folding ϕ-values are compared to experimental data for 19 proteins, and for more than half of these, experimental data are reproduced with correlation coefficients between $r = 0.41$ and 0.88; calculations of transition state free energy barriers correlate with rates measured for 37 single domain proteins ($r = 0.69$). The model provides insight into the contribution of alternative-folding pathways, the validity of quasi-equilibrium treatments of the folding landscape, and the magnitude of the Arrhenius prefactor for protein folding. Finally, we discuss the limitations of simple native-state-based models, and as a more general test of such models, provide predictions of folding rates and mechanisms for a comprehensive set of over 400 small protein domains of known structure.

Keywords: protein folding; transition state; kinetics; ϕ-values; master equation

Introduction

Recent theoretical models by Alm & Baker,[1] Galzitskaya & Finkelstein,[2] and Munoz & Eaton[3] have focused on the importance of topology in determining protein-folding mechanisms, using simple free energy functions to make predictions about folding rates and transition state (TS) structures. All three groups used a simplified approach in which each residue is considered to be either ordered as in the native state or completely disordered, with ordered residues occurring in one or more contiguous segments of the protein chain (the multiple sequence approximation). The models balance the entropic cost of ordering residues against the free energy decrease associated with making native interactions. Munoz & Eaton scaled the strength of native interactions on the basis of protein stability, a first-order approximation that allowed for the calculation of relative folding rates. Galzitskaya & Finkelstein, and Alm & Baker considered interactions between ordered segments, and were successful at predicting the distribution of structure in the folding TS for a limited number of proteins. In contrast to these models, work by Portman *et al.* introduces the use of a local order parameter to bypass the requirement that individual residues be either completely ordered or disordered.[4–6] Studies by Clementi *et al.*, and more recently by Koga & Takada continue the theme of native topology-based models, but relax the multiple sequence approximation by performing off-lattice simulations using a simplified representation of the protein chain.[7,8]

The correlation between folding rates and the simple topological measure, average contact order, suggests that such models may be sufficient to explain folding rates.[9] However, recent experimental data indicate that proteins with similar

E.A. & A.V.M. contributed equally to this work.

Abbreviations used: TS, transition state; TST, transition state theory.

E-mail address of the corresponding author: dabaker@u.washington.edu

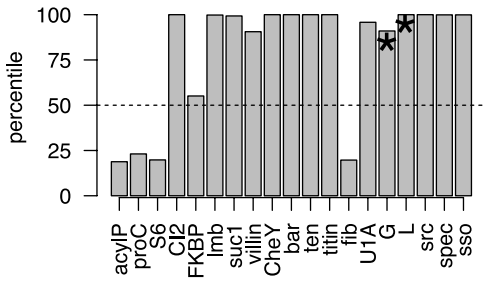

Figure 1. Statistical test of predicted ϕ-values. To assess the accuracy of ϕ-value predictions, predicted values were compared to experimentally measured values and to 1000 random permutations of the measured values as described in the text. Bars show the percent of randomly permuted values that do not correlate as well with predictions as the measured values (lower correlation coefficient). Predictions were made using the basic free energy function, except for starred proteins, for which the full free energy function including hydrogen bonding and torsion strain terms was used. In the following Figures, protein names are abbreviated as follows: acylP, muscle acylphosphatase; proC, procarboxypeptidase; S6, ribosomal protein S6; CI2, chymotrypsin inhibitor 2; FKBP, FK501-binding protein; lmb, lambda repressor; suc1, the protein product of the cell cycle gene p13suc1; villin, the N-terminal domain of the villin headpiece; CheY, bacterial chemotactic protein CheY; bar, barnase; ten, tenascin; fib, the tenth type III domain repeat of fibronectin; U1A, U1A spliceosomal protein; G, protein G; L, protein L; src, the src SH3 domain; spec, the α-spectrin SH3 domain; sso, the Sso7d SH3 domain.

topologies, such as proteins G and L, or the src and spectrin SH3 domains and Sso7d, can fold *via* different mechanisms, suggesting that some topologies allow multiple, nearly isoenergetic folding pathways, particularly when there is a symmetry in the native-state structure.[10–12] Protein design studies that successfully switch the native-folding pathway to an alternative-folding pathway in proteins L and G variants by strengthening or weakening specific intramolecular interactions, while not substantially changing the folding rate, further support this hypothesis.[13,14] These results point to the need for a more accurate free energy function to characterize folding in some proteins, and motivate the addition of terms that reflect hydrogen bonding and backbone torsion strain to our simple model. Guerois & Serrano recently used such an energy function to predict folding mechanisms successfully within the SH3 family.[15]

Simple theoretical models can be used to test our understanding of the folding process. Before any conclusions can be drawn, however, the validity of the model must be checked by comparison with experimental data. In principle, the results of any experimental measurement of folding kinetics or thermodynamics can be derived from a complete model of the folding landscape. In practice, the folding of many proteins has been characterized in terms of the folding rate of the naturally occurring protein, and by using site-directed mutagenesis (ϕ-value analysis) to probe the distribution of structure in the folding TS.[16] If the results from model calculations and experimental measurements are in agreement, then our basic understanding of the processes that underlie protein folding is validated. Additionally, there are several ways that these simple theoretical models can be used to provide information inaccessible by standard experimental techniques. In particular, unstable excited states such as the TS can be directly characterized, the energy landscape can be perturbed, and alternative folding pathways higher in free energy than the experimentally observed pathway can be identified.

We extend our previous model of protein-folding mechanisms by (1) investigating the addition of hydrogen bonding and backbone torsional strain terms to the free energy function, (2) computing folding rates and comparing these to experimental data for 37 proteins, (3) computing ϕ-values and comparing these to experimental data for 19 single domain proteins, (4) making available predictions of folding rates and TS structures for a comprehensive set of small protein domains of known structure, and (5) comparing the transition state theory (TST) approximation for folding rates with a more rigorous approach based on the kinetic master equations, and estimating the Arrhenius prefactor for the folding of small proteins.

Basic assumptions

As in previous models,[1–3] the folding free energy landscape is simulated by enumerating all configurations available to a protein chain. A configuration is uniquely defined by the state of all residues in the protein, where each residue is taken to be in one of two states: ordered as in the folded structure, or completely disordered. Ordered residues are required to occur in one or two contiguous groups in the linear protein sequence. Protein configurations are considered to be linked kinetically if they differ with respect to the state of exactly one residue.

In the basic free energy function, attractive interactions are taken to be proportional to the surface area buried by the ordered regions, and where noted, the energy function includes terms that reflect hydrogen bond strength and backbone torsion strain as described in Methods. Energetically favorable attractive interactions are offset by the entropic cost of ordering residues and of closing loops between ordered regions.

Results and Discussions

Comparing predicted and observed folding mechanisms

We compare model calculations of protein-folding pathways with experimentally observed

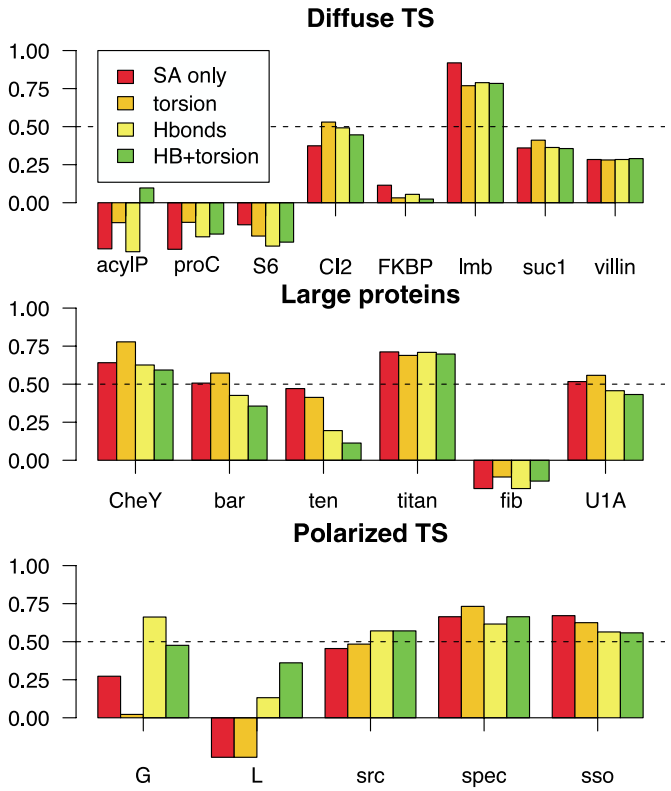

Figure 2. Comparison of four free energy functions. Free energy functions were compared based on the accuracy of ϕ-value predictions taken from calculations in which each was used. Bars indicate linear correlation coefficient, *r*, between model predictions and measured values. Four sets of predictions are shown: basic model using surface area and entropy terms only, basic + backbone torsion strain, basic + hydrogen bonding, basic + backbone torsion strain and hydrogen bonding terms.

ϕ-values by generating a simulated TS ensemble that consists of the 100 lowest free energy TS configurations on the model landscape (each configuration in the ensemble is weighted according to its free energy). A computed ϕ-value for a given residue is taken to be the frequency with which that residue is ordered in the TS ensemble. An implicit assumption is that each residue is ordered because of its favorable interactions with other ordered residues.

To test the statistical significance of the model ϕ-value predictions, the correlation of predictions to experimental data was compared to the correlation of predictions to randomized data in the following manner: 1000 decoy data sets were generated for each protein by randomly permuting the measured ϕ-values. The sequence position for each measurement was kept constant while the values were permuted to insure that results were not biased by the choice of experimentally probed sites. Next, the linear correlation coefficient, *r*, between the predicted values and the 1001 total data sets was assessed, and the rank of the correlation to the actual data set was computed as a percentile as shown in Figure 1. For over half of the proteins examined, the correlation to the actual data was above the 99th percentile.

Experimentally characterized TSs can be divided into three categories based on the distribution of ϕ-values observed: (1) polarized TSs in smaller proteins, (2) compact subdomains within larger proteins, and (3) diffuse TSs in which the observed structure is not as well-formed as in the native state and is distributed across much of the protein. As shown in Figure 2, calculations for proteins in the first two categories seem to be more reliable than those for proteins in the third.

Testing the free energy function

To test the value of adding additional terms to our model, calculations were performed using several variants of our free energy function: (1) the basic function including only the entropic and surface area burial based terms; (2) the basic function plus a term reflecting backbone torsion strain (non-glycine residues with positive ϕ torsion angles penalized +0.5 kcal/mol upon ordering; 1 cal = 4.184 J); (3) the basic function plus a term reflecting hydrogen bonding (backbone–backbone and side-chain–backbone hydrogen bonds between ordered residues assigned a free energy between 0 and −0.5 kcal/mol each); (4) the basic function plus both the backbone torsion strain and hydrogen bonding terms. As shown in Figure 2, the additional terms did not greatly affect the accuracy of ϕ-value predictions for most proteins. For some proteins, however, the additional terms did make a difference as described in detail in following sections. In particular, for proteins L and G, differences in hydrogen bond strength and backbone torsion strain play a crucial role in determining the folding pathway in model simulations. Additional terms were tested, such as amino acid and backbone torsion angle dependent free energies for ordering residues, but did not improve results enough to justify the added number of free parameters. Model calculations that allowed for

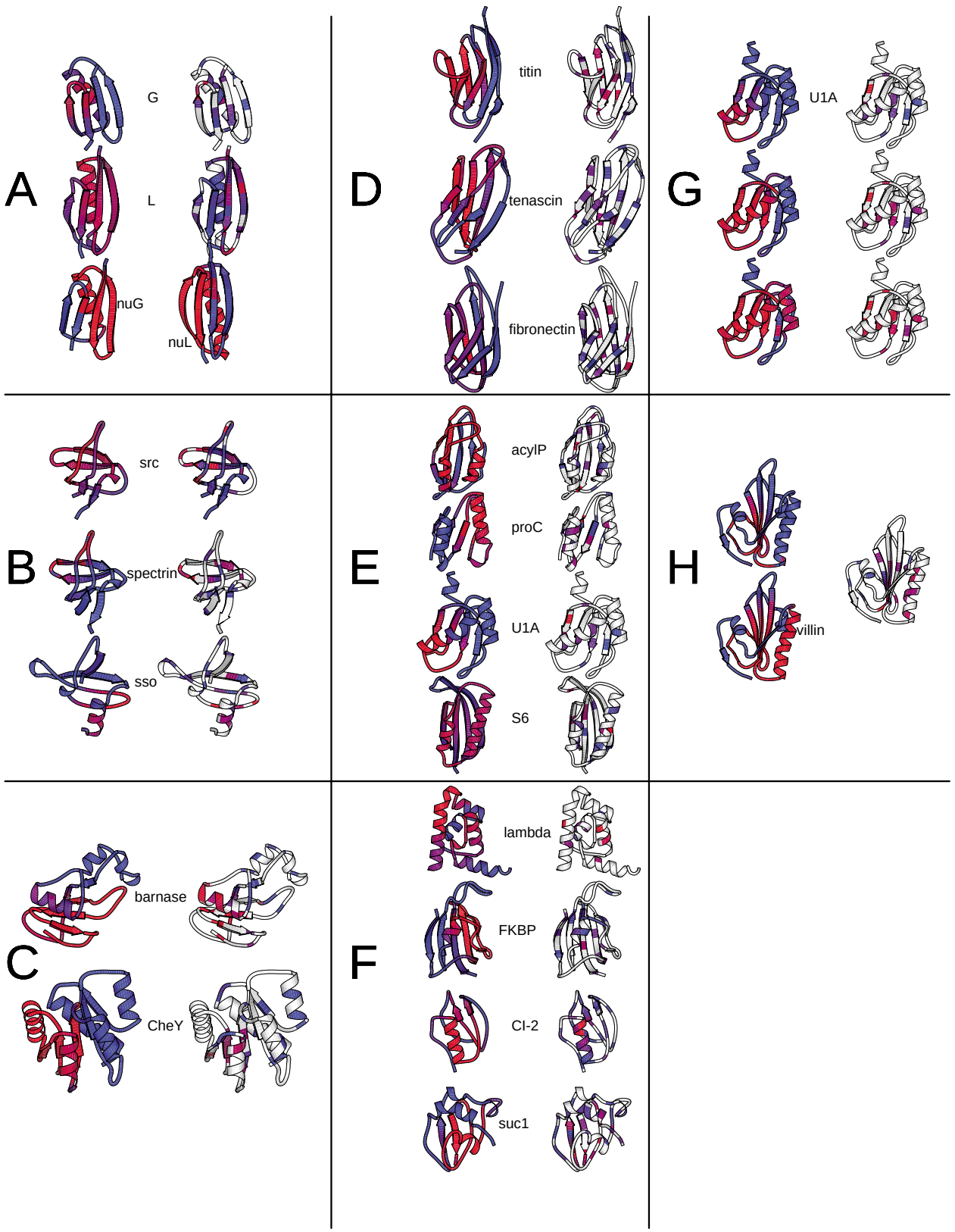

Figure 3. Predicted *versus* observed ϕ-values. ϕ-Value predictions were computed for 19 proteins as described in the text. ϕ-Value predictions are shown (left column) next to experimentally measured ϕ-values (right column). ϕ-Values for each residue are displayed on the protein using a gradient from blue ($\phi = 0$) to red ($\phi = 1$), white regions indicate residues without experimental data. (a) Proteins L and G, and redesigned proteins L and G (labeled nuL and nuG), note that since comprehensive ϕ-value analysis has not been carried out on the two redesigned proteins, only predictions are shown. (b) SH3 fold: src, spectrin, and Sso7d. (c) Large proteins with compact subdomains: CheY and barnase. (d) IgG fold: titin, tenascin, and fibronectin domain 10. (e) Ferredoxin fold: acylphosphatase, procarboxypeptidase,

three contiguous groups of ordered residues required that 2–5 residues be linked together kinetically in order to reduce the computational difficulty, and as a result, did not perform as well at reproducing ϕ-values and folding rates (results not shown).

In the discussion and Figures that follow, ϕ-values were computed using the basic free energy function except where indicated.

Small proteins with polarized transition states

Proteins L and G

Proteins L and G provide a good example of how model calculations can be used to interpret experimental results. These two proteins have a very similar topology as shown in Figure 3(a). Moreover, the symmetry of their native-state topology suggests that for every folding pathway, there is a complementary pathway, which begins ordering residues at the opposite end of the chain, and indeed the two proteins have complementary distributions of structure in their folding transition state ensembles. As might be expected, model calculations based on surface area burial and entropic considerations alone (the basic free energy function) do not accurately reproduce experimentally observed TS structure. Adding the hydrogen bond and torsion strain terms to the free energy function seems to capture the experimentally observed TS structure, prompting us to investigate in greater detail the factors that lead to this apparent symmetry breaking on the model folding landscape.

ϕ-Value analysis of protein L indicates that the N-terminal hairpin is formed, and the C-terminal hairpin remains disordered in the folding TS.[11] As shown in Figure 3(a), model ϕ-value calculations reproduce this asymmetry. In model calculations, the symmetry breaking is due to favorable side-chain–main-chain hydrogen bonding in the N-terminal hairpin and two non-glycine residues with unfavorable positive ϕ-angles in the C-terminal β-turn. In the case of protein G, the C-terminal hairpin rather than the N-terminal hairpin is ordered in the TS,[12] and has been shown to be stable in isolation.[17] In model calculations, the differences in hairpin stability and the asymmetric ϕ-value distribution result from extensive hydrogen bonding in the C-terminal hairpin and a greater total amount of surface area burial. Thus, for these two proteins, differences in hydrogen bond strength and backbone torsion strain, which account for a very small part of the total free energy function, are sufficient to change the distribution of structure at the rate-limiting step.

As an experimental test of the hypothesis that changes in the relative stabilities of local structural elements can change the folding pathway, computational protein design methods have been used to stabilize the first β-hairpin in protein G and the second hairpin in protein L. The folding TSs of the two redesigned proteins were found to be reversed: in contrast to their wild-type counterparts, the first hairpin is ordered in the TS of the redesigned protein G and the second hairpin is largely ordered in the TS of the redesigned protein L.[13,18] As shown in Figure 3(a), model calculations on crystal structures of these mutants using the full free energy function seem to capture the energetic differences between the mutant and wild-type proteins, and predict that the complementary pathways should be observed for these mutants. In the redesigned protein L, the non-glycine residues with positive ϕ torsion angles were replaced by a canonical type I′ turn, lowering the free energy of the second turn. For the redesigned protein G, hydrophobic packing is optimized in the N-terminal hairpin and a key hydrogen bond is deleted in the C-terminal hairpin.

SH3 fold

The SH3 fold is another case for which the model reproduces the observed TS structure. The fold consists of two opposing three-stranded β-sheets. In the src and spectrin SH3 domains, the local three stranded sheet comprised of the well-packed distal loop and the n-src loop is found to be ordered at the folding TS, while the sheet comprised of the RT loop and C-terminal strand is mostly disordered.[10,19] In model calculations (Figure 3(b)), this asymmetric distribution of structure in the TS can be attributed in part to more contacts in the compact distal loop hairpin compared to the relatively disordered RT loop, but is likely related to topological considerations as well. The three stranded sheet that includes the distal loop consists of three local β-strand pairs, while the opposing sheet contains a β-strand pair separated in sequence by nearly the length of the protein. When the model is used to simulate kinetics for a circularly permuted variant of the spectrin SH3 domain in which the N and C termini are joined, and the chain is cleaved between the two strands of the distal loop, this region is no longer predicted to be ordered at the rate limiting step, although the strength of the interactions in this region are unchanged.[10] Consistent with this

and U1A. (f) λ-repressor, FKBP, CI-2, and suc1. (g) U1A TS placement at different stabilities: left column shows predicted ϕ-values at 0 kcal/mol (top), +3 kcal/mol unstable (middle), +5 kcal/mol (bottom); right column shows experimental ϕ-values at m_f/m_{eq} values of: 0.5 (top), 0.7 (middle), and 0.85 (bottom). (h) ϕ-Value predictions for the villin headpiece at two stabilities: 0 kcal/mol (top), +5 kcal/mol (bottom); right column shows experimental values.

result, experimental φ-value analysis of this circular permutant indicates a different distribution of structure at the TS.[20]

The Sso7d domain differs notably from the other SH3 domains considered in that it contains only one three-stranded β-sheet, as the C-terminal region is helical. Although the distal loop β-strands are similar in structure to those of the other SH3 domains, the turn region is presumably less favorable due to the presence of five glycine residues. Both experimentally and in the model calculations, the C-terminal helix and nearby regions of the n-src loop are found to be ordered in the TS. Similar results on these SH3 folds were obtained by Guerois & Serrano who found hydrogen bonding and torsion angle dependent terms necessary to reproduce the experimental results,[15] and by Koga & Takada.[8]

Large proteins with compact subdomains

CheY and barnase

For CheY and barnase, results from the folding model point to a mechanism that is derived from first-order "topological" features. In fact, model calculations including only the topologically based terms, the surface area burial and entropy, perform slightly better than those using additional terms at reproducing the experimentally observed TS structure. The folding model identifies the compact subdomain at the CheY N-terminus and the compact core made by packing the C-terminal β-sheet against the N-terminal helix in barnase as optimal solutions that maximize surface area burial while ordering the fewest number of residues (Figure 3(c)). Consistent with these results, the N-terminal domain of CheY has been shown to be structured in the folding TS,[21] and for barnase, the C-terminal β-sheet and the N-terminal α-helix account for most of the structure observed in the folding TS.[22]

Immunoglobulin fold

The three immunoglobulin proteins studied have two minicore regions that correspond to the strands on the left and right half of the proteins shown in Figure 3(d), either of which could be used to nucleate the folding process. From a purely topological perspective, local interactions make the strands on the left more favorable as a folding nucleus, while the strands on the right form a minicore that is stabilized by interactions that span most of the length of the linear protein sequence. Model predictions for all three proteins suggest that the local core should comprise the folding nucleus; experimental data confirm this prediction for two of the three proteins (titin and tenascin), but also suggest that, for fibronectin, interactions in the non-local core are sufficiently strong to shift the folding nucleus to this alternative site.[23–25]

Proteins with diffuse transition states

Of the proteins studied by φ-value analysis to date, about half display a diffuse pattern of largely intermediate φ-values distributed across most of the protein. Model calculations reproduce these distributions in some cases, but are generally less accurate for proteins in this class compared to those described above. This may reflect the inherent limitation of the model that residues are either completely ordered or disordered, not partially ordered.

Ferredoxin-like fold

Two interwoven βαβ motifs form the scaffold for the ferredoxin-like fold, and as a result, the protein core includes mostly non-local interactions between both helices and all four strands. Compared to the immunoglobulin fold, this fold does not appear to have a clear topological bias toward a specific nucleus. However, for one protein in this set, U1A, Ternstrom *et al*. report a polarized TS centered around a locally stabilized minicore.[26] In particular, the asymmetry in the helix placement in U1A results in a very local minicore including the first helix and the adjacent beta-hairpin. Experimentally, these differences are realized in the folding TSs for the four proteins: acylphosphatase, procarboxypeptidase and ribosomal protein S6 have diffuse TSs with intermediate φ-values spread across most of the length of the proteins;[27–29] for the asymmetric U1A, the TS is polarized, and high φ-values are clustered in the local minicore surrounding the first helix. For this reason, we have included U1A in the class of large proteins with compact subdomains in Figure 2.[26] Model calculations on this set succeed at picking up on low resolution topological biases in the case of U1A, but do not perform as well on the other more symmetric proteins which may include more non-local interactions as well as multiple folding pathways (Figure 3(e)). Interestingly, Koga & Takada recently reported similar results on this group of proteins using an off-lattice model,[8] suggesting that these results reflect an inherent limitation of Go-type models rather than an artifact of our particular approach.

CI-2, FKBP12, λ-repressor, Suc1

The remaining proteins have TSs that can be described as delocalized (Figure 3(f)). CI-2 is an example of a protein with a low average φ-value, and the residues ordered in the TS structure are spread across part of the α-helix and some residues on the nearby β-strands.[30] This may indicate that multiple pathways are accessible for this protein, or a partial ordering of many residues in a nucleation–condensation type reaction.[31] Model calculations for CI-2 identify the α-helix and nearby β-strands as a compact local minicore, but overestimate the total amount of structure in the

TS. The FKBP12 TS has been compared to that of CI-2, and is mostly delocalized with low ϕ-values spread throughout the large β-sheet.[32] Model calculations, however, predict high ϕ-values in a small somewhat compact region of the protein, failing to reproduce the pattern of low and delocalized ϕ-values, suggesting that there may be multiple folding pathways available to this protein while only one is identified by model calculations. For Suc1, much of the observed structure in the TS is found in the β sheet, particularly the center strands.[33] Strands two and four make extensive contacts including a network of electrostatic interactions, but strand two is not predicted to be structured in model calculations. The lack of structure in helix 1, despite the strong helical propensity of its amino acids, observed both experimentally and in model calculations may be due in part to the relatively small surface area buried in this region of the native structure as evaluated by our energy function. In the case of λ-repressor, the model correctly identifies the helices determined by experiment to be important for stability at the rate-limiting step, and these helices form a well-packed interface that buries a substantial amount of surface area.[34]

Limitations of the model

There are three key limitations of the simple model that taken together may account for the disagreement between calculated and experimentally measured ϕ-values observed for some proteins: (1) the free energy function used is of limited accuracy; (2) configurations are taken to consist of residues that are either fully ordered or disordered, although actual protein conformations may include partially ordered structure; (3) only native interactions are considered. Figure 2 shows how the calculated TS for protein G and protein L can be sensitive to relatively small changes in the free energy function. Since changes as small as 1 kcal/mol in conformations near the rate-limiting step can dramatically change the amount of flux through different pathways, a good free energy function should be able to predict protein stability to within 1 kcal/mol, well beyond the reach of currently available potential functions. Proteins with diffuse TSs illustrate the second limitation of the model: experimentally much of the protein chain is observed to be partially ordered at the rate limiting step of folding for this group, but the model allows only for residues to be fully ordered or fully disordered. The ACBP folding TS (not shown), which includes many non-native interactions,[35] highlights the third limitation: non-native interactions cannot be accounted for by a model which considers only native structure. The assumption that only native interactions contribute can also be complicated by relatively small shifts that could occur late in folding, for example, an α-helix could shift only slightly relative to an adjacent β-sheet, but dramatically change the distribution of contacting residues.

Transition state placement

The Hammond postulate from organic chemistry suggests that lowering the free energy difference between the TS and folded state on the folding pathway should increase their structural similarity.[31] Applying this postulate to model calculations, the TS should have a higher value of N_f, the number of residues ordered, as the $\Delta\Delta G_{folding}$ is lowered. Experimentally, the TS should have a higher m_f/m_{eq} value, which measures the fraction of surface area buried at the rate-limiting step. The m-values are denaturant dependencies for the folding rate constant (m_f) and the equilibrium constant (m_{eq}). The folding pathway for U1A has been characterized extensively at different values of m_f/m_{eq}.[26] For comparison, the $\Delta\Delta G_{folding}$ of model calculations was adjusted by scaling the surface area burial term in our free energy function. Figure 3(g) shows a series of ϕ-value predictions using the full free energy function (including hydrogen bonding and torsion strain terms) in which the $\Delta\Delta G_{folding}$ is scaled to: 0, +3, and +5 kcal/mol; experimental measurements that correspond to m_f/m_{eq} values of: 0.5, 0.7, and 0.85 are shown for comparison.[26] Interestingly, calculations using our basic free energy function (data not shown) resulted in a sharp transition between a TS that is mostly disordered at low $\Delta\Delta G_{folding}$ to one that is mostly ordered at high $\Delta\Delta G_{folding}$, while calculations with the full free energy function (shown in Figure 3(g)) produced partially ordered TS configurations at intermediate stabilities. One explanation for these results is that the full free energy function maps out a more complicated free energy landscape with many local minima that have the potential to become saddle points as the overall stability is changed. The calculations shown reproduce the order in which structure is formed as the rate-limiting step is pushed toward the fully ordered state. Model calculations on other proteins suggest that similar changes in TS structure accompanying changes in protein stability are quite common, and for the Sso7d SH3 domain, the predicted TS can change qualitatively as the stability is changed; zero stability TSs (such as the one depicted in Figure 3(b)) for the Sso7d SH3 domain consist of order mainly in the n-src loop and the C-terminal helix, while earlier TSs (for stabilized landscapes that favor folded configurations) are structured mainly in the N-terminal RT loop (not shown). Thus, the quality of ϕ-value predictions may be dependent on accurate placement of the TS along the reaction coordinate. In some cases, the model may be able to determine the relative free energy difference between configurations at the same position of the reaction coordinate, N_f, but not between configurations at different values of N_f using the default scaling factor for attractive interactions, γ: for

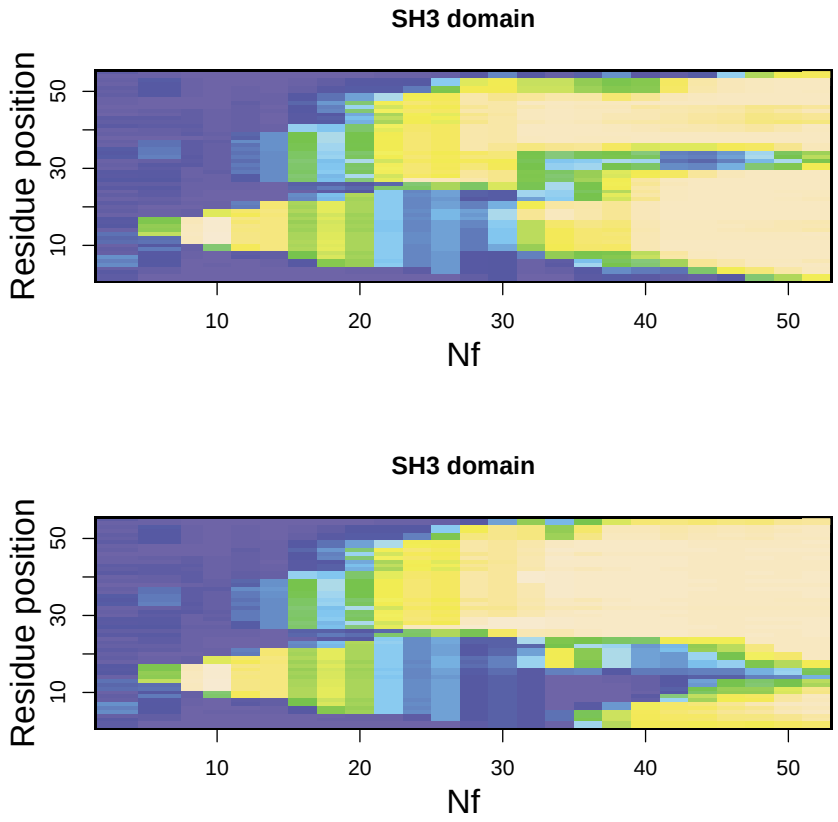

Figure 4. Deviation of steady-state distribution from equilibrium due to depletion near the folding free energy barrier. The frequency with which a particular residue is ordered in the ensemble of structures with a given value of N_f (the number of residues ordered) is shown. Upper plot: equilibrium distribution for the α-spectrin SH3 domain, where each configuration is weighted by its Boltzmann factor. Lower plot: steady-state distribution, where each configuration is weighted according to its steady-state concentration obtained *via* the master equation. The populations on both graphs are shown with the following color scheme (on a scale from 0 to 1): 0.01–0.35, hues of blue; 0.35–0.63, hues of green; 0.63–0.9, hues of yellow; 0.9–1.0, hues of light brown.

example, experimental results for the villin headpiece correlate significantly better with predictions for a +5 kcal/mol destabilized protein ($r = 0.59$) than with predictions for a zero stability landscape ($r = 0.24$) as shown in Figure 3(h).[36]

Characterizing the transition state ensemble

An approximation commonly employed in the study of protein-folding landscapes is that of thermodynamic equilibrium across the landscape. The master equation approach[37–39] recently employed by Cieplak *et al*. and Finkelstein *et al*. is a more rigorous treatment which includes flux over suboptimal-folding pathways, as well as deviations from equilibrium concentrations near the TS barrier.[2,40] We pursue this approach further and directly compare the concentration and flux through excited states with those predicted from a simpler model that assumes equilibrium across the folding landscape. In addition, we test the basic assumptions of the transition state theory (TST) by comparing folding rates calculated using the two methods. To reduce the number of states considered by the computationally more expensive master equation approach, we introduce a hybrid approach in which only states in the vicinity of the folding barrier are treated explicitly. Configurations distant from the barrier are treated as a thermally weighted source population (on the unfolded side) or a sink (on the folded side). Further, multiple residues are combined into links that are ordered or disordered simultaneously.[2] In all master equation calculations we employ the basic free energy function without hydrogen bonding or backbone torsion strain terms.

Figure 4 shows the relative population of all configurations on the folding landscape of the α-spectrin SH3 domain as a function of the number of residues ordered at equilibrium and for the steady-state case. There is markedly less ordering of residues in the N-terminal part of the protein just after the TS barrier ($N_f \simeq 30$) in the steady-state (lower plot) compared to the equilibrium distribution (upper plot) because there is little barrier crossing in this region (the lowest energy transition states are ordered in the C-terminal part of the protein). Conformations with the N-terminal part of the protein ordered are lower in free energy at $N_f \simeq 35$, as evident in the equilibrium distribution, but are not kinetically accessible.

The master equation approach was used to determine protein-folding rates, and Figure 5 compares these rates to those calculated from a simpler TST rate expression (the Arrhenius rate law):

$$k_f = k_0 \exp\left(-\frac{\Delta G^{TS}}{RT}\right) \quad (1)$$

where k_0 is the intrinsic kinetic rate (Arrhenius prefactor), ΔG^{TS} is the free energy of the TS, R is the gas constant, and T is the temperature. To facilitate comparison, configurations higher in free energy than the lowest TS were excluded in both

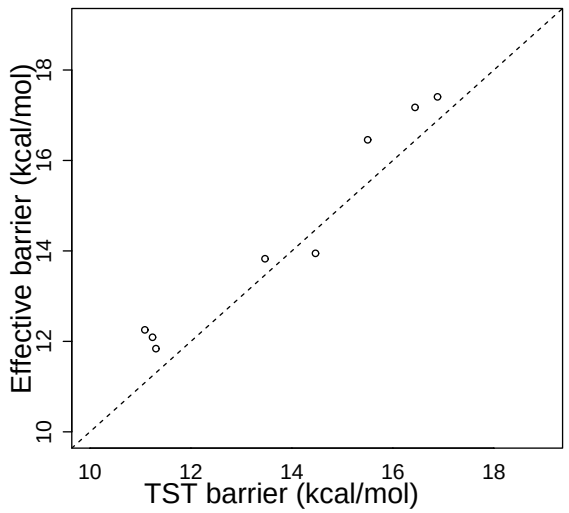

Figure 5. Comparison of folding rates obtained from master equation approach to TST. For each protein in the test set, the effective free energy barrier for the steady-state approach (calculated as $-RT\ln(k_f/k_0)$) is plotted *versus* the free energy of the single lowest free energy TS. The free energy cutoff for the master equation is set to +0.03 kcal/mol above the lowest free energy TS, such that only contributions from the lowest free energy path are included for comparison with the TST prediction. The dashed line shown is $y = x$. The linear correlation coefficient between the two methods is $r = 0.98$. The TST barrier can be higher than the effective master equation barrier because the master equation approach allows flux from a single TS into multiple kinetically connected configurations, thus increasing the total flux.

calculations. We expect the TST to overestimate the folding rates, since it neglects population depletion close to folding barriers. Nonetheless, Figure 5 shows an excellent agreement between master equation and TST rates (with correlation coefficient $r = 0.98$). Therefore, the simpler TST rate expression is used for comparison across the larger set of folding rates in the following section.

To probe the contribution of suboptimal paths to folding dynamics, configurations above a threshold free energy cutoff over the lowest free energy TS were removed from the landscape (Figure 6). Exclusion of paths with free energy maxima more than 2 kcal/mol greater than the lowest free energy TS did not appreciably change the rate of flux across the landscape, suggesting that in the model, the dominant contribution to folding dynamics comes from pathways within 2 kcal/mol of the lowest free energy path.

Folding rates

Model calculations of TS free energy were compared with experimentally measured folding rates for a set of 39 proteins. The experimental rates were taken at the midpoint of the denaturation curves where proteins are 50% folded, to be consistent with the model calculations in which the protein stability was set to zero. TS free energies were computed from the 100 lowest free

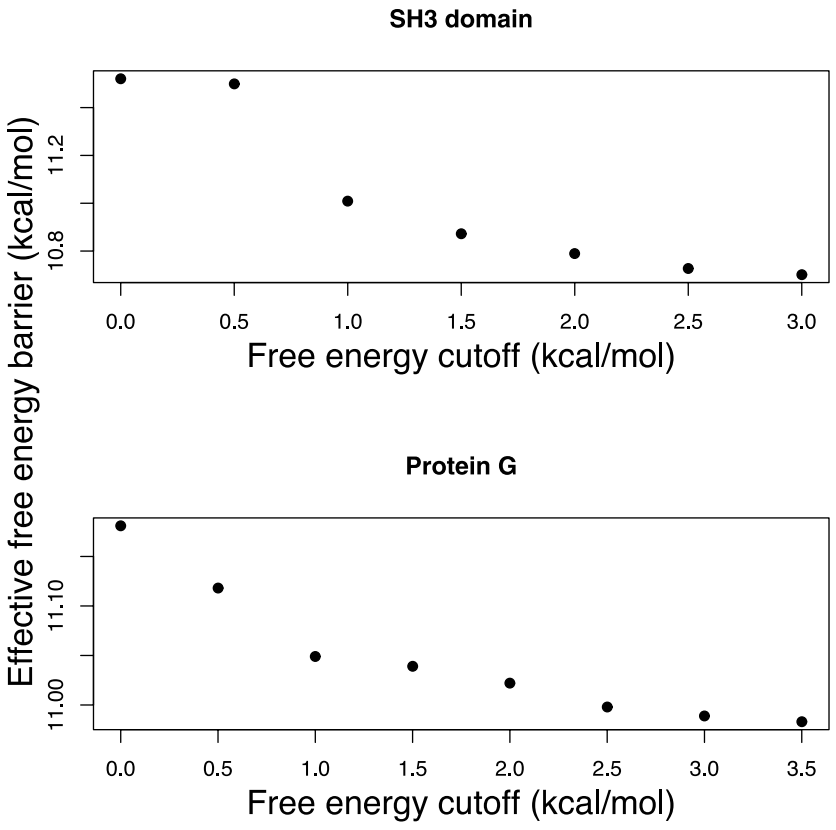

Figure 6. Contribution of high energy states to protein-folding kinetics. The calculated effective free energy barriers (calculated as $-RT\ln(k_f/k_0)$) for protein G and the src SH3 domain model landscapes are shown as a function of the threshold above which protein configurations were excluded from the calculation.

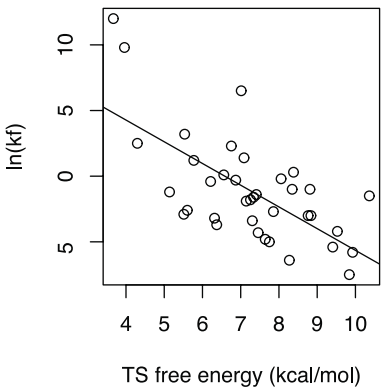

Figure 7. Comparison of calculated TS energy to measured folding rates. The *x*-axis shows the free energy of the TS ensemble computed from model calculations using the basic free energy function including only surface area burial and entropy terms. The free energy function is scaled by a factor of 0.54 such that the slope of the best fit line is $-1/RT$. TS free energies are calculated by taking the logarithm of the partition function that includes the 100 lowest free energy TSs. The linear correlation coefficient between predicted free energies and log folding rates is $r = 0.67$, with a *p*-value of 7×10^{6}.

energy TSs, using:

$$\Delta G^{\mathrm{TS}} = -RT \log\left\{ \sum_{i \in \mathrm{TS\ ensemble}} \exp\left(-\frac{\Delta G_i}{RT}\right) \right\} \quad (2)$$

where ΔG_i is the energy of the ith TS in the ensemble.

Using the simple free energy function without contributions from hydrogen bonding or torsional strain the correlation between experimental and computed rates is 0.67 (Figure 7). To be consistent with equation (1), the free energies ΔG_i in equation (2) were scaled by a factor of 0.54 to make the slope of the best fit line in Figure 7 equal to $-1/RT$. The scale factor may be viewed as a correction for the overestimation of the TS barriers, which occurs in our model since some residues are only partially ordered and some interactions are only partially formed at the TS. The incorporation of partial structure into such models has been addressed by Portman *et al.*, and is an important area for future study.[4–6] The correlation of the model TS free energies with observed folding rates does not change significantly using a free energy function that includes hydrogen bonding and torsion strain corrections (data not shown). ϕ-value distributions are likely to be more sensitive to such high resolution details than are folding rates, because given a set of pathways with roughly equal free energy barriers, changes of 1–2 kcal/mol can have large effects on the level of flux through the different pathways while having relatively little effect on folding rates.

The prefactor in the Arrhenius rate law can be estimated by extrapolation to a zero free energy barrier. The result for k_0 is approximately $10^5\ \mathrm{s}^{-1}$, in agreement with earlier semi-empirical and theoretical estimates.[3,41] Since the Arrhenius prefactor provides an upper bound for protein-folding rates in the absence of a free energy barrier, it is interesting to note that naturally occurring proteins span an almost entire range of *in vitro* folding rates, from the fastest physically allowed to the slowest biologically relevant.

Conclusions

Previous studies by our group and others have shown the applicability of simple models based on native structure to modeling protein-folding kinetics. We have further tested and extended this class of models by making predictions of folding rates and ϕ-values for a comprehensive set of proteins that have been experimentally characterized.

We find that experimentally probed TS structures generally fall into three categories: (1) small proteins with polarized TSs, (2) large proteins with compact subdomains, and (3) proteins with diffuse TSs. Our results suggest that an approach based on native state topology performs better at reproducing ϕ-values for proteins in the first two categories than for the third, and that adding additional terms such as hydrogen bonding and backbone torsion strain does not greatly affect model accuracy for most proteins. This may result from a fundamental limit of such models: two structurally distinct pathways that differ in free energy by as little as 1–2 kcal/mol can produce rates almost indistinguishable from those obtained from the lower energy pathway alone. This suggests that proteins with a single observable folding mechanism might have alternative-folding pathways close in free energy that are never observed experimentally. An ideal model must in principle be able to detect such small differences, otherwise a nearly isoenergetic alternative pathway can be mistaken for the lowest in free energy. The scatter in Figure 7, however, indicates that a simple free energy function based on the fraction of native contacts formed is not sufficiently accurate to discriminate between alternative-folding pathways separated by only a few kcal/mol. Additional limitations include the absence of non-native and partially formed interactions. In particular, partially formed interactions may be partly responsible for the poor performance of the model on proteins with diffuse TS structure.

Despite these limitations, we find that such a model can reproduce ϕ-value distributions for many proteins, and that a simple model with no adjustable parameters predicts TS energies that correlate reasonably well with experimentally measured folding rates. On the basis of these results, we use our model to make some general observations about protein-folding landscapes, which are not straightforward on the basis of the available experimental data alone. First, we note that simply changing the strength of stabilizing

interactions is enough to change the TS structure, particularly in proteins with symmetric native-state topologies such as proteins G and L. Quantitatively, we observe that most of the configurations that contribute to flux through the TS ensemble are within 2 kcal/mol of the lowest free energy TS configuration. This result is not trivial, since there are a very large number of higher energy configurations that could contribute because of their higher collective entropy. Although we observe some depletion of states near the folding free energy barrier relative to the populations expected at equilibrium, TST provides a very good approximation to kinetics on our model landscape, implying that the folding free energy barrier is large compared to the ruggedness of the landscape, and is relatively symmetric. Finally, since our model allows for the direct calculation of the TS free energy, we can estimate the magnitude of the Arrhenius prefactor term to be about $10^5\ s^{-1}$, in agreement with earlier theoretical and semi-empirical estimates.[3,41]

Since our model was developed to account for the data available to the authors, the key test of its validity will be its ability to predict the outcome of future protein-folding experiments. To facilitate such testing, we provide predictions† of folding rates and mechanisms for a comprehensive set of protein domains of known structure under 100 residues in length. The simple physical basis of our model allows for predictions to be used to interpret new experimental results when the two are in agreement, and for other cases should help point out limitations of Go-type models, and areas for further improvement.

Methods

Identifying transition-state configurations

A TS is defined as the highest energy state on the lowest energy path from the unfolded to the folded state. Additional transition states are defined as the highest free energy states on the lowest free energy paths which do not include previously identified transition states. Transition states are identified using an algorithm described previously.[1]

Free energy function

The full free energy function including both hydrogen bonding and backbone torsion strain terms is given by:

$$F = -\gamma\Delta\mathrm{SA} - H(\mathrm{config}) + \sum_{r\in\mathrm{residues}} \mathrm{Ord}(r)F_{\mathrm{local}}(r) + 1.8RT\sum_{\mathrm{loops}} \ln\left(\frac{L}{L_0}\right) \quad (3)$$

where

$$\Delta\mathrm{SA\ (surface\ area)} = \mathrm{SA}_{\mathrm{folded}} - \mathrm{SA}_{\mathrm{unfolded}}$$

is the difference between folded and unfolded solvent-accessible surface areas,

$$F_{\mathrm{local}}(r) = \begin{cases} 2.3\ \mathrm{kcal/mol} & \text{if } \phi > 0 \text{ and not glycine,} \\ 0.8\ \mathrm{kcal/mol} & \textit{otherwise} \end{cases}$$

$$\mathrm{Ord}(r) = \begin{cases} 1 & \text{if residue } r \text{ is ordered} \\ 0 & \text{if residue } r \text{ is disordered} \end{cases}$$

Here, ϕ is a backbone torsion angle. The first two terms represent the interactions that contribute to protein stability: surface area burial and hydrogen bonding. The constant, γ, which controls the free energy associated with surface area burial, is fixed for each protein such that the stability of the folded state ensemble (defined as all configurations with at least half of their residues ordered) is equal to that of the unfolded state ensemble (all configurations with less than half of their residues ordered), which is approximately true for most proteins (the models described by Galzitskaya & Finkelstein, and Munoz & Eaton also scale native interactions to adjust protein stability[2,3]). To compute buried surface area, ordered residues are modeled using their native state atomic coordinates and the unfolded state is modeled as an extended chain. Disordered residues are not considered in surface area calculations. Calculations are carried out using the method of LeGrand & Merz.[42]

Energies of backbone–backbone, side-chain–backbone and side-chain–side-chain hydrogen bonds (H (config)) were determined using an empirical potential function described elsewhere (T.K., A.M. & D.B., unpublished results). Briefly, the potential requires explicit placement of polar hydrogen atoms, and is dependent on (a) the distance between the hydrogen (H) and the acceptor (A) atoms, (b) the angle at the hydrogen atom (D–H···A) (D, donor atom), and (c) the angle at the acceptor atom (H···A–AB) (AB, heavy atom bound to the acceptor atom). The distance dependence was described by a 10–12 potential with a minimum at a distance of 1.9 Å between acceptor and hydrogen atoms. The angle-dependent terms of the hydrogen bonding potential were derived from hydrogen bond geometries observed in high-resolution (2.0 Å or better) protein crystal structures. Only hydrogen bonds with proton positions given by the chemistry of the donor group were considered in the derivation of the potential. The observed angle-dependent probabilities for side-chain–side-chain hydrogen bonds in the database of protein structures showed maxima at 180° (for the angle at the hydrogen atom) and 120°–140° (for the angle at a donor with sp^2 hybridization, with a slightly sharper distribution around a maximum of 120° for a donor with sp^3 hybridization), and were used to derive interaction energies for side-chain–side-chain and side-chain–backbone hydrogen bonds (backbone–backbone hydrogen bonds displayed a slightly different geometry, presumably due to steric constraints). To apply the derived hydrogen bonding function to hydrogen bond scoring in the experimentally determined structures used here, explicit hydrogen atoms were placed either according to the known geometry around the donor group (for N, Q, R, and W), or by optimization of the hydrogen bonding network involving all polar groups, using a scoring function which included van der Waals interactions and solvation terms as well as the hydrogen bonding function described above. Finally, hydrogen bonds were scaled between 0 and -0.5 kcal/mol. The hydrogen bonding term is absent from all calculations using the basic free energy function.

† http://www.bakerlab.org

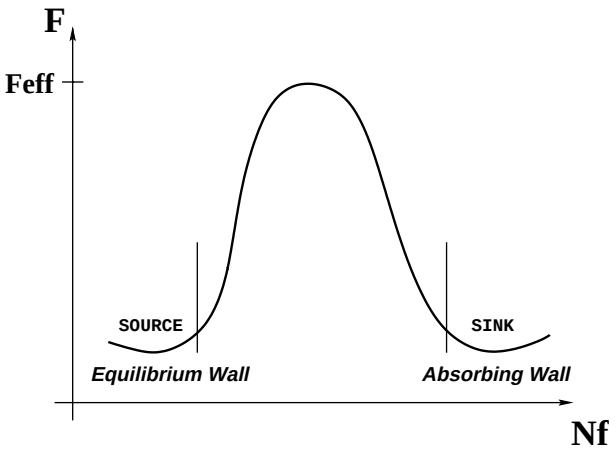

Figure 8. Schematic demonstration of the master equation approach using a one-dimensional free energy landscape. Master equations are solved in the region between the source (equilibrium wall) and the sink (absorbing wall). At the source, thermodynamic equilibrium is imposed; a particle (protein) is committed to the final state once it reaches the sink.

The third term represents the cost of ordering a single residue and is taken to be 1.8 kcal/mol as in our previously described model.[1] For calculations including the torsion strain penalty, 0.5 kcal/mol was added to the cost of ordering non-glycine residues with positive ϕ-angles.

The last term represents the entropic cost of closing a loop between two ordered segments, where L is the length of the loop ($L_0 = 0.15$ Å). This entropy was estimated from simulations of loop closure frequencies in polypeptide chains,[43] and has the same functional form as the Jacobsen–Stockmayer expression used in polymer chemistry.[44]

Free parameters

The free energy associated with surface area burial was scaled for each protein such that the unfolded and folded state were of equal stability, and the entropy cost of ordering a single residue was taken to be 1.8 kcal/mol.[1] The entropy loss associated with loop closure was taken directly from off-lattice simulations.[43] The strength of hydrogen bonds and the penalty for non-glycine residues with positive ϕ-angles were chosen such that they improved the ability of the model to accurately predict ϕ-values, and did not significantly affect TS energies. For the calculation of folding rates, the basic free energy function without hydrogen bonding or backbone torsion strain terms was used, and the free energies were multiplied by a factor of 0.54 such that model predictions could be directly compared with experimental measurements (observed folding rates indicate that free energy barriers to folding span about 9 kcal/mol).

Master equation approach

In order to investigate deviations from a simple TST and their effect on protein folding, we solve a system of master equations given by:

$$\frac{dn_i}{dt} = \sum_j (k_{ji} n_j - k_{ij} n_i) \tag{4}$$

where i denotes a single protein conformation, and Metropolis rules are used to describe an elementary kinetic step from i to j:

$$k_{ij} = k_0 \times \begin{cases} 0 & \text{if } ij \text{ impossible} \\ 1 & \text{if } F_i \geq F_j \\ e^{(F_i - F_j)/RT} & \text{if } F_i < F_j \end{cases}$$

Here, k_0 is the intrinsic kinetic rate (crossing attempt frequency).

In the case of steady-state protein diffusion across the free energy landscape, $dn_i/dt = 0$, and the system of equations becomes algebraic. We introduce a hybrid approach in which we consider protein folding in microscopic detail close to the barrier while assuming Boltzmann distribution of states well before the barrier. This allows the treatment of the important features at the barrier to be without approximation while retaining computational tractability. The approach is implemented through two boundary conditions: a source at thermal equilibrium, and an irreversible sink (Figure 8):

$$\begin{cases} n_i = n_0 e^{-F_i/RT} & i \in \text{equilibrium wall} \\ n_i = 0 & i \in \text{absorbing wall} \end{cases}$$

where n_0 sets the total number of particles (proteins) in the ensemble. The equilibrium wall is defined as the set of configurations directly connected to the source population at equilibrium, and the absorbing wall is defined as the set of configurations directly connected to the irreversible sink.

The folding rate is given by:

$$\text{rate} = \frac{\text{flux into absorbing wall}}{n_0}$$

The nonequilibrium steady-state solution corresponds to creating a source of particles in the denatured well and a sink in the native well. By moving the equilibrium wall further away from the location of the TSs, we can relax the assumption of equilibrium in the vicinity of the folding barriers. On the other hand, assuming equilibrium among low free energy conformations distant from the barriers helps reduce the computational complexity of the problem. Note also that by moving the absorbing wall further away, we allow a particle to recross the barrier multiple times.

In order to make the system of algebraic equations well-defined, we identified clusters of kinetically connected nodes on the free energy landscape. Only clusters stretching from the source to the sink can carry flux across; isolated groups of nodes can only equilibrate within themselves. The number of nodes included in the flux-carrying cluster depends on the free energy cutoff imposed on every node; setting the cutoff lower than the lowest free energy TS makes all clusters disconnected, and setting the cutoff higher includes effects from multiple TSs.

Finally, we used a reduced landscape in which several residues are added or removed in a single kinetic step whenever the number of conformations was too great for the master equation approach to be feasible.[2]

Folding rates

To investigate the relationship between predicted TS free energies and experimentally measured folding rates, a set of 37 small proteins was used. The data for this set were taken from Grantcharova *et al.*,[45] Jackson[46] and from Schymkowitz *et al.*,[33] and Guerois & Serrano,[15] and measured at the midpoint transition where the protein stability is zero. A complete list of the proteins used and their corresponding PDB identifiers is included in the following section.

PDB files

To build model landscapes for ϕ-value and folding rate predictions the following proteins and PDB files were used: cytochrome b562, 256b; barstar, 1a19; ACBP, 2abd; tendamistat, 2ait; acylphosphatase, 1aps; procarboxypeptidase, 1aye; Sso7d SH3 domain, 1bf4; α-spectrin SH3 domain, 1bk2; CheY, 2chf; chymotrypsin inhibitor 2, 2ci2; cspB cold-shock protein, 1csp; ribosomal protein L9, 1div; FK506-binding protein, 1fkb; src SH3 domain, 1fmk; fibronectin type III repeat 9, 1fnf; fibronectin type III repeat 10, 1fnf; HPR, 1hdn; CD2 lymphocyte adhesion protein, 1hng; horse heart cytochrome *c*, 1hrc; colicin E9 immunity protein Im9, 1imq; λ-repressor, 1lmb; cspA cold-shock protein, 1mjc; dihydrolipoamide acetyltransferase from pyruvate dehydrogenase, 2pde; protein G, 1pgx; protein G C-terminal beta-hairpin, 1pgx; PI3-kinase SH3 domain, 1pks; ribosomal protein S6, 1ris; barnase, 1rnb; fyn SH3 domain, 1shf; tenascin, 1ten; titin, 1tit; ubiquitin, 1ubq; U1A spliceosomal protein, 1urn; villin headpiece, 2vik; twitchin, 1wit. The coordinates for the suc1 monomer were obtained from Jane Endicott. The coordinates for protein L were obtained from Jason O'Neill prior to release in the PDB as 1hz5.

Acknowledgments

This work was supported by the Howard Hughes Medical Institute. T.K. was supported by post-doctoral fellowships of the European Molecular Biology Organization and the Human Frontier Science Program Organization. E.A. was also supported by a Molecular Biophysics Training Grant from the NIH.

References

1. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA*, **96**, 11305–11310.
2. Galzitskaya, O. & Finkelstein, A. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA*, **96**, 11299–11304.
3. Munoz, V. & Eaton, W. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.
4. Portman, J., Takada, S. & Wolynes, P. (1998). Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Lett.* **81**, 5237–5240.
5. Portman, J., Takada, S. & Wolynes, P. (2001). Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach. *J. Chem. Phys.* **114**, 5069–5081.
6. Portman, J., Takada, S. & Wolynes, P. (2001). Microscopic theory of protein folding rates. II. Local reaction coordinates and chaindynamics. *J. Chem. Phys.* **114**, 5082–5096.
7. Clementi, C., Nymeyer, H. & Onuchic, J. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.
8. Koga, N. & Takada, S. (2001). Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* **313**, 171–180.
9. Plaxco, K., Simons, K. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
10. Grantcharova, V., Riddle, D., Santiago, J., Alm, E., Tsai, J. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024.
11. Kim, D., Fisher, C. & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971–984.
12. McCallister, E., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein g folding. *Nature Struct. Biol.* **7**, 669–673.
13. Nauli, S., Kuhlman, B. & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nature Struct. Biol.* **8**, 602–605.
14. Kuhlman, B., O'Neill, J., Kim, D., Zhang, K. & Baker, D. (2001). Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc. Natl Acad. Sci. USA*, **98**, 10687–10691.
15. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
16. Matouschek, A., Kellis, J. T. J., Serrano, L. & Fersht, A. (1989). Mapping the transition-state and pathway of protein folding by protein engineering. *Nature*, **340**, 122–126.
17. Blanco, F., Rivas, G. & Serrano, L. (1994). A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nature Struct. Biol.* **1**, 584–590.
18. Kuhlman, B., O'Neill, J., Kim, D., Zhang, K. & Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J. Mol. Biol.* **315**, 471–477.
19. Martinez, J., Pisabarro, M. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721–729.
20. Viguera, A., Blanco, F. & 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.
21. Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* **1**, 43–55.

22. Serrano, L., Matouschek, A. & Fersht, A. (1992). The folding of an enzyme. III. Structure of the transition-state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805–818.
23. Fowler, S. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. *Structure*, **9**, 355–366.
24. Cota, E., Steward, A., Fowler, S. & Clarke, J. (2001). The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. *J. Mol. Biol.* **305**, 1185–1194.
25. Hamill, S., Steward, A. & Clarke, J. (2000). The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165–178.
26. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc. Natl Acad. Sci. USA*, **96**, 14854–14859.
27. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005–1009.
28. Villegas, V., Martinez, J., Aviles, F. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027–1036.
29. Otzen, D. & Oliveberg, M. (1999). Salt-induced detour through compact regions of the protein folding landscape. *Proc. Natl Acad. Sci. USA*, **96**, 11746–11751.
30. Itzhaki, L., Otzen, D. & Fersht, A. (1995). The structure of the transition-state for folding of chymotrypsin inhibitor-2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288.
31. Matthews, J. & Fersht, A. (1995). Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of hammond behavior for the gross structure of the transition-state and anti-hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry*, **34**, 6805–6814.
32. Fulton, K., Main, E., Daggett, V. & Jackson, S. E. (1999). Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445–461.
33. Schymkowitz, J., Rousseau, F., Irvine, L. & Itzhaki, L. (2000). The folding pathway of the cell-cycle regulatory protein p13suc1: clues for the mechanism of domain swapping. *Struct. Fold. Des.* **8**, 89–100.
34. Burton, R., Huang, G., Daugherty, M., Calderone, T. & Oas, T. (1997). The energy landscape of a fast-folding protein mapped by Ala-Gly substitutions. *Nature Struct. Biol.* **4**, 305–310.
35. Kragelund, B., Osmark, P., Neergaard, T., Schiodt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. M. (1999). The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nature Struct. Biol.* **6**, 594–601.
36. Choe, S., Matsudaira, P., Wagner, G. & Shakhnovich, E. (2000). Differential stabilization of two hydrophobic cores in the transition state of the villin 14t folding reaction. *J. Mol. Biol.* **304**, 99–115.
37. Delbrück, M. (1940). Statistical fluctuations in autocatalytic reactions. *J. Chem. Phys.* **8**, 120–124.
38. van Kampen, N. G. (1981). *Stochastic Processes in Physics and Chemistry*, North-Holland, New York.
39. Hänggi, P., Talkner, P. & Borkovec, M. (1990). Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* **62**, 251–341.
40. Cieplak, M., Henkel, M., Karbowski, J. & Banavar, J. (1998). Master equation approach to protein folding and kinetic traps. *Phys. Rev. Lett.* **80**, 3654–3657.
41. Eaton, W. (1999). Searching for downhill scenarios in protein folding. *Proc. Natl Acad. Sci. USA*, **96**, 5897–5899.
42. Le Grand, S. & Merz, K. (1993). Rapid approximation to molecular surface area *via* the use of boolean logic and look-up tables. *J. Comput. Chem.* **14**, 349–352.
43. Yi, Q., Scalley-Kim, M., Alm, E. & Baker, D. (2000). NMR characterization of residual structure in the denatured state of protein L. *J. Mol. Biol.* **299**, 1341–1351.
44. Jacobson, H. & Stockmayer, W. (1950). Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* **18**, 1600–1606.
45. Grantcharova, V., Alm, E., Baker, D. & Horwich, A. (2001). Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* **11**, 70–82.
46. Jackson, S. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91.

Edited by C. R. Matthews

(Received 6 March 2002; received in revised form 2 July 2002; accepted 9 July 2002)