Landscape approaches for determining the ensemble of folding transition states: Success and failure hinge on the degree of frustration

Hugh Nymeyer*†, Nicholas D. Socci§, and José Nelson Onuchic*†

*Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319; and §Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10021

Edited by R. Stephen Berry, University of Chicago, Chicago, IL, and approved November 5, 1999 (received for review July 2, 1999)

We present a method for determining structural properties of the ensemble of folding transition states from protein simulations. This method relies on thermodynamic quantities (free energies as a function of global reaction coordinates, such as the percentage of native contacts) and not on ''kinetic'' measurements (rates, transmission coefficients, complete trajectories); consequently, it requires fewer computational resources compared with other approaches, making it more suited to large and complex models. We explain the theoretical framework that underlies this method and use it to clarify the connection between the experimentally determined Φ value, a quantity determined by the ratio of rate and stability changes due to point mutations, and the average structure of the transition state ensemble. To determine the accuracy of this thermodynamic approach, we apply it to minimalist protein models and compare these results with the ones obtained by using the standard experimental procedure for determining Φ values. We show that the accuracy of both methods depends sensitively on the amount of frustration. In particular, the results are similar when applied to models with minimal amounts of frustration, characteristic of rapid-folding, single-domain globular proteins.

protein folding | Φ values | folding funnels | folding landscapes

Energy landscape theory and the funnel concept have provided a theoretical framework for understanding protein folding (1–7), which is an alternative to the earlier idea that there is a single pathway for the folding event comprising uniquely defined structural intermediates (8, 9). The connection between the landscape theory and real proteins is best established in the context of small fast folding proteins, which fold on millisecond time scales and have a single folding domain; i.e., they are two-state folders with a single, well defined funnel (10). In addition to the theoretical literature describing this theory and its applications (see, for example, the citations above, refs. 5 and 6, and references therein), a new generation of clever experiments [NMR dynamic spectroscopy, protein engineering, laser initiated folding, and ultrafast mixing (see, for example, refs. 11–27)] are providing the temporal and spatial detail needed to extend and elaborate on it.

A central result of this theory is that proteins with funneled landscapes have population dynamics that can be understood as the diffusion of an ensemble of configurations over a low-dimensional free energy surface (1, 3, 4). This energy surface may be constructed by using many different order parameters. The primary requirements are that they distinguish native-like and non-native-like structures and that they group together conformations with similar energies;¶ i.e., the dispersion in energies of states with similar values of these parameters is small. Many simple order parameters that are computationally convenient, like the number of tertiary contacts Q, or experimentally convenient, like the radius of gyration, satisfy these requirements. A successful description of the folding mechanism using only a few of these parameters is possible because the funnel of a protein landscape is deep compared with its energetic ruggedness, which is produced by frustration (1, 2). The rate of local motion on the landscape is set by the reconfigurational diffusion coefficient. Free energy barriers that separate the unfolded and native state ensembles impede this population diffusion and set crossing times that can be determined (for weak frustration) from a Kramer's-like equation.‖ When these crossing times are rate-limiting, the free energy of activation for folding is set by the height of these barriers on the landscape. The structural properties of the transition state ensemble can then be determined by sampling the thermal distribution of states in the dominant free energy barrier rather than by making ''kinetic'' measurements (e.g., measuring folding rates and transmission coefficients, or simulating complete folding pathways). The elimination of these kinetic measurements allows the folding transition state to be rapidly computed (compared with alternate methods), especially in large and complex systems (31–35) that cannot be studied by other means (29, 36–39).

Some kinetic methods have been used by others to compute Φ values from atomically detailed simulations. In one method (40, 41), strong unfolding biases are put to the native protein by raising the temperature, and the subsequent unfolding trajectory is interpreted as the reverse of a typical folding trajectory. The trajectories found by this method do show interesting correlation with the Φ values of experiments, but the unphysically high temperatures needed to force this transition in a reasonable time introduce large distortions into the landscape, creating difficulties in quantitatively connecting these results with proteins at physiological temperatures.

In this manuscript, we introduce the landscape-based method and validate its use by using a minimalist lattice protein model. By studying several different potentials, we determine the amount of frustration for which this approach can be reliably applied to determine transition state structure.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: 3LC, three-letter code; 2LC, two-letter code.

†To whom reprint requests should be addressed at: University of California at San Diego, Department of Physics 0319, Urey Hall 7210, 9500 Gilman Drive, La Jolla, CA 92093-0319. E-mail: hnymeyer@ucsd.edu or jonuchic@ucsd.edu.

¶These energies are actually free energies averaged over the degrees of freedom not explicitly included in the structure of the landscape: e.g., solvent degrees of freedom (5).

‖The accuracy of this method in predicting folding rates is demonstrated in the context of a simple lattice model in ref. 4. Equilibrium sampling with a simple order parameter is used to determine the height of the free energy barrier, and the autocorrelation time in the order parameter is used to estimate the reconfigurational diffusion constant. A Kramer's equation constructed from these two parameters provides an estimate of folding rates around the folding temperature correct to within a factor of order unity. This is expected because, although several recent studies have shown that a large fraction of free energy barrier states are actual microscopic transition states in moderately designed systems (28–30) and nearly all transition states lie in the barrier region, some of the configurations in this region are not real microscopic transition states (due, for example, to topological constraints).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

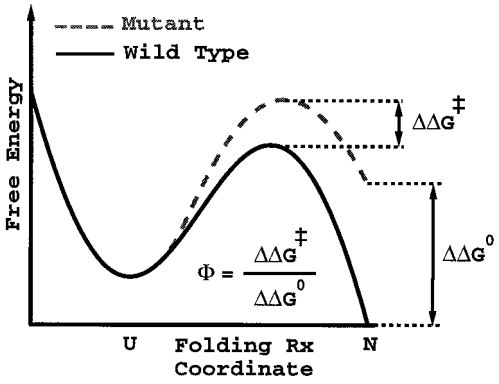

Fig. 1. Approximate schematic of a Φ value, which holds for proteins that are weakly frustrated. The solid curve is a schematic free energy profile for a wild-type protein; the dashed curve is for a suitable mutant. The free energy profile is drawn against a single order parameter for folding. For many small proteins, many simple global structural measures of nativeness may be used for this free energy projection. $\Delta\Delta G^0$ and $\Delta\Delta G^{\ddagger}$ are, respectively, the change in native state stability and change in activation free energy upon mutation. A Φ value near 1 suggests that the local environment of the mutated residue is native-like in the transition state; a Φ value near 0 suggests that the local environment of the mutated residue is unfolded-like in the transition state. This interpretation of Φ values becomes less valid as the frustration increases.

Methods and Simulations

Spatially localized transition state structure in proteins can be inferred by measuring the effect that point mutations have on the native stability and the folding (unfolding) rate. These effects can be used to compute the Φ value (42), an approximate measure of the protein structure around the site of the mutation:

$$\Phi \equiv \frac{-RT\ln(k_{\mathrm{mut}}/k_{\mathrm{wt}})}{\Delta\Delta G^0}. \qquad [1]$$

k_{mut} and k_{wt} are the mutant and wild-type folding rates, and $\Delta\Delta G^0$ is the change in total stability of the folded state when the mutation is made.

As we have discussed in the introduction, the crossing rate in minimally frustrated systems can be described via a Kramer's-like equation,

$$k = k_0\exp[-\Delta G^{\ddagger}/RT], \qquad [2]$$

where k_0 depends on the reconfigurational diffusion coefficient and the geometric shape of the barrier.** If this prefactor is insensitive to the specific amino acid sequence [this is expected in weakly frustrated systems and has been observed in some simple models (10)], then

$$-RT\ln(k_{\mathrm{mut}}/k_{\mathrm{wt}}) = \Delta G^{\ddagger}_{\mathrm{mut}} - \Delta G^{\ddagger}_{\mathrm{wt}} = \Delta\Delta G^{\ddagger}, \qquad [3]$$

and the Φ value is a measure of the ratio of the change in activation free energy of the folding barrier to the change in total stability upon making a particular mutation; i.e.,

$$\Phi = \frac{\Delta\Delta G^{\ddagger}}{\Delta\Delta G^0}, \qquad [4]$$

as illustrated in Fig. 1. If the amount of frustration is greatly increased, the folding times become controlled by long lived traps, and the assumption of a single k_0 that depends only on the average ruggedness of the landscape breaks down. When such traps are dominant, the ensemble of states in the top of the free energy barrier cannot be associated with transition states anymore.

If the free energy change created by the mutation is related in a simple way to some structural quantity, then the Φ value is a measure of the average difference in that quantity between the unfolded and transition state ensembles, expressed as a fraction of the average unfolded to native difference—it is in this sense that the Φ value tells us the "location" of the barrier via a number that varies between 0 and 1 (44). [But because Φ values only measure relative free energy changes, they cannot be directly interpreted in terms of structure without knowledge of the typical unfolded state structure (27).] Intermediate range Φ values can arise as an average over a heterogeneous transition state ensemble with different amounts of local structure (for experimental support for this, see ref. 45). Comparisons of Φ values from different point mutations at the same sequence position have generally supported the simple interpretation of the Φ value as an indication of local structure (46), although exceptions have been observed in at least one small protein (13).

Unfortunately, an average transition state ensemble structure cannot for most systems be practically determined from simulations by computing changes in folding rates. A 3% error in the folding rate requires >1,000 separate folding simulations, and it is impossible to perform this many simulations except for the simplest models. It will certainly remain an infeasible computation in all atom models for many decades because most protein folding times are in the millisecond time range or greater, and simulations are currently limited to a few tens or hundreds of nanoseconds of real time. As discussed in the introduction, the landscape framework provides an alternate computational method for studying the transition state ensemble. By using a small number of order parameters to define a landscape, we can identify the dominant barrier separating the unfolded and native basins. The change in the folding rate when a mutation is made can then be computed by finding the change in the activation free energy of this barrier; i.e., we assume that the transition states are the thermally occupied states in the free energy barrier region. In systems in which Φ values are accurate probes of the average transition state structure, we show that this method is equally valid as the experimental technique.

To demonstrate the applicability of this computational approach, we use a model that retains only the most essential characteristics of small globular proteins. We represent a protein as a short chain of monomers that are constrained to the vertices of a three-dimensional cubic lattice. The nonbonded interactions between the monomers are contact interactions between neighboring lattice sites with an AB type of potential. These models, pioneered by Lau and Dill (47) and extensively characterized by others (see refs. 5, 6, and 48 for relevant citations), capture many of the general features of real proteins. The particular lattice polymers used in this manuscript are 27 monomers long with compact $3 \times 3 \times 3$ native states. Details of this model and its behavior can be found in refs. 4, 49, and 50.

The landscape theory has shown that energetic frustration can have profound effects on folding dynamics and on the nature of the transition state ensemble. To study these effects, we perform the calculations for the three different sequences shown in Fig. 2. Although these sequences have the same native structure and native energy, they have different levels of frustration, which is reflected in their respective (T_f/T_g) ratios of folding to glass transition temperatures. The first sequence is a Gō-like sequence, which is nearly maximally unfrustrated for a given native structure. In this model, only native interactions are attractive; all others are simply excluded volume interactions. $T_f/T_g > 2$ in this system. The second system has a three-monomer type AB

**For real proteins, measurements of the rate of intrachain diffusion have put a lower bound on the rate prefactor of ≈10 μs (43), although its exact value is unknown.

BIOPHYSICS

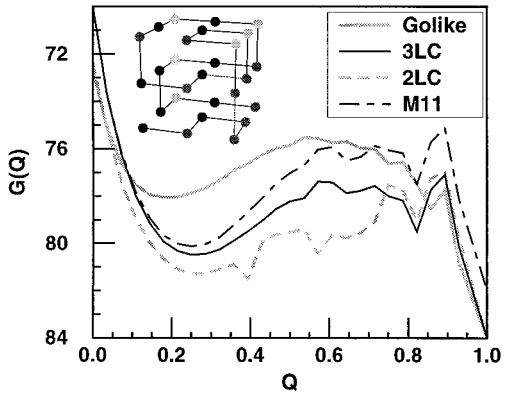

Fig. 2. The free energy profile $G(Q)$ shown at T_f for the three models studied in this paper and one of their mutants. The three models have the same structure but different potentials: a Gō-like potential where every bead is a different type, and only beads adjacent in the native structure are attractive; a 3LC sequence **ABABBBCBACBABABACACBACAACAB,** where contacts of identical type have energy of −3 units and contacts of different type have energy of −1 units (arbitrary energy units are used to have a folding temperature and glass temperature of order unity); and a 2LC sequence **ABABBBB-BABBABABAAABBAAAAAAB.** The mutant (M11) is from the 3LC system, in which the native interaction between beads 6 and 13 (numbered starting from the lower left corner of the structure shown in the inset) is reduced in energy from −3 to −1. All mutants show two-state behavior. (*Inset*) The native $3 \times 3 \times 3$ structure for all of the different wild-type and mutant sequences studied in this paper.

potential: i.e., there are three types of monomers (A, B, and C), identical monomer types are strongly attractive (energy = −3), and nonidentical monomer types are weakly attractive (energy = −1). We refer to this system as the three-letter code system (3LC). The introduction of attractive non-native interactions increases the frustration. $T_f/T_g \approx 1.6$. Of the three systems, this most resembles the frustration level and configurational entropy of small, fast-folding proteins (51). The third system has a two-monomer type AB potential: i.e., the same potential as before, but with only two monomer types. We refer to this as the two-letter code system (2LC). Additional frustration created by the reduction in the number of monomer types reduces the T_f/T_g ratio to ≈1.3. We observe that the amount of frustration in the system can drastically change the nature of folding and the effectiveness of using simple reaction coordinates and equilibrium sampling to probe folding transition state structure.

We begin by determining Φ values through the standard experimental procedure: namely, by measuring the change in the stability and folding rate created by different mutations. The Φ values for these mutants are computed by using Eq. **1**. We then determine the Φ values by using the landscape approach (Eq. **4**). In this method, the free energy as a function of a folding order parameter—in this instance, Q, the fraction of native ("tertiary") contacts formed in a state—is used (rather than changes in the folding rate) to determined the change in activation free energy of different mutations.

Equilibrium constants are not determined by exponential fitting of the relaxation in Q of a population versus time and the use of a two-state approximation as in the experimental manner. Instead, a thermal ensemble at each value of Q is computed from a long trajectory, and changes in free energy are computed by using the well known free energy perturbation method (52), which gives the change in free energy of a system as

$$\Delta F = -RT\langle e^{-\Delta H/RT}\rangle_{H_0}, \qquad [5]$$

where $\langle \ldots \rangle_{H_0}$ is a thermal average of the unperturbed (wild-type) system (indexed by Q). Every equilibrium quantity we need is then rapidly computed from a single wild-type simulation. Representative examples of wild-type and perturbed free energy profiles are shown in Fig. 2.

We produce mutants by weakening one of the native interactions (from an energy of −3 to −1), so the Φ value is measuring the amount of formation of the weakened bond. One could refer to these Φ values as "bond" Φ values instead of "residue" Φ values like those in typical protein experiments. Bond mutations are preferred over residue mutations because they provide a more detailed structural picture of the transition state ensemble. Similar values can be computed by measuring the rate and stability changes of double and single mutants and subtracting the one-body, single-mutant changes from the full double-mutant changes (53). Because residue Φ values are (to lowest order in perturbation theory) an average over bond Φ values, the range of variation of residue Φ values with position is much smaller than the range of variation of bond Φ values with position; consequently, bond Φ values are more useful for comparing various methods of calculation, even though these methods can be used to calculate residue Φ values with equal validity under identical conditions.

Results and Discussion

There are 28 possible bond mutants in our models. The Φ values for these contacts are computed from folding simulations by using Eq. **1** and from equilibrium sampling and free energy perturbation by using Eq. **4**. The values computed via the two methods are compared in Fig. 3. To perform the free energy perturbation calculation, we assume that the folding transition state structures are the ensemble of thermally occupied microstates that have a value of Q with the highest free energy between $5/28$ and $23/28$. (The upper cutoff is used because lattice artifacts can produce sharp free energy peaks above $Q = 23/28$ that are not actual folding barriers.) Taking the highest free energy point rather than a fixed point allows the transition state to shift along the Q coordinate. To compute changes in stability, we take all states with $Q < 16/28$ as the unfolded ensemble and $Q = 1$ as the native state. The comparison of Φ values from the two methods is shown for the different systems with differing levels of frustration.

How well the Φ values computed via the two methods agree depends strongly on the degree of frustration in the system. The sequence with the least amount of frustration (Gō-like) has the best agreement between the two Φ values (normalized correlation coefficient 0.86 and a slope close to 1). The 3LC, a sequence with greater frustration ($T_f/T_g \approx 1.6$), shows a slightly larger variation between the two sets of Φ values (normalized correlation 0.84) and a slope around 1.6; namely, the Φ values computed from changes in the free energy barrier seen with Q are consistently overestimated by about a factor of 1.6. The 2LC, the sequence with the greatest amount of frustration ($T_f/T_g \approx$ 1.3), shows no agreement between the two sets of Φ values (normalized correlation −0.49). Clearly, the effective use of a small number of global order parameters as reaction coordinates depends critically on the degree of frustration. This degree of frustration in real proteins will clearly determine how effective the use of simple order parameters is in interpreting real data or studying more detailed protein models.

Different contacts (bonds) can have different Φ values not only because of energetic heterogeneity but also because of topological factors that arise from a combination of the polymeric nature of the chain and the structure of the native state. In Fig. 4, we compare the Φ values computed for the same structure with two different potentials (Gō-like and 3LC). The strong correlation between these two different sequences indicates the central role of topology in determining the folding mechanism and Φ values as long as the energetic frustration is not too large. However, already for the systems with energetic

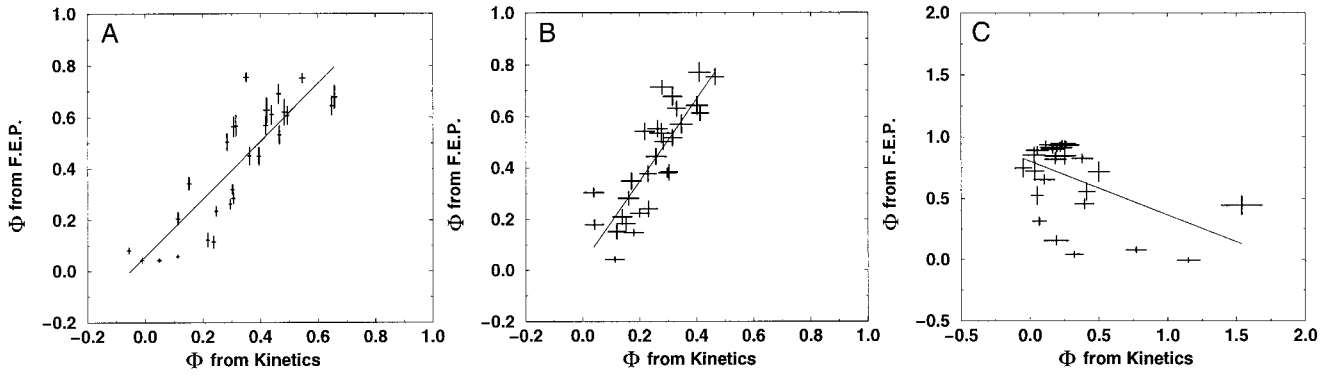

Fig. 3. A comparison of the kinetic and free energy perturbation methods for inferring folding transition state structure. Each panel shows the comparison for a different sequence (and Hamiltonian): *A* is the Gō-like sequence, *B* is the 3LC sequence, and *C* is the 2LC sequence. The Gō-like sequence has the least amount of energetic frustration, and *C* the most. The native structure and potential for these sequences are shown in Fig. 2. Mutants are made by weakening specific nonbonded interactions between beads that are adjacent in the native structure. The Φ values for these contacts are then computed by the standard experimental procedure (ordinate) and by a free energy perturbation technique (abscissa). Agreement is good (normalized correlations 0.86 and 0.84) for the models in *A* and *B*, which have energetic frustration less than or comparable to small, fast-folding globular proteins. Because the Gō sequence in *A* has no energetic frustration, the heterogeneity of the Φ values is mostly determined by topological factors attributable to a combination of the polymeric nature of the chain and the structure of the native state. More frustrated sequences, such as in *C*, show no agreement (normalized correlation −0.49) between the two methods and out-of-range Φ values, which suggests that the assumption of a Kramer's type of rate with a fixed rate prefactor is not valid. In the experimental method (ordinate), the folding rate of the wild type (k_{wt}) and the mutant (k_{mut}) as well as the change in native stability under mutation are measured and used to compute a Φ value as $\Phi \equiv -RT\, ln(k_{mut}/k_{wt})/\Delta\Delta G^0$. These Φ values should be similar to the measure of $\Delta\Delta G^0/\Delta\Delta G^{\ddagger}$, the ratio of the change in the folding activation free energy to the change in native stability, when the assumption that folding follows a Kramer's type of equation with a fixed rate prefactor is valid. In the free energy perturbation method, we determine the free energy as a function of a folding reaction coordinate—in this instance, *Q*, the fraction of formed native nonbonded interactions. The barrier height is defined as the difference in free energy between the highest free energy point along *Q* between 5/28 and 23/28 and the free energy of states, with $Q < 16/28$. The Φ value is then computed directly from $\Delta\Delta G^0/\Delta\Delta G^{\ddagger}$ by taking $Q < 16/28$ as the unfolded conformations. For both methods, the unfolded state is defined as all conformations with $Q < 16/28$. Error bars show 68% confidence limits calculated from 1,000 bootstrapping simulations.

frustration comparable to the 3LC, the impact of energetic effects is noticeable in the sensitivity of the choice of barrier. Recent theoretical (54, 55) and experimental (56, 57) work supports this idea that much of the transition state ensemble structure in real proteins is determined by the topology of the particular protein under consideration. This agreement is another indication that the amount of energetic frustration in real single-domain fast-folding proteins is similar to or less than the 3LC sequence. A comparison of the Φ values of the Gō-like and 2LC model shows no significant correlation (data not shown).

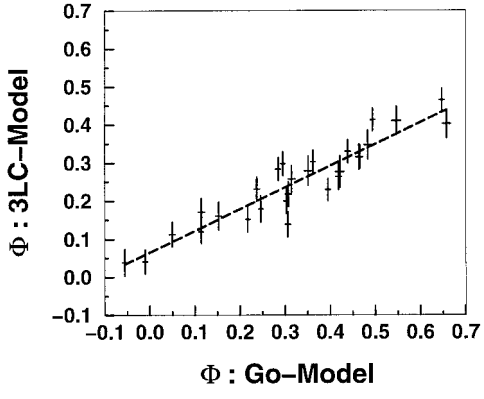

Fig. 4. A comparison of the 28 bond Φ values—produced by mutating a single native interaction, decreasing its energy from −3 to −1—from the model with a Gō-like potential and a 3LC potential. (The sequence and native structure are shown in Fig. 2.) The agreement shows that, for sequences with reduced frustration, the native structure (topology) has a large role in determining the Φ values. Details of the potential interactions may not be as critical. Φ values are determined here from the rates by using Eq. **1**.

Fig. 3 also shows that several Φ values for the 2LC sequence computed by the experimental procedure lie well outside the range of 0–1—values for which a simple explanation of Φ values in terms of structural changes becomes problematic. The sudden appearance of these large Φ values indicates that the conventional interpretation is no longer valid for systems with this higher level of frustration. In this regime, the experimental method of using rate changes to infer the transition state structure is not significantly more accurate than using a free energy function of one- or two-order parameters. This suggests that, in large, slow folding proteins—those that might contain a significantly higher level of frustration than in the smaller single domain proteins—Φ values may not be reliable probes of transition state structures as in the smaller, faster folding proteins.

Why does the agreement break down with increasing frustration? Clearly, one or more of the assumptions made are invalid. Either folding cannot be described via a Kramer's-like rate equation in which the prefactor is identical for both mutant and wild-type proteins, or the transition state ensemble cannot accurately be approximated by using simple reaction coordinates such as *Q* (that are effective in less frustrated models). We have not precisely quantified the level of frustration at which these two assumptions become invalid. The appearance of Φ values outside the range of 0 and 1 and the sudden loss of correlation between the Φ values of the Gō-like and 2LC sequences suggests that the entire description of folding in terms of diffusion along a macroscopic coordinate becomes invalid for sequences with frustration levels comparable to the 2LC.

Although the Φ values computed from changes in folding rates correlate well with Φ values computed from the free energy surface, it is clear that the latter values are generally overestimated. This is visible in the best fit line of the 3LC, where the slope is ≈1.6 instead of near unity (Fig. 3*B*). This discrepancy can

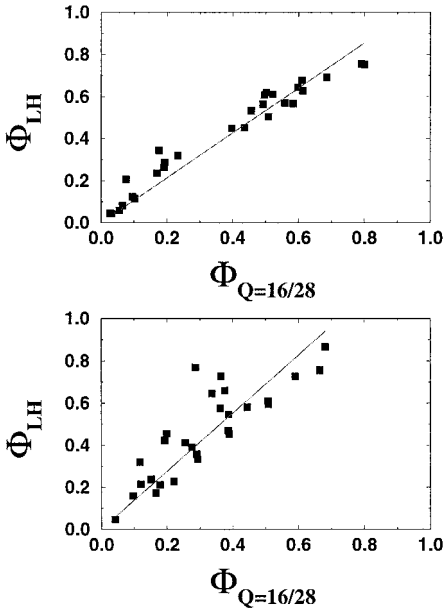

Fig. 5. (*Upper*) A comparison of the Φ values computed for the Gō-like model from the free energy perturbation formula (based on Eq. **4**) by using two different formulas for estimating the activation free energy $\Delta\Delta G^{\ddagger}$. The abscissa shows the Φ values computed by assuming a fixed transition state location at $Q = 16/28$; the ordinate shows the Φ values computed by assuming that the activation free energy is equal to the variation between the unfolded free energy minimum and the maximum barrier point. The solid line is a least squares fit constrained to pass through the origin with a slope of 1.07. (*Lower*) The same plot but shown for the 3LC model. The solid line has a slope of 1.38. These two plots demonstrate that the less frustrated models have less sensitivity of their Φ values to the assumed position of the barrier. The overestimation of the Φ values computed from free energy perturbation in the 3LC model is apparent from the lower panel. Because the top of the free energy barrier is very broad, variations in the reconfigurational diffusion coefficient with Q and the existence of fundamental motions that allow jumps of several units in Q can shift the actual barrier location. In this instance, the actual location is close to $Q = 16/28$, which is at the lower Q position of the barrier.

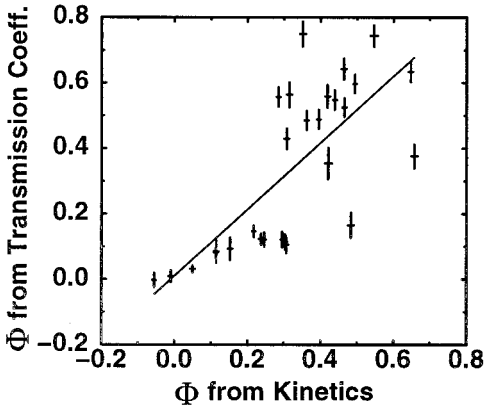

Fig. 6. A comparison of the Φ values computed by measuring changes in the folding rates and by averaging over the thermally weighted population of conformations with transmission coefficients between 0.4 and 0.6. Comparison is done for the minimally frustrated Gō sequence and structure shown in Fig. 2. The two results have a normalized correlation coefficient of 0.73—agreement is thus of comparable quality to using our method with a single order parameter Q but at a much greater computational cost. Ten-thousand states with Q values in the range 6/28 to 27/28 were sampled in equilibrium at ≈100,000 step intervals. Three-hundred and ninety states from this sample had transmission coefficients between 0.4 and 0.6, defined as the fraction of folding simulations that, when started in a given conformation, find the native conformation before any conformation with $Q \leq 7$ (as determined from 100 independent folding simulations). This set of structures was then used as a putative transition state ensemble for computing Φ values via the free energy perturbation equation.

be traced to inaccuracies in the method of computing the change in the free energy barrier height $\Delta\Delta G^{\ddagger}$: i.e., uncertainties in the choice of barrier location. The actual barrier location appears to be closer to $Q = 16/28$, as can be seen in Fig. 5—using a fixed location of $Q = 16/28$ to compute $\Delta\Delta G^{\ddagger}$ yields slightly more dispersion in the Φ values but a better overall slope. The greater sensitivity of the 3LC to the choice of barrier location is attributable to the larger number of traps in the barrier region, which are off-pathway as compared with the less frustrated Gō-like model. Because the free energy barriers for folding are quite broad, the inclusion of these off-pathway traps and the existence of kinetic effects such as moves that make several steps in Q and a positionally dependent configurational diffusion coefficient can shift the average barrier location by an appreciable amount—from the free energy maximum toward the lower Q end of the barrier near $Q = 16/28$.

Certainly, there are other methods that can be used to determine transition state structures. Most of these methods eschew the use of simple reaction coordinates like Q in favor of more complicated coordinates, which in principle could better identify the transition state ensemble. For example, in refs. 29 and 30, a transmission coefficient is used to identify putative transition state structures. That is, a transition state structure is identified as one for which approximately half of simulations begun in that state reach the folded state before unfolding. These approaches are much more computationally intensive, and the proposed reaction coordinates cannot be associated with any experimentally accessible measurement. More importantly, these coordinates do not in practice provide more quantitative information than simple coordinates like Q, which can be rapidly computed for any structure. As we have discussed, no single coordinate will determine an exact transition state ensemble, but for sequences with reduced frustration, many coordinates provide reasonable answers. To demonstrate this assertion, we have shown that Φ values computed with our method are similar to the ones obtained with "transmission coefficient" reaction coordinates for the minimally frustrated Gō sequence in Fig. 6. It should be noted that this manuscript uses a slightly different method for locating transition states than ref. 29: we generate a proper thermal ensemble of states with transmission coefficients near 0.5 rather than selecting one transition state from each of several folding simulations.

Conclusion

Often it is difficult (or impossible) to use a fully molecular model to interpret experimental measurements of protein properties. For example, to understand the effects that mutations have on the folding rate, and to use this information to determine the structural details of the transition state ensemble, we cannot turn to atomically detailed models. Thus, there is a need for a theoretical/computational framework that is able to determine this structural transition state ensemble without the need for fully kinetic simulations of these models. In addition, such a framework is necessary to relate the effects of local mutations observed experimentally to structural details of the protein. In this paper, we developed a landscape approach to computationally determine these structural details (computing Φ values) without the need for kinetic information, and we tested the assumptions behind this approach on a minimalist protein model.

We showed that the connection between the folding rate (kinetics) and the free energy barrier (thermodynamics) depends strongly on the degree of energetic frustration in the

protein. For good folders with sufficiently weak frustration, this connection can be expressed via a Kramer's-like equation, and the folding rates can be clearly described as a diffusive process on a low-dimensional free energy surface constructed by using simple structural measures of native state similarity. For these proteins, interpretation of the experimentally measured Φ values as changes in free energy differences of a simple transition state ensemble is accurate and useful. We also notice that, for minimally frustrated systems, geometric reaction coordinates like the percentage of native contacts Q work effectually and give similar results to those of more complex coordinates. This indicates that many features of the system are insensitive to the microscopic details. Because real proteins need to fold reliably and be robust to slight changes in environment and sequence, their frustration level is usually low enough for this approach to work.

However, the situation changes dramatically when the level of frustration increases. The equivalence of the Φ values determined from simple geometric reaction coordinates and from changes in rates does not hold. In this increased frustration regime, the folding rates are controlled by long lived traps. The folding can no longer be described as diffusion over a free energy barrier observed by using simple parameters measuring structural overlap with the native state. No general theoretical framework will be able to capture the features critical to folding. Instead, many microscopic details are relevant for understanding the folding mechanism and must be modeled precisely, and the conventional experimental interpretation of Φ values in terms of geometrical features breaks down.

Work at the University of California at San Diego was supported by the National Science Foundation (Grant MCB-9603839) and by the molecular biophysics training grant program (NIH T32 GN08326) for H.N. Work at The Rockefeller University was funded by the National Science Foundation (Grant DMR-7932803) and the Alfred P. Sloan Foundation.

1. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
2. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8721–8725.
3. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins Struct. Funct. Genet.* **21,** 167–195.
4. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996) *J. Chem. Phys.* **104,** 5860–5868.
5. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem.* **48,** 545–600.
6. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4,** 10–19.
7. Nymeyer, H., Garcia, A. E. & Onuchic, J. N. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5921–5928.
8. Englander, S. W. & Mayne, L. (1992) *Annu. Rev. Biophys. Biomol. Struct.* **21,** 243–265.
9. Kim, P. S. & Baldwin, R. L. (1990) *Annu. Rev. Biochem.* **59,** 631–660.
10. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996) *Fold. Des.* **1,** 441–450.
11. Fersht, A. R. (1997) *Curr. Opin. Struct. Biol.* **7,** 3–9.
12. Eaton, W. A., Munoz, V., Thompson, P., Chan, C. K. & Hofrichter, J. (1997) *Curr. Opin. Struct. Biol.* **7,** 10–14.
13. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. (1997) *Nat. Struct. Biol.* **4,** 305–310.
14. Riddle, D. S., Santiago, J. V., Bray, S. T., Doshi, N., Grantcharova, V., Yi, Q. & Baker, D. (1997) *Nat. Struct. Biol.* **4,** 805–809.
15. Elove, G. A., Bhuyan, A. K. & Roder, H. (1994) *Biochemistry* **33,** 6925–6935.
16. Jennings, P. & Wright, P. (1993) *Science* **262,** 892–896.
17. Plaxco, K. W. & Dobson, C. M. (1996) *Curr. Opin. Struct. Biol.* **6,** 630–636.
18. Sosnick, T. R., Mayne, L. & Englander, S. W. (1996) *Proteins Struct. Funct. Genet.* **24,** 413–426.
19. Ballew, R. M., Sabelko, J. & Gruebele, M. (1996) *Nat. Struct. Biol.* **3,** 923–926.
20. Phillips, C. M., Mizutani, Y. & Hochstrasser R. M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7292–7296.
21. Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H. & Dyer, R. B. (1996) *Biochemistry* **35,** 691–697.
22. Matthews, C. R. (1993) *Annu. Rev. Biochem.* **62,** 632–683.
23. Cordes, M. H. J., Davidson, A. R. & Sauer, R. T. (1996) *Curr. Opin. Struct. Biol.* **6,** 3–10.
24. Raschke, T. M. & Marqusee, S. (1997) *Nat. Struct. Biol.* **4,** 298–304.
25. Lin, L., Pinker, R. J., Forde, K., Rose, G. D. & Kallenbach, N. R. (1994) *Nat. Struct. Biol.* **1,** 447–452.
26. Pascher, T., Chesick, J. P., Winkler, J. R. & Gray, H. B. (1996) *Science* **271,** 1558–1560.
27. Villegas, V., Martinez, J. C., Aviles, F. X. & Serrano, L. (1998) *J. Mol. Biol.* **283,** 1027–1036.
28. Socci, N. D., Nymeyer, H. & Onuchic, J. N. (1997) *Phys. D* **107,** 366–382.
29. Pande, V. S. & Rokhsar, D. S. (1998) *Proc. Natl. Acad. Sci. USA* **96,** 1273–1278.
30. Du, R., Pande, V. S., Grosberg, A. Yu. & Tanaka, T. & Shakhnovich, E. S. (1998) *J. Chem. Phys.* **108,** 334–350.
31. Li, Z. & Scheraga, H. A. (1984) *Proc. Natl. Acad. Sci. USA* **84,** 6611–6615.
32. Boczko, E. M. & Brooks, C. L., III (1995) *Science* **269,** 393–396.
33. Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1996) *J. Mol. Biol.* **257,** 430–440.
34. Guo, Z., Brooks, C. L., III & Boczko, E. M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10161–10166.
35. Sheinerman, F. B. & Brooks, C. L., III (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1562–1567.
36. Vakhter, B. & Berry, R. S. (1999) *J. Chem. Phys.* **110,** 2195–2201.
37. Klimov, D. K. & Thirumalai, D. (1998) *J. Mol. Biol.* **282,** 471–492.
38. Elber, R., Meller, J. & Olender, R. (1999) *J. Phys. Chem. B* **103,** 899–911.
39. Dellago, C., Bolhuis, P. G. & Chandler, D. (1999) *J. Chem. Phys.* **110,** 6617–6625.
40. Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 8473–8478.
41. Lazaridis, T. & Karplus, M. (1997) *Science* **278,** 1928–1931.
42. Matouschek, A, Kellis, J. T., Jr., Serrano, L. & Fersht, A. R. (1989) *Nature (London)* **340,** 122–126.
43. Hagen, S. J., Hofrichter, J., Szabo, A. & Eaton, W. A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 11615–11617.
44. Mines, G. A., Pascher, T., Lee, S. C., Winkler, J. R. & Gray, H. B. (1996) *Chem. Biol.* **3,** 491–497.
45. Oliveberg, M., Tan, Y. J., Silow, M. & Fersht, A. R. (1998) *J. Mol. Biol.* **277,** 933–943.
46. Matouschek, A, Otzen, D. K., Itzhaki, L. S., Jackson, S. E & Fersht, A. R. (1995) *Biochemistry* **34,** 13656–13662.
47. Lau, D. F. & Dill, K. A. (1989) *Macromolecules* **22,** 3986–3997.
48. Shakhnovich, E. I (1998) *Fold. Des.* **3,** R45–R58.
49. Socci, N. D. & Onuchic, J. N. (1995) *J. Chem. Phys.* **103,** 4732–4744.
50. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1998) *Proteins Struct. Funct. Genet.* **32,** 136–158.
51. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3626–3630.
52. Zwanzig, R. (1954) *J. Chem. Phys.* **22,** 1420–1426.
53. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224,** 771–782.
54. Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D. (2000) *Adv. Protein Chem.* **53,** 87–152.
55. Shea, J. E., Onuchic, J. N. & Brooks, C. L., III (1999) *Proc. Natl. Acad. Sci. USA* **96,** 12512–12517.
56. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998) *Nat. Struct. Biol.* **5,** 714–720.
57. Martinez, J. C., Pisabarro, M. T. & Serrano, L (1998) *Nat. Struct. Biol.* **5,** 721–729.