P versus *Q*: Structural reaction coordinates capture protein folding on smooth landscapes

Samuel S. Cho*†, Yaakov Levy*‡, and Peter G. Wolynes*†‡§

*Center for Theoretical Biological Physics and Departments of †Chemistry and Biochemistry and ‡Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Contributed by Peter G. Wolynes, November 19, 2005

Minding your p's and q's has become as important to protein-folding theorists as it is for those being instructed in the rules of etiquette. To assess the quality of structural reaction coordinates in predicting the transition-state ensemble (TSE) of protein folding, we benchmarked the accuracy of four structural reaction coordinates against the kinetic measure P_{fold}, whose value of 0.50 defines the stochastic separatrix for a two-state folding mechanism. For two proteins that fold by a simple two-state mechanism, c-src SH3 and CI-2, the Φ-values of the TSEs predicted by native topology-based reaction coordinates (including *Q*, the fraction of native contacts) are almost identical to those of the TSE based on P_{fold}, with correlation coefficients of >0.90. For proteins with complex folding mechanisms that have especially broad, asymmetrical free energy barriers such as the designed 3-ankyrin repeating protein (3ANK) or proteins with distinct intermediates such as cyanovirin-N (CV-N), we show that the ensemble of structures with $P_{fold} = 0.50$ generally does not include the chemically relevant transition states. This weakness of P_{fold} limits its usefulness in protein folding studies. For such systems, elucidating the essential features of folding mechanisms requires using multiple reaction coordinates, although the number is still rather small. At the same time, for simple folding mechanisms, there is no indication of superiority for P_{fold} over structurally chosen and thermodynamically relevant reaction coordinates that correctly measure the degree of nativeness.

energy landscape theory | minimal frustration | transition-state ensemble | P_{fold} | intermediates

Proteins fold by navigating through an energy landscape that is globally funneled toward a structurally defined native state (1). The funneled nature of energy landscapes explains why protein topology so strongly determines folding and binding kinetics (Fig. 1). Much discussion has gone on among theorists regarding reaction coordinates for complex processes such as folding. Surprisingly, much of this discussion fails to recognize that reaction coordinates are ultimately just "crutches" for calculating experimental observables. In other words, to define a reaction coordinate a theorist is required also to specify a reaction rate theory to use with this coordinate to predict rates. Transition-state theory (TST) is the simplest theory of predicting reaction rates for chemical reactions dating to Wigner (2). In this well known theory, two stable states (i.e., reactant and product) are separated by an ambiguous, unstable region of phase space called the transition state. TST postulates that when a reactant crosses the transition state once, the molecule continues to the product state without recrossing (Fig. 2*a*). Later recrossing events often can be considered negligible as reflected by the robustness of the TST for predicting rates in gas-phase kinetics. In such situations, the transition-state ensemble (TSE) corresponds to the free energy barrier peak for an appropriately chosen reaction coordinate. For natural proteins, the unfolded and folded states also are separated by at least one bottleneck or transition state. In protein folding processes, however, the recrossing events are nontrivial because frictional effects, arising from the solvent collisions, dihedral angle barriers, and forming adventitious nonnative contacts, can exert forces on the reaction coordinates that alter the direction of motion (3). A protein crosses the transition state multiple times before reaching the folded state (Fig. 2*b*), as was analytically predicted by Bryngelson and Wolynes (4) and later observed in simulations (5). The TST, therefore, overestimates the rate coefficient, which only counts the number of forward trajectories, neglecting any recrossing events. Frictional effects grow as the glass transition from landscape ruggedness is approached. When friction is large, the transition state generally does not correspond to the peak of the free-energy barrier (3–5). There is much evidence, however, that real proteins are far from this glassy limit. In the simplest case, folding kinetics can be interpreted by using a single transition state that separates the unfolded and folded states. Protein engineering allows the structures in the TSE for these systems to be probed (6). In a strictly two-state situation, the TSE would be reasonably defined by a single stochastic separatrix and corresponds to that set of structures having an equal probability of first completing the folding process before unfolding to a completely denatured state (7).

Motivated by this observation, the quantity P_{fold} has been defined. It is the probability that a given structure will reach a decidedly folded state before reaching the unfolded state (7). For a protein that undergoes a two-state folding mechanism, the P_{fold} of the TSE members should be 0.50. To compute P_{fold} for a given structure, one starts several independent trajectories at the folding temperature (T_f) from that structure until the protein reaches either the unfolded or the folded state, and then one calculates the appropriate average. Although the concept of P_{fold} is rather simple, unfortunately, it is computationally intensive to evaluate. To be statistically meaningful, tens to hundreds of simulations starting from each conformation are needed. Further, the simulation time required for a single trajectory, starting from a candidate transition-state conformation, to commit to unfolding or folding can be >100 ns when using all-atom simulations with an empirical force field (8). The parallelizable nature of the problem has motivated some to use distributed computing approaches to carry out such computations (9). Even in the most rigorous studies carried out so far, however, the computation of P_{fold} is limited to an exceedingly small set of conformations. Computing P_{fold} also requires the precise knowledge of T_f because P_{fold} is highly sensitive to temperature (see Fig. 7, which is published as supporting information on the PNAS web site). Determining the value of T_f from simulations, however, is not always possible. In all-atom simulations of proteins it has seldom been possible, if ever, to observe transitions between the folded and unfolded states, even for the simplest proteins. Thus, T_f is uncertain for these models. Although approximating the folding temperature from experiments for a well-studied protein (8, 10, 11) may be acceptable, an arbitrary choice of temperature (12) is clearly inadvisable because the

Conflict of interest statement: No conflicts declared.

Abbreviations: TST, transition-state theory; TSE, transition-state ensemble.

§To whom correspondence should be addressed. E-mail: pwolynes@chem.ucsd.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0509768103

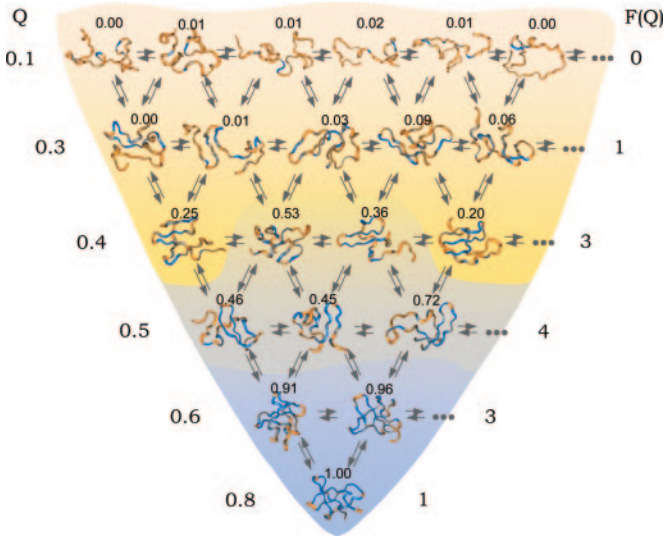

Fig. 1. A schematic representation of the ordering of protein structures as they descend a funneled energy landscape. The structures of c-src SH3 with varying Q are colored according to the degree of local structural order, Q_i, in residue i, ranging from low Q_i (orange) to high Q_i (blue). The P_{fold} of each structure is denoted above each protein structure. The regions of the energy landscape corresponding to the unfolded, the folded, and the transition states based on Q are colored as yellow, blue, and gray regions, respectively. The free energy with respect to Q [$F(Q)$] is also shown.

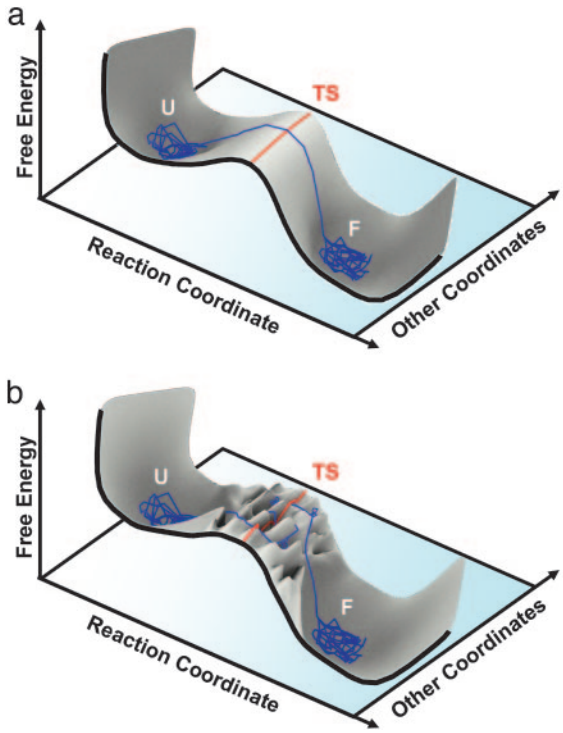

Fig. 2. Schematics depicting two possible trajectories of protein folding. (*a*) A single crossing of the transition state as predicted by TST. (*b*) Multiple crossings of the transition state in the limiting case of high friction due to ruggedness.

slightest deviation from T_f can significantly misplace the TSE. A more troubling aspect of P_{fold}, beyond the practical burdens of computing it, is that P_{fold} does not have any direct relationship to the observables measured in experiments or used to perturb folding thermodynamically. One can only fantasize about the improvements in single-molecule technologies needed to experimentally measure P_{fold} for a given conformation because rigorously such a protocol would entail the exact replication of the protein conformation for multiple trials (13). Thus, although P_{fold} identifies members of the TSE in a strict sense, the severe practical drawbacks of P_{fold} would demand finding reliable alternatives without these handicaps. Also, the appropriateness of P_{fold} for proteins with complex mechanisms (i.e., with intermediates) has not been quantified until now, and as we shall see presents its own difficulties.

Fortunately, for natural proteins it is possible to replace the kinetically defined P_{fold} with one or more reasonably accurate structurally defined reaction coordinates that accurately predict and characterize the TSE. A key idea of energy landscape theory is that this should be possible whenever the energy landscape is not very frustrated. One study illustrating this was already carried out by Onuchic and coworkers (14), which showed that thermodynamic reaction coordinates predict the measurable structural features of a TSE well when the landscape is strongly funneled by comparing directly computed Φ-values with those inferred from the TSE (14). In keeping with Bryngelson and Wolynes's theory (4), these results show that that when the landscape is glassy or frustrated, thermodynamic coordinates fail to describe the structural ensemble as measured by Φ-values (14). Thus, general arguments and these specific results have encouraged the use of calculating Φ-values based on unfrustrated models (15, 16). Simulations based on unfrustrated landscapes using native-structure-based reaction coordinates also predict many qualitative experimental observations of protein folding and binding (17–19). The predicted folding rates of many small proteins agree well with experimental observations (20, 21), and the Φ-values usually agree with experimental values (19, 22). Despite these successes, and ignoring the capability of rate theory to use a variety of reaction coordinates so long as the results are properly corrected by Kramers-like transmission factors (4), some researchers have vigorously questioned whether structural reaction coordinates like Q, the fraction of native contacts, are appropriate for describing the TSE even on unfrustrated landscapes (7, 23). It has been argued *a priori* that structural coordinates like Q will fail to identify the transition state, even for funneled landscapes (24, 25). Some maintain that P_{fold} is the only reliable reaction coordinate for real proteins (26). This assertion contradicts other studies showing that Q gives acceptable results as a reaction coordinate for model proteins (6, 13, 14). Studies on all-atom models show a clear correlation between P_{fold} and Q (11). Q is not the only possible structural coordinate. Shoemaker *et al.* (27) showed that reaction coordinates measuring only a handful of contact areas function equally well, if they are chosen appropriately post hoc. Alternative structural quantities have also been used as reaction coordinates. $\langle L \rangle$, the mean shortest path length, has been reported to characterize the TSE better than other order parameters (28). In general, the fact that the structures in TSEs as defined by P_{fold} are found to have high structural similarity to each other in at least one case (29) indicates again that some geometric measures should be sufficient for structurally describing the TSE.

To address the issues highlighted above, we examine two experimentally well studied proteins that, in the laboratory, fold with a two-state folding mechanism (c-src SH3 and CI-2) (17, 30, 31) to quantify rigorously the accuracy of native-structure-based reaction coordinates in describing the TSE as probed by Φ-value analysis (32). The TSE structures obtained by using several different reaction coordinates are compared with the one based on P_{fold}. All these ensembles are found to be essentially the same in structure. We then extend our study to a more complex system, where the concept of P_{fold} itself is suspect. In a system that is

thermodynamically two-state but with a broad, asymmetrical free energy barrier (3ANK), the folding mechanism was found to actually involve two sets of competing folding routes. One of the transition states was indeed not detected by using P_{fold}. Finally, we study a protein having a clear three-state folding mechanism (CV-N). In this case, P_{fold} fails to detect either of the appropriate transition states.

Results and Discussion

Although the folding of proteins is complex with numerous degrees of freedom, possibly involving multiple transition states and competing routes, for smooth landscapes of minimally frustrated natural proteins, the energy landscape theory suggests only a few collective coordinates should be necessary to describe the kinetics. In some cases, even a single reaction coordinate may suffice.

Structural Reaction Coordinates Identify and Describe the TSE as Well as P_{fold}. For two-state proteins, if friction effects are small, the peak of the free-energy barrier, as described by the structural reaction coordinate, must reasonably correspond to the TSE found by using P_{fold}. Structures in the TSE as predicted by the structural reaction coordinate should have approximately equal probabilities to fold or unfold. To evaluate whether the reaction coordinate Q in this sense reliably predicts the TSE, we simulated the two-state folders c-src SH3 and CI-2. For these proteins, we also calculated the P_{fold} of structures over a range of Q between the unfolded and folded states to determine which values of Q correspond to the putative TSE, i.e., $P_{fold} = 0.50 \pm 0.10$. Those structures whose Q is $1k_BT$ from the barrier top of the free energy profile are considered to form the "predicted TSE." A free energy profile with respect to Q and its corresponding P_{fold} (Fig. 3 *a* and *b*) shows that for both proteins, the peak of the barrier, as defined by Q, corresponds to $P_{fold} = 0.50$. That is, the TSE according to Q agrees reasonably well with the TSE according to P_{fold} (Fig. 3 *a* and *b*). Although Q, on average, is able to identify the TSE, there exists some structures whose P_{fold} lie outside of the range $P_{fold} = 0.50 \pm 0.10$, even though Q predicts them to be members of the TSE. To assess whether these and similar outliers significantly taint the predicted TSE, we compared the two TSEs using the Kolmogorov–Smirnov test (33), a well established statistical test that determines whether two overlaps distributions can be taken as subsets chosen from the same underlying distribution. According to this test, the TSEs derived by P_{fold} and Q are equivalent (see Fig. 8 *a* and *b*, which is published as supporting information on the PNAS web site). We see then that in this exhaustive survey one cannot distinguish these ensembles in terms of pair-structural patterns.

We now compute the experimentally accessible quantities, Φ-values, according to the four chosen reaction coordinates, Q, Q_S (similarity of native distances of natively contacting residues), $\langle L \rangle$, and R_g, and compare them with the Φ-values of the TSE as defined by P_{fold}. To make a quantitative comparison between the Φ-values determined by P_{fold} and those predicted by the structural coordinates, we used the linear correlation coefficient, r, and the slope of the best-fit line, m (Fig. 3 *c* and *d*). For both proteins, the Φ-values as determined by the reaction coordinates Q, Q_S, and $\langle L \rangle$ agree strikingly well with those of the TSE described by P_{fold} with correlation coefficients ≈0.90 and 0.95 for c-src SH3 and CI-2, respectively. The slopes of the correlations are ≈0.70 and 0.80 for c-src SH3 and CI-2, respectively, indicating that the Φ-values are slightly underestimated using these structural reaction coordinates as compared with P_{fold}. Evidently, there exist only minor differences between the TSE as determined by P_{fold} or using any of the reaction coordinates studied that are based on the protein native topology (i.e., Q, Q_S, and $\langle L \rangle$). R_g, however, generally grossly underestimates the Φ-values. $\langle L \rangle$ turns out to be at best comparable with Q and Q_S when describing the TSE via Φ-value analysis, contrary to a previous suggestion (28). The difference between Q and Q_S is modest, as is reflected in their equivalent characterizations of the TSE.

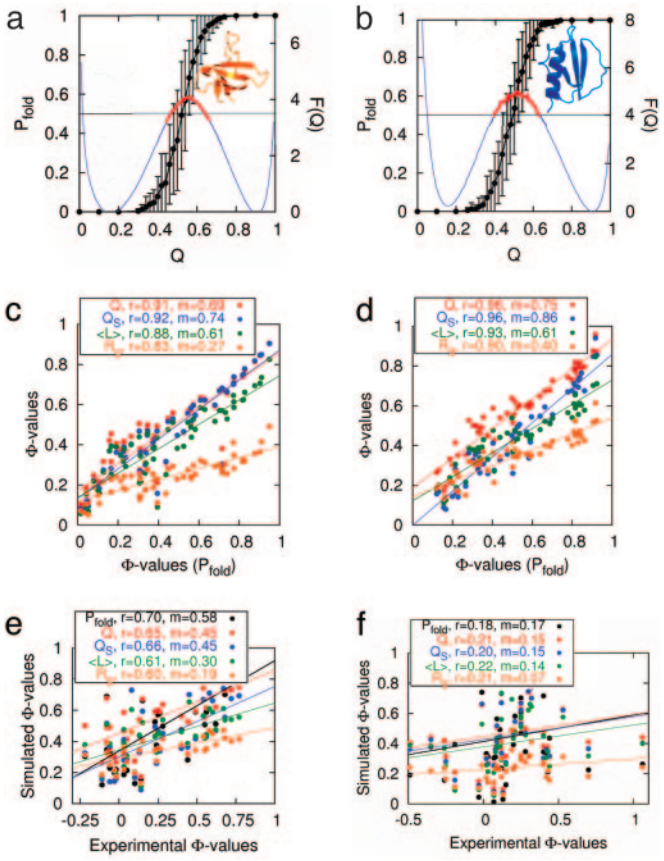

Fig. 3. Comparing the TSE obtained from P_{fold} and the structural reaction coordinates of two-state folding proteins. (*a* and *b*) For both c-src SH3 (*a*) and CI-2 (*b*), the free-energy profile using Q as a reaction coordinate is overlaid with the average P_{fold} of structures (with error bars indicating 1 SD) over the range $Q = 0.30–0.80$. The putative TSE corresponds to $P_{fold} = 0.50 \pm 0.10$, whereas the TSE predicted by Q is $1k_BT$ from the peak of the free energy profile. (*c* and *d*) The Φ-values of the TSE as predicted by Q, Q_S, $\langle L \rangle$, and R_g are compared with the putative TSE for c-src SH3 (*c*) and CI-2 (*d*). (*e* and *f*) The simulated Φ-values as calculated using the aforementioned measures are compared with the experimentally observed Φ-values for c-src SH3 (*e*) and CI-2 (*f*). The correlation coefficient, r, and the slope of the best-fit line, m, are used for quantitative comparisons.

We compare the Φ-values observed from experiments to the Φ-values as determined by P_{fold} and the structural coordinates. For c-src SH3, the correlation coefficient between the experimental Φ-values and the calculated ones with Q has been reported previously to be ≈0.60 (22). We found a correlation coefficient of 0.65. In our analysis, the highest correlation coefficient is observed when the Φ-values are based on P_{fold} with $r = 0.70$, but other reaction coordinates performed similarly well (Fig. 3*e*). We note that the difference between the correlation coefficients using P_{fold} and Q is 0.05. There is thus only a miniscule improvement when using P_{fold}. The correlation between experimentally determined Φ-values and those obtained by simulating c-src SH3 using an all-atom model with an empirical force field, where nonnative interactions are considered, is only 0.74. That correlation would be improved to 0.93 if the Φ-values of the hydrophilic residues were excluded from comparison (8). Plotkin and coworkers (22) have shown that the correlation between simulated Φ-values for CI-2 with experi-

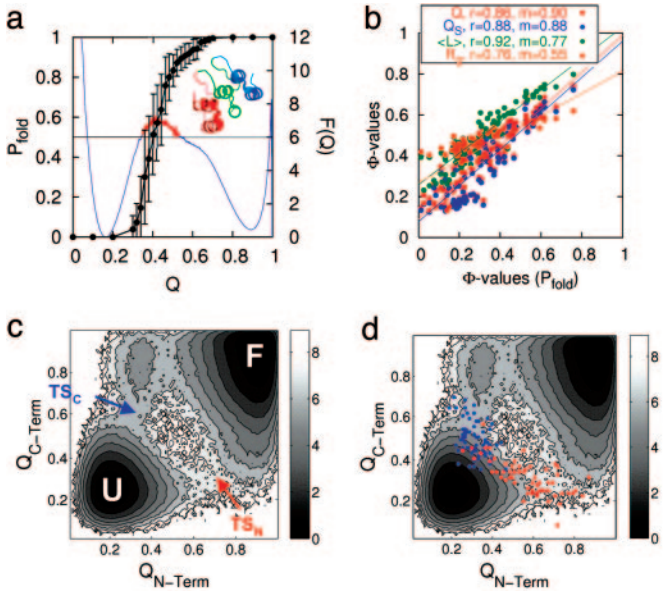

Fig. 4. Comparing the TSE obtained from P_{fold} and structural reaction coordinates for 3ANK, a protein with a broad, asymmetrical free-energy barrier. (*a*) The free energy profile of 3ANK using Q as a reaction coordinate is overlaid with the average P_{fold} of structures (with error bars indicating 1 SD) over the range $Q = 0.30–0.80$. (*b*) The Φ-values of the TSE as predicted by Q, Q_S, $\langle L \rangle$, and R_g are compared with the putative TSE. (*c*) The free energy surface projected onto the N-terminal ($Q_{N\text{-Term}}$) and C-terminal ($Q_{C\text{-Term}}$) halves of 3ANK with the unfolded, transition, intermediate, and folded states indicated for the two competing nucleating routes. (*d*) The two clusters of structures in the putative TSE (i.e., $P_{fold} \sim 0.50$) are overlaid on the free energy profile projected onto $Q_{N\text{-Term}}$ and $Q_{C\text{-Term}}$.

mental values is improved by including nonadditive energetic terms, which arise from solvent and side-chain effects. Of course, such a nonadditive model still corresponds to a perfectly funneled landscape. Our simulation model is purely additive, so $r = 0.65$ is likely the best achievable correlation. The agreement between the simulated Φ-values using reaction coordinates with experiment is not precise. This lack of precision is found to be equally true for the Φ-values coming from the P_{fold} TSE as well as the others (Fig. 3*f*). Clearly, the source of disagreement between the experimental and simulated Φ-values is not the inadequacy of the reaction coordinate, but rather the lack of nonadditive energetic terms in the model. We note that although supplementing the pairwise model with nonadditive interactions increases the correlation between experimental and simulated Φ-values for many proteins, SH3 is an exception showing almost no effect on Φ-values from increased nonadditivity (22). For both proteins, describing the TSE using P_{fold} rather than any of the native topology-based reaction coordinates results in no appreciable improvement of agreement with experiment.

Broad Free Energy Barrier Masks a Competition Between Two- and Three-State Transitions. We next used the same protocol for 3ANK folding. 3ANK is a designed ankyrin repeat protein with three repeating subunits, each with an identical consensus sequence (34). 3ANK is predicted by Q to fold by a two-state transition with a broad, asymmetrical free energy barrier (Fig. 4*a*) (35). Again, we found that $P_{fold} = 0.50$ corresponds to the Q at the peak of the free energy profile (Fig. 4*a*). This finding is remarkable considering that the free energy barrier ranges from $Q = 0.30$ to 0.70, and the peak lies far closer in Q to the unfolded than the folded state. The Φ-values determined by the reaction coordinates Q, Q_S, and $\langle L \rangle$ agree with those of the TSE based on P_{fold} (Fig. 4*b*). Why is the free energy barrier broad? To answer this question, we divided the 3ANK in half and projected the free energy profile onto two coordinates, $Q_{N\text{-Term}}$ and $Q_{C\text{-Term}}$, the fraction of native contacts of the N- and C-terminal halves. This approach was motivated by the earlier predictions of Ferreiro *et al.* (35) that the folding nucleus of ankyrin repeat proteins corresponds to ≈1.5 repeats. The resulting free energy profile exhibits a competition between a N-terminal nucleating two-state transition and a C-terminal nucleating three-state transition (Fig. 4*c*). In a recent experimental study, a 3-ankyrin repeat protein with a similar sequence to 3ANK exhibited equilibrium intermediates (three-state folding mechanism) at high temperatures but not at low temperatures (two-state folding mechanism) (36). The discrepancy in the observed folding behaviors can be rationalized by a competition of folding mechanisms similar to that found in the simulations.

Examining the Φ-values of the transition states and intermediate in the free-energy landscape reveals the existence of two parallel sets of routes to the folded state. The free energy barrier for N-terminal nucleation is higher than the C-terminal counterpart. The N-terminal transition state (TS_N) is folded in the first repeat and the N-terminal half of the second repeat (see Fig. 9*a*, which is published as supporting information on the PNAS web site). In the case of C-terminal nucleation, the folding pathways include a high-energy intermediate. The first transition state (TS_C) consists of a folding nucleus that is folded in the second repeat and the N-terminal half of the third repeat (Fig. 9*b*).

How do we reconcile the complex mechanism that we can ferret out with multiple structural coordinates, and that also is supported by experimental evidence, with an analysis using P_{fold}? We clustered the structures with $P_{fold} = 0.50$ according to the similarity measure, q (see *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site). The resulting tree yields predominantly two sets of clusters. These clusters correspond to either N or C terminus nucleation (TS_N and TS_C, respectively), again implying that there are two parallel routes of nucleation in the folding of 3ANK (Fig. 4*d*). Unfortunately, these structural clusters correspond to the N-terminal transition state as they should, but they only contain the first C-terminal transition state. There is no indication from the TSE predicted by P_{fold} of the second transition state along the C-terminal nucleation route, although it clearly exists.

P_{fold} Fails When There Are Intermediates. To test fairly whether P_{fold} can identify folding through multiple transition states, we simulated a protein that does not have competing pathways but has an intermediate according to free energy profiles based on Q. We selected CV-N, a single-chain protein composed of two domains with high sequence and structure similarity to each other. Laboratory experiments have classified wild-type CV-N as a two-state folder, yet a mutation can stabilize an intermediate (37). Go model simulations of CV-N showed previously a three-state folding transition with a high-energy intermediate (38). A two-state folding transition occurs when the Go model is constrained by disulfide bonds present in the protein (38). For our test, we modeled CV-N without considering disulfide bonds. The choice of a protein system with a high-energy intermediate allows a rigorous analysis of such an intermediate case with minimal computations. The high-energy intermediate is easily seen by projecting the free energy along Q, along with the Q of the N and C termini and their interface (Fig. 5). The two domains depend on one another to fold.

When we analyzed the region between the folded and unfolded states, $P_{fold} = 0.50$ clearly corresponds to the intermediate. The two actual transition states are barely represented in the ensemble of structures with $P_{fold} = 0.50$ (Fig. 5*a*). It is easy to see that using P_{fold} to identify and distinguish multiple transition states is generally impossible. When an intermediate occurs in the folding, there are three possible situations with regard to P_{fold}

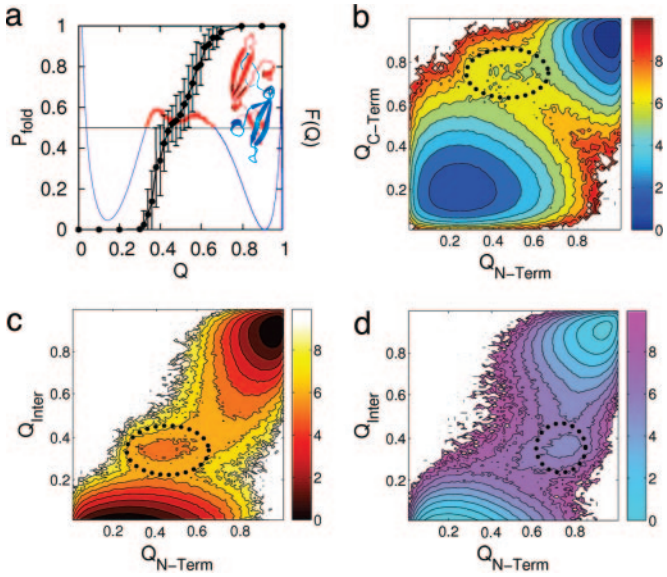

Fig. 5. Comparing the TSE obtained from P_{fold} and structural reaction coordinates for CV-N, a protein that is simulated to fold with a three-state folding mechanism. (*a*) The free energy profile of CV-N using Q as a reaction coordinate is overlaid with the average P_{fold} of structures (with error bars indicating 1 SD) over the range $Q = 0.30–0.80$. (*b*) The free energy profile is projected onto the N-terminal ($Q_{N\text{-}Term}$) and C-terminal ($Q_{C\text{-}Term}$) halves of CV-N, corresponding to the two domains in the protein. (*c*) The free energy profile is projected onto $Q_{N\text{-}Term}$ and the interface between the two domains (Q_{Inter}). (*d*) The free energy profile is projected onto $Q_{C\text{-}Term}$ and Q_{Inter}.

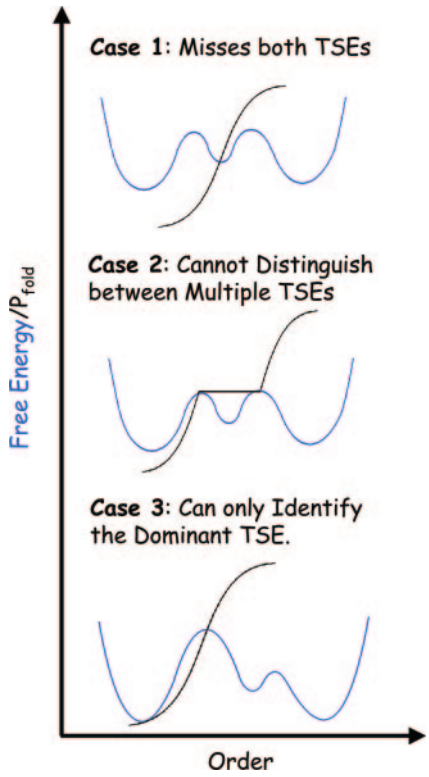

Fig. 6. A schematic depicting the three possible relationships between P_{fold} and free energy profiles for protein systems with two folding transition states.

(Fig. 6). The first possibility is that P_{fold} will miss all of the transition states. The $P_{fold} = 0.50$ ensemble will correspond to another part of the free energy surface, usually an intermediate, as is the case of CV-N. The P_{fold} of the individual transition states never equal 0.50 but will have higher or lower values. Sometimes the $P_{fold} = 0.50$ ensemble will correspond to several different transition states of the free energy surface. In this case, as illustrated by 3ANK, one must use clustering algorithms to differentiate the chemically distinct TSEs. The very meaning of the TSE must again involve other measures that capture this clustering. Finally, in favorable situations the $P_{fold} = 0.50$ ensemble will correspond to only a single dominant transition state but will ignore others that may be important upon mutation. In every case where the folding mechanism involves more than one transition state, we have found that using P_{fold} alone cannot describe even the basic features of the folding process, at least for a minimally frustrated system. It is much better then to use direct structure-based reaction coordinates.

Conclusions

Protein folding has long been viewed as being rich in complexities. With the development of the energy landscape theory, our view of protein folding, however, has greatly simplified from the hopelessly complex one first presented by Levinthal's paradox. Because of their funneled energy landscapes, global structural measures of similarity to the native state are clearly adequate for describing the folding progression for most natural proteins. P_{fold} may be used unambiguously to characterize a TSE for a simple two-state folding processes, but it is unnecessary for the minimally frustrated case. The high computational demands of determining P_{fold} can be avoided by the use of native structure-based reaction coordinates. These coordinates predict the TSE for minimally frustrated systems just as well as P_{fold} does. Our study shows that the global reaction coordinates based on the native topology of a protein, such as Q, Q_S, and $\langle L \rangle$, fully satisfy the criteria needed to accurately identify and describe of the TSE. The Φ-values of the TSE as determined by P_{fold} and the thermodynamic reaction coordinates are nearly identical. They are, therefore, equally accurate descriptors of the TSE as probed by current experiments.

Understanding the folding of larger, more complex proteins, even if unfrustrated, requires generally the use of several reliable reaction coordinates that can distinguish the multiple transition states and/or parallel routes that are present in the folding process. For such cases, no single global measure of protein-folding progression will ever be adequate. Thus, even P_{fold}, often invoked as the standard by which all reaction coordinates should be judged, is itself still insufficient for describing even the qualitative features of folding mechanisms when they are complex enough to have fine intermediates. By using multiple, and possibly local, reaction coordinates and a reasonably intuitive understanding of the principles of protein-folding science, however, a complete picture of protein folding can be obtained.

As we take a step back from our calculations, it is impossible not to marvel at how simple protein folding actually is, at least in comparison with our fears. One must keep in mind that the simplest protein-folding processes are enormously complicated chemical reactions involving very many degrees of freedom. Yet, evolution has led to the global organization of the landscape of proteins into a funnel. The funnel concept allows us to obtain much information about the folding process using only a few coordinates for folding progression. Even the most complex folding processes found in natural proteins seem to require only a handful of reaction coordinates.

Materials and Methods

We first simulated native topology-based (Go) models (17) of c-src SH3 [Protein Data Bank (PDB) ID code 1SRL] and CI-2 (PDB ID code 2CI2). These models both fold by a simple

two-state mechanism. We then examined systems with more complicated mechanisms: 3ANK (PDB ID code 1N0Q), which exhibits a two-state folding transition with a broad, asymmetrical free energy barrier, and CV-N (PDB ID code 2EZM), which displays a three-state folding transition. In these models, the native topology alone is used as input. These models, thus, have perfectly funneled energy landscapes. Because there is no friction from adventitious contacts, the systems are far from glassy. The details of the model have been described elsewhere (17) and are described in the *Supporting Materials and Methods*. Multiple trajectories with numerous unfolding/folding transitions were collected and analyzed by using the weighted histogram analysis method (WHAM) to calculate the free energy surface projected onto the various reaction coordinate(s) of interest. The folding temperature (T_f) was identified as the peak of a specific heat vs. temperature profile.

Identification of TSE by P_{fold}. For each protein, P_{fold} was calculated for structures with Q over the range 0.30–0.80, which is between the unfolded and folded states. This region of Q was divided into 24 bins, and 100 conformations were randomly selected for each bin to evaluate the correlation between P_{fold} and Q. To determine the P_{fold} of each conformation, 100 independent runs (n_{runs}) were performed at each protein's respective folding temperature, ($T/T_f = 1.0$). In total, 240,000 simulations were performed for each system. In statistical terms, this is one of the most rigorous P_{fold} analyses carried out to date. Each simulation continued until it either reached the unfolded or folded state, which we defined as a Q value corresponding to at most $1k_BT$ above the unfolded and folded minima, respectively. For each conformation n_{fold} trajectories reach the folded state, whereas ($n_{runs} - n_{fold}$) trajectories reach the unfolded state. $n_{runs} = 100$ independent simulations were followed until the folded or unfolded state per structure. The P_{fold} thus is calculated as $P_{fold} = n_{fold}/n_{runs}$ has errors on the order of $1/\sqrt{n_{runs}} = 10\%$.

Comparison of Reaction Coordinates with P_{fold} via Φ-Value Analysis. For comparison with results based on P_{fold} determination, we evaluated four structural reaction coordinates to characterize ensembles: the fraction of native contacts (Q), the similarity of natively contacting residue pairs to their native distances (Q_S), the average shortest path length ($\langle L \rangle$), and the radius of gyration (R_g). The details of computing Q_S and $\langle L \rangle$ are described in *Supporting Materials and Methods*. Given the free energy profile using any of the above-mentioned reaction coordinates, the TSE structure can be quantified by Φ-value analysis. Experimentally, Φ-values are determined by changing the protein sequence to delete some of the native contacts made by a given residue. The ratio of the apparent free-energy change of the transition state (neglecting frictional factors!) and the folded state with respect to the unfolded state is calculated to yield the Φ-value for a residue (32). In simulations we also compute the Φ_{ij}-value for each native contact pair between residues i and j from the probability of formation P_{ij}

$$\Phi_{ij}^{sim} = \frac{\Delta\Delta G^{TS-U}}{\Delta\Delta G^{F-U}} \approx \frac{P_{ij}^{TS} - P_{ij}^{U}}{P_{ij}^{F} - P_{ij}^{U}}.$$

The Φ_i value of residue i, corresponding to experimentally observed values, is the average of Φ_{ij} over the interacting partners, j

$$\Phi_i^{sim} = \frac{1}{n}\sum_{j}^{n} \Phi_{ij}^{sim}.$$

We thank Diego Ferreiro for insightful discussions on the ankyrin folding mechanism and Jared Morante for his artistry in constructing Fig. 2. This work was supported by National Institutes of Health Grant 5R01GM44557 and the National Science Foundation-sponsored Center for Theoretical Biological Physics Grants PHY-0216576 and 0225630.

1. Onuchic, J. N. & Wolynes, P. G. (2004) *Curr. Opin. Struct. Biol.* **14,** 70–75.
2. Wigner, E. (1938) *Trans. Faraday Soc.* **34,** 29–41.
3. Frauenfelder, H. & Wolynes, P. G. (1985) *Science* **229,** 337–345.
4. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93,** 6902–6915.
5. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996) *J. Chem. Phys.* **104,** 5860–5868.
6. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996) *Fold Design* **1,** 441–450.
7. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1998) *J. Chem. Phys.* **108,** 334–350.
8. Gsponer, J. & Caflisch, A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6719–6724.
9. Snow, C. D., Sorin, E. J., Rhee, Y. M. & Pande, V. S. (2005) *Annu. Rev. Biophys. Biomol. Struct.* **34,** 43–69.
10. Settanni, G., Rao, F. & Caflisch, A. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 628–633.
11. Ding, F., Guo, W. H., Dokholyan, N. V., Shakhnovich, E. I. & Shea, J. E. (2005) *J. Mol. Biol.* **350,** 1035–1050.
12. Chong, L. T., Snow, C. D., Rhee, Y. M. & Pande, V. S. (2005) *J. Mol. Biol.* **345,** 869–878.
13. Best, R. B. & Hummer, G. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 6732–6737.
14. Nymeyer, H., Socci, N. D. & Onuchic, J. N. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 634–639.
15. Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 777–782.
16. Portman, J. J., Takada, S. & Wolynes, P. G. (1998) *Phys. Rev. Lett* **81,** 5237–5240.
17. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000) *J. Mol. Biol.* **298,** 937–953.
18. Levy, Y., Wolynes, P. G. & Onuchic, J. N. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 511–516.
19. Levy, Y., Cho, S. S., Onuchic, J. N. & Wolynes, P. G. (2005) *J. Mol. Biol.* **346,** 1121–1145.
20. Koga, N. & Takada, S. (2001) *J. Mol. Biol.* **313,** 171–180.
21. Chavez, L. L., Onuchic, J. N. & Clementi, C. (2004) *J. Am. Chem. Soc.* **126,** 8426–8432.
22. Ejtehadi, M. R., Avall, S. P. & Plotkin, S. S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 15088–15093.
23. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2002) *Biophys. J.* **83,** 3525–3532.
24. Hubner, I. A., Edmonds, K. A. & Shakhnovich, E. I. (2005) *J. Mol. Biol.* **349,** 424–434.
25. Li, L. & Shakhnovich, E. I. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13014–13018.
26. Mirny, L. & Shakhnovich, E. (2001) *Annu. Rev. Biophys. Biomol. Struct.* **30,** 361–396.
27. Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1999) *J. Mol. Biol.* **287,** 675–694.
28. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8637–8641.
29. Rao, F., Settanni, G., Guarnera, E. & Caflisch, A. (2005) *J. Chem. Phys.* **122,** 184901.
30. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999) *Nat. Struct. Biol.* **6,** 1016–1024.
31. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 260–288.
32. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224,** 771–782.
33. Eastwood, M. P., Hardin, C., Luthey-Schulten, Z. & Wolynes, P. G. (2003) *J. Chem. Phys.* **118,** 8500–8512.
34. Mosavi, L. K., Minor, D. L., Jr., & Peng, Z. Y. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16029–16034.
35. Ferreiro, D. U., Cho, S. S., Komives, E. A. & Wolynes, P. G. (2005) *J. Mol. Biol.* **354,** 679–692.
36. Devi, V. S., Binz, H. K., Stumpp, M. T., Pluckthun, A., Bosshard, H. R. & Jelesarov, I. (2004) *Protein Sci.* **13,** 2864–2870.
37. Barrientos, L. G., Lasala, F., Delgado, R., Sanchez, A. & Gronenborn, A. M. (2004) *Structure (London)* **12,** 1799–1807.
38. Cho, S. S., Levy, Y., Onuchic, J. N. & Wolynes, P. G. (2005) *Phys. Biol.* **2,** S44–55.