Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics

Samuel S. Cho[a,b,1], Yaakov Levy[a,c,2], and Peter G. Wolynes[a,b,c,3]

[a]Center for Theoretical Biological Physics and Departments of [b]Chemistry and Biochemistry and [c]Physics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Contributed by Peter G. Wolynes, October 13 , 2008 (sent for review September 10 , 2008)

Energy landscape theory requires that the protein-folding mechanism is generally globally directed or funneled toward the native state. The collective nature of transition state ensembles further suggests that sufficient averaging of the native interactions can occur so that the knowledge of the native topology may suffice for predicting the mechanism. Nevertheless, while simple homogeneously weighted native topology-based models predict the folding mechanisms for many proteins, for other proteins knowledge of the native topology, by itself, seems not to suffice in determining the folding mechanism. Simulations of proteins with differing topologies reveal that the failure of homogeneously weighted topology-based models can, however, be completely understood within the framework of a funneled energy landscape and can be quantified by comparing the fluctuation of entropy cost for forming contacts to the expected fluctuations in contact energy. To be precise, we find the transition state ensembles of proteins with all-α topologies, which are more uniform in the specific entropy cost of contact formation, have transition state ensembles that are more readily perturbed by differences in energetic weights than are the transition state ensembles of proteins with significant amounts of β-structure, where the specific entropy costs of contact formation are more widely distributed. This behavior is consistent with a random-field Ising model analogy that follows from the free energy functional approach to folding.

energy landscape theory | random-field Ising model | native topology-based models

Evolution has sculpted the energy landscapes of natural proteins to be globally directed toward the native state (1). More precisely, sequences have evolved so that the interactions present in the native state are more stabilizing than extreme value statistics would lead one to expect for forming random contact interactions (2). In this case, nonnative interactions provide primarily a frictional influence (3, 4). If, furthermore, the nonnative interactions were so weak that they can be completely neglected in comparison to those in the native structure, the balance between the chain entropy and native interaction energies alone would determine the folding mechanism. Reflecting this balance, even crude topology-based measures, such as contact order, provide rough estimates of the folding rates of two-state folding proteins (5). Because many contacts are formed in the transition state ensemble, it further becomes reasonable to simplify the model by replacing individual contact energies with an average value, neglecting sequence variability. The resulting energy landscape is perfectly funneled, but now encodes only the native topology (6). Such averaged contact energy models predict the folding rate of proteins in many cases (7), even when the simple contact order estimate is not very accurate (8). Many studies have shown that a wide range of details of folding and binding mechanisms, such as whether specific intermediates form or not, are also correctly predicted by such native topology-based models in many cases (6, 9). In some circumstances where seemingly minor differences of topology are involved, even predicting mechanistic subtleties is possible from this homogeneous model (10). More quantitative features about the structure of the transition state ensembles, such as the Φ values, are also generally well predicted by pure topology models (11–13), but at this level more discrepancies appear (13). These discrepancies caution us that while the successes of pure native topology-based models are impressive, one must examine the homogeneity assumption that is made in topology-based modeling, which averages the native contact energies. In quantitative terms, can we determine when the homogeneity assumption will suffice and when it will not?

Failures of the contact averaging approximation were first noted in studying structurally homologous proteins having disparate sequences but essentially the same topology. According to the averaging ansatz, even if such proteins are distantly related in sequence, they should exhibit similar folding mechanisms because they share the same native contact pattern. A striking example of the seeming validity of the averaging approximation occurs in the folding of the src and spectrin SH3 domains, which both have the same all-β topology. Even though they have low sequence homology (27%), they are experimentally observed to exhibit very similar transition state ensembles, and this behavior is also seen in simulations (14, 15). The structure of the transition state ensemble is also robust to changes in environmental conditions for these systems (14). Another example is provided by comparing the folding of acylphosphatase with the folding of human procarboxypeptidase A2 activation domain. These proteins both have similar α/β topologies and folding mechanisms while sharing only 13% sequence identity (16), again indicating that the native topology suffices to determine the folding mechanism. Other sets of proteins with nearly identical α/β topologies and low sequence similarity, however, do sometimes exhibit different folding mechanisms, but this often involves symmetry-breaking between two essentially isomorphic folding routes (17–19). The small differences of free energy between two possible routes can easily be determined by just a few contacts. The most dramatic differences in the folding mechanism for topologically equivalent proteins are seen in sets of all-α structural homologues. For Im7 and Im9, both nearly identical 4-helix bundles, the folding mechanism of Im7 involves a populated intermediate, whereas Im9 folds by a 2-state manner, even though there is 60% sequence identity between the two proteins (20). Interestingly, the main transition states still have similar Φ values (21). Recently, Clarke and co-workers (22) showed that the folding rates of α-spectrin repeats of similar topology can vary over several orders of magnitude. Although the native topology clearly plays a critical role in the protein-folding mechanism, these examples imply that energetic weights of the specific residue interactions can sometimes be important as well.

Author contributions: S.S.C. and P.G.W. designed research; S.S.C. performed research; S.S.C. contributed new reagents/analytic tools; S.S.C., Y.L., and P.G.W. analyzed data; and S.S.C. and P.G.W. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742.

[2]Present address: Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel.

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

This article contains supporting information online at www.pnas.org/cgi/content/full/0810218105/DCSupplemental.

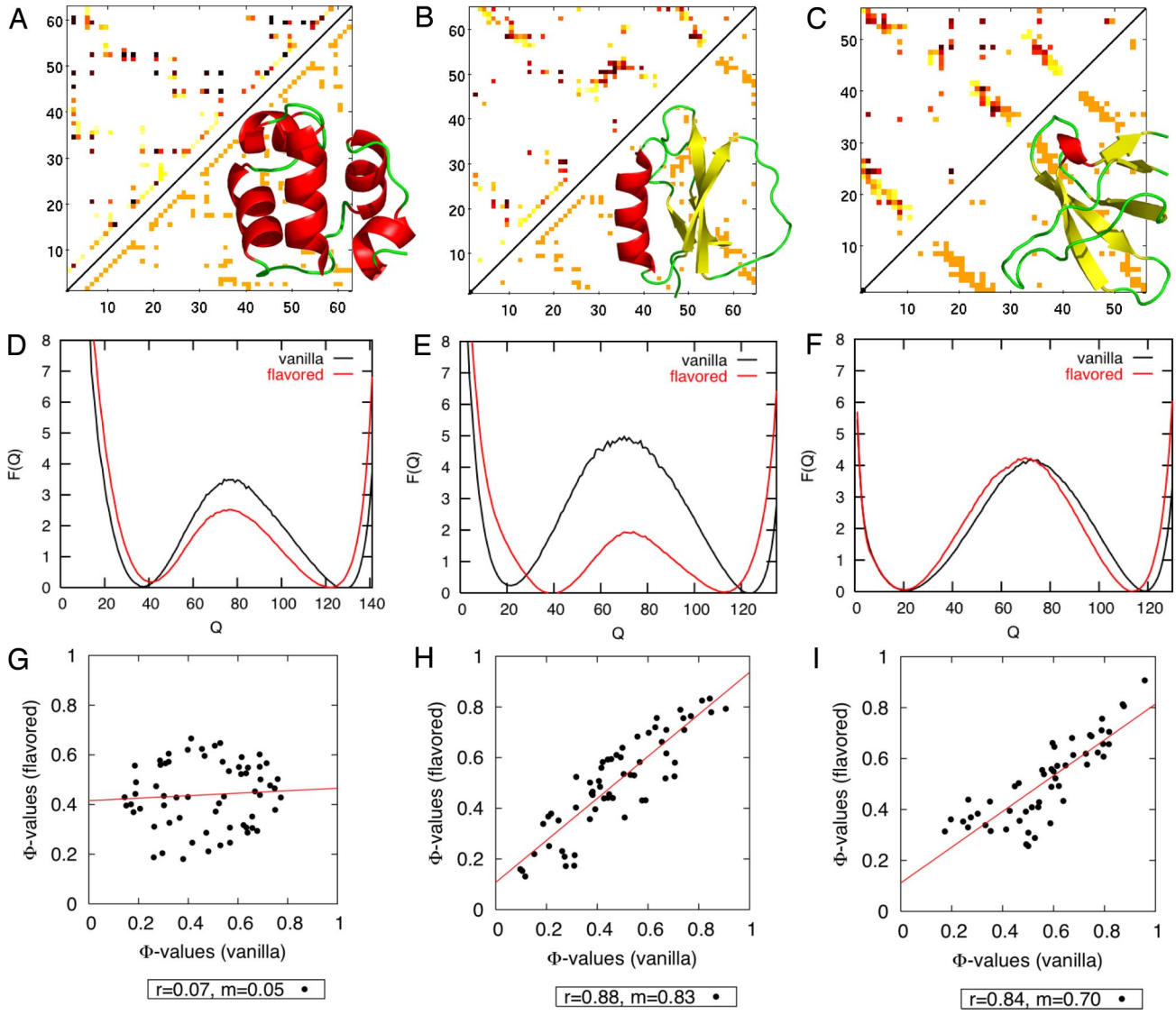

Fig. 1. The folding mechanisms of the all-α lambda repressor [Protein Data Bank (PDB) ID code 1R69] (*A*, *D*, and *G*), the α/β CI2 (PDB ID code 2CI2) (*B*, *E*, and *H*), and all-β src-SH3 domain (PDB ID code 1SRL) (*C*, *F*, and *I*). (*A–C*) The matrices of the interaction energies in the vanilla and flavored native-topology-based models are plotted below and above the diagonal, respectively, with darker colors representing stronger interactions. The corresponding native structures are also shown. (*D–F*) From simulations of the vanilla and flavored models, the free energy profiles were generated with respect to the order parameter Q. (*G–I*) The Φ values from the vanilla and flavored models are compared in a plot with a best-fit line; *m* is the slope of the line.

The effects of energetic heterogeneity of the native interactions on the folding mechanism have already been addressed by using analytical energy landscape theory. Using the free energy functional approach (23, 24), Plotkin and co-workers found that introducing energetic heterogeneity to native interactions in a minimally frustrated system lowers the free energy barrier until it vanishes with a sufficiently large dispersion of native contact energies, and similar behaviors were seen in simulations on lattices (25–27). From the free energy functional perspective, the effects of contact heterogeneity are very much analogous to the well-known phase transition in the random field ferromagnet when the dispersion of site energies become large (28). Sometimes, with sufficiently large dispersion of the native contact energies, the Φ values becomes bimodal, with extreme values close to 0 or 1 (26, 27). Recently, in the context of the α/β CI2 and the all-β src-SH3 domain, Suzuki and Onuchic (29) have shown that the structure of the transition state ensemble is robust and insensitive to energetic details.

In principle, we will directly compare the analytical results of free energy functional approaches with those of native topology-based model simulations. We have carried out simulations that show that, in keeping with the expectations from analytical theory, homogeneously weighted native topology-based models (based on the averaging approximation) determine the folding mechanism of proteins when the entropy costs of contact formation are widely distributed, but that such models fail when the native contact heterogeneity is sufficiently large, even for the α/β and all-β-protein. For the latter, however, the necessary heterogeneity for the breakdown of the averaging ansatz is larger than seems physically reasonable. This explains why homogeneously weighted native topology-based models with large contact entropy dispersion readily reproduce the folding mechanisms of some proteins, whereas the folding mechanisms of proteins with too narrow a distribution of contact entropies cannot be so easily predicted.

Results and Discussion

Homogeneous Versus Heterogeneous Contact Energies in Funneled Landscapes We begin by comparing simulations of the simple homogeneously weighted C_α models to corresponding simulations

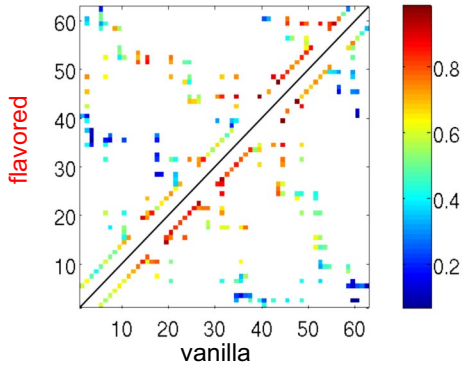

Fig. 2. The probability of a contact in the transition state of the lambda repressor, an all-α protein, with the vanilla and flavored models.

having energetic heterogeneity based on the 20-letter Miyazawa–Jernigan (MJ) contact potential (30). Although the degree of heterogeneity of the MJ potential may be too large, its native contact dispersion is similar to what is predicted by other more refined contact potentials (31). For linguistic simplicity, we will refer to these two variants, both describing perfectly funneled landscapes, as "vanilla" and "flavored" models, respectively. As a starting point, we surveyed several 2-state folding proteins that have been studied previously by both simulations and in the laboratory. We chose the all-α lambda repressor, the α/β CI2, and the all-β src-SH3 domain. In all three proteins, the contact energies in the flavored models seem evenly distributed, with no immediately obvious clusters of either high or low energetic weights (Fig. 1 *A–C*). To quantitatively characterize the folding mechanism, we performed the weighted histogram analysis method (WHAM) to calculate thermodynamic quantities with respect to the order parameter, Q, the fraction of native contacts. We recently showed that Q is one of several simple structural reaction coordinates that captures the folding mechanism on smooth landscapes, even for complicated folding mechanisms (32). In the case of the lambda repressor and CI2, a decrease in the free energy barrier is observed (Fig. 1 *D* and *E*), as predicted analytically (25, 26). We also note that the unfolded basin free energy minimum occurs at a higher Q (the fraction of native contacts) in the flavored model than in the vanilla model, whereas conversely the folded basin has lower Q. For src-SH3, however, the free energy barrier does not change when the energetic heterogeneity is introduced (Fig. 1*F*). For both CI2 and src SH3, the Φ values derived from the simulations of vanilla and flavored models are very similar, with correlation coefficients greater than 0.70 (Fig. 1 *H* and *I*). In contrast, the Φ values for the lambda repressor predicted by the vanilla and fully flavored models are essentially uncorrelated with each other (Fig. 1*G*). A closer analysis of the transition state ensemble for the vanilla model reveals that the folding nucleus consists of structured second and third helices with largely unformed long-range interactions (Fig. 2). In contrast, the transition state ensemble of the flavored model predominantly includes structured long-range interactions between the second and fourth helices (Fig. 2). Oas and co-workers (33) performed NMR spectroscopy of 7 alanine-to-glycine mutants of the lambda repressor, and their limited observations indicated that the first and fourth helices are most populated in the transition state ensemble, whereas the second, third, and fifth helices are less populated. It seems that, although not precisely reproducing the experiment, the flavored model agrees more with the pattern of experimental results than does the vanilla model. To determine whether the short-range interaction energies are the source of the discrepancy between the folding mechanisms observed in the vanilla and flavored models, an inhomogeneous model was also simulated where only the contact energies of the short-range interaction energies of the flavored model were changed back to those of the vanilla model. Now, the free energy barrier becomes about the same as that for the vanilla model [supporting information (SI) Fig. S1*A*], but one still finds the poor correlation between the Φ values in this partially flavored model and the homogeneous case (Fig. S1*B*).

We also simulated several other representative all-α protein domains that we selected from the CATH database (34) (CATH IDs: 1v54E0, 1f6vA0, and 1cy5A0). We chose these proteins because they capture a diverse range in the degree of short-range vs. long-range interactions, as well as helical content (Fig. 3 *A–C*). The contact map of 1v54E0 contains mostly relatively short-range interactions (Fig. 3*A*), whereas 1cy5A0 has a large number of long-range interactions (Fig. 3*C*). 1f6vA0 has an intermediate number of long-range interactions (Fig. 3*B*). Again, the energetic weights seem to be evenly distributed across all of the native interactions (Fig. 3 *A–C*). In all three cases, the flavored model yields a lower free energy barrier than the vanilla model and the folded basin has a lower Q for the flavored model (Fig. 3 *D–F*). For 1f6vA0, the peak of the free energy barrier occurs at a lower Q in the flavored model (Fig. 3*E*). In each case, the Φ values predicted by the vanilla and flavored models for these all-α proteins exhibit no significant correlation (Fig. 3 *G–I*).

Energetic and Entropic Fluctuations in the Folding Mechanism. The differences in the topologies of all-α and all-β proteins can be quantified by the ratio of the number of long-range interactions versus short-range interactions ($N_{\text{long}}/N_{\text{short}}$). Three different peaks appear in the distribution of $N_{\text{long}}/N_{\text{short}}$ for the nonredundant set of the PDB, corresponding to the all-α, α/β, and all-β topologies (Fig. S2*A*). These peaks are also observed when only proteins that have been shown to be 2-state folders are included (Fig. S2*B*). All-α proteins have proportionally the lowest number of long-range interactions because the intrahelical interactions stabilize the secondary structure. For all-β proteins, numerous long-range interactions must form between individual sheets.

To examine the interplay between energetic and entropic contributions to folding, the energy and entropy lost upon formation of native contacts is calculated for the lambda repressor, CI2, and src-SH3 domain (Fig. 4). The energy, $E(Q)$, can be readily calculated as a summation of the inhomogeneous energetic weights, ε_{ij}, of the native interactions (i,j) for the native contacts made (Q_{ij}):$E(Q) = +\Sigma_{ij}\ \varepsilon_{ij}Q_{ij}$. Similarly, the entropy, $S(Q)$, can be represented approximately as a summation of the entropy (S_{ij}) lost upon forming native contacts in the context of an already partially formed ensemble of structures: $S(Q) = +\Sigma ij\ S_{ij}Q_{ij}$. A reasonable approximation to S_{ij} can be found by following an approach similar to that of Shoemaker *et al*. (24). They suggested that initially the entropy lost in sequentially forming short-range interactions can be approximated by the Jacobson–Stockmayer formula (35), $S_{ij} = +k_{\text{B}}\log[\Delta V/|i-j|^{3/2}]$.

Assuming that the denatured protein can be modeled as a random flight chain, the quantity $\Delta V = [(3/2)\pi]^{3/2}\Delta\tau/l_0^3)$, where $\Delta\tau$ is the volume of the interaction range and l_0 is the persistence length. But Shoemaker *et al.* (24) also argued that if some structure is already formed, any entropy lost will continue to make sequentially distant interactions and saturate to that of a typical fluctuating segment of a chain, as introduced by Flory in the mean field theory of rubber vulcanization. This yields $S_{ij} = +k_{\text{B}}\log[\Delta V/(\mu/N)^{3/2}]$ where μ is the number of contacts made and N is the number of contacts in the native state. Interpolating between the two extremes, Shoemaker *et al.* arrived at the following mean field approximation to the contact entropy loss in a partially structured folding ensemble:

$$S_{ij} = +k_{\text{B}}\log[\Delta V/(|i-j|^{-3/2} + (\mu/N)^{-3/2})].$$

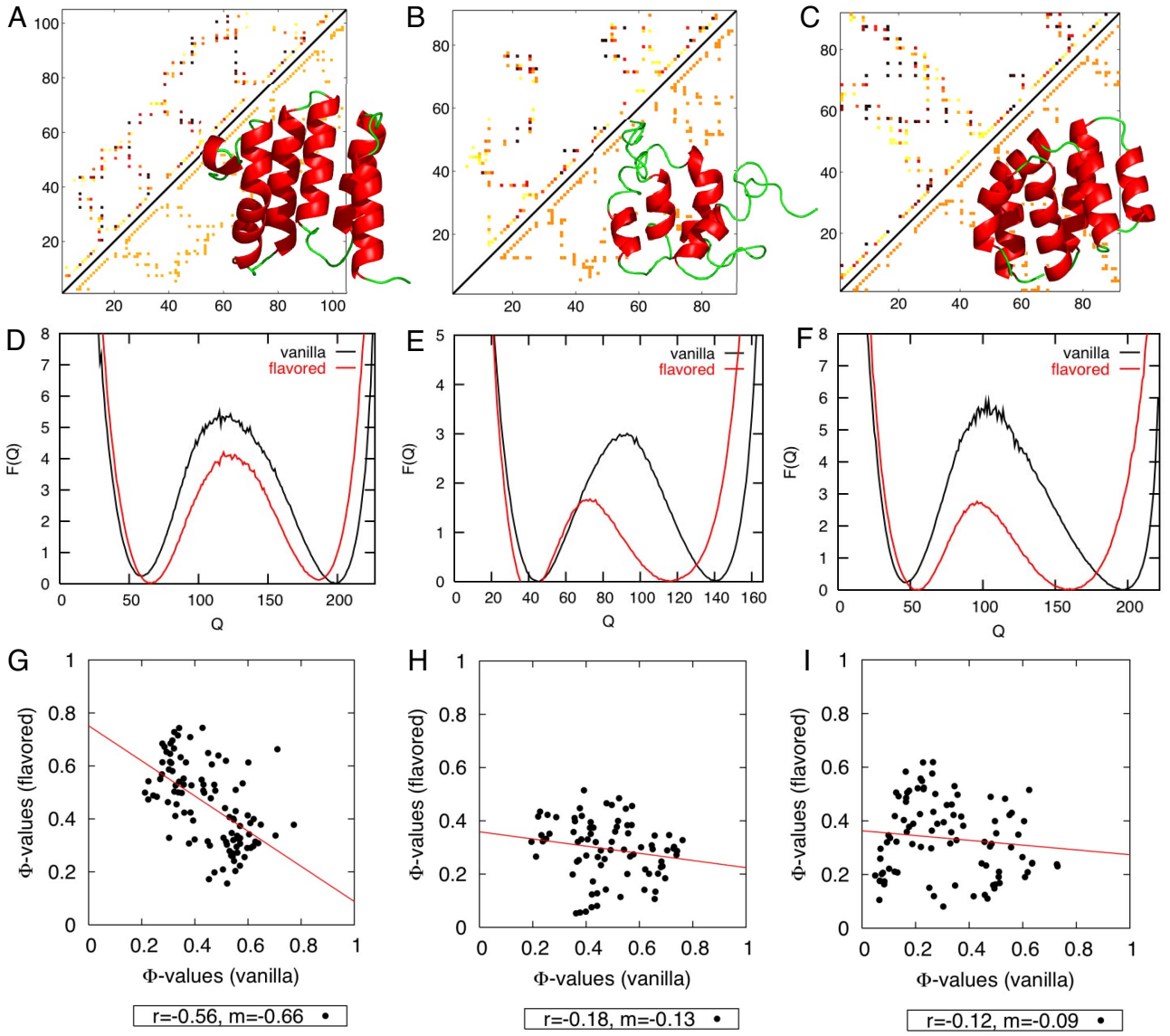

Fig. 3. The folding mechanisms of three all-α proteins (from *Left* to *Right*, 1v54E0, 1f6vA0, and 1cy5A0) selected from the CATH database. (*A–C*) The matrices of the interaction energies in the vanilla and flavored models are plotted below and above the diagonal, respectively, with darker colors representing stronger interactions. The corresponding native structures are also shown. (*D–F*) From simulations of the vanilla and flavored models, the free energy profiles were generated with respect to the order parameter *Q*. (*G–I*) The Φ values from the vanilla and flavored models are compared in a plot with a best-fit line.

The resulting free energy functional takes the form:

$$F(Q_{ij}(\mu)) = \sum_{ij} \varepsilon_{ij} Q_{ij}(\mu)$$

$$+ T\left(\sum_{ij} S_{ij} Q_{ij}(\mu) + \sum_{\mu'=1}^{\mu} \sum_{ij} (\partial S_{ij}(\mu')/\partial \mu') \delta Q_{ij}(\mu') + N \log(\nu) \right)$$

$$+ T\left(\sum_{ij} Q_{ij} \log(Q_{ij}(\mu)) + (1 - Q_{ij}(\mu)) \log(1 - Q_{ij}(\mu)) \right)$$

where $\delta Q_{ij}(\mu') = Q_{ij}(\mu') - Q_{ij}(\mu' - 1)$, and the final term accounts for the different ways of forming a contact in a partially ordered protein. The entropy lost as the chain goes from the unfolded to folded states is estimated as $N \log(\nu)$, where ν is the number of conformations per residue. This is essentially the free energy function of an inhomogeneous-field Ising magnet. The inhomogeneity contains both an entropic and an energetic part.

When the mean-field expressions for the energy and entropy of ensembles from the simulations are stratified with respect to Q, both the entropy and energy, on average, are nearly linearly related to Q (Fig. 4 *A* and *B*) for both proteins. On the other hand, the fluctuations, as quantified by the variance, of the entropy costs of forming contacts $\langle \delta S^2 \rangle$ at Q value and the energies of the formed contacts $\langle \delta \varepsilon^2 \rangle$ show different trends for each protein (Fig. 4 *C* and *D*). By comparing the quantity $\langle \delta S^2 \rangle / \langle \delta \varepsilon^2 \rangle$ at the transition state for each of the proteins, we can quantify which of the two contributions to the "random" fields will dominate the pattern of contacts formed. The ratio determines whether the entropic or energetic fluctuations dominate the folding mechanism. A high (low) value indicates that entropic (energetic) fluctuations determine the structure of the transition state ensemble. It is noteworthy that the ratio $\langle \delta S^2 \rangle / \langle \delta \varepsilon^2 \rangle$ is strongly correlated with the above-mentioned $N_{\text{long}}/N_{\text{short}}$, with a correlation coefficient of 0.90 (Fig. 5). Therefore, for

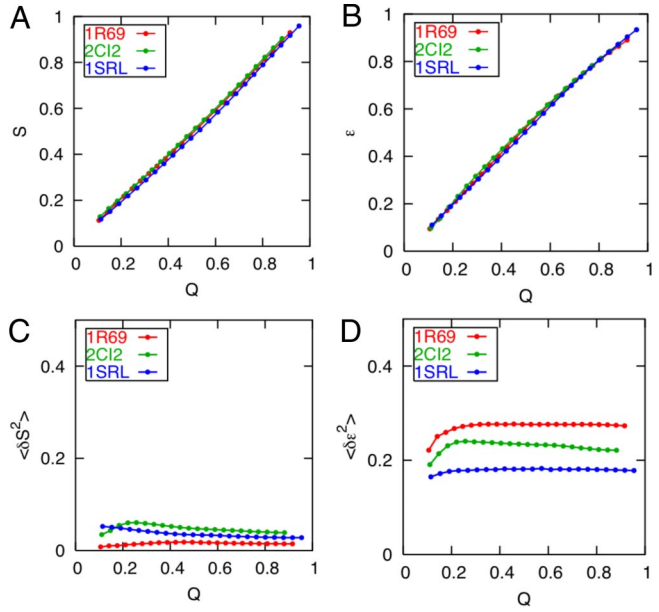

Fig. 4. The entropy and energy lost from the formation of native contacts for all-α (red), α/β- (green), and all-β (blue) proteins. Shown are the entropy (*A*) and energy (*B*), as well as the variance of the entropy (*C*) and energy (*D*), all plotted with respect to the order parameter, *Q*.

a protein with a high number of long-range contacts (e.g., all-β protein), the entropic fluctuations will tend to dominate the folding mechanism, whereas for proteins with a low number of long-range contacts (e.g., all-α protein), the folding mechanism should be susceptible to energetic fluctuations.

Testing the Criterion for When Energetic Heterogeneity Plays a Significant Role. The above observations suggest the sensitivity in the Φ values to the energetic details between the vanilla and flavored models depends largely on the value of $\langle\delta S^2\rangle/\langle\delta\varepsilon^2\rangle$ for each protein system. To confirm this, we studied a series of models where $\langle\delta\varepsilon^2\rangle$ is varied over a range, but $\langle\delta S^2\rangle$, of course, remains constant for each given protein topology. We expect that once $\langle\delta\varepsilon^2\rangle$ increases sufficiently (and thereby decreasing $\langle\delta S^2\rangle/\langle\delta\varepsilon^2\rangle$), large deviations in the Φ values from those of the homogeneous vanilla model will occur. Using this reasoning, a key simulation test of the argument becomes possible: in the simulation world (if not in the laboratory!), we can design an all-β protein, such as the src-SH3 domain, to have a transition state ensemble that is sensitive to energetic fluctuations, such as an all-α protein, by using an unrealistically large variation in the native contact energy.

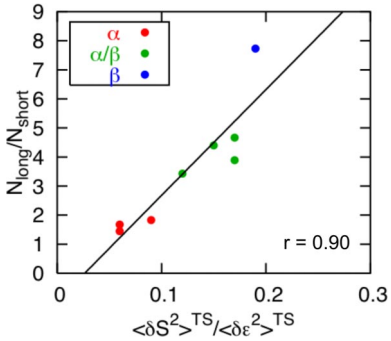

Fig. 5. The relationship between the ratio of the entropic and energetic fluctuations at the transition state with the ratio between long- and short-range native interactions for well-studied two-state folding proteins.

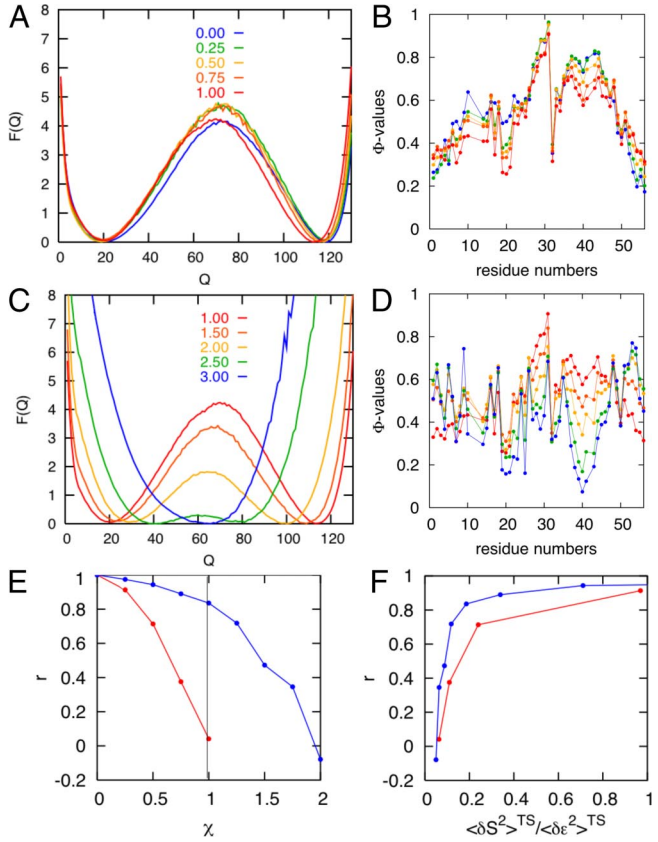

Fig. 6. Flavored model simulations of src-SH3 domain protein with a range of distributions of the Miyazawa–Jernigan contact energies. The free energy profiles (*A*) and Φ values (*B*) are shown for simulations using the varying parameter, χ, in a range where the folding mechanism does not change significantly. The free energy profiles (*C*) and Φ values (*D*) are shown for simulations using the varying parameter, χ, in a range where the folding mechanism does change significantly. The dependence of the correlation between the Φ values of the vanilla model versus the flavored models, *r*, with a range of χ (*E*) and $\langle\delta S^2\rangle/\langle\delta\varepsilon^2\rangle$ (*F*) is shown in blue for the all-β src-SH3 domain protein and in red for the all-α lambda repressor, for comparison.

To construct models with varying $\langle\delta\varepsilon^2\rangle$, we studied variable sets of interresidue energetic weights, $\varepsilon_{i,j}^{\text{new}}$ which can interpolate between the vanilla and the flavored models and that can furthermore extrapolate past the usual flavored model in energetic heterogeneity linearly: $\varepsilon_{i,j}^{\text{new}} = \chi(\varepsilon_{i,j}^{\text{MJ}} - \bar{\varepsilon}^{\text{MJ}}) + \bar{\varepsilon}^{\text{MJ}}$. Here $\varepsilon_{i,j}^{\text{MJ}}$ is the original MJ weight for a given residue pair (i,j), $\bar{\varepsilon}^{\text{MJ}}$ is the mean value of the entire set of MJ weights, and χ is a parameter that can be varied. The value of χ equal to 0 and 1 corresponds to the vanilla and flavored models, respectively. Values of χ between 0 and 1, inclusive, have distributions of energetic weights with the variance $(\delta\varepsilon^2)$ ranging from 0 (i.e., vanilla model) to that of the fully flavored model. The variance can be increased even further by choosing values of χ greater than 1.

For the all-β protein, src-SH3 domain, we first calculated the free energy profile and the Φ values over the range of χ between 0 and 1 (Fig. 6 *A* and *B*). Very little difference is observed between the results of the vanilla and flavored models, as well as the intermediate models. However, when χ is increased past 1, the free energy barrier begins rapidly to decrease, whereas the unfolded state becomes more structured and the folded state becomes less structured, as is seen for all-α proteins (Fig. 6*C*). The free energy barrier height decreases with increasing χ until the free energy profile contains only a single minimum, corresponding to a downhill folding scenario (36, 37). While this physically unrealistic regime

cannot be achieved in the laboratory, these general trends agree with the arguments based on the free-energy functional of a β-protein with enhanced native contact heterogeneity (26). Also, a marked difference in the Φ values exists (Fig. 6*D*), as was seen earlier only for the all-β proteins. Therefore, with a sufficiently large $\langle\delta\varepsilon^2\rangle$, albeit in an unrealistic regime, the entropy costs intrinsic to forming the topology of the protein are no longer the sole significant factors in folding. The correlation between the Φ values of the vanilla as compared to those of the various flavored models disappears at a lower value of χ in the lambda repressor than the src-SH3 domain (Fig. 6*E*). In both proteins, the Φ values of the flavored models remain close to that of the vanilla model if $\langle\delta S^2\rangle/\langle\delta\varepsilon^2\rangle$ is greater than around 0.60 (Fig. 6*F*).

Conclusions

Near the end of the movie "Magnum Force" (1973), Dirty Harry famously tells the villain, "A man's got to know his limitations." Likewise, a good theoretician must understand the limitations of models and appreciate the regimes where they will fail. Toward that end we hope to have clarified when folding mechanisms can be predicted from simple, homogeneously weighted native structure-based models and when the details of the energetics must be better understood. Indeed, the dispersion of entropy needed to form the transition state is very often dominant. In other cases, these entropy cost fluctuations can be overcome by the fluctuations in energetic reward for forming specific contacts for many proteins, usually having all-α topologies. It seems that the folding mechanisms of all-α proteins generally are more sensitive to energetic heterogeneity than those of more long-range topologies. The details of the folding of all-α proteins cannot always be very accurately predicted from homogeneously weighted native-topology-based models because the strength of the individual interactions can dramatically change the folding mechanism. Nevertheless, even when the details are hard to predict, the large differences in the folding mechanisms often found for all-β structural homologues of nearly identical structures can be fully understood within the framework of the energy landscape theory. Because of the sensitivity of the folding mechanism to energetic heterogeneity, the detailed mechanistic predictions are not trivial for these systems. Although strong nonnative contacts slow the kinetics of proteins because of friction, the presence of weak nonnative interactions has been shown analytically to increase the folding rate (38) by reducing the entropy of the unfolded state by collapse, and such interactions may generally play a significantly greater role in all-α as compared to other proteins (39). The same may be true for nonadditive interactions that arise from the presence of water and side chains that are absent in our models but have been shown to be important in determining the transition state ensemble for the last assembly events (12). Regardless, energy landscape theory explains why, in very many cases, folding is not as difficult to understand as some still fear (40) and even gives us a quantitative understanding of the limitations of the simplest versions of the folding funnel.

Materials and Methods

In our study, we used a C_α native-topology-based model where a single bead centered on the C_α position represents a residue, as described previously (6) with homogeneous native contact energies ("vanilla model"). The set of energetic weights of the Miyazawa–Jernigan potential (30) was the basis for introducing native energetic heterogeneity to the model ("flavored model"). A detailed description of both models is presented in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We are grateful to José Onuchic for critically reading the manuscript. This work was funded by the National Institutes of Health (Grant 5R01GM44557) and the National Science Foundation-sponsored Center for Theoretical Biological Physics (Grants PHY-0216576 and 0225630). S.S.C. is supported by Ruth L. Kirschstein National Research Service Awards from the National Institutes of Health.

1. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14:70–75.
2. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA* 84:7524–7528.
3. Bryngelson JD, Wolynes PG (1989) Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J Phys Chem* 93:6902–6915.
4. Hardin C, Luthey-Schulten Z, Wolynes PG (1999) Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides. *Proteins* 34:281–294.
5. Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277:985–994.
6. Clementi C, Nymeyer H, Onuchic JN (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.
7. Chavez LL, Onuchic JN, Clementi C (2004) Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J Am Chem Soc* 126:8426–8432.
8. Gosavi S, Chavez LL, Jennings PA, Onuchic JN (2006) Topological frustration and the folding of interleukin-1 beta. *J Mol Biol* 357:986–996.
9. Levy Y, Wolynes PG, Onuchic JN (2004) Protein topology determines binding mechanism. *Proc Natl Acad Sci USA* 101:511–516.
10. Clementi C, Garcia AE, Onuchic JN (2003) Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L. *J Mol Biol* 326:933–954.
11. 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.
12. Ejtehadi MR, Avall SP, Plotkin SS (2004) Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc Natl Acad Sci USA* 101:15088–15093.
13. Levy Y, Cho SS, Onuchic JN, Wolynes PG (2005) A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *J Mol Biol* 346:1121–1145.
14. Martinez JC, Serrano L (1999) The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol* 6:1010–1016.
15. Hubner IA, Edmonds KA, Shakhnovich EI (2005) Nucleation and the transition state of the SH3 domain. *J Mol Biol* 349:424–434.
16. Chiti F, *et al.* (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat Struct Biol* 6:1005–1009.
17. Gunasekaran K, Eyles SJ, Hagler AT, Gierasch LM (2001) Keeping it in the family: Folding studies of related proteins. *Curr Opin Struct Biol* 11:83–93.
18. Zarrine-Afsar A, Larson SM, Davidson AR (2005) The family feud: Do proteins with similar structures fold via the same pathway? *Curr Opin Struct Biol* 15:42–49.
19. Karanicolas J, Brooks CL, III (2003) Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J Mol Biol* 334:309–325.
20. Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE (1999) Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J Mol Biol* 286:1597–1608.
21. Friel CT, Capaldi AP, Radford SE (2003) Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: Similarities and differences in the folding of homologous proteins. *J Mol Biol* 326:293–305.
22. Scott KA, Batey S, Hooton KA, Clarke J (2004) The folding of spectrin domains I: Wild-type domains have the same stability but very different kinetic properties. *J Mol Biol* 344:195–205.
23. Bohr HG, Wolynes PG (1992) Initial events of protein folding from an information-processing viewpoint. *Phys Rev A* 46:5242–5248.
24. Shoemaker BA, Wang J, Wolynes PG (1997) Structural correlations in protein folding funnels. *Proc Natl Acad Sci USA* 94:777–782.
25. Plotkin SS, Wang J, Wolynes PG (1997) Statistical mechanics of a correlated energy landscape model for protein folding funnels. *J Chem Phys* 106:2932–2948.
26. Plotkin SS, Onuchic JN (2000) Investigation of routes and funnels in protein folding by free energy functional methods. *Proc Natl Acad Sci USA* 97:6509–6514.
27. Plotkin SS, Onuchic JN (2002) Understanding protein folding with energy landscape theory - Part II: Quantitative aspects. *Q Rev Biophys* 35:205–286.
28. Villain J (1985) Equilibrium critical properties of random field systems - new conjectures. *J Phys (Paris)* 46:1843–1852.
29. Suzuki Y, Onuchic JN (2005) Modeling the interplay between geometrical and energetic effects in protein folding. *J Phys Chem* 109:16503–16510.
30. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644.
31. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG (2004) Water in protein structure prediction. *Proc Natl Acad Sci USA* 101:3352–3357.
32. Cho SS, Levy Y, Wolynes PG (2006) P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc Natl Acad Sci USA* 103:586–591.
33. Burton RE, Huang GS, Daugherty MA, Calderone TL, Oas TG (1997) The energy landscape of a fast-folding protein mapped by Ala→Gly substitutions. *Nat Struct Biol* 4:305–310.
34. Pearl FMG, *et al.* (2003) The CATH database: An extended protein family resource for structural and functional genomics. *Nucleic Acids Res* 31:452–455.
35. Jacobson H, Stockmayer WH (1950) Intramolecular reaction in polycondensations. I. The theory of linear systems. *J Chem Phys* 18:1600–1606.
36. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins* 21:167–195.
37. Cho SS, Weinkam P, Wolynes PG (2008) Origins of barriers and barrierless folding in BBL. *Proc Natl Acad Sci USA* 105:118–123.
38. Clementi C, Plotkin SS (2004) The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci* 13:1750–1766.
39. Sutto L, Latzer J, Hegler JA, Ferreiro DU, Wolynes PG (2007) Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proc Natl Acad Sci USA* 104:19825–19830.
40. Whitesides GM, Krishnamurthy VM (2005) Designing ligands to bind proteins. *Q Rev Biophys* 38:385–395.