Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection

Faruck Morcos[a], Nicholas P. Schafer[a,b], Ryan R. Cheng[a], José N. Onuchic[a,b,c,d], and Peter G. Wolynes[a,b,c,d,1]

[a]Center for Theoretical Biological Physics and Departments of [b]Chemistry, [c]Physics and Astronomy, and [d]Biochemistry and Cell Biology, Rice University, Houston, TX 77005

Contributed by Peter G. Wolynes, July 17, 2014 (sent for review June 30, 2014)

The energy landscape used by nature over evolutionary timescales to select protein sequences is essentially the same as the one that folds these sequences into functioning proteins, sometimes in microseconds. We show that genomic data, physical coarse-grained free energy functions, and family-specific information theoretic models can be combined to give consistent estimates of energy landscape characteristics of natural proteins. One such characteristic is the effective temperature T_{sel} at which these foldable sequences have been selected in sequence space by evolution. T_{sel} quantifies the importance of folded-state energetics and structural specificity for molecular evolution. Across all protein families studied, our estimates for T_{sel} are well below the experimental folding temperatures, indicating that the energy landscapes of natural foldable proteins are strongly funneled toward the native state.

energy landscape theory | information theory | selection temperature | funneled landscapes | elastic effects

The physics and natural history of proteins are inextricably intertwined (1, 2). The cooperative manner in which proteins find their way to a folded structure is the result of proteins having undergone natural selection and not typical of random polymers (3, 4). Likewise, the requirement that most proteins must fold to function is a strong constraint on their phylogeny. The unavoidable random mutation events that proteins have undergone throughout their evolution have provided countless numbers of physicochemical experiments on folding landscapes. Thus, the evolutionary patterns of proteins found through comparative sequence analysis can be used to understand protein structure and energetics. In this paper, we compare the information content in the correlated changes that have occurred in protein sequences of common ancestry with energies from a transferable energy function to quantify the influence of maintaining foldability on molecular evolution.

Funneled Folding Landscapes from Evolution in Sequence Space

The key to our analysis is the principle of minimal frustration (3, 5), which states that, for quick and robust folding, the energy landscape of a protein must be dominated by interactions found in the native conformation. This native conformation is, therefore, separated by an energy gap from other compact structures that otherwise might act as kinetic traps (6, 7). These kinetic traps might appear on the folding landscape during evolution if a random mutation was to stabilize a conformation distinct from the functional one, leading to unviability. In this way, evolution and physical dynamics are coupled. A funneled, minimally frustrated landscape can be achieved if the sequence of the protein evolves to stabilize the native state while not increasing the landscape ruggedness.

If folding were the only physicochemical constraint on evolution, the ensemble of naturally observed sequences would correspond to the set of sequences that has a solvent-averaged free energy for the native conformation below a threshold set by the expected ground-state energy for a random sequence. Because sequence space is vast, the usual arguments showing the equivalence of microcanonical and canonical ensembles in statistical mechanics suggest that this evolutionary ensemble characterized by a threshold energy would be equivalent to a canonical distribution of sequences characterized by a Boltzmann probability: $e^{(-\Delta E/k_B T_{sel})}$. This Boltzmann-like probability contains the energy gap between the folded configuration and the compact misfolded configurations along with an appropriate selection temperature (T_{sel}) (4, 8–10) quantifying how strong the folding constraints have been during evolution. T_{sel} is the apparent temperature at which sequences were selected by evolution for a particular protein family or fold. It does not correspond to a critical temperature in the laboratory but can, nonetheless, still be usefully compared with other measurable temperatures, such as the glass transition temperature and folding temperature. Of course, other constraints on molecular evolution exist, including the maintenance of the ability of a protein to bind to appropriate partners (11, 12), catalyze appropriate reactions as for the serine proteases with their famous catalytic triad (13, 14), undergo allosteric changes (15), and avoid aggregation (16). All of these factors potentially enter the quantitative statistical theory of molecular evolutionary outcomes.

Under the quasiequilibrium selection hypothesis based on folding energy alone, given the physical free energy function E, the probability of any given sequence having attained a given fold can be computed in principle. For a single structural family, finding this probability essentially corresponds with

Significance

Natural protein sequences, being the result of random mutation coupled with natural selection, have remarkable properties that are not typical of unselected random sequences, including the ability to robustly fold to an organized structure that is needed to function. We estimate the selection temperature, the effective temperature at which sequences were selected by evolution, for eight protein families and compare these values with experimental data for folding temperatures of proteins in each family. The selection temperature measures the importance of maintaining the stability and structural specificity of the folded state on the evolutionary process. For all families, the selection temperature is below physiological temperature, indicating that maintaining the structural integrity of the folded state is an important constraint on evolution.

Author contributions: F.M., N.P.S., R.R.C., J.N.O., and P.G.W. designed research; F.M., N.P.S., and R.R.C. performed research; F.M., N.P.S., and R.R.C. contributed new reagents/analytic tools; F.M., N.P.S., R.R.C., J.N.O., and P.G.W. analyzed data; and F.M., N.P.S., R.R.C., J.N.O., and P.G.W. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: pwolynes@rice.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413575111/-/DCSupplemental.

simulating a Potts model with the Potts variables representing the possible amino acid types being placed on the sites in a representative average native structure, because 3D native structure is largely conserved (9, 17). At the same time, various averages over the allowed natural sequence variation can also be found using mean field theory (9). Thus, from a sufficiently reliable physical energy function, one should be able to predict the distribution of amino acids and the local evolutionary sequence entropy at any given location in the protein. The covariance of the amino acids in the sequence at different loci also would follow from this Potts model (Fig. 1).

By comparing such predictions starting from a physical energy function with the same quantities inferred from observed phylogenetic sequence comparisons, the strength of the selection, as quantified by T_{sel}, can be found. A statistically more robust way of making the quantitative connection between the physical and evolutionary landscapes is to use inverse statistical mechanics algorithms to infer an information theoretic energy function for the sequence distribution specifically for a single structural family. Such an inferred energy function has been used to predict the common backbone structure of a protein family starting from multiple sequences (18, 19). The energy gaps between random globules and native states from the information theoretic single-family Hamiltonian can then be compared with the gaps for the same native structure that are found from the physical, transferable, free energy function that is ordinarily used for structure prediction.

If the transferable energy function used is sufficiently close to the actual free energy function under which the proteins have evolved, the gap comparison allows us to quantify the strength of the folding constraint. In some cases, we can also estimate T_{sel} in a different way by computing the mutational stability changes predicted by the information theoretic energy function and comparing the predicted changes directly with measured values. This alternate approach provides a consistency check on the results from the transferable predictive energy landscape, which can be used more widely.

Simple statistical models of protein energy landscapes have been very helpful in understanding folding (3, 20, 21). Similar statistical ideas can also be used to describe protein evolution under the assumption that thermodynamic stability of the native state is the dominant constraint (22). To make this paper self-contained, in *SI Text*, we review the corresponding statistical analyses of folding and evolution based on uncorrelated energy landscapes. (The energies of structurally related states of proteins are, in fact, correlated. The effects of adding pairwise correlations on the energy landscapes of random heteropolymers have been examined in ref. 23. The change in the estimate of the glass transition temperature on adding pairwise correlations was found to be small.) These analyses connect quantitatively the problems of evolving in sequence space on geological timescales and folding in configuration space on laboratory timescales, resulting in a relation between two physicochemical quantities—

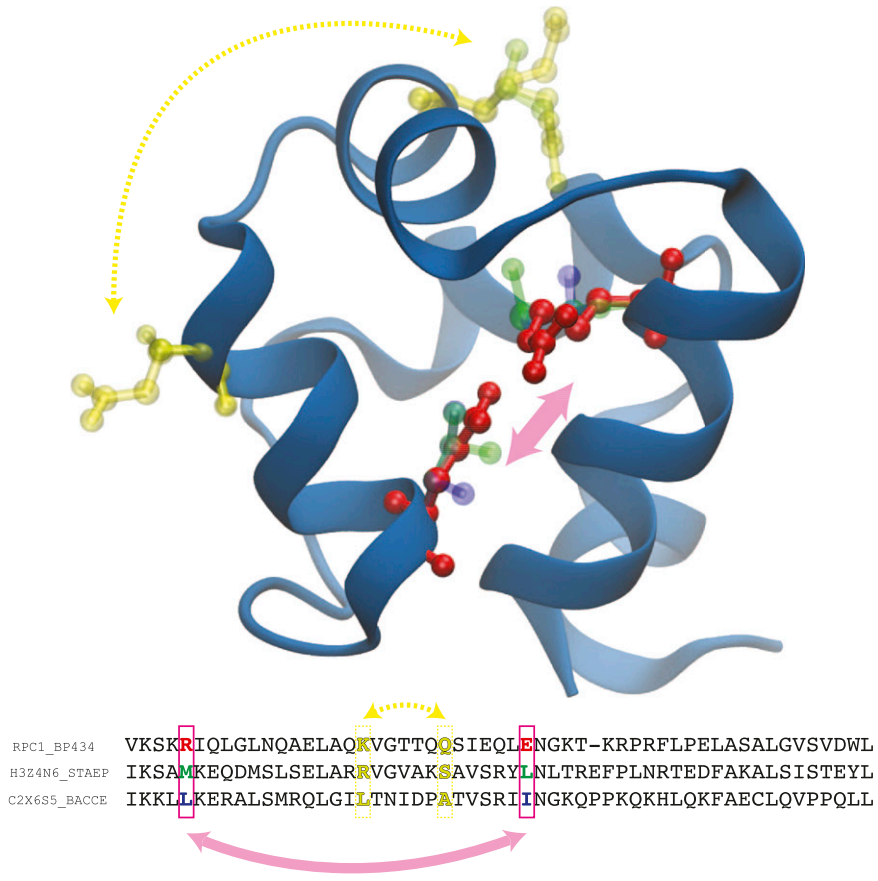

Fig. 1. Shown is the structure of the repressor protein CI (32) (Uniprot identifier: RPC1_BP434). Side chains from two additional sequences from the protein family are superimposed on the template structure. The multiple sequence alignment of the three proteins is also shown. DCA is able to quantify the statistical couplings between coevolving residues, which tend to be large when the residues are in contact and weak but potentially significant between residues that are spatially distant. Here, two examples of pairs of coevolving residues are shown: one example where the residues are in direct contact (pink arrows) and one example where they are not (yellow arrows). These pairs of residues have the same color in the 3D structure diagram as they do on the alignment.

the folding temperature T_f and the glass transition temperature T_g—with the evolutionary effective temperature T_{sel} (22):

$$\frac{2}{T_f T_{sel}} = \frac{1}{T_g^2} + \frac{1}{T_f^2}. \quad [1]$$

The folding temperatures from the experiment along with selection temperatures obtained by comparing physical and information theoretic Hamiltonians allow us to obtain T_g in absolute units as well as the dimensionless ratio T_f/T_g. The physical model used in this study assumes that the effective interactions between amino acid residues are temperature-independent, an approximation that breaks down because of solvent effects (24). It has been suggested, therefore, that these temperatures might be usefully interpreted as effective interaction strengths (25). Because of the temperature-dependent nature of intraprotein interactions, the T_g values given here should be understood as measures of landscape ruggedness related to the trap/decoy energy rather than precise determinations of experimental glass transition temperatures. Likewise, the dimensionless ratio of the folding temperature T_f to the glass temperature T_g measures how funneled the landscape is, with high values corresponding to nearly ideal funnels. The evolutionary inferred ratios turn out to be fairly close to the values inferred earlier based on purely physical arguments that set up correspondences between three-letter code lattice models and real proteins by making use of experimental information about residual structure and dynamics in the molten globules of helical proteins (26). The coevolution-based analysis suggests that protein landscapes are actually somewhat more funneled than was originally inferred. In some cases, the ratio approaches the higher estimates for T_f/T_g arrived at by two distinct sets of physical arguments: one set by Kaya and Chan (27) is based on observed high cooperativity of calorimetric folding transitions, and the other set by Clementi and Plotkin (28) is based on matching observed folding kinetics.

Results

We studied eight different protein families (defined by Pfam) (29). Each of these contains more than 4,500 sequences. All have at least one experimentally determined structure. The protein lengths range from 60 to 286 aa. Each of the families represents a distinct tertiary structure. A list of the specific proteins considered and their respective families is provided in Table S1.

For each family of proteins, we use direct coupling analysis (DCA) to infer a global statistical model for sequences in that family. DCA takes as input a multiple sequence alignment of sequences belonging to a single-protein family. Using a maximum entropy approach, DCA infers an effective energy function consisting of single-site fields and pairwise couplings that is able to approximately reproduce the empirically observed single-site and pairwise amino acid frequencies from the input sequence alignment. This energy function can also be used to estimate the probability (P_{DCA}) that an arbitrary sequence (not necessarily present in the input alignment) is part of the family. From this probability, a unitless energy can be defined by $H_{DCA} = log(P_{DCA})$. For the corresponding physical energy function, we use an estimate from a successful structure prediction model, $E = H_{AWSEM}$; to be precise, we use the energy of the sequence in its native structure according to the associative memory, water-mediated, structure, and energy model (AWSEM) (30). Details of how these quantities are computed are in *SI Text*. We use the probabilities and energies of random sequences having the composition of natural proteins as a reference state to get $\overline{E}$, and therefore, the selection temperature is obtained as a ratio of energy gaps from the two Hamiltonian values:

$$T_{sel} = \frac{-H^{nat}_{AWSEM} + H^{mg}_{AWSEM}}{k_B \log(P^{nat}_{DCA}/P^{mg}_{DCA})} = \frac{-H^{nat}_{AWSEM} + H^{mg}_{AWSEM}}{k_B (H^{nat}_{DCA} - H^{mg}_{DCA})}. \quad [2]$$

The *nat* superscript indicates that the quantity is evaluated for native sequences, whereas the *mg* superscript indicates that quantities are evaluated for random (molten globule) sequences: $H^{nat}_{DCA} - H^{mg}_{DCA} = \log(P^{nat}_{DCA}/P^{mg}_{DCA})$. We then perform a linear least squares fit to the combined set of native and random ordered pairs (H_{DCA} and H_{AWSEM}) to find the slope of the line and thus, T_{sel}. This formulation gives a single value of T_{sel} for each protein family. The result of this analysis is shown for the PDZ family [protein tyrosine phosphatase; Protein Data Bank (PDB) ID code 1GM1] (31) in Fig. 2*A*. H_{AWSEM} is plotted vs. H_{DCA} for 26,099 sequences in the PDZ family as well as an equal number of random sequences having amino acid compositions typical of natural proteins.

The global correlation between the two landscapes, one landscape obtained from a transferable energy function useful for structure prediction (AWSEM) and the other landscape inferred from coevolutionary information for each family (DCA), is high (R = 0.924 for the PDZ family; $\overline{R} \approx 0.9$, on average, across all eight protein families). The slope of the best fit line by Eq. **2** corresponds to a selection temperature of 124 K, well below the

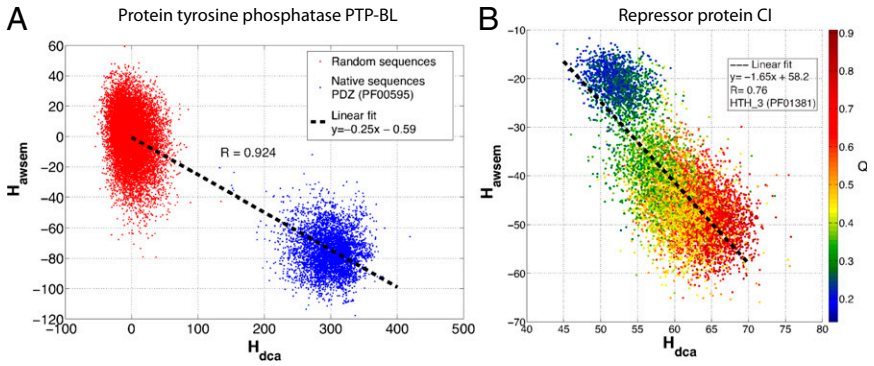

Fig. 2. (*A*) Correlation of H_{AWSEM} and H_{DCA}. The points corresponding to sequences in the PDZ family are shown in blue, and the points corresponding to an equal number of molten globule sequences are shown in red. The centers of the distributions are well-separated along both coordinates, indicating that both models are able to distinguish native sequences from molten globule sequences. The correlation coefficient between the two models is R = 0.924, indicating that, for these sets of sequences, the models are very well-correlated. The slope of this best fit line is −0.25, which corresponds to a selection temperature of 124 K. (*B*) Correlation between AWSEM and DCA Hamiltonian values for thermally occupied structures with different values of the fraction of native contacts formed Q, indicated by the color bar, from a molecular dynamics simulation of the Repressor protein CI (PDB ID code 1R69) (32) using the AWSEM potential. The two Hamiltonian values are highly correlated when evaluated over structures with a wide range of Q values.

folding temperature of most proteins. For seven other protein families, the estimated selection temperatures are also well below physiological temperature. These values indicate that the landscapes have evolved to be quite funneled and that specificity of structure and not mere stability plays an important part in selection. If T_{sel} were to equal the folding temperature, the landscapes would be rugged, and we would be forced to say that energy gap selection played no role in evolution. We also compared the two Hamiltonian values as a function of the fraction of native contacts, Q, for partially folded structures from a folding simulation of the repressor CI protein (PDB ID code 1R69) (32) sampled using AWSEM. Fig. 2*B* shows that the two landscapes are also highly correlated for the thermal ensembles ($R = 0.76$).

When mutational stability data are available, T_{sel} can be found without using the transferable energy function by comparing the mutational stability predictions from DCA with experimental data for single-site mutations ($\Delta\Delta G$). If we assume that the changes in the entropy and energy of the molten globule states are negligible for a single-point mutation, and assuming no residual structure in the denatured state, then scaling the energy change $\Delta(E-\overline{E})$ to be equal to $\Delta\Delta G_{exp}$ implies that $T_{sel} = -\Delta\Delta G / k_B \Delta H_{DCA}$. In this equation, $\Delta H_{DCA} = H_{DCA}^{(mutant)} - H_{DCA}^{(WT)}$. We can do this calculation for the PDZ family where experimental data exist and find $T_{sel} = 116$ K (Fig. S1). This estimated temperature agrees well with the value of 124 K obtained using the comparison with the transferable energy function, which we use for the other families.

Using the estimates for T_{sel} from the coevolutionary analysis for a family along with the experimental T_f for a member of each family, Eq. **1** yields estimates for T_g and thus, also T_f/T_g for typical family members. Fig. 3 summarizes the calculated T_{sel} and T_g values for all of the protein families studied here. Fig. S2 shows the dependence of the estimated value of T_{sel} on the distance threshold used to determine which pairwise interactions are to be summed in obtaining H_{DCA}. Fig. S3 also shows the pairwise distance dependance of the mean energy of a DCA residue–residue coupling. Fig. 4, *Upper* displays the T_f/T_g ratios for all protein families calculated using Eq. **1** when a distance threshold of 16 Å is used in calculating the DCA energy. Fig. S4 shows how the ratio of T_f/T_g depends on this distance threshold. The resulting T_f/T_g values are in the range of previous purely physicochemical estimates. The previous estimates were based on generic considerations for all proteins, but this approach yields T_f/T_g values for individual protein families. Another quantity that can be used

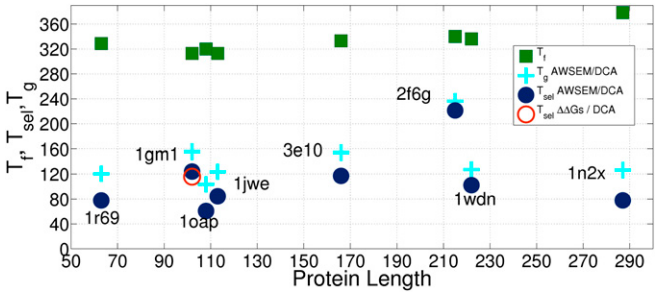

Fig. 3. T_{sel}, T_g, and T_f values in Kelvin for all protein families included in this study (denoted by the PDB ID codes of the representative structures used) are plotted vs. protein length. The names of the proteins and a list of references for the experimentally obtained T_f values are given in Table S1. The value of T_{sel} obtained by comparing stability changes predicted using DCA with experimental $\Delta\Delta G$ values directly is also given for the one family for which data are available (PDZ). In all cases, the experimental folding temperatures are above physiological temperature (~310 K), whereas the glass transition and selection temperatures are well below physiological temperature, indicating that selection of protein sequences by evolution leads to funneled folding landscapes for natural proteins.

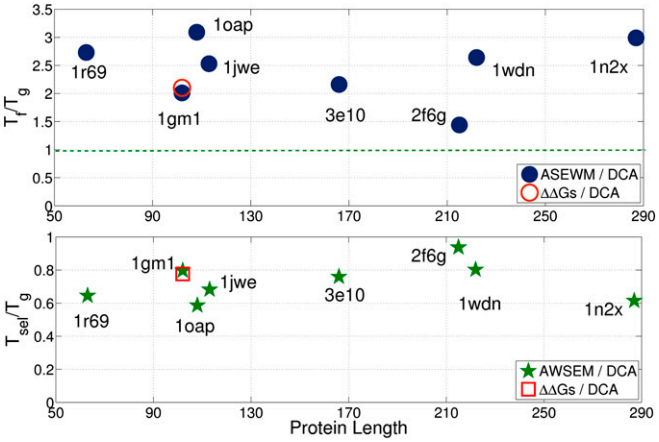

Fig. 4. (*Upper*) T_f/T_g ratios for all protein families studied. (*Lower*) T_{sel}/T_g ratios for all protein families studied. The families are denoted by the PDB ID codes of the representative structures used, and the names of the proteins are available in Table S1. T_f/T_g is used to quantify the degree of funnelness of a folding landscape, with higher values corresponding to more ideal funnels. The estimated T_f/T_g ratios for all natural protein families studied here fall above the threshold for a landscape to be considered funneled, $T_f/T_g = 1$, which is plotted as a green dashed horizontal line. Several of the estimates are clustered around the value of $T_f/T_g = 2.5$ estimated by Clementi and Plotkin (28). T_{sel}/T_g is used to quantify the degree of evolutionary optimization, with lower values corresponding to more highly optimized sequences. Most of the T_{sel}/T_g ratios for individual families are below the generic estimate of $T_{sel}/T_g = 0.85$ given by Pande et al. (21).

to quantify the degree of evolutionary sequence optimization is the ratio T_{sel}/T_g, which should be less than one for a funneled landscape. Pande et al. (21, 22) estimated T_{sel}/T_g by noting that the elements of the Miyazawa–Jernigan interaction matrix, being based on a quasichemical approximation, could be interpreted as pairwise interaction energies for pairs of amino acid types scaled by the selection temperature T_{sel} for all natural proteins considered as a single group (33). Combining this observation with the estimated entropy of the disordered collapsed globule inferred by Luthey-Schulten et al. (34) using the theory of secondary structure formation in globules, energy landscape theory arguments then give their estimated value of ~ 0.85 for T_{sel}/T_g for the set of all natural proteins (21). This estimate for T_{sel}/T_g leads to an estimate of $T_f/T_g = 1.6$, quite close to the value obtained in the work by Onuchic et al. (26) on purely physical grounds without using sequence information. Both the estimates by Onuchic et al. (26) and Pande et al. (21, 22) for T_f/T_g turn out to be on the low end of the values found here for individual families.

The energy gap between the folded and unfolded states and the folding temperature T_f also allows an estimate of the entropy of the unfolded state using the first-order transition equation $\Delta F(T_f) = \Delta E(T_f) - T_f \Delta S(T_f) = 0$. The resulting entropy per residue $S(T_f)/N$ for each family is given in Table S2. Most of these entropy values fall into the range of 0.7–1.1 k_B per residue, consistent with but a bit larger than the entropy estimates for the collapsed state by Luthey-Schulten et al. (34) that were used to give the original physical estimate for T_f/T_g ~ 1.66. We see that these estimates of the entropy of the unfolded state using coevolutionary data agree quite well with the earlier numbers for the two all α-helical proteins [repressor protein CI (PDB ID code 1R69) (32) and Dnab Helicase (PDB ID code 1JWE) (35)] but do tend to be somewhat higher for families with structures containing β-secondary structure elements.

Discussion

The early attempt by Onuchic et al. (26) to quantify the funneled nature of the landscape set up a correspondence between

the thermodynamics and dynamics of optimized two- and three-letter lattice model proteins and natural proteins. The T_f/T_g ratios found from coevolutionary analysis are higher than those first estimates. This difference suggests that evolution uses a (somewhat) more complex code than the three-letter coding that gave $T_f/T_g \sim 1.6$. Clementi and Plotkin (28) arrived at another purely physics-based estimate for T_f/T_g by asking how much a structure-based folding model, with a perfect funnel landscape, could be perturbed by the addition of nonnative interactions but nevertheless, recapitulate experimental kinetics that are usually consistent with nearly perfectly funneled landscapes, which are known to be well-predicted based on the idealized pure funnel limit (27). By tuning the strength of the nonnative interactions and calculating the corresponding folding and glass transition temperatures for this worst tolerable case, Clementi and Plotkin (28) determined that a degree of frustration corresponding to $T_f/T_g \sim 2.5$ would be a lower limit for maintaining consistency with the laboratory observations of kinetics of real proteins. Another estimate for T_f/T_g uses the fact that both theory and simulations agree that the degree of cooperativity in equilibrium folding depends on T_f/T_g. Noting this agreement and using experimental input about the sharpness of thermal unfolding, Kaya and Chan (27) estimated that the ratio T_f/T_g is probably greater than six for calorimetrically two-state proteins. These estimates, based on an optimized physical energy function and an information theoretic model for the global sequence probability derived from multiple sequence alignments, fall within the middle of the range of these previous physically based estimates.

DCA is a global statistical model for the sequences of a given protein family that allows the possibility of pairwise interactions between all residues in the protein, not just those pairs in physical contact in the native state. The correlation between experimental $\Delta\Delta G$ values with those predicted by DCA is best when interactions between residues separated by up to 16 Å in the native state are included (see Fig. S5). This distance is beyond the range of the mediated contacts used in AWSEM (9.5 Å). One possible explanation of this correlation from apparently long-range interactions is that DCA is not perfect in finding the true direct interactions, because it is based on statistical mechanical approximations and not an exact solution of the sequence Potts model, which is currently computationally intractable. At the same time, we must entertain the notion that these distant interactions are not artifacts but are real.

Several studies note that current force field-based methods for predicting $\Delta\Delta G$ on mutation using fixed backbones suffer from mediocre performance. We found that ΔH_{AWSEM}, like other fixed backbone methods, correlates reasonably well but not perfectly with a large database of experimental $\Delta\Delta G$ data (Fig. S6). DCA is a fold-specific model of the energy and therefore, poised to detect forms of energy that are particular to the symmetry-broken native state, which much like a crystal responds to interstitials, can respond collectively to site mutations. Elastic effects coming from harmonic deviations of the structures of a particular protein from the mean family structure may, thus, be important. If so, predicting the effect of mutations on the relative stability of the folded and unfolded states starting from any fixed backbone structure will be inadequate. The limitations of the fixed backbone approximation in predicting the natural covariation of amino acids have recently been noted (36). The long-range interactions inferred from DCA may be relics of these elastic effects. If so, such elastic effects may be crucial to correct the prediction of the effects of mutation on protein stability when using even highly accurate coarse-grained potentials. It is also possible that DCA captures mutational changes of residual structure in the denatured state, a possibility neglected by the assumed complete mixing approximation for the unfolded compact states. All of these effects could potentially contribute to the high correlation between DCA and experimental $\Delta\Delta G$ data; comparisons of the correlations of both DCA ($R = 0.84$) and AWSEM ($R = 0.73$) with experimental $\Delta\Delta G$ data for the PDZ family are shown in Fig. S1.

We have shown that genomic data, accurate coarse-grained free energy functions, and family-specific information theoretic models can be combined to give consistent estimates of energy landscape characteristics of natural proteins. These estimates invariably indicate that the energy landscapes of natural foldable proteins are highly funneled. The degree of funneling found by these methods is consistent with previous estimates based on general physicochemical considerations. Comparing the details of the physical and information theoretic models has already suggested ways of improving the prediction of mutational effects on the stability of protein sequence/structure pairs. Knowing the degree of funneling of natural proteins will be helpful to protein design practitioners who wish to mimic natural proteins (37). Additional application, development, and comparison of physical and information theoretic models of protein energy landscapes will greatly enhance our understanding of these critical biological macromolecules and the part that folding physics has played in their evolutionary history.

ACKNOWLEDGMENTS. This research was supported by National Science Foundation INSPIRE Award MCB-1241332, National Institutes of Health Grant R01 GM44557, and the Center for Theoretical Biological Physics sponsored by National Science Foundation Grants PHY-1427654 and MCB-1214457 and the Cancer Prevention and Research Institute of Texas. Additional support was also provided by the D. R. Bullard-Welch Chair at Rice University.

1. Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA* 96(19): 10689–10694.
2. Zeldovich KB, Shakhnovich EI (2008) Understanding protein evolution: From protein physics to Darwinian selection. *Annu Rev Phys Chem* 59(2008):105–127.
3. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA* 84(21):7524–7528.
4. Ramanathan S, Shakhnovich E (1994) Statistical mechanics of proteins with evolutionary selected sequences. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 50(2):1303–1312.
5. Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. *Science* 267(5204):1619–1620.
6. Mirny LA, Abkevich V, Shakhnovich EI (1996) Universality and diversity of the protein folding scenarios: A comprehensive analysis with the aid of a lattice model. *Fold Des* 1(2):103–116.
7. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14(1): 70–75.
8. Finkelstein AV, Badretdinov AY, Gutin AM (1995) Why do protein architectures have boltzmann-like statistics? *Proteins: Struct Funct Bioinform* 23(2):142–150.
9. Saven JG, Wolynes PG (1997) Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J Phys Chem B* 101(41):8375–8389.
10. Meyerguz L, Grasso C, Kleinberg J, Elber R (2004) Computational analysis of sequence selection mechanisms. *Structure* 12(4):547–557.
11. Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc Natl Acad Sci USA* 102(31):10930–10935.
12. Lovell SC, Robertson DL (2010) An integrated view of molecular coevolution in protein–protein interactions. *Mol Biol Evol* 27(11):2567–2575.
13. Fersht A (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Macmillan, New York).
14. Yomo T, Saito S, Sasai M (1999) Gradual development of protein-like global structures through functional selection. *Nat Struct Mol Biol* 6(8):743–746.
15. Süel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Mol Biol* 10(1):59–69.
16. Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep* 8(8):737–742.
17. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5(4):823–826.
18. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109(26):10340–10345.
19. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30(11):1072–1080.
20. Plotkin SS, Wang J, Wolynes PG (1997) Statistical mechanics of a correlated energy landscape model for protein folding funnels. *J Chem Phys* 106(7):2932–2948.
21. Pande VS, Grosberg AY, Tanaka T (2000) Heteropolymer freezing and design: Towards physical models of protein folding. *Rev Mod Phys* 72(1):259–314.

22. Pande VS, Grosberg AY, Tanaka T (1997) Statistical mechanics of simple models of protein folding and design. *Biophys J* 73(6):3192–3210.
23. Plotkin SS, Wang J, Wolynes PG (1996) Correlated energy landscape model for finite, random heteropolymers. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 53(6):6271–6296.
24. Shimizu S, Chan HS (2000) Temperature dependence of hydrophobic interactions: A mean force perspective, effects of water density, and nonadditivity of thermodynamic signatures. *J Chem Phys* 113(11):4683–4700.
25. Chan HS, Zhang Z, Wallin S, Liu Z (2011) Cooperativity, local-nonlocal coupling, and nonnative interactions: Principles of protein folding from coarse-grained models. *Annu Rev Phys Chem* 62(2011):301–326.
26. Onuchic JN, Wolynes PG, Lutheyschulten Z, Socci ND (1995) Toward an outline of the topography of a realistic protein-folding funnel. *Proc Natl Acad Sci USA* 92(8):3626–3630.
27. Kaya H, Chan HS (2000) Polymer principles of protein calorimetric two-state cooperativity. *Proteins* 40(4):637–661.
28. Clementi C, Plotkin SS (2004) The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Sci* 13(7):1750–1766.
29. Finn RD, et al. (2010) The pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211–D222.
30. Davtyan A, et al. (2012) AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J Phys Chem B* 116(29):8494–8503.
31. Walma T, et al. (2002) Structure, dynamics and binding characteristics of the second pdz domain of ptp-bl. *J Mol Biol* 316(5):1101–1110.
32. Mondragn A, Subbiah S, Almo SC, Drottar M, Harrison SC (1989) Structure of the amino-terminal domain of phage 434 repressor at 2.0 a resolution. *J Mol Biol* 205(1):189–200.
33. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* 18(3):534–552.
34. Luthey-Schulten Z, Ramirez BE, Wolynes PG (1995) Helix-coil, liquid crystal, and spin glass transitions of a collapsed heteropolymer. *J Phys Chem* 99(7):2177–2185.
35. Weigelt J, Brown SE, Miles CS, Dixon NE, Otting G (1999) NMR structure of the n-terminal domain of e. coli dnab helicase: Implications for structure rearrangements in the helicase hexamer. *Structure* 7(6):681–690.
36. Ollikainen N, Kortemme T (2013) Computational protein design quantifies structural constraints on amino acid covariation. *PLoS Comput Biol* 9(11):e1003313.
37. Shakhnovich EI, Gutin AM (1993) A new approach to the design of stable proteins. *Protein Eng* 6(8):793–800.