UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Energy landscapes for protein folding, binding, and aggregation : simple funnels and beyond

Permalink

https://escholarship.org/uc/item/90t943q5

Author

Cho, Samuel Sung-Il

Publication Date

2007

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Energy Landscapes for Protein Folding, Binding, and Aggregation: Simple Funnels and Beyond

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Chemistry

by

Samuel Sung-Il Cho

Committee in charge:

Professor Peter G. Wolynes, Chair
Professor Elizabeth Komives
Professor Katja Lindenberg
Professor J. Andrew McCammon
Professor José N. Onuchic

2007

Copyright ©

Samuel Sung-Il Cho, 2007

All rights reserved.

The dissertation of Samuel Sung-Il Cho is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2007

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Science in its purest form is completely objective, so one may think that it is devoid of any feeling. Thankfully, my graduate school experience has been filled with warm memories of many who continue to encourage and guide me, offering me far more opportunities to be successful than anyone deserves. In fact, I have come to think of my friends and teachers here as family members. My advisor, Peter Wolynes, was a "father" to me who took me under his wing by providing overwhelming support and guidance through my development into a mature scientist. Regrettably, there are sometimes indications that his work is not yet finished, but I like what I see so far. Above all, I cannot thank him enough for insisting on rigorous standards for performing research that often required looking at a far bigger picture than I was accustomed. My "big brother", Koby Levy showed me the ropes early on in my research, and I am so proud of the work we accomplished together once I got the hang of things. I am especially thankful for his sometimes unsolicited thoughts, especially the harshly critical ones that often turn out to be the most valuable ones. My other "big brother", Diego Ferreiro, often insists on referring to me as "sensei" because of our relationship when he first joined the group, but I certainly learned more from our collaborations. I am also grateful to my "mother", Katja Lindenberg, for providing sage advice and warm encouragement, beginning when I was her TA and continuing well beyond. Also, I have enjoyed the fruitful discussions from my "uncles" José Onuchic and Andy McCammon, and my "aunt" Betsy Komives, who turned out to be far more than just my thesis committee, sometimes even playing the role of coauthors. My interactions with my "immediate family" in the Wolynes and Onuchic groups, as well as my "relatives" in the Center for Theoretical Biological Physics and the Molecular Biophysics Training Program provided the multidisciplinary approach to science I always appreciated and pursued. Charlie Brooks and his group at TSRI, who are my "distant relatives", generously assisted me in setting up some of the all-atom MD simulations that complemented my studies, and I am

honored for being included in their “CHARMM family”.

Of course, I thank my actual family who has patiently encouraged me throughout my studies. I have always been amazed by the strong work ethic both of my parents modeled for me to follow. My brother, John, has always been there through all the good times, but I am most grateful that he firmly stood with me through the difficult times as well. Finally, I thank my wife, Sandra, for her unbending support and love while insisting that I set very high standards for myself. My research may have been possible without you, but it would be meaningless without you.

Chapter 2 is reprinted from Cho SS, Levy Y, Wolynes PG. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. Proc. Natl. Acad. Sci., USA. 2006, 103: 586-591.

Chapter 4 is reprinted from Cho SS, Levy Y, Onuchic JN, Wolynes PG. Overcoming residual frustration in domain-swapping: The roles of disulfide bonds in dimerization and aggregation. Phys. Biol. 2005, 2: S44-S55.

Chapter 5 is reprinted from a manuscript in press from Cho SS, Levy Y, and Wolynes PG Quantitative Criteria for Native Energetic Heterogeneity Influences in the Prediction of Protein Folding Kinetics. Proc. Natl. Acad. Sci., USA. 2007.

VITA

April 24, 1977	Born, Washington, DC, USA
2001	University of Maryland, Baltimore County B.S., Biochemistry and Molecular Biology B.S., Computer Science
2003	University of California, San Diego M.S., Chemistty
2007	University of California, San Diego Ph.D., Chemistry (Physical)

PUBLICATIONS

1. Yang SC, **Cho SS**, Levy Y, Cheung MS, Levine H, Wolynes PG, Onuchic JN. Domain swapping is a consequence of minimal frustration. Proc. Natl. Acad. Sci., USA., 2004, 101 (38): 13786-13791.

2. Levy Y, **Cho SS**, Onuchic JN, Wolynes PG. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. J. Mol. Biol., 2005, 346 (4): 1121-1145.

3. Levy Y, **Cho SS,** Shen T, Onuchic JN, Wolynes PG. Symmetry and frustration in protein energy landscapes: A near degeneracy resolves the Rop dimer-folding mystery. Proc. Natl. Acad. Sci., USA., 2005, 102 (7):2373-2378.

4. **Cho SS**, Levy Y, Onuchic JN, Wolynes PG. Overcoming residual frustration in domain-swapping: The roles of disulfide bonds in dimerization and aggregation. Phys. Biol. 2005, 2: S44-S55.

5. Ferreiro DU, **Cho SS**, Komives EA, Wolynes PG. The energy landscape of modular repeat proteins: Topology determines folding mechanism in the ankyrin family. J. Mol. Biol. 2005, 354: 679-692.

6. **Cho SS**, Levy Y, Wolynes PG. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. Proc. Natl. Acad. Sci., USA. 2006, 103: 586-591.

7. Ferreiro DU, Cervantes CF, Truhlar SME, **Cho SS**, Wolynes PG, Komives EA. Stabilizing IκBα by 'consensus' design. J. Mol. Biol. 2007, 365:1201-1216.

8. **Cho SS**, Levy Y, Wolynes PG. Quantitative Criteria for Native Energetic Heterogeneity Influences in the Prediction of Protein Folding Kinetics. Proc. Natl. Acad. Sci., USA. 2007. (in press)

ABSTRACT OF THE DISSERTATION

Energy Landscapes for Protein Folding, Binding, and Aggregation: Simple Funnels and Beyond

by

Samuel Sung-Il Cho

Doctor of Philosophy in Chemistry

University of California, San Diego, 2007

Professor Peter Wolynes, Chair

The Funneled Energy Landscape Theory is currently the most widely accepted theory of protein folding. In this dissertation, the basic concepts of the Energy Landscape Theory are introduced, highlighting some of its major successes in the studies of protein folding and binding kinetics. In particular, the focus is on an idealized native-topology based (Go-type) model that corresponds to a perfectly funneled energy landscape. This simple model has proven to accurately predict the folding mechanism of many proteins, even when simplifying approximations are made.

While there exists much evidence that models based on perfectly funneled energy landscapes are sufficient in many cases, there are indications that in other cases the idealized view needs some added complexity to faithfully represent folding and binding mechanisms. By exploring experimentally studied systems where the simplest Go-type models are insufficient, new paradigms and concepts add to our current understanding of protein folding, binding, and aggregation.

1. Theory of Protein Folding

The question of how proteins fold into a well-defined native state by discriminating against countless alternatives is widely recognized to be one of the most important unsolved mysteries in science. How do proteins find its folded state in a relatively short time? This seemingly simple question is particularly perplexing when one considers the inherent complexity of protein molecules. Assuming that the search for the folded state is random, one can naively estimate that even the simplest proteins should reach the folded state from the unfolded state on a timescale of an astronomical number of years when in fact real proteins fold far more quickly. Of course, Nature has already solved the "protein folding problem", and She continues to do so in an incredible fraction of that time, on the order of milliseconds to days. The real problem lies in our own difficulties in understanding how She repeatedly accomplishes this incredible feat so elegantly and simply.

Currently, the leading theory of protein folding is that natural proteins have evolved to have sequences that result in energy landscapes that are globally directed or "funneled" towards a uniquely structured folded state, while generally discriminating against misfolded states (1). The folding process is not a sequential list of requisite steps, but rather, there are many different ways that an unfolded protein can reach the native state through many competing pathways (Figure 1-1). The Energy Landscape Theory is the result of both rigorous theoretical analysis and careful experimental observations, and there is now much evidence to support these basic ideas. Recent work strongly suggests

that even binding mechanisms are encoded into the sequence of proteins (2-4).

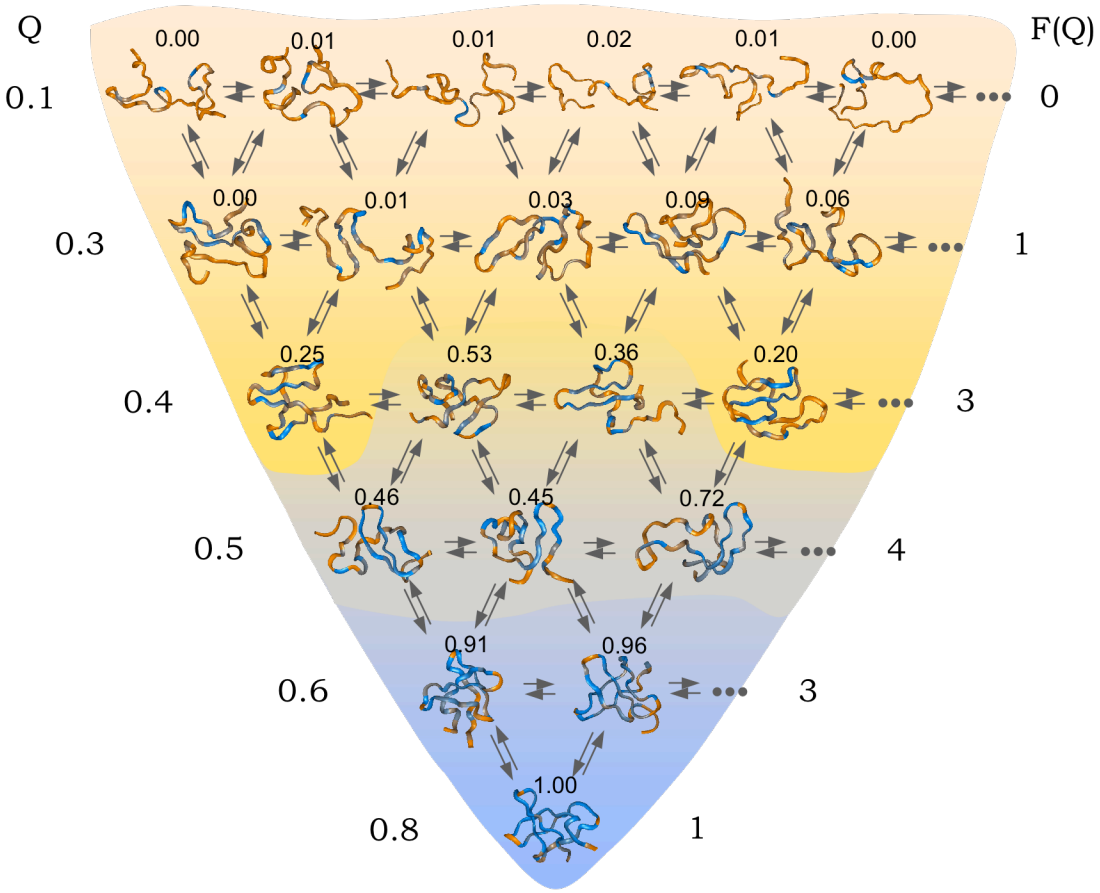


Figure 1-1: A schematic representation of the ordering of proteins in a funneled energy landscape. The structures of c-src SH3 with varying Q are colored according to the degree of local structural order, Q_i, in residue i, ranging from low Q_i (orange) to high Q_i (blue). The P_{fold} of each structure 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

1.1 The Dark Ages of Protein Folding

When the term "protein" was first coined in 1838, the importance of these substances was immediately recognized, largely because of their close connections with

life processes (5). In fact, the word "protein" comes from the Greek word "protos", which means "primary" (5). Even today, it is difficult to refrain from being in awe of how proteins dictate almost every aspect of life processes. Proteins are the building blocks of the cytoskeleton that gives cells their shapes, they dramatically increase the rate of biochemical reactions, and they control genetic transcription, just to name a few roles. Indeed, proteins are critical components of almost all cellular structure and processes. The failure of a given protein to properly fulfill its function(s) can lead to a host of diseases and even death. Further, proteins can sometimes fold into an incorrect structure, leading to Alzheimer's, Parkinson's, and variant Creutzfeldt-Jacob disease (6). The importance of understanding the fundamental framework of protein folding cannot be overstated.

We have long known that most proteins must fold into a specific and unique three-dimensional native conformation to perform their functions. In 1961, Christian Anfinsen demonstrated that when ribonuclease became unfolded via denaturation, it folds back into its native conformation and preserves its enzymatic activity without the assistance of any helpers (7). That is, the determinants of protein folding are encoded in the protein sequence itself. Therefore, the process of protein folding can be naively described as a thermodynamic search for a given protein's most stable structure. It was later discovered by others that some proteins are unable to independently find their native states, so they require the assistance of chaperones that prevent misfolding (8). There are still others that seem to have multiple relatively stable states, where one of the states

leads to aggregation and disease (9). However, it is generally accepted that most proteins falls into the first category.

Although proteins are able to assume their native states in a short time, this process was feared to be hopelessly complex to understood from a simple theoretical framework. This notion is reflected in what is now known as "Levinthal's Paradox". To point out the impossibility of a random search in protein folding, in 1969 Cyrus Levinthal posited an argument similar to the following (10). Let us take a relatively small protein of about 100 amino acids, with about 10 accessible states per amino acid. That would mean that the total number of accessible states of the protein is 10^{100}. Even if each state takes 1 ps to access, it will take about 10^{81} years (much, much longer than the age of the universe!) for the protein to randomly search its conformational space and find its lowest energy state. Obviously, proteins do not take anywhere near that long to fold or life itself would never have existed. Therefore, the solution to Levinthal's Paradox is that protein folding is not a random search, but a directed and biased one. But how?

1.2 Funneled Energy Landscape Theory of Protein Folding

In the late 1980's, Wolynes and co-workers introduced a well-defined possible solution to Levinthal's paradox (11, 12). Through careful mathematical analyses and incorporation of general experimental observations of protein folding, Bryngelson and Wolynes introduced the "principle of minimal frustration", which states that natural

proteins are the product of evolutionary selection such that the amino acid residues interact in such a way as to be globally attracted toward the native state (11). Any interaction that does not contribute towards the native state (i.e. non-native interactions) is mostly repulsive, providing primarily a frictional influence. The non-native interactions that compete with the native interactions are said to be frustrated. By minimizing frustration and thereby pruning frustrated interactions, evolution has generally selected against trap states that do not correspond to the native state. A minimally frustrated protein sequence has an energy landscape that resembles a partially rugged funnel with a single lowest energy basin that corresponds to the native state (Figure 1-2a), resulting in relatively fast folding kinetics as expected in natural proteins. In such an energy landscape, if any misfolded trap state exists, the energy for the misfolded state is high and the barrier to leave the trap state is very small. On the other hand, a frustrated amino acid sequence (i.e. a random sequence of amino acids without the benefit of evolutionary selection) has an energy landscape with multiple minima with energies close to one another (Figure 1-2b). Due to the competition between each of the low energy minima with high barriers between them, the protein with the frustrated sequence will exhibit slow or "glassy" folding kinetics. When the energetic frustration is effectively removed, the only remaining determinant of protein folding kinetics is the topology of the protein.

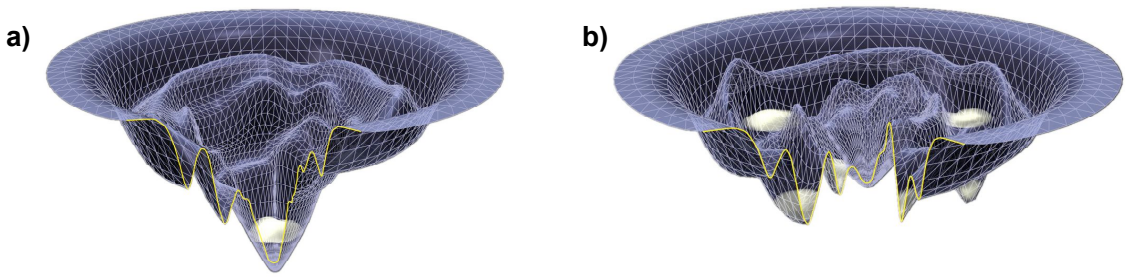


Figure 1-2: Schematic representations of protein energy landscapes. For natural proteins, the protein is globally attracted, or funneled, to a unique native state, leaving a single minimum whose energy is far less than all others (a). This is in contrast to a random sequence of amino acids where multiple minima have about equal energies, leading to misfolding and slow kinetics (b). By evolving sequences that lead to proteins with funneled energy landscapes, proteins very quickly reach the native state.

A model that quantitatively captures these Energy Landscape Theory ideas is a Go-type model (13), which takes the minimal frustration principle further by ideally assuming that proteins have no frustration. The hypothesis being tested is to determine whether the level of frustration present in proteins is negligible. In the model, every amino acid residue pair interacts favorably if the interaction is found in the native state, and they are repulsive (or not represented at all) otherwise, resulting in a perfectly funneled energy landscape with no trap states. The concept of a Go-type model is akin to the ideal gas law in physical chemistry in that it is a simplifying approximate representation of real systems. It is not clear whether it is even possible for any protein to have a sequence such that there are no frustrated interactions at all, but this approximation seems to be a good starting point in that it captures many of the qualitative, and even quantitative, properties of real proteins (13-15). In the original lattice Go model and the simplest off-lattice Go-type models, the energies of the

interacting residues were treated equally regardless of the sequence identity of the interacting residues, making this a purely a topological model in that only the structure of the native state is used as input to construct a Go-type model (13, 16). Structural homologues with nearly identical structures do not always exhibit similar mechanisms, so the validity of this approximation must have some significant limits, as we will explore in detail later (see Chapter 5). Further, the presence of misfolded intermediates in the folding mechanism is not represented at all. Despite the incredible simplicity, Go-type models have proven for numerous proteins to reliably predict whether the folding mechanism involves an intermediate (13), the folding rate (15), and oftentimes the transition state structure at a residue-level resolution (3, 14, 17), although not always (14).

Another representation of proteins, based on empirical force fields (18), is sometimes cited in the literature as being a more "realistic" representation of proteins because they include non-native interactions that Go-models inherently lack (19). Developed and optimized to study the dynamics of proteins in the native basin, empirical force fields are typically constructed by generating many parameters for use in a relatively simple energy function that is general for any sequence of amino acids. These parameters are predominantly derived from quantum mechanical calculations, but experimental quantities from IR or microwave spectra are used whenever possible (18). The approach is very sensible, and the resulting force fields are often unquestionably accepted as being an accurate representation of proteins that can generally be applied to

study the entire folding process. However, it is likely that these force fields have significantly more non-native interactions than are present in natural proteins, resulting in a rugged energy landscape with trap states that do not actually exist, even for simple peptides (20). It is further unknown whether all of the significant non-native interactions are captured by these force fields. At present, it is very difficult to ascertain how well these empirical force fields perform because simulations of protein folding are extraordinarily computationally intensive so computational protein folding studies using these force fields have been limited to small peptides or very small proteins. Further, these simulations rarely, if ever, are performed long enough to observe more than one transition, so calculating thermodynamic quantities is extraordinarily difficult (21). Also, it is not immediately clear how one would "fix" an error in a force field because many, if not all, of the parameters are dependent on each other, although it is not impossible. Recently, MacKerell, Feig, and Brooks systematically changed the phi and psi dihedral angles to reduce the transformations of α-helices to π-helices, which are extraordinarily rarely observed in nature but observed in simulations far more often than expected (22). Their work shows that a general improvement to the force field is possible, although it is certainly not trivial.

Still, there is some evidence that these force fields have some promise in predicting the folding mechanism of proteins. Pande and coworkers carried out the simulations of the folding of many "mini-proteins" (i.e., peptides comprised of less than 50 amino acids) using the Folding@Home distributed computing approach, which takes advantage of idle

computer time of many hundreds of thousands of volunteers (23). While folding simulations of small peptides (about 30-60 residues) have been carried out using this brute-force strategy with remarkable success in the prediction of folding rates (24), it is not clear whether the same can be done for larger proteins. Further, it is not yet clear how well these simulations reproduce the finer details of protein folding mechanisms. More studies are necessary to quantify how funneled empirical force fields are, and improvements must be made to remove unrealistic trap states. The empirical force field and energy landscape approaches are summarized in Figure 1-3. Clearly, no protein representation is flawless, and the limits of the various models must be appreciated. By making reasonable approximations that capture many of the qualitative and quantitative properties of protein folding processes, however, we can learn much, and a Go-type model is such an approach.

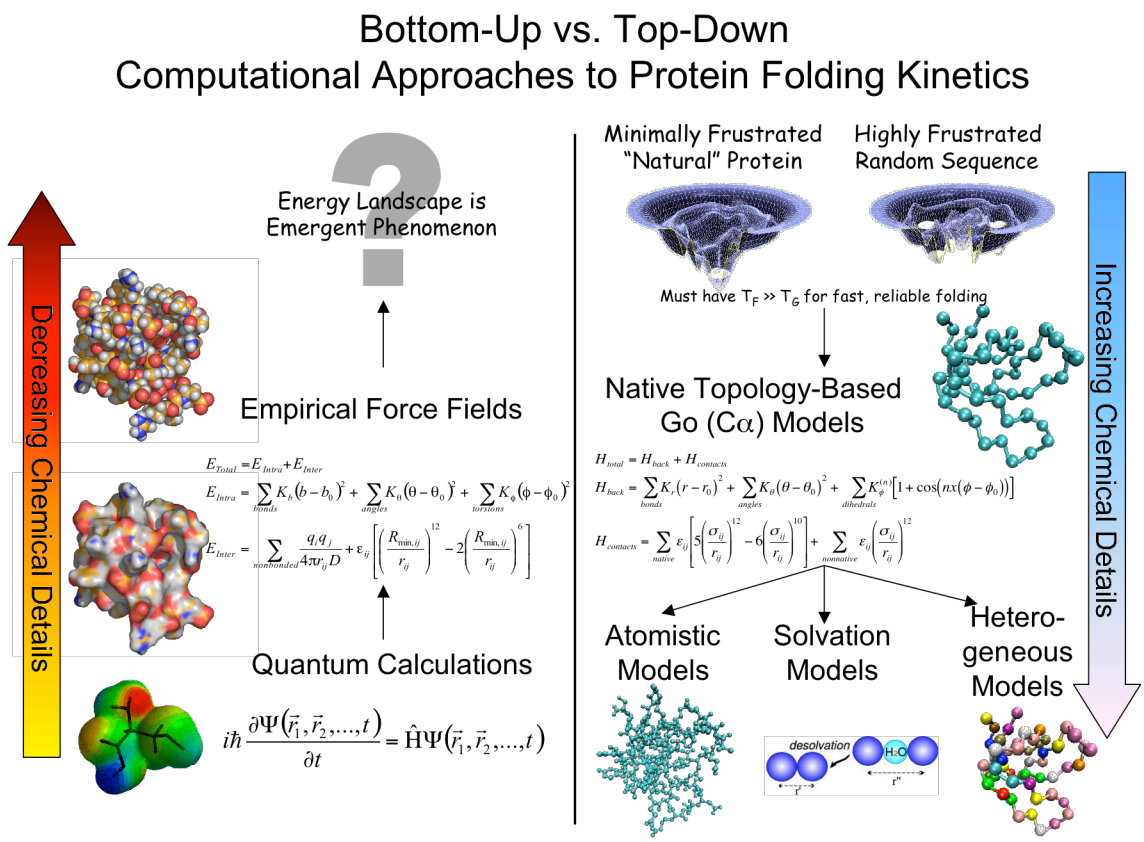


Figure 1-3: Two independent approaches to computationally study protein folding kinetics. In the "bottom-up" approach, one starts from detailed quantum calculations of each residue of a protein to develop parameters for use in a coarse-grained and simple energy function (left). Simulations from this approach yields an energy landscape that is emergent from first principles. Alternatively, if one can make a reasonable assumption about the energy landscape *a priori*, based on broad experimental observations and analytical arguments, a model that captures the basic principles of the energy landscape theory can be constructed. In the "top-down" approach, chemical details are progressively added to the simplest model until it captures the underlying physics to address the question at hand (right).

1.3 Perfectly Funneled Energy Landscapes: Go-type Models

Here, we describe a typical off-lattice C_α Go-type model, as was described by Clementi and coworkers (13). In the simplest variant, a single bead centered on the C_α

position represents a residue. Bond and angle potentials string together the beads to their neighbors along the protein chain. The dihedral potential encodes the secondary structure. The defining characteristic of a Go-type model is that the protein's native topology determines the network of favorable long-range tertiary interactions while all other non-bonded interactions are repulsive. The Go-type model Hamiltonian for a protein with configuration Γ is as follows (see also Figure 1-4):

$$H(\Gamma,\Gamma_0) = H_{backbone} + H_{nonbonded}$$

$$H_{backbone} = \sum^{bonds} K_r(r-r_0)^2 + \sum^{angles} K_\theta(\theta-\theta_0)^2 + \sum^{dihedrals} K_\phi^{(n)}\left[1-\cos(n(\phi-\phi_0))\right]$$

$$H_{nonbonded} = \sum_{i<j-3}^{native} \varepsilon_1(i,j)\left[5\left(\frac{\sigma_{i,j}^{nat}}{r_{i,j}}\right)^{12} - 6\left(\frac{\sigma_{i,j}^{nat}}{r_{i,j}}\right)^{10}\right] + \sum_{i<j-3}^{non-native} \varepsilon_2(i,j)\left(\frac{\sigma^{non}}{r_{i,j}}\right)^{12}$$

The K_r, K_θ, and K_φ are the force constants of the bonds, angles and dihedral angles, respectively. The r, θ, and φ are the bond lengths, the angles, and the dihedral angles, with a subscript zero representing the corresponding values taken from the native configuration, Γ_0. The non-bonded contact interactions, $H_{nonbonded}$, contain Lennard-Jones 10-12 terms for the non-local native interactions and a short-range steric repulsive term for the non-native pairs, corresponding to a perfectly funneled energy landscape. We chose as parameters of the energy function $K_r=100\varepsilon$, $K_\theta=20\varepsilon$, $K_\varphi^{(1)}=1.0\varepsilon$, $K_\varphi^{(3)}=0.5\varepsilon$. In the homogeneous representation of inter-residue interactions, every native interaction energy has the same value and thus $\varepsilon_1=\varepsilon_2=\varepsilon$. The interaction energies can alternatively be made sequence dependent by having the value of $\varepsilon_1$be dependent on the identity of the interacting residues, i and j, and thereby introduce native energetic heterogeneity. $\sigma_{i,j}^{nat}$ is

the distance between the C_α atoms of the residues (i,j) in the native configuration and $\sigma^{non} = 4.0\,\text{Å}$ for all non-native residue pairs. The network of native contact pairs was determined using the CSU (Contacts of Structural Units) software (25). Thermodynamic quantities are readily obtained by collecting multiple constant temperature simulations and analyzing them via the weighted histogram analysis method (WHAM). Despite the simplicity of this model, it has proven to accurately reproduce many qualitative and quantitative details of the folding mechanism. So, if one knows the final native state, we can quantitatively describe the many different paths that the protein can take to get there.

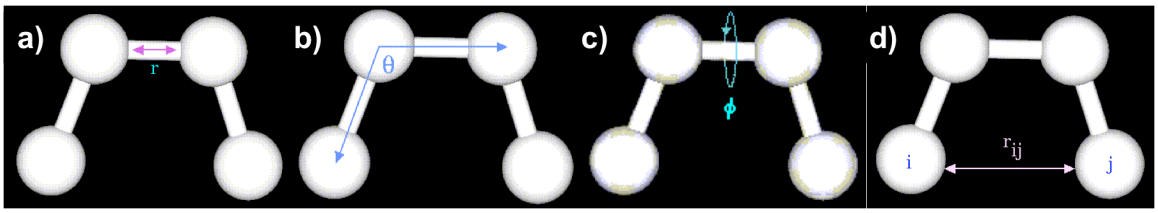


Figure 1-4: A schematic reflecting the different terms that define a Go-model. The short-range interaction terms include the bond, angle, and dihedral terms (a-c), which are defined by harmonic wells whose respective minimum correspond to their values in the native state. The long-range interactions between two residues that are in native contact (d) are described by a Lennard-Jones type 10-12 potential.

1.4 Folding on Perfectly Funneled Landscapes

Since the introduction of off-lattice Go-type models, simulations of many proteins have been carried out, and much has been learned about folding mechanisms. In particular, many of the general features of the protein folding process are captured from Go-type models. The unfolded and the folded states are separated by at least one free energy barrier that corresponds to a transition state (Figure 1-5). When using appropriate

reaction coordinates, the structure of the transition state can be characterized for comparison to experimental observables. Further, Go-type models for many proteins largely capture the rates of folding and the nature of any productive intermediate that may exist in the folding process. A complete survey of the successes (and limitations) of Go-type models would be very lengthy but we will highlight some of the major studies using Go-type models.

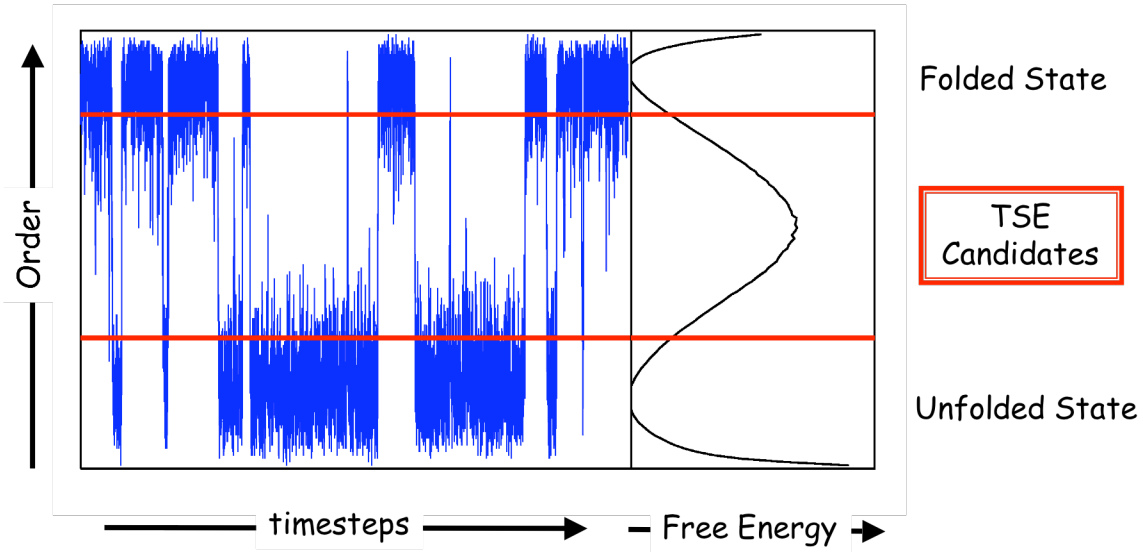


Figure 1-5: A typical trajectory from a Go-type model used in our studies (left) with the corresponding free energy profile generated using WHAM (right).

The success of any model representing the protein folding process must be measured not only by the qualitative observables but also quantitative ones as well. That the pattern of folding rates of many proteins can be well-predicted by Go-type models was demonstrated by a survey of many globular proteins undertaken by Takada and coworkers (14). Clementi and coworkers showed that the models not only discriminated

between two- and three-state folding proteins, but that the predicted structure of the transition state and intermediates were similar to those that were observed experimentally (13). When the structures of the transition state were characterized quantitatively by calculating Φ-values, which is at a residue-level resolution, it was clear that a substantial improvement in the agreement between simulations and experiments could still be made (14). Clearly, the purely additive model is limited in reproducing the transition state, and Plotkin and coworkers showed that the structure of the transition state as characterized by Φ-values agrees better with experiments when many-body interactions that implicitly include the effects of sidechains and waters are included (17).

It is important to note that Go-type models are not limited to proteins with compact geometry. While most Go-type simulations have been performed for globular proteins, Ferreiro and coworkers showed that Go-type models of linear ankyrin repeat-containing proteins also show remarkable success in reproducing and even predicting experimental observations (26), demonstrating that the simple topology-based model accurately captures the folding mechanism of these proteins just as well as that of globular proteins. While the models faithfully reproduced many experimental observations of the ankyrin repeat proteins' folding mechanism, perhaps their greatest success in the study was the prediction of an intermediate in the case of Notch ankyrin repeat domain that was later verified by experiment (27).

Despite the many successes, the simplest Go-type model cannot be expected to be without limitations. As such, one can improve upon the simplest variant by

introducing complexity into the model so that the underlying physics is better represented. We note two general directions that are not mutually exclusive. Karanicolas and Brooks introduced a heterogeneous Go model with energies of the long-range residue-residue interactions depending on the identity of the individual residues (28). Another direction was to increase the chemical detail of the protein representation by involving additional atoms (29, 30) or explicitly including water molecule interactions (31).

2. Identifying and Characterizing the Transition State Ensemble

Transition state theory (TST) is the simplest theory of predicting reaction rates for chemical reactions dating to Wigner. 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. The assumption that later recrossing events often can be considered negligible seems quite reasonable, 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. TST as a general concept is applicable to protein folding when it is a strongly cooperative process.

2.1 Transition State of Protein Folding

For natural proteins, the unfolded and folded states are usually 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, from dihedral angle barriers, and from forming adventitious non-native contacts, can exert forces on the reaction coordinates that alter the direction of motion. A protein may cross the transition state multiple times before reaching the folded state, as was analytically predicted by Bryngelson and Wolynes and later observed in simulations.

TST, therefore, overestimates the rate coefficient, which only counts the number of forward trajectories, neglecting any recrossing events (Figure 2-1a). 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 (Figure 2-1b). There is much evidence, however, that real proteins are far from this glassy limit. In the simplest case, folding kinetics can be interpreted 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. In a strictly two-state situation, the TSE would be reasonably defined by a single stochastic separatrix and corresponds to that set of structures that have an equal probability of first completing the folding process before unfolding to a completely denatured state.

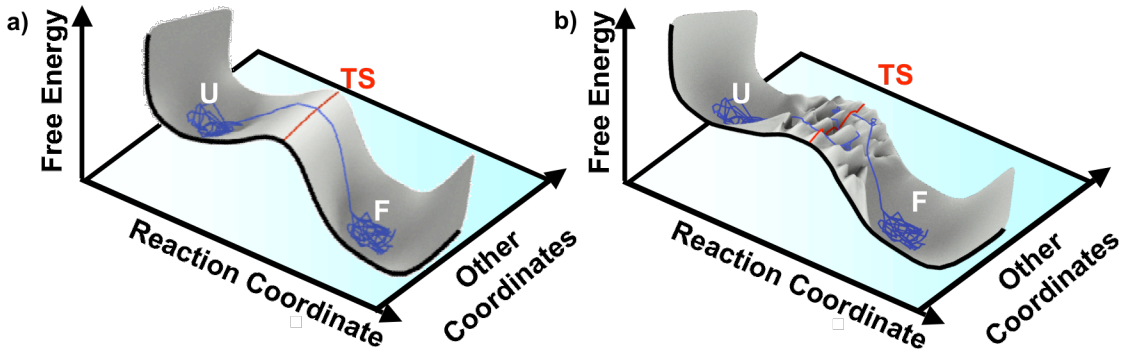


Figure 2-1: Schematics depicting two possible trajectories of protein folding: (a) a single crossing of the transition state as predicted by TST on smooth landscapes and (b) multiple crossings of the transition state in the limiting case of high friction due to ruggedness in a frustrated landscape.

2.2 P_{fold} as the definition of the TSE for Proteins

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 (32). 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 (Figure 2-2).

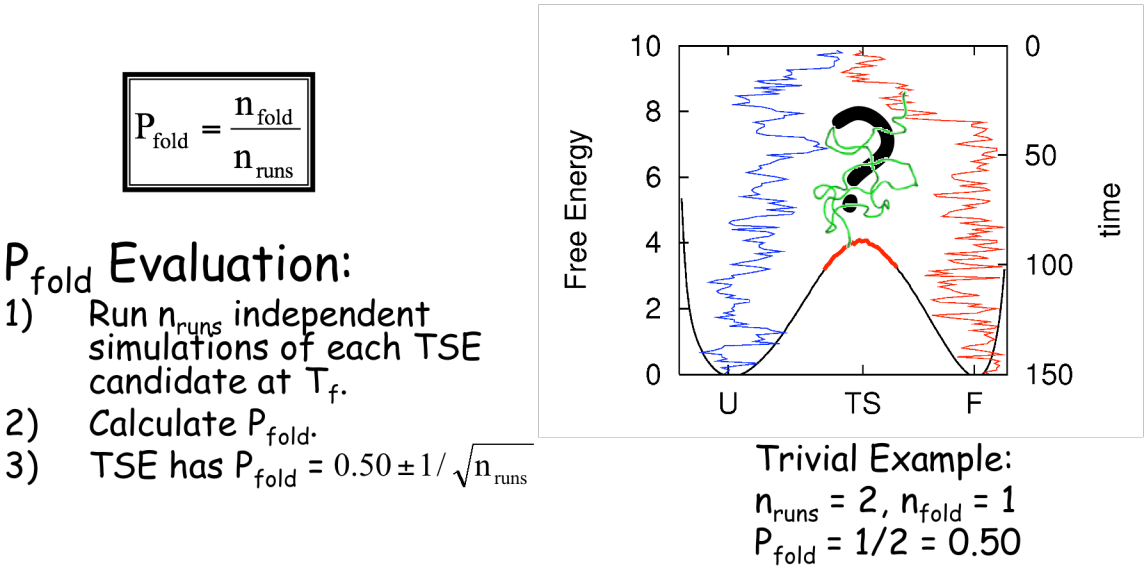


Figure 2-2: To ascertain whether a given structure is a member of the transition state ensemble, one computes its P_{fold} by running many independent simulations, each starting from the same conformation. Its progress is monitored to see whether it reaches the folded state before reaching the unfolded state. That probability is defined as P_{fold}, which is equal to 0.50 (within statistical error) if it is a member of the transition state ensemble.

While 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 longer than 100ns when using all-atom simulations with an empirical force field (33). The parallelizable nature of the problem has motivated the use of distributed computing approaches to carry out such computations (21). Even in the most rigorous studies carried out so far, however, the computation of P_{fold} is limited to a rather small set of conformations. Computing P_{fold} also requires the precise knowledge of T_f since P_{fold} is highly sensitive to temperature (Figure 2-3). Determining the value of T_f from simulations, however, is not always possible. In all-atom simulations of proteins it has been seldom 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. While approximating the folding temperature in a simulation with the laboratory value from experiments for a well-studied protein (33-35) may be acceptable, an arbitrary choice of temperature (36) is clearly inadvisable because the slightest deviation from T_f can significantly misplace the TSE.

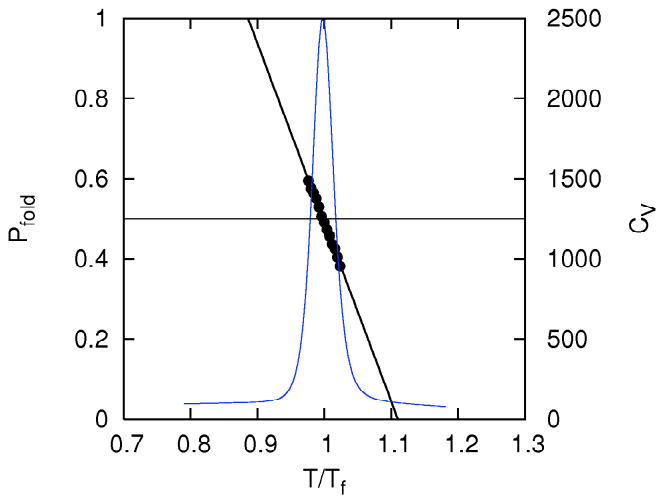


Figure 2-3: The temperature sensitivity of P_{fold}. The specific heat for the folding of c-src SH3 protein as a function of temperature in units of T_f is shown with the average P_{fold} of the same set of structures with the same set of initial velocities at different temperatures.

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. Presently, 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 (37). Thus, although P_{fold} identifies members of the TSE in a strict sense, the severe practical drawbacks of using P_{fold} 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 presently in this situation using P_{fold} has 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 et al., 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 (38). In keeping with Bryngelson and Wolynes's theory, they also showed that that when the landscape is glassy or frustrated, thermodynamic coordinates by themselves fail to describe the structural ensemble as measured by Φ-values (38). Thus general arguments and these specific results have encouraged the use of calculating Φ-values based on unfrustrated models (39, 40). Simulations based on unfrustrated landscapes using native-structure based reaction coordinates also predict many qualitative experimental observations of protein folding and binding (2, 3, 13). The predicted folding rates of many small proteins agree well with experimental observations (14, 15), and the Φ-values usually agree with experimental values (3, 17). 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 (12), some researchers have argued that structural reaction coordinates like Q, the fraction of native contacts, are inappropriate for describing the TSE even on unfrustrated landscapes (32, 41). They have argued *a priori* that structural

coordinates like Q will fail to identify the transition state, even for funneled landscapes (29, 42). Some further maintain that P_{fold} is the only reliable reaction coordinate for real proteins (43). Such extreme views conflict with numerous other studies showing that Q gives acceptable results as a reaction coordinate for model proteins (37, 38, 44). Studies on all-atom models show a clear correlation between P_{fold} and Q (35). We should recognize, however, that Q is not the only possible structural coordinate that can be used for kinetics. Shoemaker et al. showed that reaction coordinates measuring only a handful of contact areas function equally well, if they are chosen appropriately *post hoc* (45). Alternative structural quantities have also been used as reaction coordinates. <L>, the mean shortest path length has been reported to characterize the TSE better than other order parameters (46). 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 (47), indicates again that some *a priori* 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) (13, 48, 49). We will quantify rigorously the accuracy of native structure based reaction coordinates in describing the TSE as probed by Φ-value analysis (50). The TSE structures obtained using several different reaction coordinates are compared to 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 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.

2.3 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 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 1 k_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} (Figure 2-5a,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} (Figure 2-5a,b). While Q, on average, is able to identify the TSE, there do exist some structures in the Q ensembles whose P_{fold} lies 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 (51), 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 according to P_{fold} and Q are equivalent. We see then that in this exhaustive survey one cannot distinguish these ensembles in terms of pair structural patterns.

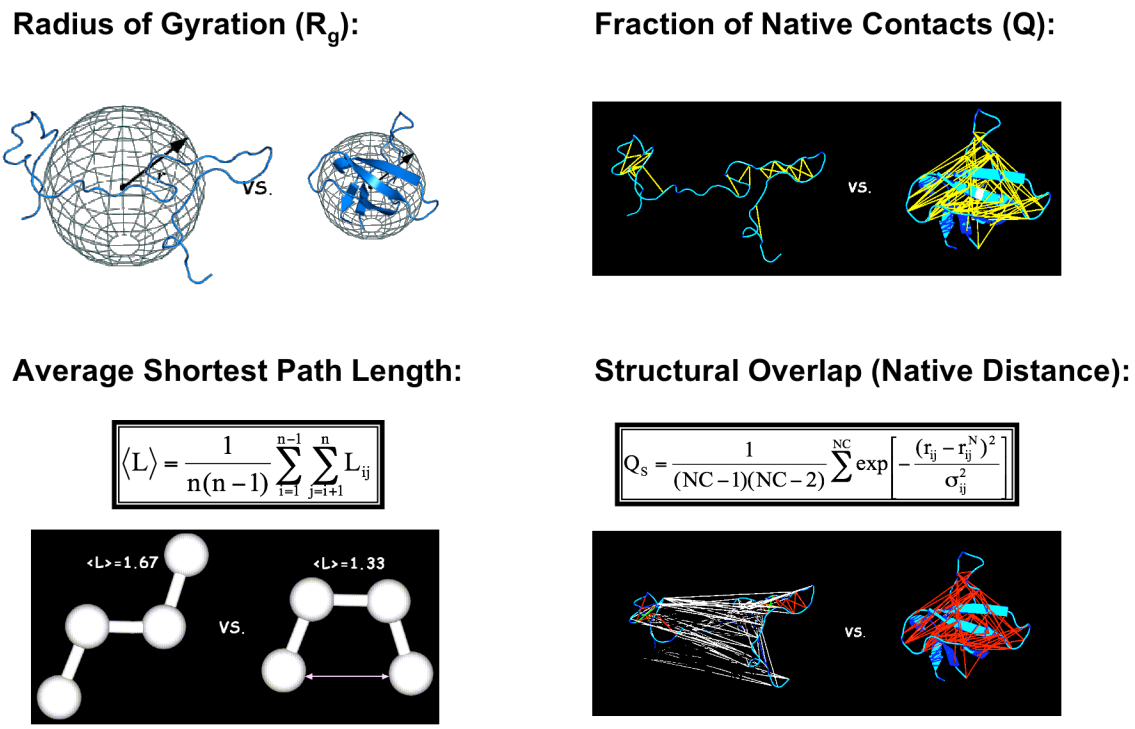


Figure 2-4: Structural order parameters selected for evaluation as reaction coordinates. The radius of gyration, (R_g) and fraction of native contacts (Q) are commonly used in the study of protein folding kinetics. The average shortest path length (<L>) has recently been cited in the literature as being better than Q. The measure of the structural overlap of the native distances (Q_S) is more precise than Q because it considers not only whether a native contact is made, it also imposes a Gaussian penalty as the interactions becomes far from its native distance.

We now compute the experimentally accessible quantities, Φ-values, according to the four chosen reaction coordinates, Q, Q_S, <L>, and R_g, and compare them to the Φ-values of the TSE as defined by P_{fold} (Figure 2-4). 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. 10c,d). For both proteins, the Φ-values as determined by the reaction coordinates Q, Q_S, and <L> agree strikingly well to those of the TSE described by P_{fold} with correlation coefficients around 0.90 and 0.95 for c-src SH3 and CI-2, respectively. The

slopes of the correlations are about 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 <L>). R_g, on the other hand, generally grossly underestimates the Φ-values. <L> turns out to be at best comparable to Q and Q_S when describing the TSE via Φ-value analysis, contrary to a previous suggestion (46). The difference between Q and Q_S is modest, as is reflected in their equivalent characterizations of the TSE.

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 previously reported to be around 0.60 (17). 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 (Figure 2-5e). We note that the difference between the correlation coefficients computed using P_{fold} and Q is 0.05. There is thus only a miniscule improvement of predictability 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 non-native 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 (33). Plotkin and coworkers have shown that the correlation between simulated Φ-values for CI-2 with experimental values is improved by including non-additive energetic terms, which arise from solvent and sidechain effects (17). Of course, such a non-additive model still corresponds to a perfectly funneled landscape. Our simulation model is purely additive, so this 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 (Figure 2-5f). 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 non-additive energetic terms in the model. We note that while supplementing the pairwise model with non-additive interactions increases the correlation between experimental and simulated Φ-values for many proteins, SH3 proves an exception showing almost no effect on Φ-values from increased non-additivity (17). 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.

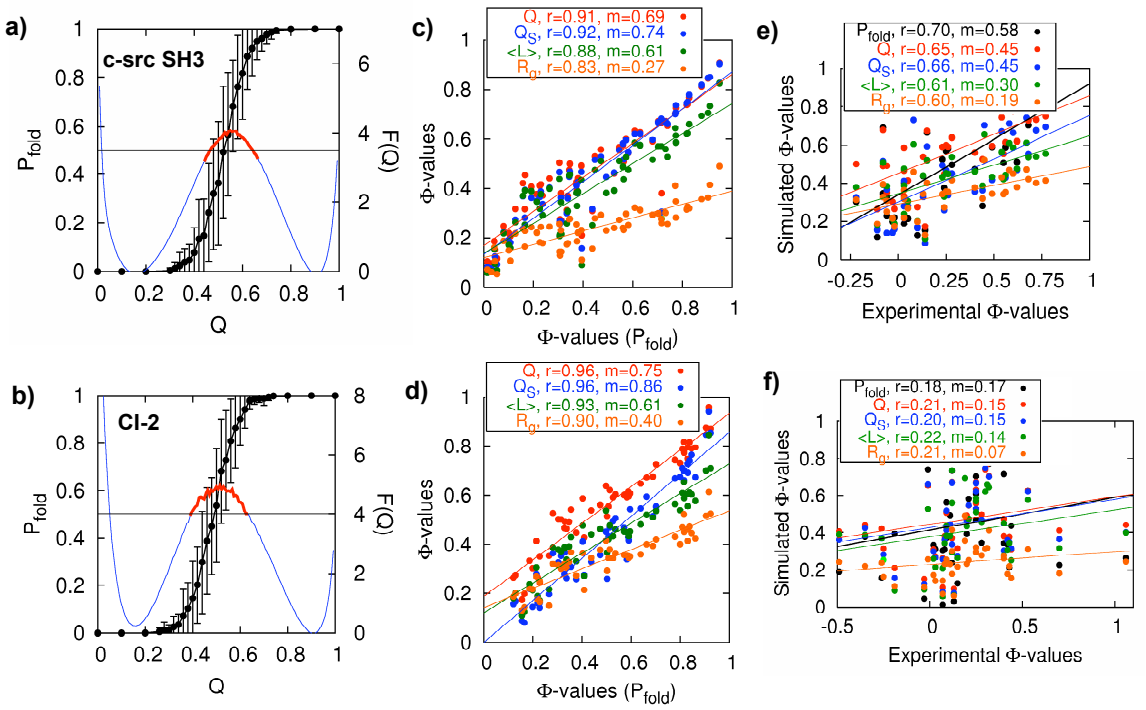


Figure 2-5: Comparing the TSE obtained from P_{fold} and the structural reaction coordinates of two-state folding proteins. (a and b) For both c-src SH3 (a) and CI-2 (b), the free-energy profile using Q as a reaction coordinate is overlaid with the average P_{fold} of structures (with error bars indicating 1 SD) over the range Q = 0.30–0.80. The putative TSE corresponds to P_{fold} =0.50 ± 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, <L >, 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.

2.4 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 (52). 3ANK is predicted by Q to fold by a two-state transition with a broad, asymmetrical free energy barrier (Figure 2-6a). Again, we found that P_{fold}=0.50

corresponds to the Q at the peak of the free energy profile (Figure 2-6a). This 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 state than the folded state. The Φ-values determined by the reaction coordinates Q, Q_S, and <L> agree with those of the TSE based on P_{fold} (Figure 2-6b). Why is the free energy barrier broad? To answer this, 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. that the folding nucleus of ankyrin repeat proteins corresponds to approximately 1 ½ repeats (26). The resulting free energy profile exhibits a competition between a N-terminal nucleating two-state transition and a C-terminal nucleating three-state transition (Figure 2-6c). In a recent experimental study, a 3-ankyrin repeat protein with a similar sequence to 3ANK exhibited equilibrium intermediates (3-state folding mechanism) at high temperatures but not at low temperatures (2-state folding mechanism) (53). The discrepancy in the observed folding behaviors can be rationalized by a competition of folding mechanisms similar to that found in the simulations.

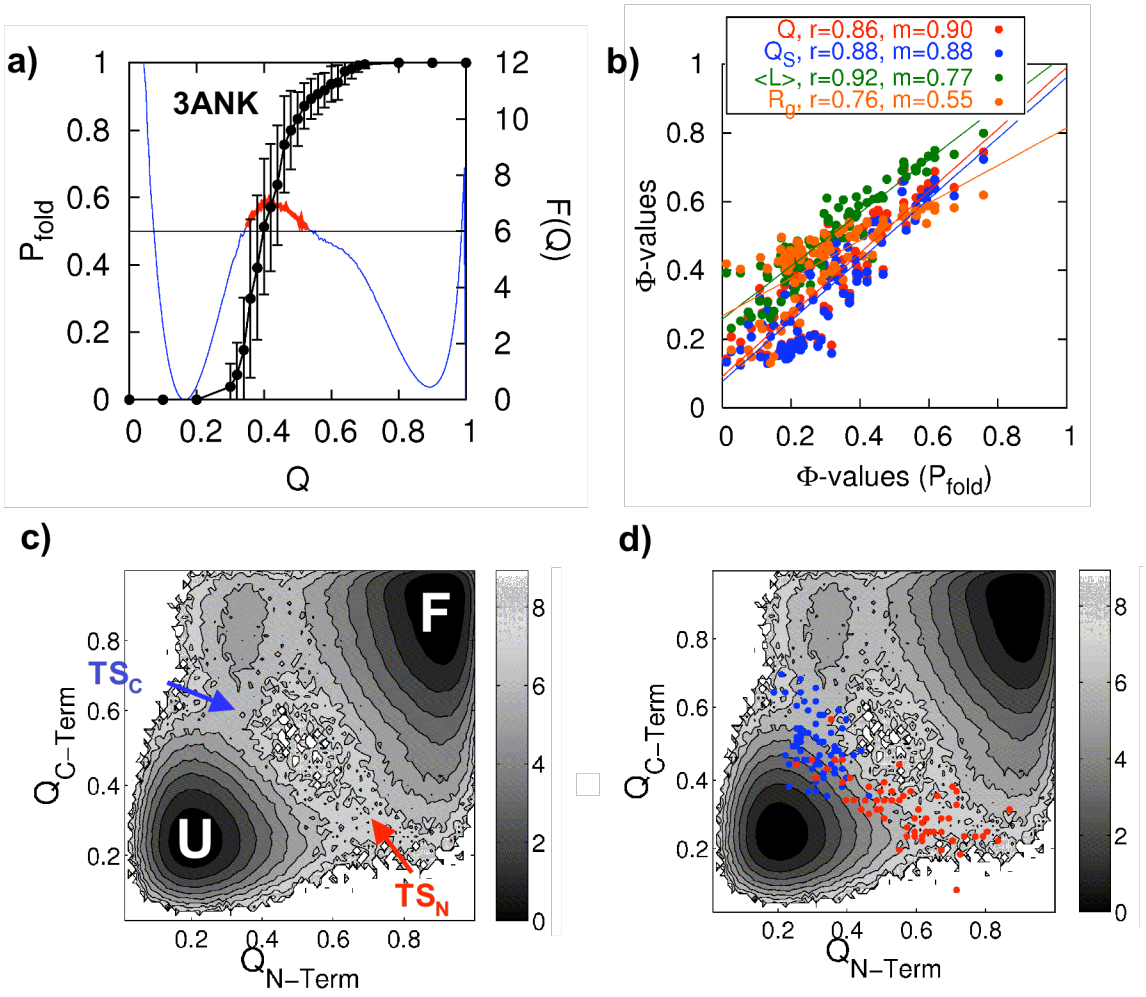


Figure 2-6: Comparing the TSE obtained from P_{fold} and structural reaction coordinates for 3ANK, a protein with a broad, asymmetrical free-energy barrier. (a) The free energy profile of 3ANK using Q as a reaction coordinate is overlaid with the average P_{fold} of structures (with error bars indicating 1 SD) over the range Q = 0.30–0.80. (b) The Φ-values of the TSE as predicted by Q, QS, <L>, 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} ~ 0.50) are overlaid on the free energy profile projected onto $Q_{N\text{-}Term}$ and $Q_{C\text{-}Term}$.

How do we reconcile the complex mechanism that we can ferret out with multiple structural coordinates, and also 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. This 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 (Figure 2-6d). 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.

2.4 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 (54). Go model simulations of CV-N showed previously a three-state folding transition with a high-energy intermediate (4). A two-state folding transition occurs when the Go-model is constrained by disulfide bonds present in the protein (4). 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 (Figure 2-7a-d). The two domains depend upon one another to fold.

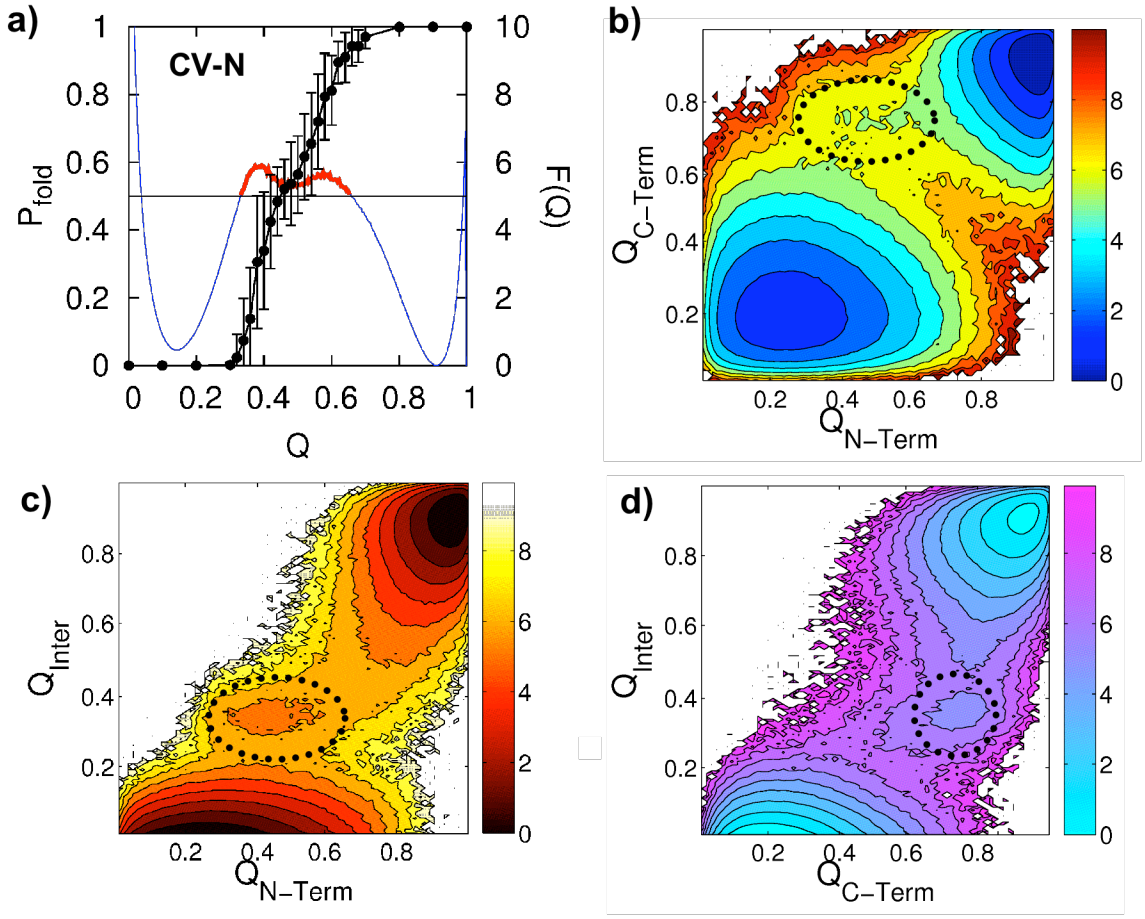


Figure 2-7: Comparing the TSE obtained from P_{fold} and structural reaction coordinates for CV-N, a protein that is simulated to fold with a three-state folding mechanism. (a) The free energy profile of CV-N using Q as a reaction coordinate is overlaid with the average P_{fold} of structures (with error 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}.

When we analyzed the region between the folded and unfolded states, P_{fold}=0.50 clearly corresponds to the intermediate and not to either of the transition states! The two actual transition states are barely represented at all in the ensemble of structures with P_{fold}=0.50 (Figure 2-7a). 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} (Figure 2-8). 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 corresponds 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. We find it is much better to use direct structure based reaction coordinates.

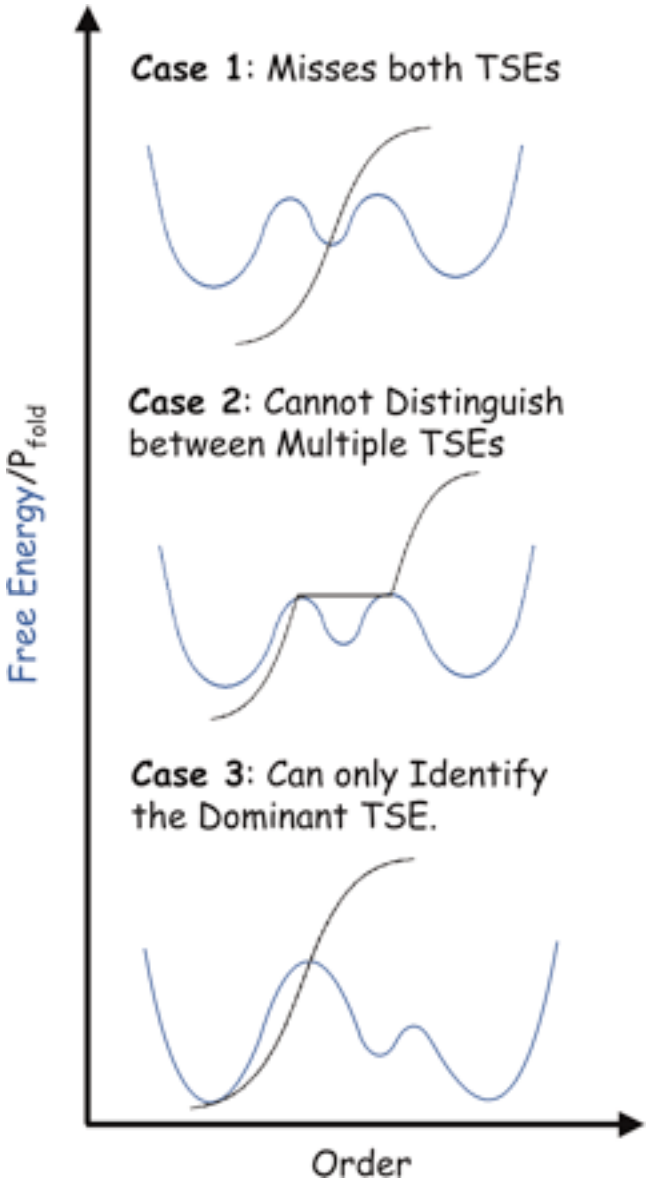


Figure 2-8: **A schematic depicting the three possible relationships between P_{fold} and free energy profiles for protein systems with two folding transition states.**

2.5 "Minding your p's and q's" in Protein Folding Kinetics

Protein folding has long been viewed as being rich in complexities. With the development of the energy landscape theory, our view of protein folding has, however,

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 quite 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 to carry out this expensive procedure 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 predict the TSE for minimally frustrated systems just as well as P_{fold}. Our study shows that global reaction coordinates based on the native topology of a protein, such as Q, Q_S, and <L>, 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, generally requires 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. 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 to 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.

Reprinted from:

Cho SS, Levy Y, Wolynes PG. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. Proc. Natl. Acad. Sci., USA. 2006, 103: 586-591.

3. Funneled Energy Landscapes for Binding Mechanisms

Funneled energy landscapes are now well accepted as the foundation for unimolecular folding, but most proteins interact with partners in the cell. The dynamic principles of protein association mechanisms are fundamental to protein networking, protein function, and pathogenic aggregation. Once the basic principles have been established, we may be able to design more stable complexes as pharmacological inhibitors. The remarkable efficiency of organizing many partners to yield biological functions strongly suggests a directed search in protein recognition processes that may be analogous to folding of single proteins. Can the organization of proteins into complexes be understood within the framework of the Funneled Energy Landscape Theory?

Recent work strongly suggests that we can indeed generalize the concept of funneled energy landscapes to protein-protein association mechanisms as well. In fact, there are already far too many examples of its application to protein-protein association mechanisms to adequately describe here. Instead, we will briefly highlight a few notable examples of studies using native topology-based (Go-type) models just to get a flavor of how binding mechanisms can be accurately predicted from these simple models. As we begin to explore more complicated systems, however, we cannot expect all binding processes to be neatly described by simple, idealized funneled energy landscapes. As such, it may be necessary to develop new paradigms beyond basic funneled energy landscapes to understand association mechanisms. Even experimentally observed behaviors that seem to contradict the Funneled Energy Landscape Theory must be

addressed if we are to fully understand protein folding and binding mechanisms.

3.1 Funneled Energy Landscapes of Protein-Protein Assembly Mechanisms

Protein recognition and binding, whether they result in either transient or long-lived complexes, play a fundamental role in biology. To test the applicability of the Funneled Energy Landscape Theory to oligomerization mechanisms, Levy and coworkers surveyed the association mechanisms of many protein-protein complexes (2, 3). Just as for single protein chains, a Go-type model that is globally directed towards the native state was used, thus corresponding to a perfectly funneled energy landscape. The main difference in the methodology is that the native state corresponds to the complex structure, not the individual, isolated components. Simulations of oligomers can be readily compared with experimentally observed mechanisms, such as whether a fine intermediate is populated or, whenever possible, the structure of the transition state ensemble as measured using protein engineering experiments.

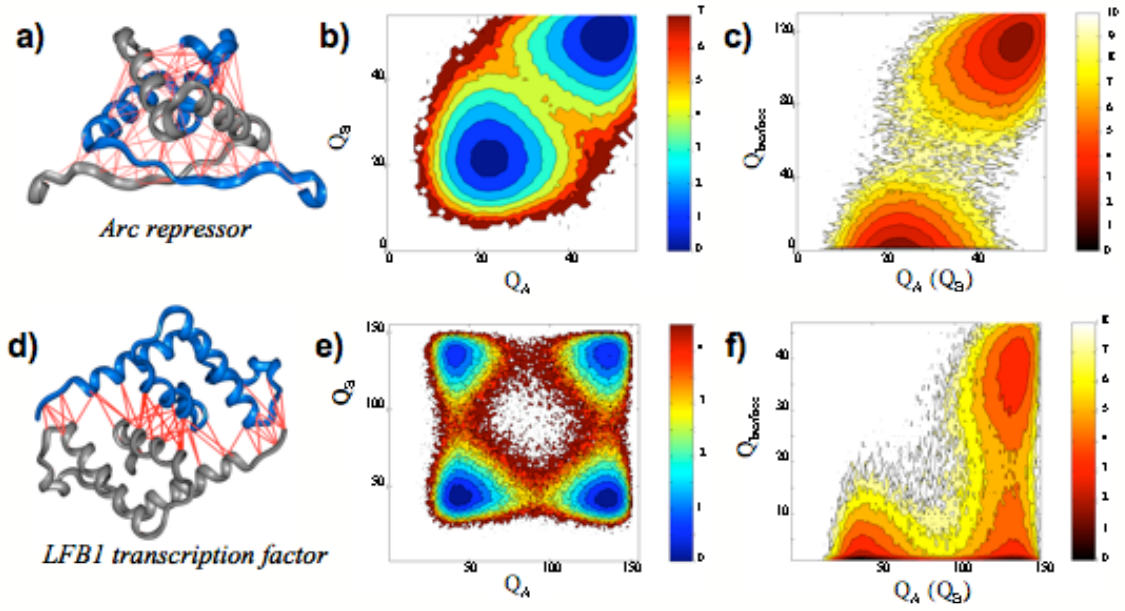


Figure 3-1: The structures and free energy surfaces of folding and binding of obligate, two-state arc repressor dimer (a-c) and nonobligate, three-state LFBI transcription factor dimer (d-f). The ribbon diagrams for the dimers consists of one monomer colored blue and the other colored grey. Red lines indicate native contact interactions that define the interface. The free energy surfaces are plotted as a function of the intramolecular native contacts (Q_A and Q_B) and that of the interface (Q_{Inter}).

The mechanism of protein assembly can be experimentally characterized as association that starts from unfolded subunits (two-state dimers) or folded subunits (three-state dimers) (55, 56). Those proteins that were experimentally observed to reach the native state cooperatively by a two-state mechanism were predicted by the Go-type model to have a coupled folding and binding mechanism. In contrast, when the routes to the native state involved an intermediate, the Go-type model predicted that the folding of the individual subunits occurred first, which corresponds to the experimentally observed intermediate, before then proceeding to the bound conformation, again in remarkable agreement with experimental observations (2, 3). In general, Go-type model simulations

predict reasonably well many of the finer features of the binding mechanism, as reflected by the prediction of Φ-values that compare with experimentally observed values (3).

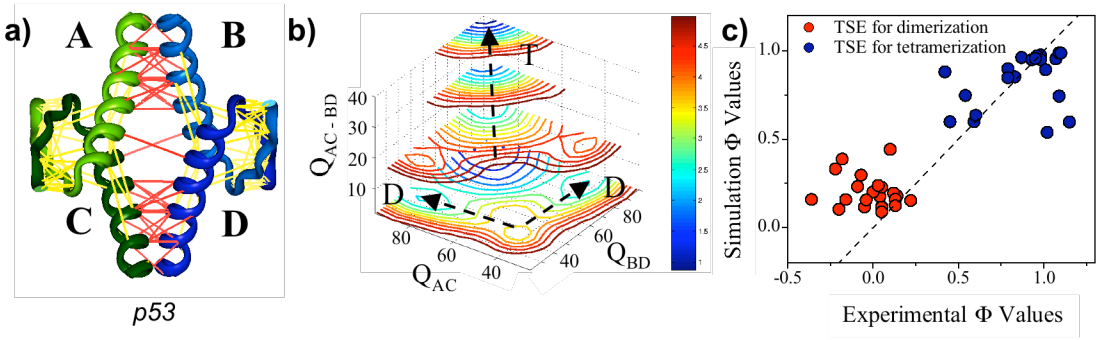


Figure 3-2: The structure and oligomerization mechanism of p53 tetramer. The ribbon diagrams for the tetramer (a) consists of a dimer colored two different shades of green and the other colored two different shades of blue. Yellow lines indicate native contact interactions between the monomers A and C or B and D, while red lines indicate native contact interactions between dimers AC and BD. The four-dimensional free energy surface (b) is plotted as a function of the formation of the dimers AC and BD and the tetramer interface (AC-BD). D and T refer to the folded dimer and the folded tetramer, respectively. The Φ-values for the transition state ensembles (TSEs) of dimerization and tetramerization as compared with experimentally observed values.

The tetramerization of p53 is an ideal system to study via simulation because its association mechanism has been extensively characterized via experiments (57). The homotetramer is formed by two sequential steps: first dimerization of two unfolded chains, which in turn further associate to form the tetramer. That is, the formation of the dimers is coupled to monomer folding, while the association of already folded dimers forms the tetramer. Consistent with experiments, formation of two dimers is obligatory before formation of the tetramer and no trimeric state is observed (Figure 3-2b). A quantitative comparison between the Φ-values obtained via simulations and experiments can be made at the dimerization transition state and the tetramerization transition state.

Consistent with experiments, we find that the dimerization transition state has significantly lower Φ-values than that of the tetramerization transition state. Beyond a qualitative agreement, however, the exact Φ-values agreement is somewhat poor. The fact that many negative Φ-values are found experimentally suggests that the system is energetically frustrated. Interestingly, Pande and coworkers studied the dimerization reaction of the tetramerization domain of p53 using all-atom MD simulations that includes non-native interactions (36), and the Φ-values for the dimerization transition state is qualitatively similar to those obtained by the Go-type model simulations. This discrepancy thus remains puzzling, but may be related to energetic heterogeneity, as we discuss later (See Chapter 5).

A particularly notable example that we highlight is the prediction by Levy and coworkers that the binding process of HIV-protease dimer involves a thermodynamic intermediate that corresponds to a monomer (58). That is, the monomer conformation may be thermodynamically populated, strongly indicating that a new target for drug design could be developed which by disfavoring the binding mechanism could prevent the function of this key enzyme in the HIV life cycle.

3.2. Challenge to the Funneled Energy Landscape Theory?: Rop Dimer

Up to this point, we have highlighted the large amount of evidence that even the simplest variant of the Funneled Energy Landscape Theory can suffice to quantitatively

predict the folding and binding mechanisms of many proteins. As such, one expects that the evolutionarily designed funneled energy landscapes of proteins robustly tolerate changes in the sequence (i.e. mutations) to yield kinetically competent folders as long as stability is maintained (1). Evidenced by structural studies of mutant proteins with single substitutions, the general experience is that the effect of mutations is typically minor and localized in structure. Some proteins even retain their global tertiary structure despite extensive redesign of their hydrophobic cores (59). The large sequence space yielding a single protein topology is well illustrated by the large families of structurally related proteins, sometimes found with very little sequence similarity (59-61). Structurally homologous proteins also sometimes have a conserved folding mechanism (35, 59), as would be expected from funneled nature of their evolved energy landscapes. It might seem that the folding and binding behavior of proteins is largely solved with little left to explore. Yet there are some minor instructive anomalies.

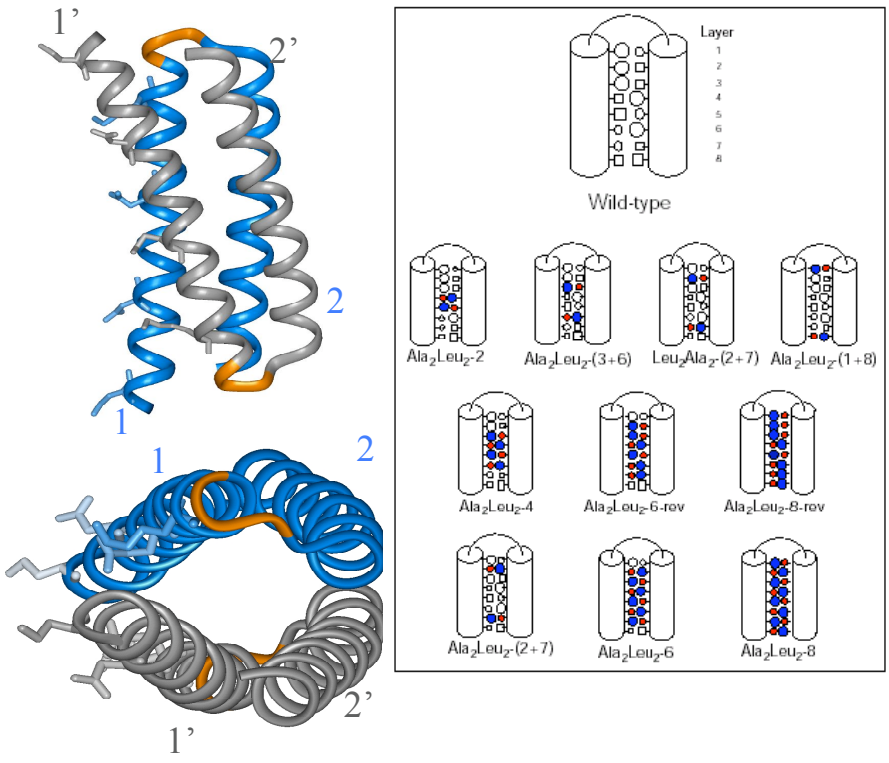


Figure 3-3: The structure of the wildtype Rop homodimer and the hydrophobic core redesign strategy undertaken by Regan and coworkers. Ribbon diagrams of the Rop dimer are shown with helices of the monomers colored blue and grey, and the turn region is colored orange (left). The Rop dimer mutations were introduced by progressively replacing the wildtype residues in the hydrophobic core with alanine and leucine residues from the middle layers towards the ends.

Among the longstanding experimental puzzles was the folding and binding behavior of the Rop (repressor of protein) homodimer. Its function is to bind two RNA hairpins in a key step regulating the replication of ColE1 plasmid in *Escherichia coli*. Although it is unstructured in the free form, the monomers associate to adopt a helix-turn helix structure finally resulting in an antiparallel coiled-coil four-helix bundle (62, 63) (Figure 3-3). Regan and coworkers systematically redesigned the hydrophobic core by mutating the "a" and "d" positions of the heptad repeats in its eight stacked layers (64-

66). The mutants were designed to differ in the number of mutated layers and the positions of the mutations (Figure 3-3). Most of the mutants are designed to have two residues with small side chains (as "a" residues) and two residues with larger side chains (as "d" residues). A set of mutants was designed in which two, four, six, or eight of the layers of the hydrophobic core were replaced by layers containing alanine at the "a" positions and leucine at the "d" positions. In other cases isoleucine, valine, and methionine were used to introduce a large side chain into the hydrophobic core instead of leucine. The antiparallel packing of the Rop monomers dictates symmetrical pattern of redesign of the core. Accordingly, redesigning layer 1 has to be accompanied with the redesign of layer 8, and the same rule applies between layers 2 and 7 and layers 3 and 6. Each of the mutants is named according to the identity of the residues at the "a" and "d" positions of the repacked layers: for example, Ala_2Leu_2-6 has the six central layers repacked with alanine in the "a" positions and leucine in the "d" positions. The "rev" suffix refers to cases in which layers 2 and 7 have reversed pattern of packing (i.e., the small and large residues are at the "d" and "a" positions, respectively) to mimic their packing in the WT dimer. The main goal of the Regan group's studies was to investigate the effect of core packing perturbations on the stability of the protein and its kinetics, and their results are summarized in Figure 3-4.

Class	No.	Mutant	*In vitro* activity	*In vivo* activity	T_m, °C	$\Delta G°$, kcal/mol	Relative k_f	Relative k_u	Structure (method)
I	1	WT	Y	Y	64	-7.7	1	1	*anti* (X-ray, NMR)
	2	Ala_2Leu_2-4	Y	P	68	-5.8	1.5	28	*anti* (*in vitro* activity)
	3	Ala_2Leu_2-2	Y	Y	72	-7.7	3.2	18	*anti* (*in vitro* activity)
	4	Ala_2Leu_2-3+6	Y	—	72	-8.4	7.5	8.3	*anti* (*in vitro* activity)
	5	Leu_2Ala_2-2+7	Y	—	85	-12.8	10	18	*anti* (*in vitro* activity)
	6	Ala_2Leu_2-6-rev	Y	—	85	-10.3	85	6.7×10^2	*anti* (*in vitro* activity)
	7	Ala_2Leu_2-8-rev	Y	N	91	-9.9	92	2.7×10^3	*anti* (*in vitro* activity)
	8	Ala_2Leu_2-2+7	Y	—	85	-8.7	120	7.1×10^3	*anti* (*in vitro* activity)
	9	Ala_2Leu_2-1+8	Y	—	54	-6.3	160	1.1×10^2	*anti* (*in vitro* activity)
	10	Ala_2Leu_2-6	Y	N	82	-8.1	310	3.1×10^4	*anti* (*in vitro* activity)
	11	Ala_2Leu_2-8	Y	N	91	-7.5	610	5.0×10^4	*anti* (*in vitro* activity)
II	12	Ala31Pro	P	—	—	—	—	—	*bisecting U* (x-ray)
III	13	Ala_2Ile_2-6	N	N	83	-5.1	—	—	*syn* (x-ray)
	14	Leu_2Ala_2-8	N	—	—	-12.8	—	—	—
	15	Ala_2Met_2-8	N	N	48	-3.1	—	—	—
IV	16	Ala_2Val_2-8	—	—	—	—	—	—	—
	17	Ala_4-8	—	—	<2	—	—	—	—
V	18	Leu_4-8	N	N	—	—	—	—	—

Figure 3-4: Summary of the kinetic studies of the Rop dimer and mutants with redesigned hydrophobic cores. The Rop variants can be classified into five classes based on their folding thermodynamics, binding activity, the dimer topology, and their folding kinetics. Based on the in vitro activity, it was concluded previously that all the mutants in class I have the anti topology. The folding rates of the Rop variants were measured at the same final fraction folded or unfolded. Class II contains the A31P mutant of Rop dimer that adopts the bisecting U topology. Class III is comprised of mutants that are highly α-helical; however, they completely lost their ability to bind RNA. The structure of Ala_2Ile_2-6 is the syn topology. The mutants in classes IV and V are less stable than the WT, and they do not bind RNA. Class IV is comprised of proteins which are underpacked (only Ala_2Met_2-8 forms dimer), and Leu_4-8 of class V is an overpacked protein that was suggested as forming a tetramer. Y, Rop protein that binds RNA; P, partial active proteins; N, no activity; —, no experimental data are reported.

This system is exceptional in that some of the mutants with the redesigned hydrophobic core (class I) both fold and unfold faster than the wildtype protein (Figure 3-4); the kinetics of the mutants in classes II-IV was not studied. This behavior is in conflict with the basic Funneled Energy Landscape Theory, where a single topology

determines the kinetics. The increases in the forward and backward rates depend on the number and position of the repacked layers in the core. For the mutant with all eight layers repacked (Ala_2Leu_2-8), the folding and unfolding rates are accelerated by more than two and four orders of magnitude, respectively. All of the mutants have similar CD spectra to those of the wildtype Rop, and they all exhibit cooperative thermal denaturation. Furthermore, the in vitro binding affinities of the mutants are comparable to that of the wildtype.

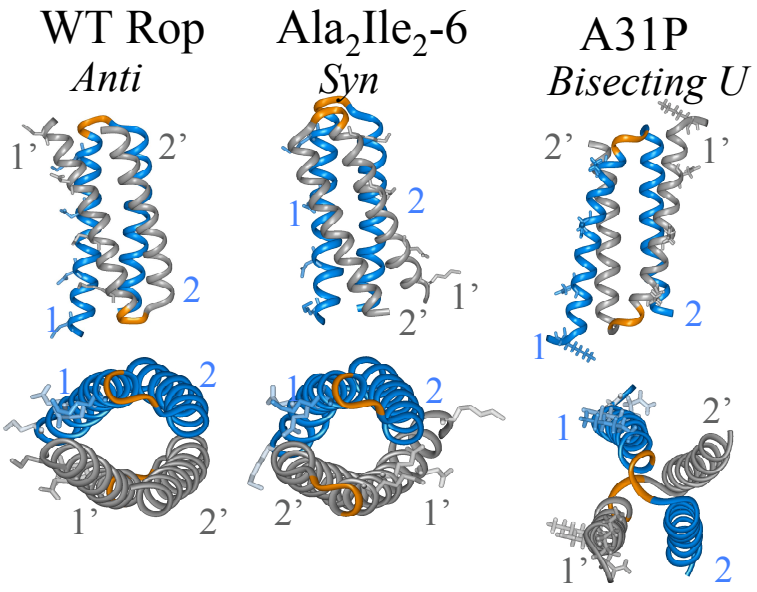


Figure 3-5: The crystal structures observed for the wildtype Rop (left) and two of its mutants (center and right). Ribbon diagrams of the Rop dimers are shown with helices of the monomers colored blue and grey, and the turn region is colored orange.

To further complicate the issue, there actually exist two alternative crystal structures sometimes found for the mutants. In one case, Ala-31, which is located in the turn between the helices of each monomer, was replaced by a proline residue (A31P;

Figure 3-5, right). The result is a dramatic conformational change that would not be expected from a single amino acid substitution, resulting in a "bisecting U" topology. This structure has been suggested to actually be a molten globule based on its thermodynamic properties that include low stability and reduced ellipticity, as well as its fluctuations in molecular dynamic simulations (67). A more dramatic change is found for a redesigned mutant that consists of two alanine and two isoleucine residues in each of the six central layers of the dimer interface (Ala_2Ile_2-6). It folds to a stable and highly α-helical structure, but has no ability to bind the RNA target of Rop. This is surprising because Ala_2Leu_2-6, which only differs by having a leucine in the "d" positions instead of isoleucine, does in fact show activity. The crystal structure of Ala_2Ile_2-6 shows that the mutant adopts the *syn* topology, a 180° flip of one monomer around an axis normal to the dimer interface (Figure 3-5, center). The reorientation of the two monomers splits the face formed by helices 1 and 1', which is essential for RNA binding.

3.3 Double-Funneled Energy Landscape Resolves the Rop Dimer Mystery

To study the folding and binding mechanisms of the wildtype Rop and the Ala_2Ile_2-6 mutant, we performed Go-type model simulations of each of the dimers. The structures of the other mutants have not yet been determined to our knowledge. The free energy profiles for the *anti* and *syn* topologies of Rop dimer were projected onto three

reaction coordinates: two corresponding to the folding of each monomer and the third corresponding to association (Figure 3-6a,b). The projected free energy surfaces for both structures show coupling between monomer folding and association. Although these plots show similar mechanisms for forming the *anti* or *syn* topologies, the binding free energies of the *syn* topology is lower than that for the *anti* topology by about 1.6 kcal/mol (Figure 3-6c). That difference becomes more pronounced when we include three-body interactions into our analysis (Figure 3-6d).

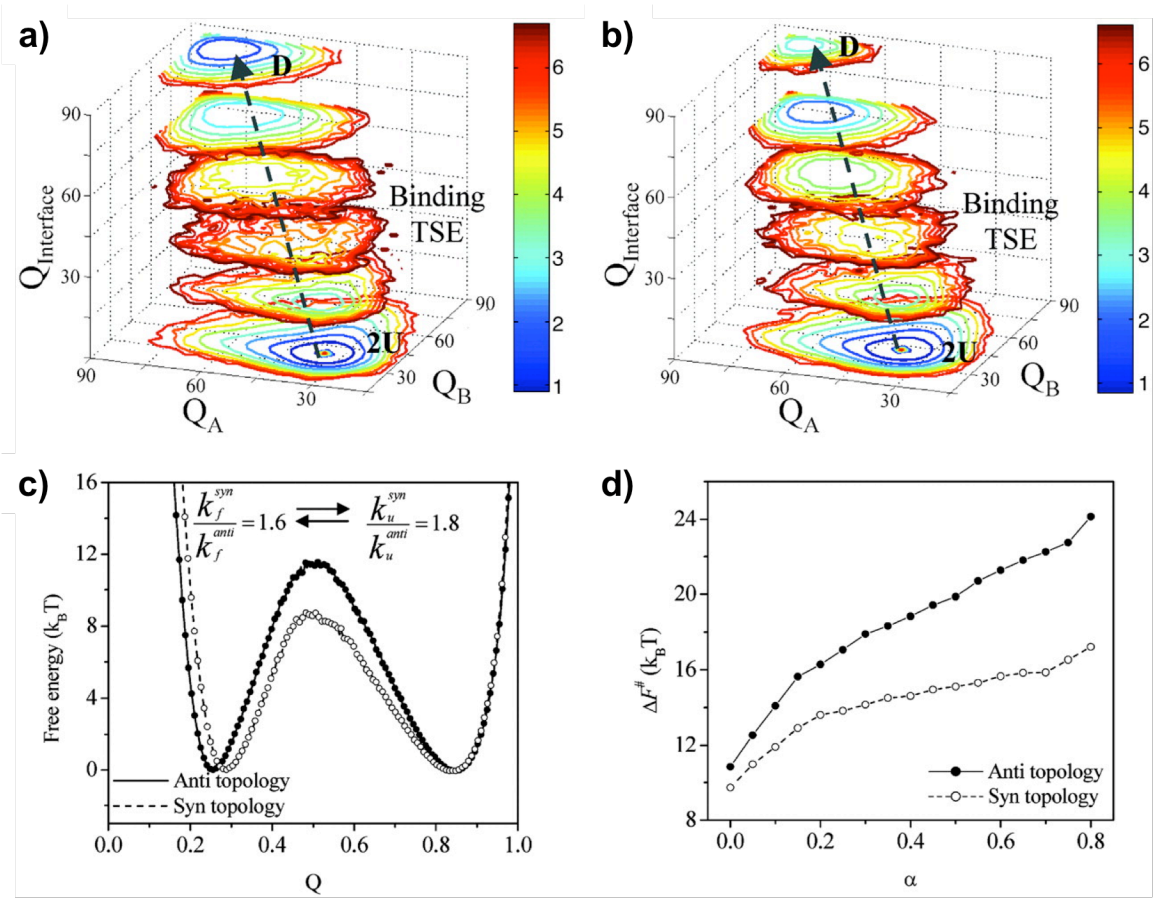


Figure 3-6: The barrier for the folding of the anti and syn forms of Rop dimer. The folding free-energy landscapes for the *anti* (A) and *syn* (B) topologies of Rop dimer are shown. The reaction coordinates are the folding of the two monomers and the formation of the interface (i.e., association). U, an unfolded monomer; D, a folded dimer. The dashed arrow illustrates the coupling between folding and association. (C). Two-dimensional free-energy profiles for the folding and association of the two forms of the Rop dimer based on the additive native topology-based simulations. The rates for folding and unfolding for each topological structure were obtained from >1,000 events (using the additive model) that were fitted to a single exponential decay. (D) The folding barrier height, $\Delta F^{\ddagger}$, as a function of α (the three-body contribution to the contact energy).

Of the 17 mutants of Rop protein that have been studied in the laboratory, the structures of only two mutants have been determined (Ala_2Ile_2-6 and A31P). Structural heterogeneity among the mutants would provide a framework for explaining the pronounced speeding up observed for both folding and unfolding rates of some mutants.

To check the assignment of a structure to each of the mutants, we threaded the sequence of each Rop mutants onto the three already observed Rop structures: *anti*, *syn*, and *bisecting U* topologies. Each resulting structure was minimized and simulated using explicit water all-atom molecular dynamics simulations. The average rmsd values of the backbone-heavy atoms of each mutant from its redesigned structure are shown in Figure 3-7. As expected, the sequence of the wildtype Rop displays a smaller rmsd for the *anti* topology than the *syn* topology. For the sequence of Ala_2Ile_2-6, a clear preference for the *syn* topology is found. The rmsd values show lower values for some of the mutants of class I when modeled as the anti topology, but for the remaining mutants, the rmsd values are lower for the syn topology. The rmsd values of the designed mutants indicate the possibility of a conformational switching for the Rop sequences. The mutants in classes II–V consistently display lower rmsd values for the syn topology. A preference for the syn topology for these mutants can explain the lack of their binding activity. We find that the mutants Ala_2Leu_2-6 and Ala_2Leu_2-8, which belong to class I, also have lower rmsd values when simulated starting from the syn form than they do from the anti form. These mutants show binding ability to RNA *in vitro* but not *in vivo*. They also have the highest folding and unfolding rates among the other mutants in class I. The preference for Ala_2Leu_2-6 and Ala_2Leu_2-8 to adopt the *syn* rather than the *anti* topology would explain their different thermodynamics and kinetics. The fast kinetics follows from the smaller barrier found in this study for the *syn* topology. We note that the rmsd values of all of the simulations starting from the *bisecting U*, regardless of sequence, were always found to be

much larger than those obtained for the *anti* and *syn* forms.

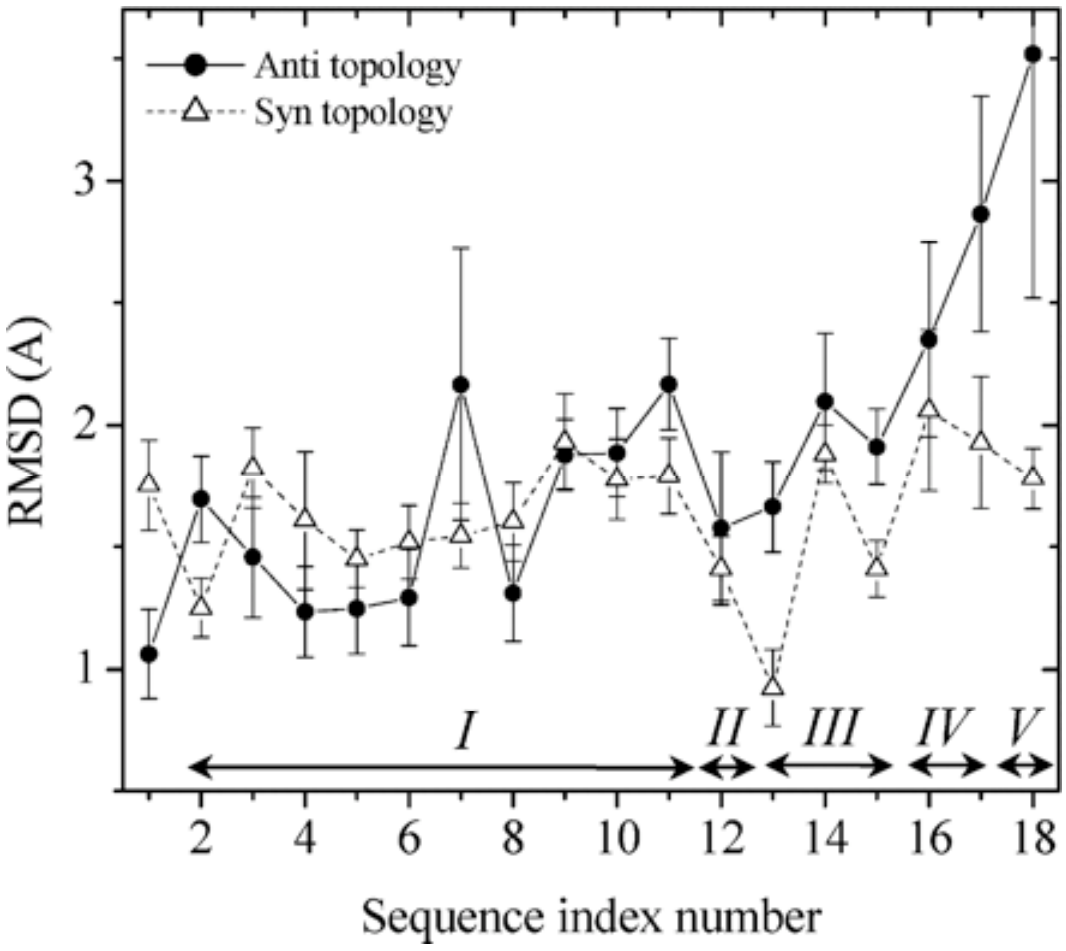


Figure 3-7: The average rmsd of each designed Rop mutant as *anti* and *syn* topologies in respect to the x-ray structures of the WT and Ala_2Ile_2-6 mutants. Each designed structure was simulated with all-atom representation of the protein with explicit solvent model for 5 ns. To account for different packing of the two monomers, the rmsd was calculated after superimposing a single monomer. The arrows indicate the mutant classes as in Figure 3-4.

We thus have proposed that the mystery could be explained by a double-funneled energy landscape where two native basins that correspond to two distinct but related structures corresponding to the wildtype Rop and the Ala_2Ile_2-6 mutant (Figure 3-8).

Arising from the near symmetry of the complexes, mutations can cause a conformational switch to a nearly degenerate but distinct topology or a mixture of both topologies. The topology predicted to have a lower free energy barrier height was further supported by all-atom simulations to give a better structural fit for those mutants that exhibited extreme folding and unfolding rates. Thus, the non-Hammond effects can be understood from Energy Landscape Theory if there are two different and distinct structures of the Rop dimer. In short, these topology-based models also were successfully used to not only solve the Rop dimer mystery in protein folding but also add a new paradigm to folding and binding mechanisms with a remarkable layer of complexity to the single funnel.

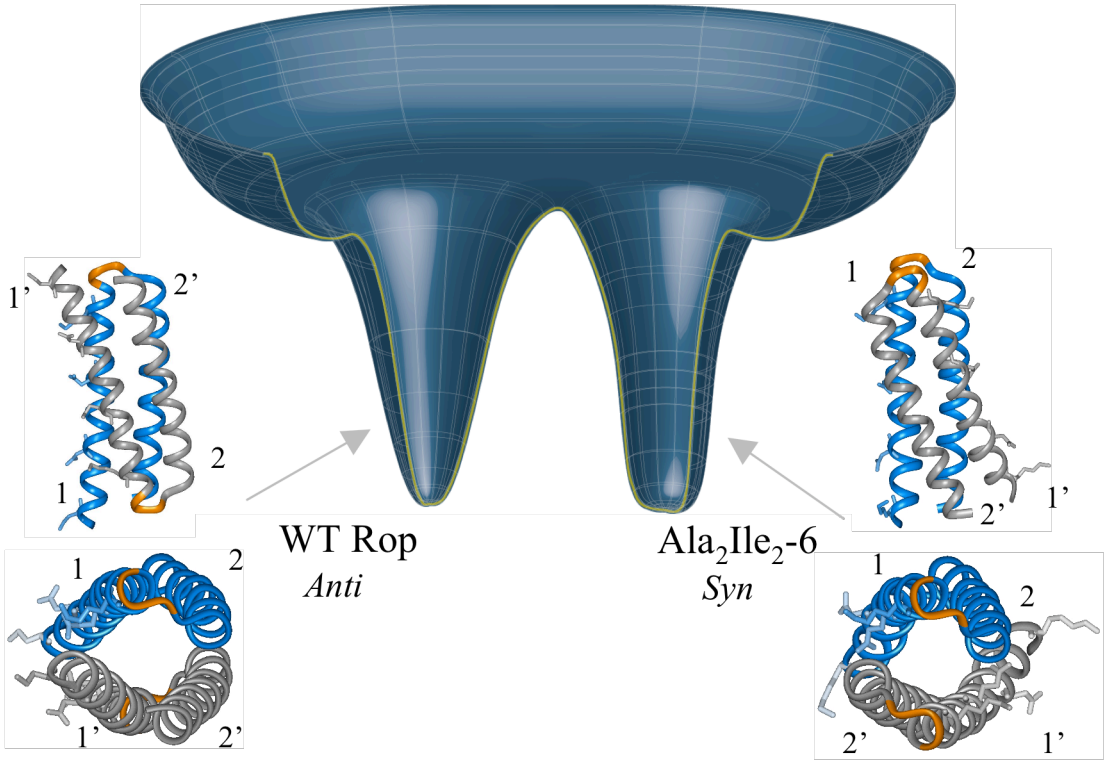


Figure 3-8: **A schematic of a double-welled funneled energy landscape for Rop dimer and ribbon diagrams of the wildtype Rop dimer and the Ala_2Ile_2-6 mutant. In these structures, one monomer is colored gray, and the other monomer is colored blue. The loop between the two helices in each monomer is colored orange. Residues Lys-3, Asn-10, Gln-18, and Lys-25 in helices 1 and 1', which constitute the binding site to the RNA, are shown by stick representation.**

4. Domain-Swapping and Protein Misfolding and Aggregation

Domain-swapping is an unconventional mechanism of oligomerization such that the structural element, or a "domain", of one chain is interchanged with a corresponding element of its partner, resulting in an intertwined homooligomer. The intramolecular interactions that would normally stabilize the monomer are thus "recruited" in swapping processes. These now become intermolecular interactions and define the interface of the complex (68, 69). Since the notion of domain-swapping was formally introduced by Eisenberg (68), about 70 proteins with domain-swapped oligomers have been characterized by X-ray crystallography and/or solution NMR. A domain-swapping event is characterized by slow kinetics with a high activation energy barrier arising from the many strong native interactions that must be rearranged. Early studies of domain-swapped proteins suggested that prolines in the hinge region connecting the swapped region with the main body of a domain-swapped oligomer could be important factors in determining whether proteins can domain-swap (70, 71). Further analysis shows, however, that this hypothesis cannot be generalized to all proteins that bind via domain-swapping. Domain-swapped proteins are typically observed and isolated under high concentration and low pH conditions (72), but in a growing number of cases swapping has been observed under physiological conditions (73, 74). This has fueled speculations about the role of domain-swapping *in vivo*. In particular, great interest has been generated by the proposal that domain-swapping is a crucial part of the mechanism for amyloid aggregation (75-77). Two amyloidogenic proteins, the human prion (78) and the human

cystatin C (79), have been observed as domain-swapped dimers, suggesting that these dimers may be the building blocks for fibers, at least in their nascent state (72, 75). The underlying mechanism and the main determinant(s) of domain-swapping can be understood in the context of the energy landscape theory. These insights may yield valuable clues about the role of domain-swapping in oligomerization and aggregation *in vivo*.

At first glance, the domain-swapping phenomenon seems to present a contradiction to the idea of a funneled energy landscape. Since domain-swapping involves homooligomers, a direct outcome of there being a funneled energy landscape for the monomer is that the very interactions that stabilize the monomeric conformation must compete with symmetrically similar ones that provide corresponding intermolecular interactions in the dimer. In other words, for any given residue *i* that is in native monomeric contact with residue *j*, there is, at first sight, nothing to prevent the same residue *i* from favorably interacting intermolecularly with residue *j'* in its partner, since the physico-chemical nature of the intramolecular interaction is the same as that of intermolecular interaction. In principle, there is, thus, no reason why one region of the protein should be more favored to swap than any other region. With many different and potentially conflicting possibilities for intermolecular interactions, a frustrated energy landscape generally results for the dimer. There would then be no preference for a single domain-swapped configuration. Even if a single domain-swapped configuration is somehow preferred at equilibrium, there would appear to be no guarantee that the most

stable structure should correspond to the one observed in nature, which might be the result of kinetic control. Therefore, a funneled energy landscape for monomeric folding generally would seem to preclude the possibility of a perfectly funneled energy landscape for oligomerization by domain-swapping. If this is the case, how do domain-swapping proteins, with the potential for a frustrated energy landscape for oligomerization, discriminate against alternatives to find their way to a unique swapped conformation?

4.1 Early Views of Domain-Swapping

It was observed, early on, from a survey of domain-swapping proteins conducted by Liu et al. that there does not appear to be sequence homology between the swapping domains that is common to all domain swapped proteins (72). Further, the secondary structure cannot be a determining factor because the swapping region can range from a single α-helix or β-sheet to an entire tertiary domain (72). One of the earliest and certainly most prominent hypotheses concerning the determinants of domain-swapping was that prolines play a pivotal role. This hypothesis was suggested largely because prolines seemed to be prevalent in the hinge regions of some of the first observed cases of domain-swapping proteins (70). The apparent line of thought was the following: The cis-trans isomerization of prolines, which has a significantly lower energetic barrier than for other natural amino acids, is the rate-determining step in the folding rate of some proteins. Owing to this, prolines often play a critical role in the observed folding kinetics, giving

rise to many long-lived intermediates (80). So, it was natural to suppose that prolines at or proximal to the hinge of domain-swapping proteins could act as local signals that would direct the global conformational change required to domain-swap with its identical partner. Experimental support for this hypothesis came in the form of a mutational study by Itzhaki et al., where it was found that two conserved prolines in the hinge region of p13suc1, a domain-swapping protein, controlled the monomer-dimer equilibrium in that system. Prolines made the hinge act like a "loaded molecular spring" that shifts towards either the monomer or the domain-swapped conformation (71). Itzhaki et al. and others suggested that prolines more generally would be levers by which naturally monomeric proteins could be re-designed artificially to stabilize the domain-swapped state. This is no doubt true. One might be tempted to go further, however, to posit that prolines in the hinge region are the main determinant of how proteins naturally oligomerize via domain-swapping.

To test this stronger hypothesis on a broader basis in the naturally occurring proteins, we asked two questions: 1) Is the prevalence of prolines in the hinge region of presently known domain-swapped proteins indeed significantly high? 2) Is the presence of prolines in or near the hinge region obligatory to oligomerize via domain-swapping? To analyze the amino acid residue prevalence in the hinge region of domain-swapping proteins, we constructed two libraries: one of domain-swapped protein structures and another of a nonredundant set of the PDB (i.e. no two proteins in the library have more than 25% sequence homology). The purpose of the latter library is to provide a baseline

containing the amino acid residue prevalences found in nature. Using PDBSelect (81), a set of 1834 nonredundant proteins (188,388 total residues) each with less than 200 amino acid residues was chosen, and the amino acid frequencies were calculated. The same calculation was carried out for the hinge region of the library of domain-swapped proteins. We used the same definition of the hinge region of domain-swapping proteins as was introduced by Eisenberg (72). In that definition, the hinge loop is defined to include those residues with a RMSD change in backbone ϕ and ψ dihedral angles of more than 20° in the domain-swapped oligomer when compared to the corresponding angles found in the monomeric configuration plus any residues in the same loop that connect secondary structure elements.

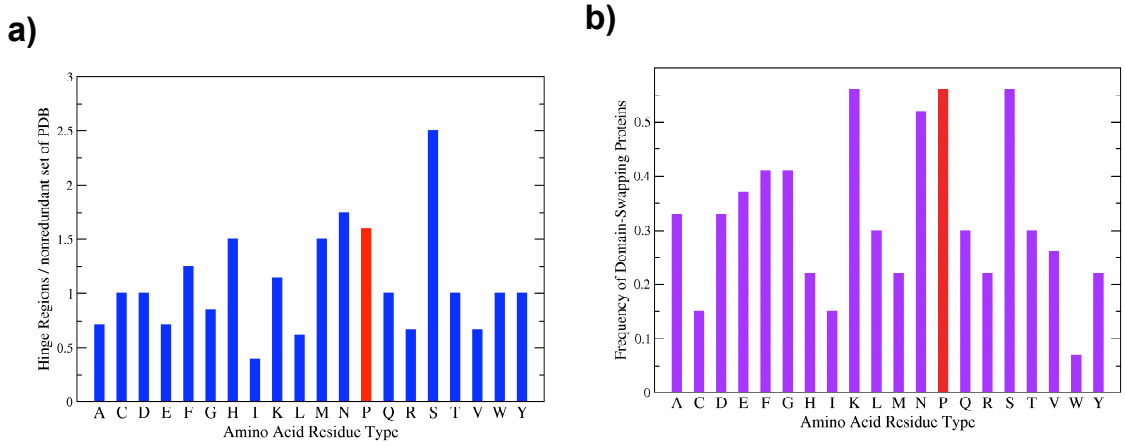


Figure 4-1: Evidence that prolines are not necessary as local signals to direct proteins to domain-swap. (a) A comparison between the distributions of amino acid residue frequency in the hinge region of domain-swapping proteins and a nonredundant set of the PDB. (b) The frequency of domain-swapping proteins with a certain residue in the hinge region.

A comparison of the amino acid frequencies of the hinge regions of proteins with the frequencies in the library of the nonredundant proteins (Figure 4-1a) shows that

prolines are not any more prevalent in the hinges than are many other residues. We found that the frequency of prolines in domain-swapping proteins (Figure 4-1b) is comparable to other kinds of residues. In fact, only about 50% of domain-swapping proteins have any prolines in their hinge region at all. In Figure 4-2, we show two examples of domain-swapping proteins that do not contain prolines in their hinge regions (Figure 4-2a) and two examples that have prolines in the hinge region (Figure 4-2b). In both of the examples in Figure 4-2a, the molecules do possess prolines that are absent from the hinge region, and indeed are distant from the hinge. For many domain-swapping proteins with prolines in the hinge region (Figure 4-2b), as is the case of p13suc1, numerous prolines can also be found dispersed throughout the sequence, again even at positions very distant from the hinge region. There is, of course, no reason to challenge the contention that prolines significantly control the monomer-dimer equilibrium in p13suc1. It remains likely in our view that some proteins could be designed, by the addition of prolines, to favor domain-swapping. However, the examples explicitly show that the presence of prolines in the sequence does not dictate whether a protein oligomerizes into a domain-swapped conformation.

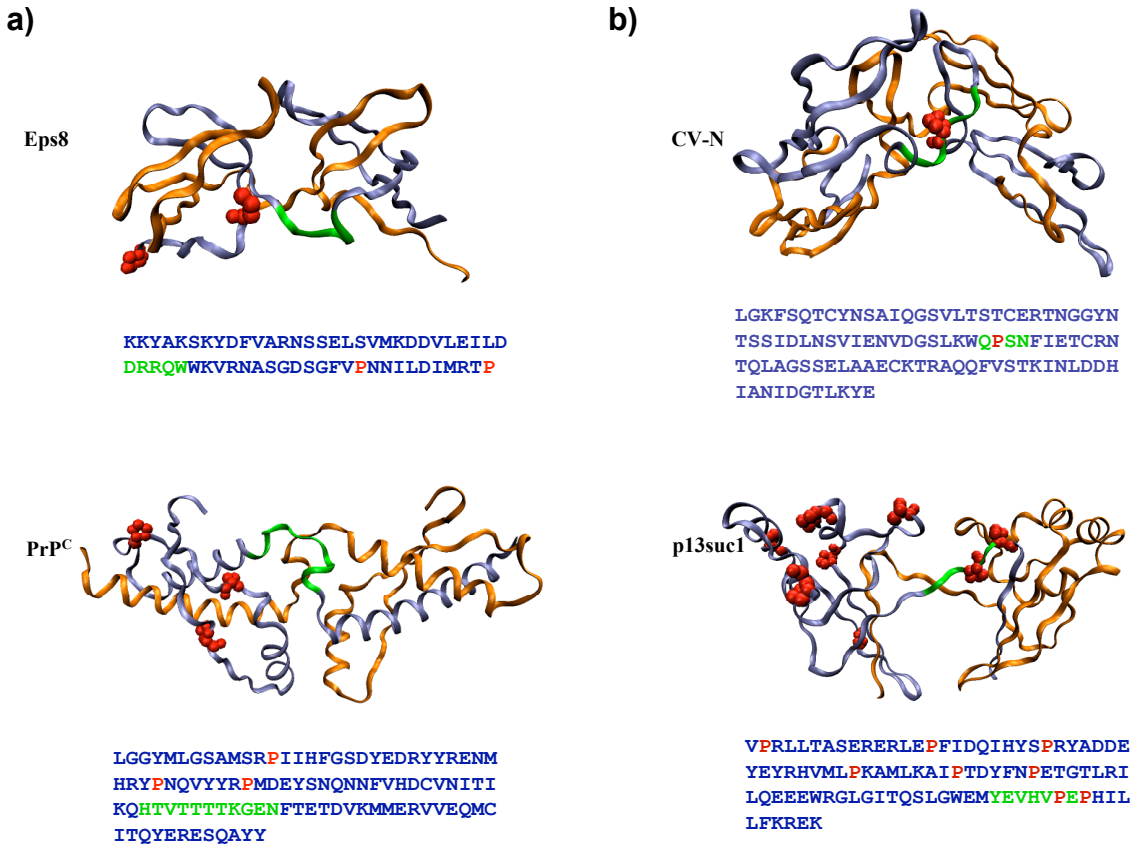


Figure 4-2: Examples of domain-swapping proteins and the proximity of the prolines to the hinge region. The structures of domain-swapping proteins without (a; Eps8 and PrP) and with (b; CV-N and p13suc1) prolines in the hinge region are shown in a ribbon representation, with each monomer colored orange, or blue, and the hinge region colored green. The prolines found in the blue chain are shown in a red space-filled representation. The sequences of the proteins are shown below each structure, in which the prolines are colored red, the hinge region residues are colored green, and the rest are colored blue.

To date, the primary strategy to engineer a protein to domain-swap has been to modify the hinge regions via mutations, additions, or deletions. Two specific examples, however, also highlight the need to look outside of the hinge region. Mutagenic studies of BS-RNase demonstrated that Pro19, located in the hinge region, is not a significant factor in the domain-swapping mechanism. Instead, Leu28, which is located outside of the hinge region, shifts the equilibrium towards the domain-swapped dimer by stabilizing the

interface (82). The sequences of two closely homologous proteins, the monomeric γB-Crystallin and the obligatory domain-swapped dimeric βB2-Crystallin, differ by the domain-swapped dimer having an acidic electrostatic repulsion between a residue in the hinge loop and a residue in the main body of the protein that prevents the formation of the monomer species (83). Clearly, the network of interactions as a whole, not just those in the hinge region, must be considered in describing domain-swapping.

To rigorously test the hypothesis that prolines are the main determinant of domain-swapping, we asked two questions: (I) Is the prevalence of prolines in the hinge region of presently known domain-swapped proteins indeed significantly high? (II) Is the presence of prolines in or near the hinge region obligatory to oligomerize via domain-swapping? To answer these questions, we constructed two libraries: one of domain-swapped protein structures and another set of non-redundant set of the PDB (i.e., no two proteins in the library have more than 25% sequence homology). The purpose of the latter library is to provide a baseline containing the amino acid residue prevalences found in nature. A comparison of the amino acid frequencies of the hinge regions of proteins with frequencies in the non-redundant proteins shows that prolines are not prevalent in the hinges when compared to many other residues.

4.2 Symmetrized-Go Model for Domain-Swapping

To address the domain-swapping mechanism, we developed a simple model, which we call the "Symmetrized-Go model". As described before, the original Go model

takes into account only contacts that exist in the native structure, and thus corresponds to a perfectly funneled energy landscape. The Symmeterized-Go model allows each intramolecular interaction found in the monomer conformation to favorably interact intermolecularly, resulting in multiple funnels. That is, there is a perfect competitive balance between intramolecular and intermolecular interactions. An important point to note is that, in principle, this model allows any region of the protein to swap with its partner and accordingly may serve as a tool to predict the domain swapped oligomeric conformation of a given protein because we use only the monomeric conformation as input. Each residue is described as a single bead, centered on the Cα position. The beads in an intact protein chain are connected to adjacent beads by bond, angle, and dihedral potentials. Simulation of the resulting simplified model of a protein allows the observation of slow conformational changes, and usually provides an accurate description of the intermediate and transition states of the folding mechanism observed experimentally. The network of favorable tertiary interactions is defined by the protein's native topology while all other non-bonded interactions are repulsive. In the Symmetrized-Go potential for a two-chain system, each individual protein chain is represented likewise using a series of single beads, each centered on the Cα position, and the native monomeric configuration is still used to define the intramolecular interactions. However, the Symmetrized-Go potential for the two chain system also contains intermolecular interactions. In the symmetrized potential, the observed intramolecular interactions of the monomer also introduce the favorable possible intermolecular

interactions. No other interactions are introduced. This model can be readily generalized to study aggregation. Indeed, this model had been previously used by Ding et al. to study the aggregation of SH3 (84). This protocol introduces intermolecular energetic frustration into the energy function, making it a predictor of not only the mechanism of domain-swapping but also allows it to make predictions of the domain-swapped structure (if unique) using only the monomer structure as input.

The energy function for the Symmetrized-Go potential for a two-chain system (designated chain A and chain B) with configuration Γ can be written explicitly:

$$H(\Gamma,\Gamma_0) = H_{backbone} + H_{\text{intra}chain} + H_{\text{inter}chain}$$

$$H_{backbone} = \sum_{bonds} K_r (r - r_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi^{(n)} \left[1 - \cos(n(\phi - \phi_0))\right]$$

$$H_{\text{intra}chain} = \sum_{chainA}^{i<j-3} \left\{ \varepsilon_1(i,j) \left[5\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - 6\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{10} \right] + \varepsilon_2(i,j)\left(\frac{\sigma_0}{r_{i,j}}\right)^{12} \right\}$$

$$+ \sum_{chainB}^{i'<j'-3} \left\{ \varepsilon_1(i',j') \left[5\left(\frac{\sigma_{i',j'}}{r_{i',j'}}\right)^{12} - 6\left(\frac{\sigma_{i',j'}}{r_{i',j'}}\right)^{10} \right] + \varepsilon_2(i',j')\left(\frac{\sigma_0}{r_{i',j'}}\right)^{12} \right\}$$

$$H_{\text{inter}chain} = \sum_{A\to B}^{i<j'-3} \left\{ \varepsilon_1(i,j') \left[5\left(\frac{\sigma_{i,j'}}{r_{i,j'}}\right)^{12} - 6\left(\frac{\sigma_{i,j'}}{r_{i,j'}}\right)^{10} \right] + \varepsilon_2(i,j')\left(\frac{\sigma_0}{r_{i,j'}}\right)^{12} \right\}$$

$$+ \sum_{B\to A}^{i'<j'-3} \left\{ \varepsilon_1(i',j) \left[5\left(\frac{\sigma_{i',j}}{r_{i',j}}\right)^{12} - 6\left(\frac{\sigma_{i',j}}{r_{i',j}}\right)^{10} \right] + \varepsilon_2(i',j)\left(\frac{\sigma_0}{r_{i',j}}\right)^{12} \right\}$$

The local backbone interactions are contained in $H_{backbone}$, which applies to both chains. K_r, K_θ, and K_ϕ are the force constants of the bonds, angles, and dihedral angles, respectively. The r, θ, and ϕ variables are the bond lengths, the angles, and the dihedral angles. The same quantities with a subscript zero represent the corresponding values

taken only from the native monomer configuration, Γ_0. The non-bonded contact interactions, $H_{intrachain}$ and $H_{interchain}$, contain Lennard-Jones 10-12 terms for the non-local "native" intrachain and interchain interactions and a short-range repulsive term for the "non-native" pairs. Strictly speaking, the "native" interchain interactions that result from symmetrizing the intrachain interactions include not only interactions that are present in the experimentally observed dimer (i.e. native), but also interchain interactions that are not present (i.e. non-native or frustrated), which is why this is nontrivial model when it comes to predicting the domain-swapped structure.

We chose as parameters of the energy function $K_r=100\varepsilon$, $K_\theta=20\varepsilon$, and $\varepsilon_1=\varepsilon_2=\varepsilon$. Forming disulfide bonds is effectively irreversible. Such bonding interactions were incorporated into the energy function by setting $\varepsilon_1=10\varepsilon$ or $\varepsilon_2=10\varepsilon$ for intramolecular or intermolecular disulfide interactions, respectively. The secondary structure biases are set as $K_\phi^{(1)}=\varepsilon$ and $K_\phi^{(3)}=0.5\varepsilon$ if the residue was either α-helical or β-sheet in character according to the DSSP definition (85) and $K_\phi^{(1)}=0.25\varepsilon$ and $K_\phi^{(3)}=0.12\varepsilon$ otherwise. Using higher flexibility for all of the turns in the proteins allows for changes in the dihedral angles of hinge regions without biasing any specific region of the protein to swap. $\sigma_{i,j}$ is the distance between the pair of residues (i,j) in the native monomeric configuration and $\sigma_0=4.0$Å for all non-native residue pairs.

A total of N native contact pairs for the monomeric conformation was determined using the CSU (Contacts of Structural Units) software (25). Only native contact pairs with sequence distance $|i-j| > 3$ were used because any three or four contiguous residues already interact through the angle and dihedral terms. The 2N intramolecular interactions

(N native interactions for each monomer) also define the 2N intermolecular interactions as follows: for each *i* and *j* intramolecular interaction that is native in the monomeric conformation we also define equal intermolecular interactions between *i* and *j'=j*. In total, 4N interactions are thus represented in the model. Therefore, there exists an energetic competition as to whether the pair of molecules should make any given contact intra- or inter-molecularly. Interchain interactions between helical residues where the sequence distance $|i-j|$ equals 4 were ignored because helical contacts are not expected to be involved in swapping.

We performed constant temperature molecular dynamics simulations of the protein systems with the Symmetrized-Go potential. We imposed an interchain center of mass constraint $E_{cons}=K(R-R_0)^2$ that becomes effective only when $R > R_0$. The minimum of the constraint, R_0, was set to the radius of gyration of the monomer conformation.

4.3 Domain-Swapping is Encoded in the Monomer Topology for Some Proteins

The first clue to direct the search for a unifying view of the domain-swapping mechanism is the somewhat tautological observation that the conformation of the swapped subunits in a domain-swapped oligomer bears a striking resemblance to the unswapped monomeric conformation (Figure 4-3a,b). Did evolution not only encode into the sequence information to fold a protein into its monomeric conformation but also instructions about whether it would oligomerize into a specific domain-swapped

conformation? To ask whether the monomeric topology is sufficient for predicting how proteins oligomerize via the domain-swapping mechanism, we applied the Symmetrized-Go potential, which we described in detail in the last section. It is important to note that this model's formulation contains no information *a priori* that biases a specific swapping region. Also, the model contains no information concerning the secondary interface, i.e. there are no interactions corresponding to those new ones that would be formed upon domain-swapping that are not represented in the monomer conformation. The latter could potentially play a role in swapping. In principle, in the symmetrized model any region of the protein can exchange interactions with its partner and nothing would preclude even the possibility of there being multiple swapping regions. Does this energy function discriminate the experimentally observed domain-swapped structure from the energetic traps? If so, we can say that there already exists, encoded in the monomer topology, sufficient information to intrinsically choose the swapping region.

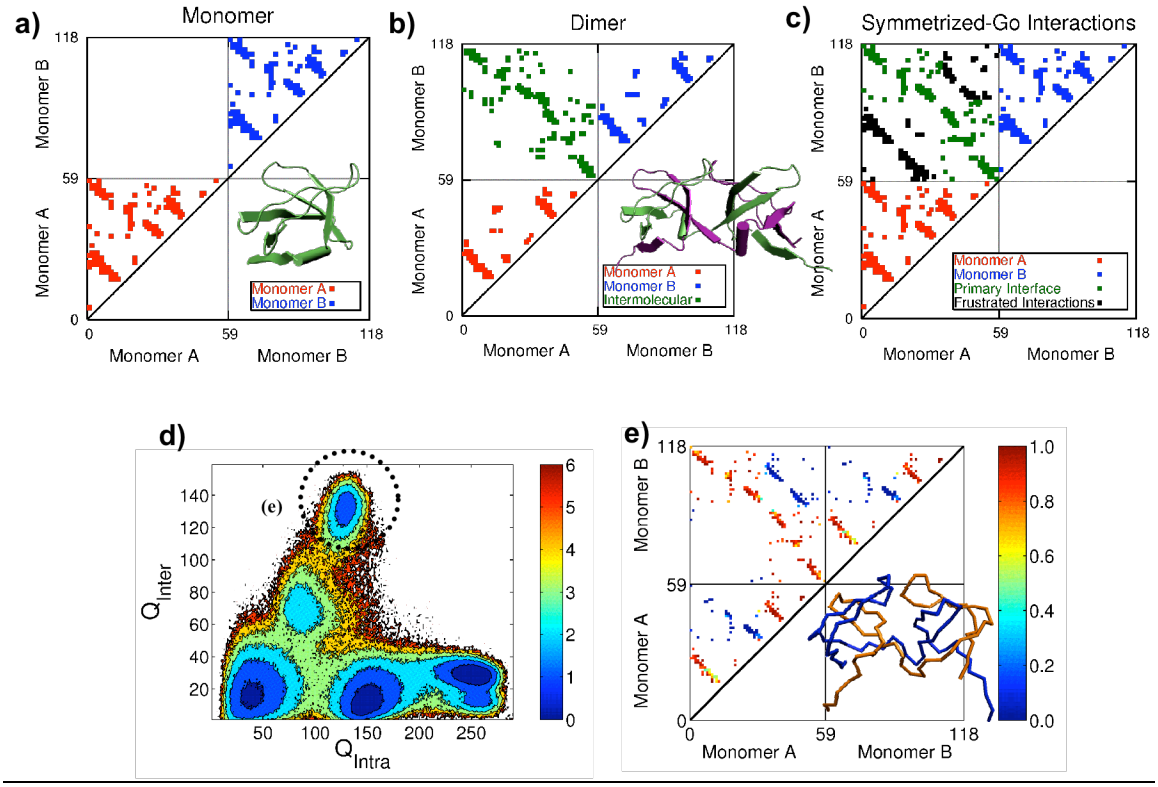


Figure 4-3: Application of the Symmetrized-Go potential to Eps8, a domain-swapping protein. The contact maps and the corresponding structures of the monomeric (a) and domain-swapped (b) Eps8 are shown. The represented favorable Symmetrized-Go interactions (c) include both the intramolecular and intermolecular interactions that have been derived from the monomeric conformation alone. The intermolecular interactions contained in the potential largely include the same interactions that are found in the experimentally observed dimer conformation (green), but there are also interactions that are not found in the experimentally observed dimer conformation (black). The free energy plot with respect to the number of intramolecular (Q_{Intra}) and intermolecular (Q_{Inter}) contacts (d) shows only a single stable domain-swapped conformation with an open-ended intermediate. The contact distribution plot of the minimum of the domain-swapped conformation (e) is shown as well as a representative structure from that minimum.

When we applied the Symmetrized-Go potential to Eps8 (epidermal growth factor receptor pathway substrate 8 SH3 domain), a domain-swapping protein, we found that despite the energetically frustrated intermolecular interactions, the model led to accurate prediction of the experimentally observed domain-swapped dimer as the most stable conformation. From our simulations, we can plot a free-energy surface as a

function of the order parameters Q_{Intra} and Q_{Inter}, the number of native intramolecular and intermolecular contacts, respectively (Figure 4-3d). Q_{Intra} indicates the degree of folding of the two monomers and Q_{Inter} indicates the degree of binding via swapping. At low Q_{Inter}, we found three basins, corresponding to two unfolded monomers, one unfolded monomer and one folded monomer, and two folded monomers. The basin with the highest Q_{Inter} corresponds to the fully swapped structure found via x-ray crystallography. At intermediate Q_{Inter}, there is a basin corresponding to one swapped and one unswapped conformation (i.e. partially domain-swapped intermediate). A contact probability plot of the basin of the domain-swapped conformation (Figure 4-3e) shows that only the interactions found in the experimentally observed domain-swapped dimer are statistically favored. The other interactions, while favorable according to the symmetrized model, are either seldom represented or are not found at all. Despite the energetic frustration that is present in the model, only the experimentally observed domain-swapped structure is found to be significantly populated. Since our initial study, Dokholyian and coworkers have used the Symmetrized-Go model to predict the structures of many experimentally observed domain-swapped structures (86). The energy landscape of domain-swapping clearly consists of two funnels: one for folding and the other for oligomerization via domain-swapping (Figure 4-4).

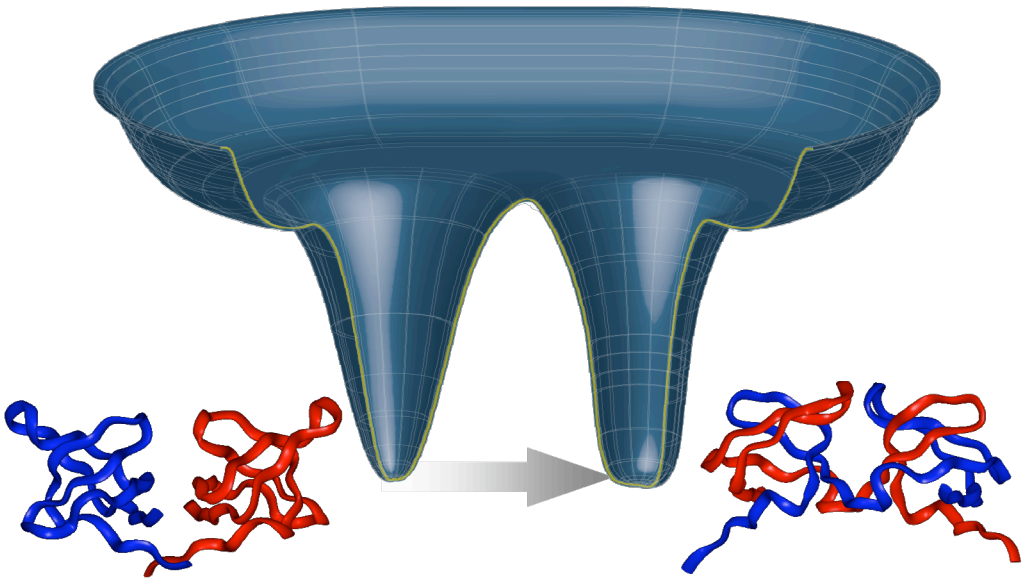


Figure 4-4: A schematic representation of a double-welled energy landscape for domain-swapping and ribbon diagrams of the monomer and domain-swapped conformations of Eps8. One monomer is colored blue and the other is colored red.

4.4 Domain-Swapping is not Encoded in the Monomer Topology for Other Proteins

We applied the Symmetrized-Go model to the 434 repressor, a well-studied dimeric protein for which no evidence of a unique domain-swapped form has been found to date. Just as with Eps8, we constructed a Symmetrized-Go potential from the conformation of a single monomer (Figure 4-5a). The free-energy surface for the 434 repressor (Figure 4-5b) shows two domain-swapped basins, reflecting a frustrated competition between the two states. This clearly contrasts with the free energy plot for Eps8, which has only one domain-swapped basin. A contact probability plot of the two basins yields two distinct domain-swapped structures (Figure 4-5c,d). One may note

that these two swapped structures of the 434 repressor have a very similar number of contacts but differ in the degree of folding of the monomer and the interface size.

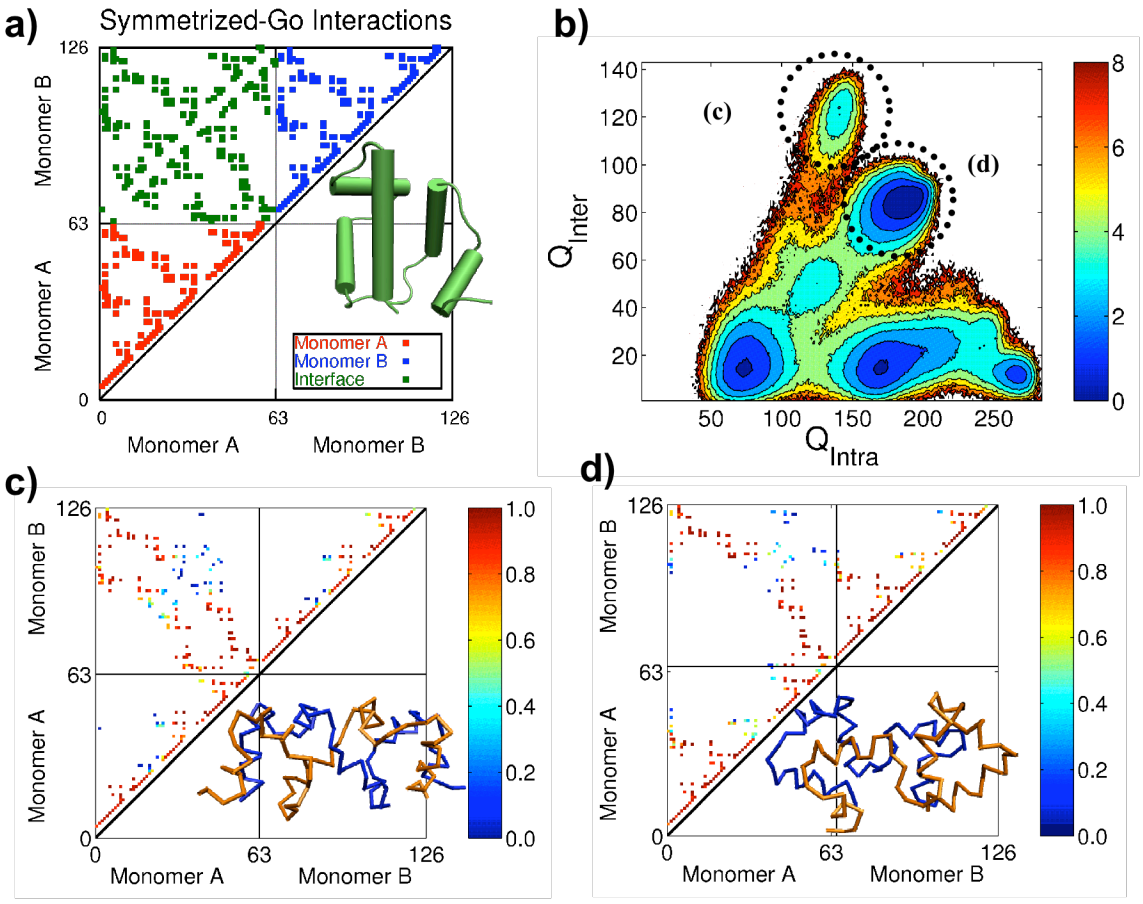


Figure 4-5: Application of the Symmetrized-Go potential to the 434 repressor, a dimeric protein showing no evidence of unique domain-swapping. The represented favorable Symmetrized-Go Interactions (a) for the 434 repressor are shown with the corresponding structure of the monomer. The free-energy plot as a function of the number of intramolecular (Q_{Intra}) and intermolecular (Q_{Inter}) contacts (b) that was derived from our simulations shows two domain-swapped minima. The corresponding contact distribution plots of the two minima from (b) are shown in (c) and (d) as well as a representative structure from its respective minimum.

We further applied the Symmetrized-Go potential to CI2 (Figure 4-6a), a protein that is found naturally as a monomer. While the wild-type protein is currently thought to be intrinsically monomeric, Perutz and colleagues have engineered a domain-swapped

dimer by the insertion of glutamate repeats in a loop within the protein (87). Similar to our study of the 434 repressor, we observed multiple minima of swapped structures when Q_{Inter} is high (Figure 4-6b). Interestingly, the most stable of the minima had the highest number of intermolecular native contacts, and the ensemble of structures for this minimum (Figure 4-6c) is similar to that structure found for the engineered domain-swapped protein (Figure 4-6d). These observations indicate that further analysis of other naturally monomeric proteins using the Symmetrized-Go potential can predict which proteins might be most amenable to re-engineering into domain-swapping oligomers by appropriate hinge mutations. We note that the observation of nonspecific domain-swapping of monomeric proteins is not simply an artifact of our model. Oliveberg observed "transient aggregates" at high concentrations that cause deviations from two-state kinetics in protein folding (88), and we believe that they are the result of the unstable domain-swapping we see in the Symmetrized-Go model. With multiple possibilities for domain-swapping, the protein is observed only in the monomeric conformation because of its higher specific concentration.

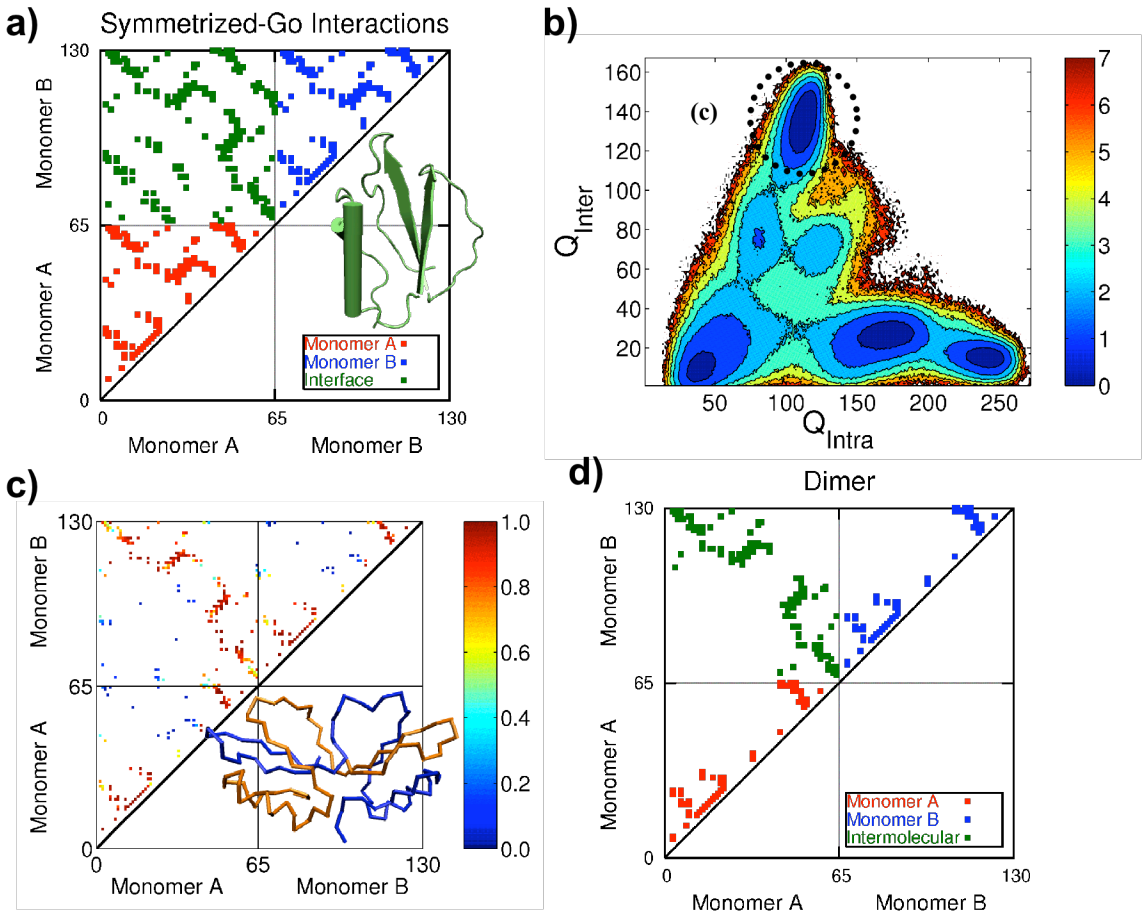


Figure 4-6: Application of the Symmetrized-Go potential to CI2, a naturally monomeric protein that has been artificially engineered to domain-swap via insertion of glutamine repeats. The represented favorable Symmetrized-Go interactions (a) for CI2 are shown with the corresponding structure of the monomer. The free-energy plot with respect to the number of intramolecular (Q_{Intra}) and intermolecular (Q_{Inter}) contacts (b) shows more than one domain-swapped minimum. The corresponding contact distribution plot of the deepest minimum from (b) is shown in (c) as well as a representative structure from its minimum. For comparison, the contact map depicting the swapping and main regions of the engineered domain-swapped of CI2 is shown in (d).

4.5 Disulfide Bonds Can Overcome Topological Insufficiencies to Undergo Domain-Swapping

We now turn our attention to two other proteins with known domain-swapped structures: CV-N (Figure 4-7a) and the human prion (PrP) (Figure 4-8a). These have

intramolecular and intermolecular disulfide bonds, respectively. CV-N has two intramolecular disulfide bonds: Cys 8 - Cys 22 and Cys 58 - 73. The intramolecular disulfide bonds of CV-N are important for the stabilizing the monomeric structure of CV-N. They are also critical to the anti-HIV activity of CV-N (89, 90). The domain-swapped structure of CV-N has been resolved by both X-ray crystallography (89) and solution NMR (91). The introduction of mutations to CV-N changed the energy landscape for folding to stabilize an intermediate (54). Our Go-model simulations of wild-type CV-N as a monomer also revealed the existence of a high-energy intermediate. We had initially thought that this result indicated an actual intermediate that was, however, not able to be observed by current experimental techniques in the wild-type but was stabilized by incorporating mutations. However, when we introduced disulfide bonds into the topology of the Go-model, the high-energy intermediate was no longer in the free-energy profile. Retaining the disulfides changes the mechanism of folding.

How does the inclusion of disulfide bonds affect the energy landscape for domain-swapping? In the domain-swapped dimer conformation of CV-N, the disulfide bonds remain oxidized, so the conformational conversion does not require a reduction of the disulfide bonds. In Symmetrized-Go simulations of CV-N (Figure 4-7b) without modifying the energetics of disulfide bonds to reflect their greater stability, the energy landscape for domain-swapping is clearly frustrated (Figure 4-7c). However, once we included the stronger intramolecular disulfide bonds into the topology of CV-N, we found that the energy landscape for domain-swapping becomes effectively unfrustrated (Figure

4-7d). A contact probability plot of the basin of the domain-swapped conformation (Figure 4-7e) shows that only those interactions found in the experimentally observed domain-swapped dimer are now favored, just as we saw in the case of Eps8. The disulfide bonds not only act to stabilize the monomer conformation, but they also limit the possible states that are accessible for domain-swapping. With the permanent disulfide linkage, only one stable state becomes possible for the dimer.

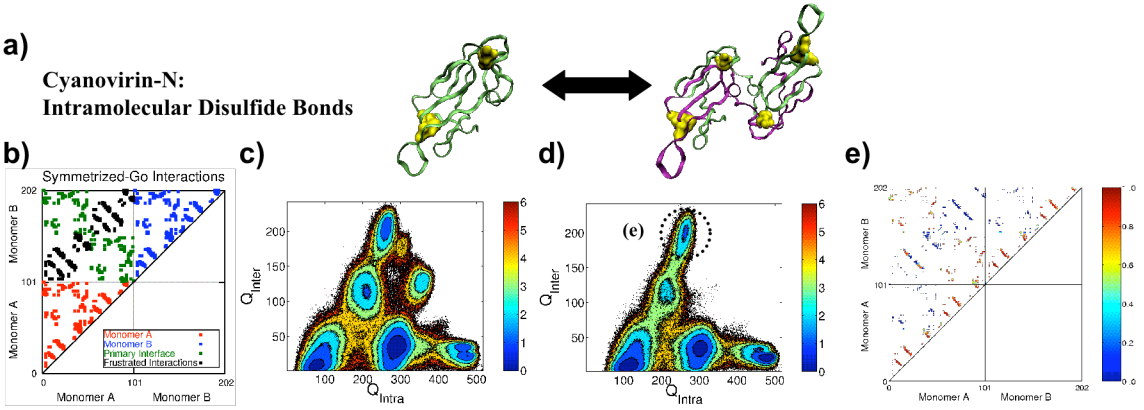


Figure 4-7: Application of the Symmetrized-Go potential to CV-N, a domain-swapping dimer with intramolecular disulfide bonds. The structures of the monomeric and domain-swapped conformations are shown (a) in a ribbon representation. The chains are colored green or purple, and the cysteine residues are shown colored yellow in a space-filled representation. The favorable Symmetrized-Go interactions of the domain-swapped dimers are shown (b). The free-energy plots as a function of the number of intramolecular (Q_{Intra}) and intermolecular (Q_{Inter}) contacts are shown, both without (c) and with (d) the explicit inclusion of disulfide bond interactions, along with a contact distribution plot of the domain-swapped basin (e).

4.6 Domain-Swapping an Early Step Towards Pathogenic Aggregation?

Despite much progress and study, the detailed mechanism for the conversion of prions (PrP) from the normal cellular form (PrP^C) to the infectious aggregate form (PrP^{Sc}) remains elusive. The structures of PrP^C for several mammal proteins have been

determined by solution NMR, and they all have the same basic monomeric structure, consisting of three long α-helices and two short β-sheet strands with a conserved disulfide bond between Cys 179 and Cys 214 that bridge helices 2 and 3. A domain-swapped dimer conformation of PrP was found experimentally in which there are intermolecular disulfide bonds between Cys 179 in one monomer and Cys 214 of its partner, bridging the helix 2 of one monomer with helix 3 in its partner (78). In contrast to the case of CV-N, the conformational change of the PrP from the monomeric to the domain-swapped dimer forms must involve the reduction of the intramolecular disulfide bond and subsequent intermolecular reoxidation. The Symmetrized-Go simulations of the PrP (Figure 4-8a) carried out without consideration of the disulfide bonds again revealed multiple possibilities for domain-swapping (Figure 4-8b). It is only upon including effectively irreversible intermolecular disulfide bonds that the energy landscape for domain-swapping becomes topologically funneled (Figure 4-8c) towards the experimentally observed domain-swapped state (Figure 4-8d,e).

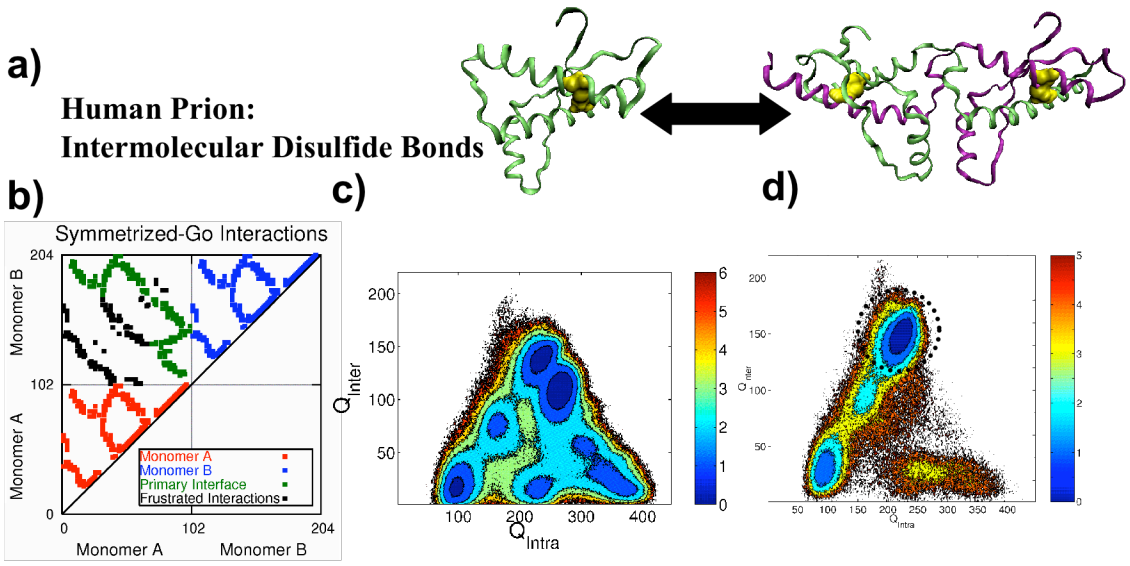


Figure 4-8: Application of the Symmetrized-Go potential to PrP, a domain-swapping dimer containing intermolecular disulfide bonds. The structures of the monomeric and domain-swapped conformations are shown (a) in a ribbon representation. The chains are colored green or purple, and the cysteine residues are shown colored yellow in a space-filled representation. The favorable Symmetrized-Go interactions of the domain-swapped dimers are shown (b). The free-energy plots as a function of the number of intramolecular (Q_{Intra}) and intermolecular (Q_{Inter}) contacts are shown, both without (c) and with (d) the explicit inclusion of disulfide bond interactions.

The pivotal role of intermolecular disulfide bonds in prion aggregation has been suggested both theoretically (92) and experimentally (93), but there is some disagreement as to whether intermolecular disulfide bonds actually do occur in the large prion aggregate (94, 95). While further study is clearly needed for a definitive answer, our present study would provide a structural basis for obligate intermolecular disulfide interactions in prion aggregation. If forming intermolecular disulfide bonds is critical for domain-swapping, these interactions may at least be transiently represented in the early stages of prion aggregation. The increase in local concentration of prion proteins caused first by domain-swapping may trigger the further conformational changes required to form PrP^{Sc}. We note

that this hypothesis does not conflict with the current understanding of the structure of the PrP^{Sc} fiber (96) in which helices 2 and 3 of PrP^{Sc} and the disulfide bond between them remains intact. It has not escaped notice that transient disulfide oxidation isomerization and reduction, perhaps in different physiological compartments or conditions, would greatly modify the kinetics of aggregate formation and fragmentation from the predictions of simpler kinetic assembly models, which currently seem unable to account fully for the quantitative details of *in vivo* pathogenesis (97-100).

Reprinted from:

Cho SS, Levy Y, Onuchic JN, Wolynes PG. Overcoming residual frustration in domain-swapping: The roles of disulfide bonds in dimerization and aggregation. Phys. Biol. 2005, 2: S44-S55.

5. Native Structural and Energetic Heterogeneity in Protein Folding

It is now clear that many protein folding and binding mechanisms can be inferred from the topology of the native state(s). Even crude topology-based measures, such as contact order, provide rough estimates of the folding rates of two-state folding proteins (101). Since many contacts form 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 (13). Such averaged contact energy models predict the folding rate of proteins in many cases (15), even when the simple contact order estimate is not very accurate (102). Many studies have shown that a wide range of details of folding and binding mechanisms, such as whether specific intermediates form or not is also correctly predicted by such native topology-based models in many cases (2, 13). In some circumstances where seemingly minor differences of topology are involved, even predicting mechanistic subtleties is possible (30). 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 (3, 14, 17), but at this level more discrepancies appear (3). 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?

5.1 Native Energetic Heterogeneity Cannot Be Ignored In Some Cases

Failures of the contact averaging approximation were first noted in studying structurally homologous proteins with 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 states, and this behavior is also seen in simulations (42, 103). The structure of the transition state ensemble is also robust to changes in environmental conditions for these systems (103). 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 (104), 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 (59, 105,

106). 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 while Im9 folds by a two-state manner, even though there is 60% sequence identity between the two proteins (107). Interestingly, the main transition states still have similar Φ-values (108). Recently, Clarke and coworkers showed that the folding rates of α-spectrin repeats of similar topology can vary over several orders of magnitude (109). While 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.

The effects of energetic heterogeneity of the native interactions on the folding mechanism have already been addressed using analytical energy landscape theory. Using the free energy functional approach first developed by Wolynes and coworkers (39, 110), Plotkin and coworkers 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 (111-113). The effects of contact heterogeneity is very much analogous to the well known phase transition in the random field ferromagnet (114). Sometimes, with sufficiently large dispersion of the native contact energies, the Φ-values becomes bimodal, with extreme values close to 0 or 1 (112,

113). Recently, in the context of the α/β CI2 and the all-β src-SH3 domain, Suzuki and Onuchic have shown that the structure of the transition state ensemble is robust and insensitive to energetic details (115). We can directly compare the analytical results of free energy functional approaches with those of native topology-based model simulations.

5.2 Homogeneous versus Heterogeneous Contact Energies in Funneled Landscapes

We began our investigation of quantifying the role of native contact energetic heterogeneity by comparing simulations of the simple homogeneously weighted C_α models to corresponding simulations having energetic heterogeneity based on the 20-letter Miyazawa-Jernigan (MJ) contact potential (116). While this degree of heterogeneity may be too large, it is similar to what is predicted by another more refined contact potential (117). 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 two-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 (Figure 5-1a-c). To quantitatively characterize the folding mechanism, we performed the weighted histogram analysis

method (WHAM) to calculate theromodynamic quantities with respect to the order parameter, Q, the fraction of native contacts. We recently showed the Q is one of several simple structural reaction coordinates that captures the folding mechanism on smooth landscapes, even for complicated folding mechanisms (118). In the case of the Lambda Repressor and CI2, a decrease in the free energy barrier is observed (Figure 5-1b,c), as predicted analytically (111, 112). We also note that the unfolded basin free energy minimum occurs at a higher Q (the fraction of native contacts) in the flavored model that in the vanilla model, while conversely the folded basin has lower Q. For src-SH3, however, the free energy barrier does not change when the energetic heterogeneity is introduced (Figure 5-1f). The Φ-values of CI2 and src-SH3 derived from the simulations of vanilla and flavored models are very similar to each other with correlation coefficients greater than 0.70 (Figure 5-1h,i). In contrast, the Φ-values for the Lambda Repressor predicted by the vanilla and fully flavored models are essentially uncorrelated with each other.

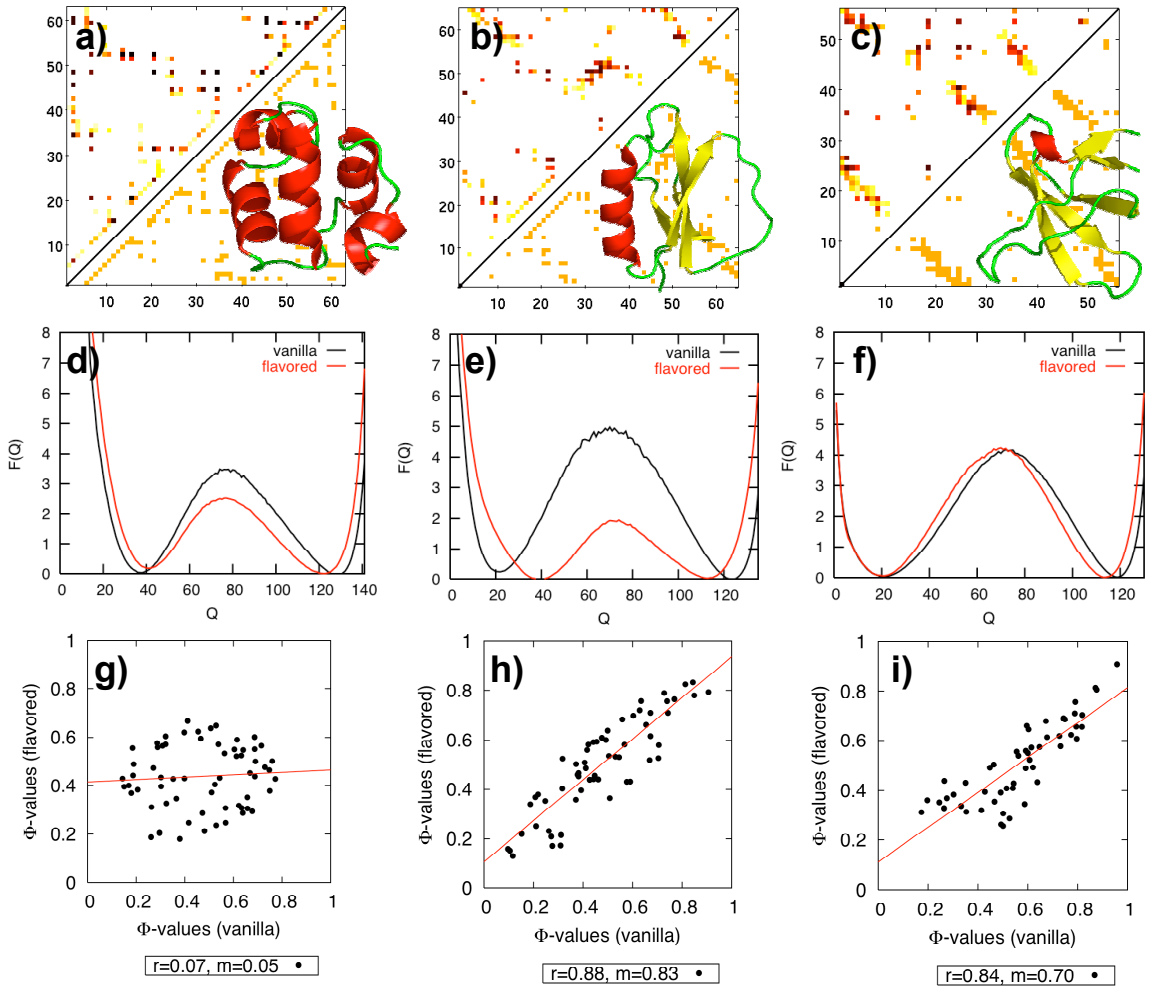


Figure 5-1: The folding mechanisms of the all-α Lambda Repressor (PDB code: 1R69), the α/β CI2 (PDB code: 2CI2), and all-β src-SH3 domain (PDB code: 1SRL). (a-c) The matrices of the interaction energies in the vanilla and flavored native topology-based models are plotted below and above the diagonal, 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.

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 (Figure 5-2). In contrast, the transition state ensemble of the flavored model predominantly includes structured long-range interactions between the

second and fourth helices (Figure 5-2). Oas and coworkers performed NMR spectroscopy of seven alanine to glycine mutants of the Lambda Repressor, and their limited observations indicate that the first and fourth helices are most populated in the transition state ensemble, while the second, third, and fifth helices are less populated (119). It seems that the flavored model agrees with the experimental results more than the vanilla model, but a clear picture is not present in either simulations or experiments.

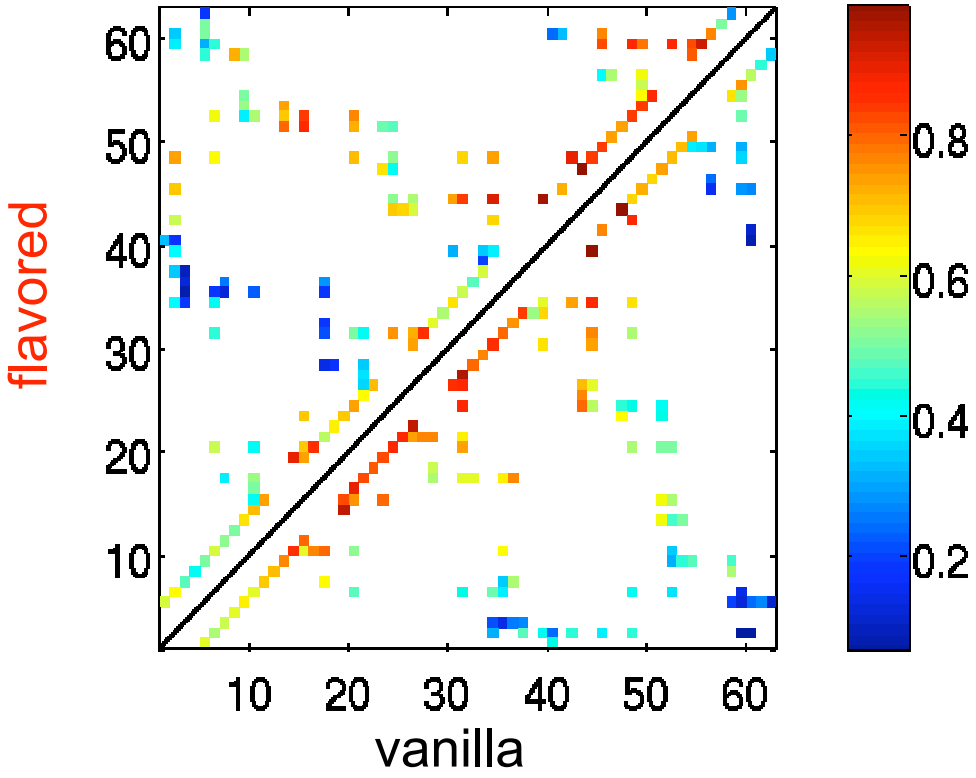


Figure 5-2: The probability of a contact in the transition state of the Lambda Repressor, an all-α protein, with the vanilla and flavored models.

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,

we also studied an inhomogeneous model 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 (Figure 5-3a), but one still finds the poor correlation between the Φ-values in this partially flavored model and the homogeneous case (Figure 5-3b).

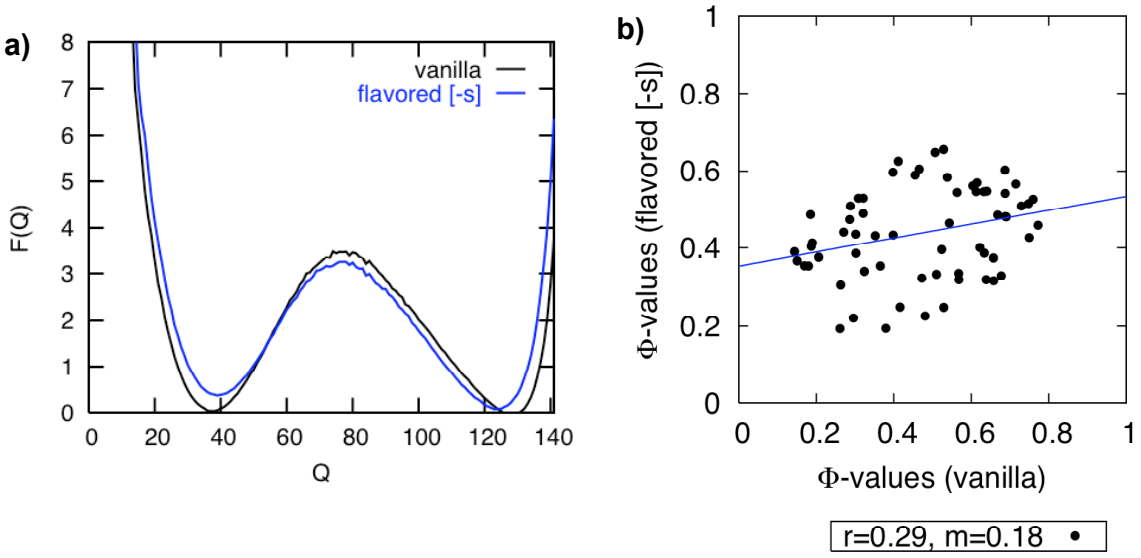


Figure 5-3: The flavored model simulation of the Lambda Repressor, an all-α protein, with the short-range interaction set at the vanilla interaction energies. (a) From simulations, the free energy profiles were generated with respect to the order parameter Q. (d) The Φ-values from the vanilla and flavored models with vanilla short-range interaction weights are compared in a plot with a best-fit line.

We also simulated several other representative all-α protein domains that we selected from the CATH database (120) (CATH ID's: 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 (Figure 5-4a-c). The contact map of 1v54E0 contains mostly of relatively short-range interactions (Figure 5-4a) while 1cy5A0

has a large number of long-range interactions (Figure 5-4c). 1f6vA0 has an intermediate number of long-range interactions (Figure 5-4b). Again, the energetic weights seem to be evenly distributed across all the native interactions (Figure 5-4a-c). In all three cases, the flavored model has a lower free energy barrier than does the vanilla model and the folded basin has a lower Q for the folded model (Figure 5-4d-f). For 1f6vA0, the peak of the free energy barrier occurs at a lower Q in the flavored model (Figure 5-4e). In each case, the Φ-values predicted by the vanilla and flavored models for these all-α proteins exhibit no significant correlation.

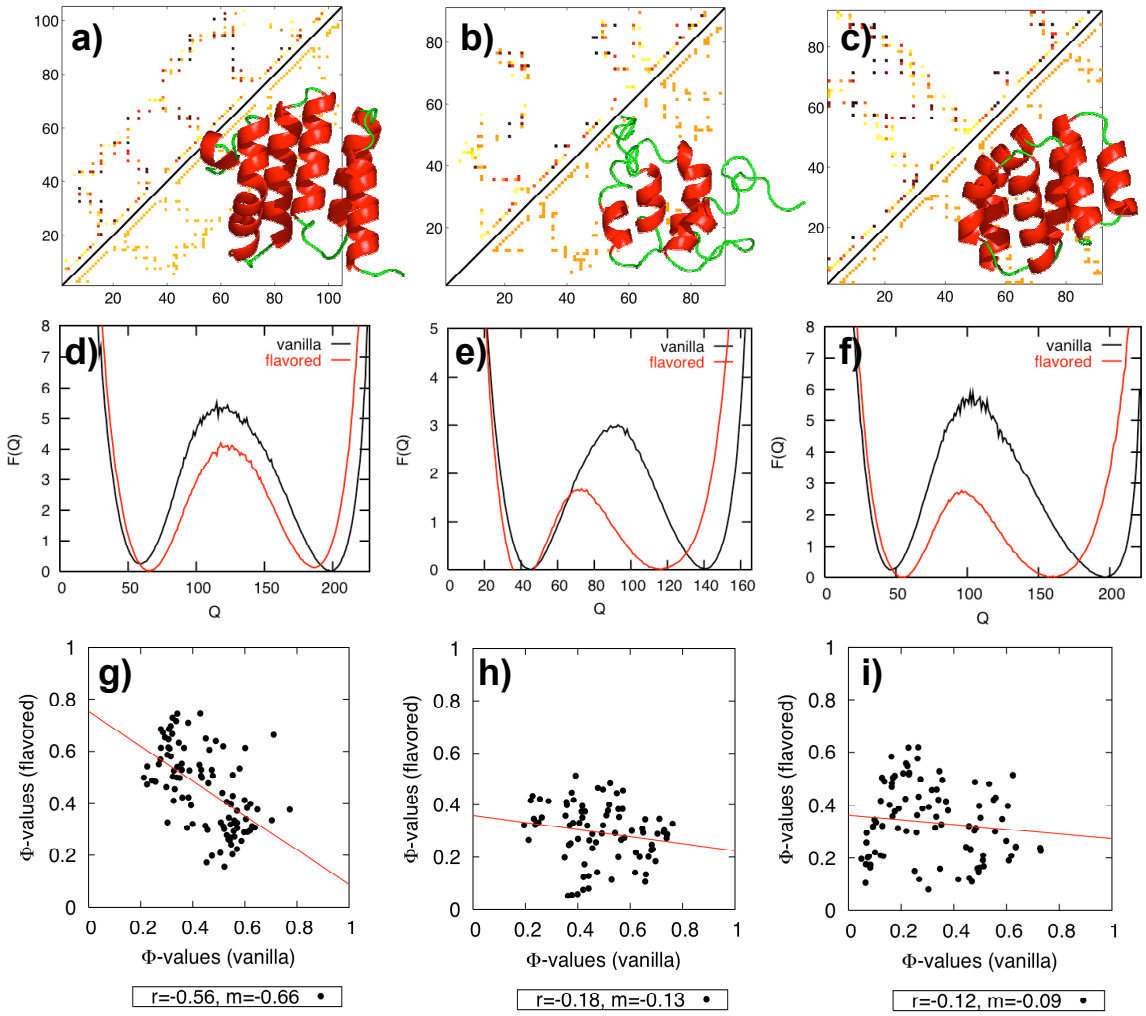
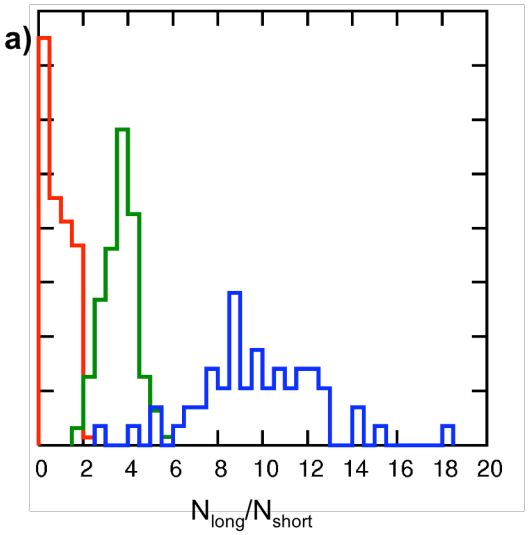


Figure 5-4: The folding mechanisms of three all-α proteins selected from the CATH database. (a-c) The matrices of the interaction energies in the vanilla and flavored models are plotted below and above the diagonal, respectively, with darker colors representing stronger interactions. The corresponding 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.

5.3 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_{long}/N_{short}). Three different peaks appear in the distribution of N_{long}/N_{short} for the nonredundant set of the PDB, corresponding to the all-α, α/β, and all-β topologies (Figure 5-5a). These peaks are also observed when proteins that have been shown to be two-state folders are only included (Figure 5-5b). All-α proteins have proportionally the lowest number of long-range interactions, because the intra-helical interactions stabilize the secondary structure. For all-β proteins, numerous long-range interactions must form between individual sheets.



b)

PDB ID	NumRes	Class	N_{long}/N_{short}
1R69	63	α	1.45
1IMQ	86	α	1.68
3GB1	56	α/β	3.42
2PTL	60	α/β	3.89
1URN	97	α/β	4.40
2CI2	65	α/β	4.67
1RIS	96	α/β	5.12
1SRL	56	β	7.73

Figure 5-5: A comparison of the ratio between long- and short-range native interactions (N_{long}/N_{short}), in sequence, across the different secondary structural classes. (a) A histogram of the long- and short-range interactions from a survey of the nonredundant set of the PDB. (b) A table of well-studied two-state folding proteins with different number of residues, secondary structure topology, and number of long- and short-range native interactions.

To examine the interplay between energetic and entropic contributions to folding, we calculated the energy and entropy lost upon formation of native contacts for the Lambda Repressor, CI2, and src-SH3 domain (Fig. 26). 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}):

$$E(Q) = +\sum_{ij} \varepsilon_{ij} Q_{ij} .$$

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:

$$S(Q) = +\sum_{ij} S_{ij} Q_{ij} .$$

A reasonable approximation to S_{ij} can be found following Shoemaker, Wang, and Wolynes (39). They suggested that initially the entropy lost in forming sequentially short-range interactions can be approximated by the Jacobson-Stockmayer formula (121),

$$S_{ij} = +k_B \log\left[\Delta V / |i-j|^{3/2}\right].$$

Assuming that the denatured protein can be modeled as a random flight chain, the quantity

$$\Delta V = \left(3/2\,\pi\right)^{3/2} \Delta\tau / l_0^3 ,$$

where $\Delta\tau$ is the volume of the interaction range and l_0 is the persistence length. But Shoemaker et al. also argued that if some structure is already formed, the entropy lost will continue to make sequentially distant interactions and saturates to that of a typical fluctuating segment of the chain, as introduced by Flory in the mean field theory of rubber vulcanization (122). This yields:

$$S_{ij} = +k_B \log\left[\Delta V / (\mu/\mathrm{N})^{3/2}\right],$$

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:

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

The resulting free energy functional takes the form:

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

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 energetic part.

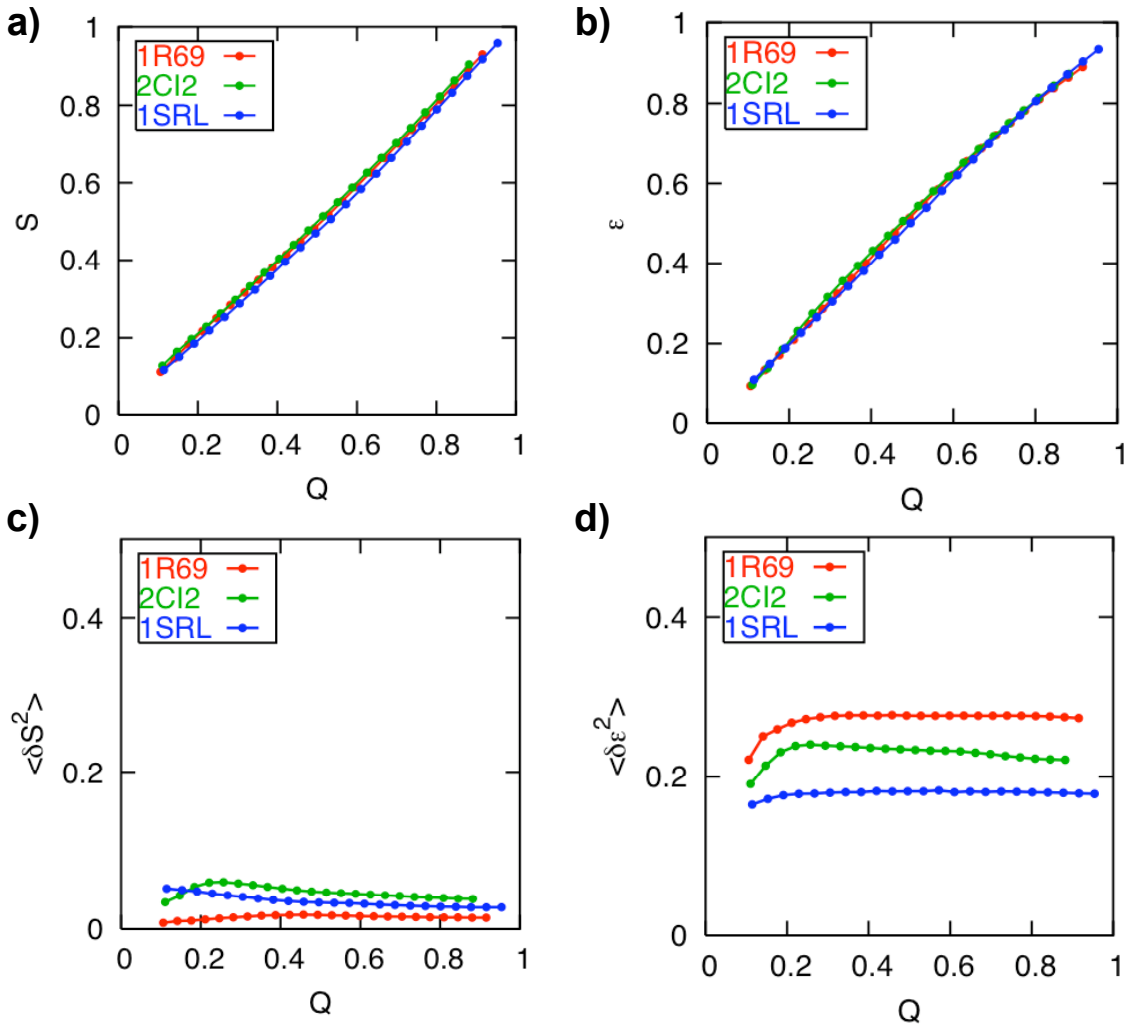


Figure 5-6: The entropy and energy lost from the formation of native contacts for all-α (red), α/β (green), and all-β (blue) proteins. Shown are the (a) entropy and (b) energy, as well as the variance of the (c) entropy and (d) energy, all plotted with respect to the order parameter, Q.

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 (Figure 5-6a,b) for both proteins. On the other hand, the fluctuations, as quantified by the variance, of the entropy costs of forming contacts ($<\delta S^2>$) at a Q value and energies of the formed contacts ($<\delta \varepsilon^2>$) show different trends

for each protein topology (Figure 5-6c,d). By comparing the quantity $<\delta S^2>/<\delta\varepsilon^2>$ at the transition state for each of the proteins, we can quantify the 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 $<\delta S^2>/<\delta\varepsilon^2>$ is strongly correlated with the abovementioned N_{long}/N_{short}, with a correlation coefficient of 0.90 (Figure 5-7). Therefore, for 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, while for proteins with a low number of long-range contacts (e.g., all-α protein), the folding mechanism should be susceptible to energetic fluctuations, if we follow the free energy functional of Shoemaker et al.

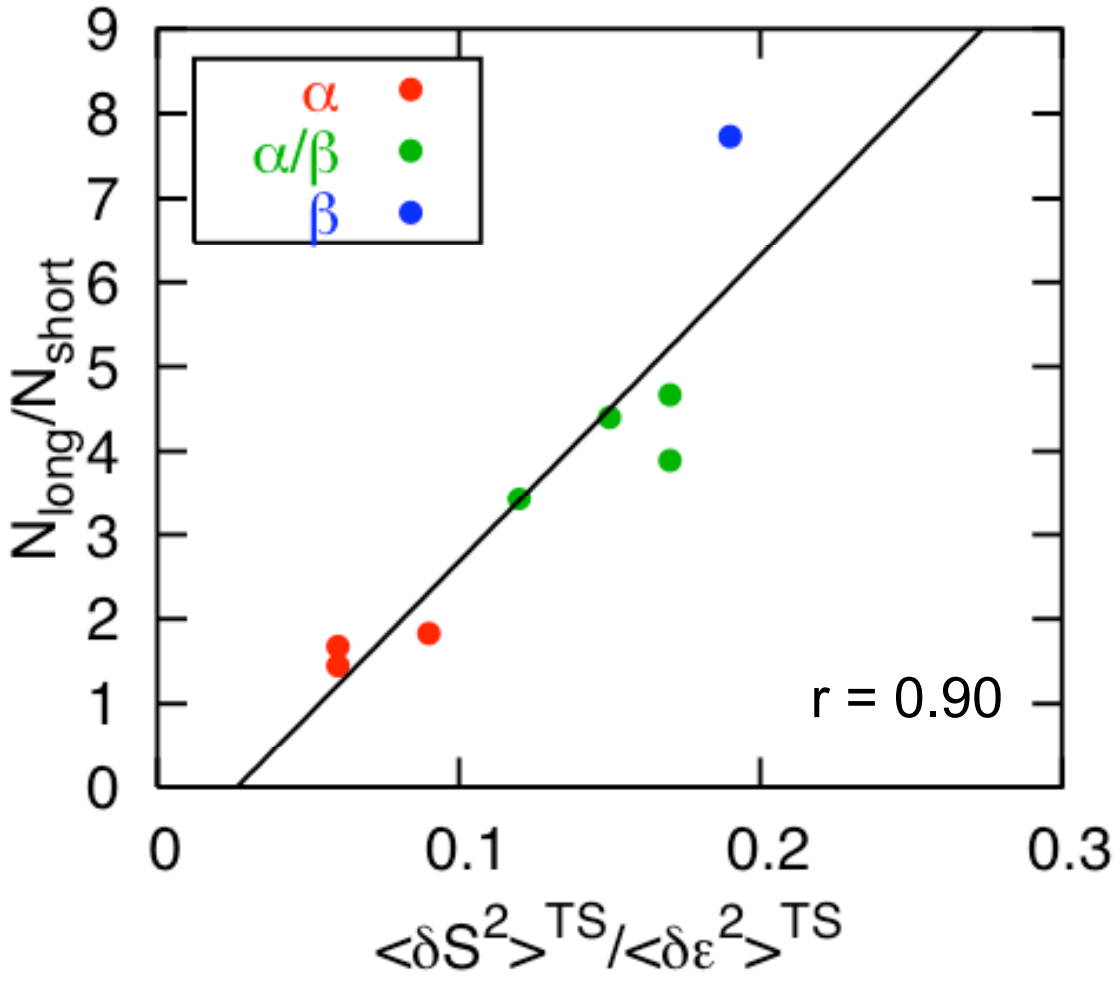


Figure 5-7: The relationship between the ratio of the entropic and energetic fluctuations with the ratio between long- and short-range native interactions for well-studied two-state folding proteins.

5.4 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 $<\delta S^2>/<\delta \varepsilon^2>$ for each protein system. To confirm this, we studied a series of models where $<\delta \varepsilon^2>$ is varied over a range, but $<\delta S^2>$, of course, remains constant for each given protein

topology. We expect that once $<\delta\varepsilon^2>$ increases sufficiently (and thereby decreasing $<\delta S^2>/<\delta\varepsilon^2>$), large deviations in the Φ-values from those of the homogeneous vanilla model will occur. Using this reasoning a key simulation test of the theory 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, like an all-α protein, by using an unrealistically large variation in the native contact energy.

To construct models with varying $<\delta\varepsilon^2>$, we studied variable sets of inter-residue energetic weights, $\varepsilon_{i,j}^{new}$, which can interpolate between the vanilla and the flavored models and that can furthermore extrapolate past the usual flavored model in energetic heterogeneity linearly: $\varepsilon_{i,j}^{new} = \chi\left(\varepsilon_{i,j}^{MJ} - \bar{\varepsilon}^{MJ}\right) + \bar{\varepsilon}^{MJ}$. Here $\varepsilon_{i,j}^{MJ}$ is the original MJ weight for a given residue pair (i,j), $\bar{\varepsilon}^{MJ}$ 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 ($\delta\varepsilon$) 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.

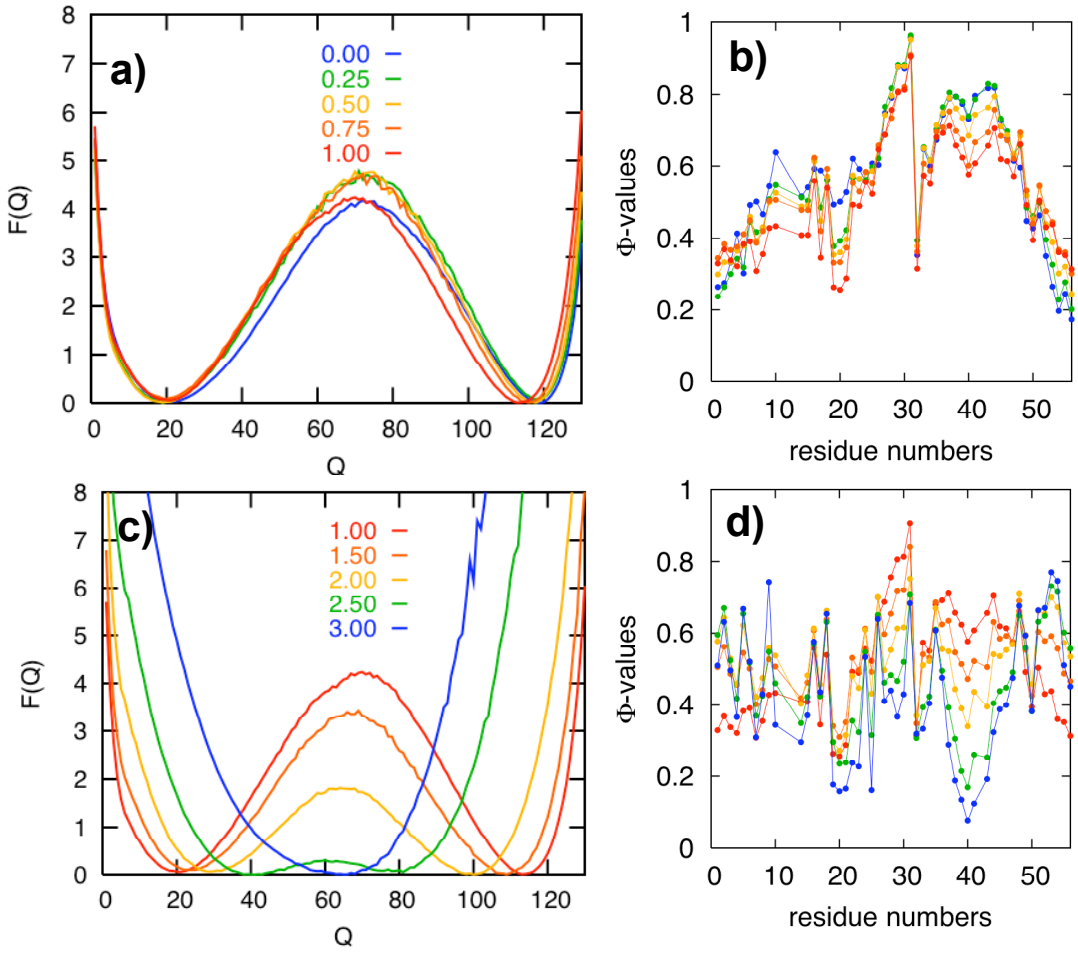


Figure 5-8: Flavored model simulations of src-SH3 domain protein with a range of distributions of the Miyazawa-Jernigan contact energies. The free energy profiles (a) and Φ-values (b) are shown for simulations using the varying parameter, χ, in a range where the folding mechanism does not change significantly. 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.

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 (Figure 5-8a,b). Very little difference is observed between the results of the vanilla and flavored models, as well as the intermediary models. However, when χ is increased past 1, the free energy barrier begins rapidly to decrease while the unfolded state becomes more structured and the folded state

becomes less structured, as is seen for all-α proteins (Figure 5-8c). 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 (123). While this physically unrealistically 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 (112). Also, a marked difference in the Φ-values exists (Figure 5-8d), as was seen earlier only for the all-α proteins. Therefore, with a sufficiently large $<\delta\varepsilon^2>$, 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. Therefore, with a sufficiently large $<\delta\varepsilon^2>$, albeit in an unrealistic regime, the entropy cost 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 Lamda Repressor than the src-SH3 domain (Figure 5-9a). In both proteins, the Φ-values of the flavored models remain close to that of the vanilla model if $<\delta S^2>/<\delta\varepsilon^2>$ is greater than around 0.20 (Figure 5-9b).

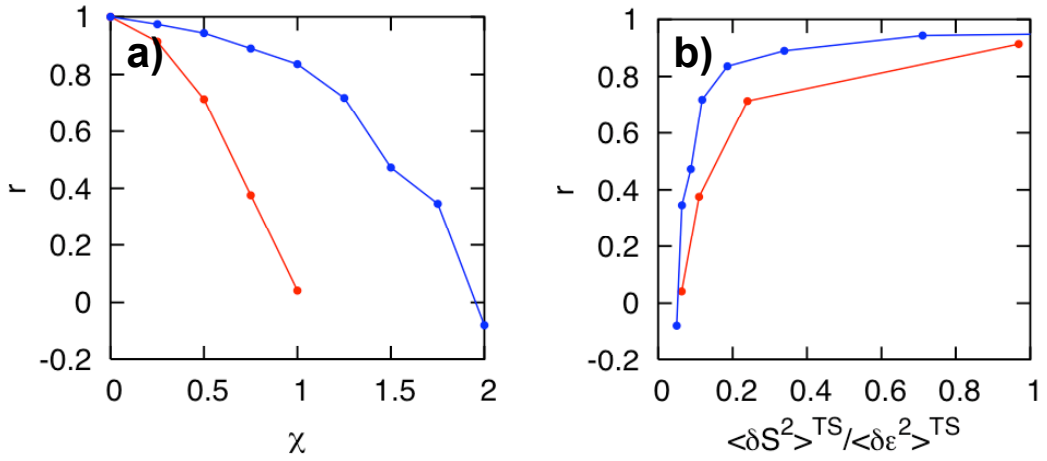


Figure 5-9: The dependence of the correlation between the Φ-values of the vanilla model versus the flavored models, r, with a range of χ (a) and <δS2>/<δε2> (b) for the all-α (1R69; red) and all-β (1SRL; blue) topologies.

Reprinted from:

Cho SS, Levy Y, Wolynes PG. Quantitative Criteria for Native Energetic Heterogeneity Influences in the Prediction of Protein Folding Kinetics. Proc. Natl. Acad. Sci., USA. 2007. (in press)

6. Looking to the Future

It is exciting to be a biophysicist these days. The sequences of entire genomes have been mapped, new biomolecular structures are resolved at an incredible rate, and the basic framework of how proteins fold is now well established. In our present work, we clearly demonstrated the simple elegance of protein folding within the framework of the Energy Landscape Theory. Indeed, simple measures are often sufficient to describe the intricacies of protein folding mechanisms. Even seemingly complicated oligomerization mechanisms, like domain-swapping, are well-described within the Energy Landscape Theory. The protein folding problem seems largely solved, at least for the simplest and idealized cases.

So where do we go from here? While the big picture of protein folding is likely solved, there are still many important exceptional questions that have yet to be resolved. There are still important details that we must know about the mechanisms of large protein complexes, the assembly of aggregates, and the interactions of proteins with other biomolecules. Also, how do natively unfolded proteins fit into the Energy Landscape Theory, if at all? Taking a step back, it is also important to note that much of the efforts of biophysicists until now have been focused on the individual components of the cell, but how they fit together *in vivo* is still far from being understood. It is clear that the next frontier is to address how they interact with one another in the cell to yield biological functions. And I look forward to see how it will all unfold!

REFERENCES

"The secret to creativity is knowing how to hide your sources." – Albert Einstein

1. Onuchic, J. N. & Wolynes, P. G. (2004) *Curr Opin Struct Biol* **14,** 70-75.

2. Levy, Y., Wolynes, P. G., & Onuchic, J. N. (2004) *Proc Natl Acad Sci U S A* **101,** 511-516.

3. Levy, Y., Cho, S. S., Onuchic, J. N., & Wolynes, P. G. (2005) *J Mol Biol* **346,** 1121-1145.

4. Cho, S. S., Levy, Y., Onuchic, J. N., & Wolynes, P. G. (2005) *Phys Biol* **2,** S44-55.

5. Tanford, C. & Reynolds, J. (2001) *Nature's robots : a history of proteins* (Oxford University Press, New York).

6. Prusiner, S. B. (1998) *P Natl Acad Sci USA* **95,** 13363-13383.

7. Anfinsen, C. B. (1973) *Science* **181,** 223-230.

8. Ellis, R. J. (2006) *Trends in biochemical sciences* **31,** 395-401.

9. Dobson, C. M. (2004) *Methods (San Diego, Calif* **34,** 4-14.

10. Levinthal, C. (1969) *Mossbauer Spectroscopy in Biological Systems***,** 22-24.

11. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc Natl Acad Sci U S A* **84,** 7524-7528.

12. Bryngelson, J. D. & Wolynes, P. G. (1989) *J Phys Chem* **93,** 6902-6915.

13. Clementi, C., Nymeyer, H., & Onuchic, J. N. (2000) *J Mol Biol* **298,** 937-953.

14. Koga, N. & Takada, S. (2001) *J Mol Biol* **313,** 171-180.

15. Chavez, L. L., Onuchic, J. N., & Clementi, C. (2004) *J Am Chem Soc* **126,** 8426-8432.

16. Go, N. (1983) *Annu Rev Biophys Bio* **12,** 183-210.

17. Ejtehadi, M. R., Avall, S. P., & Plotkin, S. S. (2004) *Proc Natl Acad Sci U S A* **101,** 15088-15093.

18. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S.*, et al.* (1998) *J Phys Chem B* **102,** 3586-3616.

19. Shakhnovich, E. (2006) *Chem Rev* **106,** 1559-1588.

20. Snow, C. D., Nguyen, N., Pande, V. S., & Gruebele, M. (2002) *Nature* **420,** 102-106.

21. Snow, C. D., Sorin, E. J., Rhee, Y. M., & Pande, V. S. (2005) *Annu Rev Biophys Biomol Struct* **34,** 43-69.

22. MacKerell, A. D., Jr., Feig, M., & Brooks, C. L., 3rd (2004) *J Am Chem Soc* **126,** 698-699.

23. Shirts, M. & Pande, V. S. (2000) *Science* **290,** 1903-1904.

24. Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J.*, et al.* (2003) *Biopolymers* **68,** 91-109.

25. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., & Edelman, M. (1999) *Bioinformatics* **15,** 327-332.

26. Ferreiro, D. U., Cho, S. S., Komives, E. A., & Wolynes, P. G. (2005) *Journal of Molecular Biology* **354,** 679-692.

27. Mello, C. C., Bradley, C. M., Tripp, K. W., & Barrick, D. (2005) *Journal of Molecular Biology* **352,** 266-281.

28. Karanicolas, J. & Brooks, C. L., 3rd (2002) *Protein Sci* **11,** 2351-2361.

29. Li, L. & Shakhnovich, E. I. (2001) *Proc Natl Acad Sci U S A* **98,** 13014-13018.

30. Clementi, C., Garcia, A. E., & Onuchic, J. N. (2003) *J Mol Biol* **326,** 933-954.

31. Cheung, M. S., Garcia, A. E., & Onuchic, J. N. (2002) *P Natl Acad Sci USA* **99,** 685-690.

32. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T., & Shakhnovich, E. S. (1998) *J Chem Phys* **108,** 334-350.

33. Gsponer, J. & Caflisch, A. (2002) *Proc Natl Acad Sci U S A* **99,** 6719-6724.

34. Settanni, G., Rao, F., & Caflisch, A. (2005) *Proc Natl Acad Sci U S A* **102,** 628-633.

35. Ding, F., Guo, W. H., Dokholyan, N. V., Shakhnovich, E. I., & Shea, J. E. (2005) *J Mol Biol* **350,** 1035-1050.

36. Chong, L. T., Snow, C. D., Rhee, Y. M., & Pande, V. S. (2005) *J Mol Biol* **345,** 869-878.

37. Best, R. B. & Hummer, G. (2005) *Proc Natl Acad Sci U S A* **102,** 6732-6737.

38. Nymeyer, H., Socci, N. D., & Onuchic, J. N. (2000) *Proc Natl Acad Sci U S A* **97,** 634-639.

39. Shoemaker, B. A., Wang, J., & Wolynes, P. G. (1997) *Proc Natl Acad Sci U S A* **94,** 777-782.

40. Portman, J. J., Takada, S., & Wolynes, P. G. (1998) *Phys Rev Lett* **81,** 5237-5240.

41. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E., & Shakhnovich, E. I. (2002) *Biophys J* **83,** 3525-3532.

42. Hubner, I. A., Edmonds, K. A., & Shakhnovich, E. I. (2005) *J Mol Biol* **349,** 424-434.

43. Mirny, L. & Shakhnovich, E. (2001) *Annu Rev Biophys Biomol Struct* **30,** 361-396.

44. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z., & Wolynes, P. G. (1996) *Fold Des* **1,** 441-450.

45. Shoemaker, B. A., Wang, J., & Wolynes, P. G. (1999) *J Mol Biol* **287,** 675-694.

46. Dokholyan, N. V., Li, L., Ding, F., & Shakhnovich, E. I. (2002) *Proc Natl Acad Sci U S A* **99,** 8637-8641.

47. Rao, F., Settanni, G., Guarnera, E., & Caflisch, A. (2005) *J Chem Phys* **122,**

184901.

48. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., & Baker, D. (1999) *Nat Struct Biol* **6,** 1016-1024.

49. Itzhaki, L. S., Otzen, D. E., & Fersht, A. R. (1995) *J Mol Biol* **254,** 260-288.

50. Fersht, A. R., Matouschek, A., & Serrano, L. (1992) *J Mol Biol* **224,** 771-782.

51. Eastwood, M. P., Hardin, C., Luthey-Schulten, Z., & Wolynes, P. G. (2003) *J Chem Phys* **118,** 8500-8512.

52. Mosavi, L. K., Minor, D. L., Jr., & Peng, Z. Y. (2002) *Proc Natl Acad Sci U S A* **99,** 16029-16034.

53. Devi, V. S., Binz, H. K., Stumpp, M. T., Pluckthun, A., Bosshard, H. R., & Jelesarov, I. (2004) *Protein Sci* **13,** 2864-2870.

54. Barrientos, L. G., Lasala, F., Delgado, R., Sanchez, A., & Gronenborn, A. M. (2004) *Structure (Camb)* **12,** 1799-1807.

55. Neet, K. E. & Timm, D. E. (1994) *Protein Sci* **3,** 2167-2174.

56. Xu, D., Tsai, C. J., & Nussinov, R. (1998) *Protein Science* **7,** 533-544.

57. Mateu, M. G. & Fersht, A. R. (1998) *Embo J* **17,** 2748-2758.

58. Levy, Y., Caflisch, A., Onuchic, J. N., & Wolynes, P. G. (2004) *J Mol Biol* **340,** 67-79.

59. Gunasekaran, K., Eyles, S. J., Hagler, A. T., & Gierasch, L. M. (2001) *Curr Opin Struc Biol* **11,** 83-93.

60. Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995) *Journal of Molecular Biology* **247,** 536-540.

61. Orengo, C. A., Jones, D. T., & Thornton, J. M. (1994) *Nature* **372,** 631-634.

62. Banner, D. W., Kokkinidis, M., & Tsernoglou, D. (1987) *Journal of Molecular Biology* **196,** 657-675.

63. Eberle, W., Pastore, A., Sander, C., & Rosch, P. (1991) *Biol Chem H-S* **372,** 648-

648.

64. Munson, M., Obrien, R., Sturtevant, J. M., & Regan, L. (1994) *Protein Science* **3,** 2015-2022.

65. Munson, M., Balasubramanian, S., Fleming, K. G., Nagi, A. D., OBrien, R., Sturtevant, J. M., & Regan, L. (1996) *Protein Science* **5,** 1584-1593.

66. Munson, M., Anderson, K. S., & Regan, L. (1997) *Folding & Design* **2,** 77-87.

67. Glykos, N. M. & Kokkinidis, M. (2004) *Proteins* **56,** 420-425.

68. Bennett, M. J., Choe, S., & Eisenberg, D. (1994) *Proc Natl Acad Sci U S A* **91,** 3127-3131.

69. Bennett, M. J., Schlunegger, M. P., & Eisenberg, D. (1995) *Protein Sci* **4,** 2455-2468.

70. Bergdoll, M., Eltis, L. D., Cameron, A. D., Dumas, P., & Bolin, J. T. (1998) *Protein Sci* **7,** 1661-1670.

71. Rousseau, F., Schymkowitz, J. W., Wilkinson, H. R., & Itzhaki, L. S. (2001) *Proc Natl Acad Sci U S A* **98,** 5596-5601.

72. Liu, Y. & Eisenberg, D. (2002) *Protein Sci* **11,** 1285-1299.

73. Park, C. & Raines, R. T. (2000) *Protein Sci* **9,** 2026-2033.

74. Botos, I., Mori, T., Cartner, L. K., Boyd, M. R., & Wlodawer, A. (2002) *Biochem Biophys Res Commun* **294,** 184-190.

75. Liu, Y., Gotte, G., Libonati, M., & Eisenberg, D. (2001) *Nat Struct Biol* **8,** 211-214.

76. Schlunegger, M. P., Bennett, M. J., & Eisenberg, D. (1997) *Adv Protein Chem* **50,** 61-122.

77. Cohen, F. E. & Prusiner, S. B. (1998) *Annu Rev Biochem* **67,** 793-819.

78. Knaus, K. J., Morillas, M., Swietnicki, W., Malone, M., Surewicz, W. K., & Yee, V. C. (2001) *Nat Struct Biol* **8,** 770-774.

79. Janowski, R., Kozak, M., Jankowska, E., Grzonka, Z., Grubb, A., Abrahamson, M., & Jaskolski, M. (2001) *Nat Struct Biol* **8,** 316-320.

80. Wedemeyer, W. J., Welker, E., & Scheraga, H. A. (2002) *Biochemistry* **41,** 14637-14644.

81. Hobohm, U. & Sander, C. (1994) *Protein Sci* **3,** 522-524.

82. Picone, D., Di Fiore, A., Ercole, C., Franzese, M., Sica, F., Tomaselli, S., & Mazzarella, L. (2005) *J Biol Chem*.

83. Lapatto, R., Nalini, V., Bax, B., Driessen, H., Lindley, P. F., Blundell, T. L., & Slingsby, C. (1991) *J Mol Biol* **222,** 1067-1083.

84. Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E., & Shakhnovich, E. I. (2002) *J Mol Biol* **324,** 851-857.

85. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577-2637.

86. Ding, F., Prutzman, K. C., Campbell, S. L., & Dokholyan, N. V. (2006) *Structure* **14,** 5-14.

87. Chen, Y. W., Stott, K., & Perutz, M. F. (1999) *Proc Natl Acad Sci U S A* **96,** 1257-1261.

88. Oliveberg, M. (1998) *Accounts of Chemical Research* **31,** 765-772.

89. Yang, F., Bewley, C. A., Louis, J. M., Gustafson, K. R., Boyd, M. R., Gronenborn, A. M., Clore, G. M., & Wlodawer, A. (1999) *J Mol Biol* **288,** 403-412.

90. Mori, T., Shoemaker, R. H., Gulakowski, R. J., Krepps, B. L., McMahon, J. B., Gustafson, K. R., Pannell, L. K., & Boyd, M. R. (1997) *Biochem Biophys Res Commun* **238,** 218-222.

91. Barrientos, L. G., Louis, J. M., Botos, I., Mori, T., Han, Z., O'Keefe, B. R., Boyd, M. R., Wlodawer, A., & Gronenborn, A. M. (2002) *Structure (Camb)* **10,** 673-686.

92. Welker, E., Wedemeyer, W. J., & Scheraga, H. A. (2001) *Proc Natl Acad Sci U S A* **98,** 4334-4336.

93. Lee, S. & Eisenberg, D. (2003) *Nat Struct Biol* **10,** 725-730.

94. Welker, E., Raymond, L. D., Scheraga, H. A., & Caughey, B. (2002) *J Biol Chem* **277,** 33477-33481.

95. May, B. C., Govaerts, C., Prusiner, S. B., & Cohen, F. E. (2004) *Trends Biochem Sci* **29,** 162-165.

96. Govaerts, C., Wille, H., Prusiner, S. B., & Cohen, F. E. (2004) *Proc Natl Acad Sci U S A* **101,** 8342-8347.

97. Eigen, M. (1996) *Biophys Chem* **63,** A1-18.

98. Masel, J., Jansen, V. A., & Nowak, M. A. (1999) *Biophys Chem* **77,** 139-152.

99. Feughelman, M. & Willis, B. K. (2000) *J Theor Biol* **206,** 313-315.

100. Tompa, P., Tusnady, G. E., Friedrich, P., & Simon, I. (2002) *Biophys J* **82,** 1711-1718.

101. Plaxco, K. W., Simons, K. T., & Baker, D. (1998) *J Mol Biol* **277,** 985-994.

102. Gosavi, S., Chavez, L. L., Jennings, P. A., & Onuchic, J. N. (2006) *J Mol Biol* **357,** 986-996.

103. Martinez, J. C. & Serrano, L. (1999) *Nat Struct Biol* **6,** 1010-1016.

104. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M., & Dobson, C. M. (1999) *Nat Struct Biol* **6,** 1005-1009.

105. Zarrine-Afsar, A., Larson, S. M., & Davidson, A. R. (2005) *Curr Opin Struct Biol* **15,** 42-49.

106. Karanicolas, J. & Brooks, C. L., 3rd (2003) *J Mol Biol* **334,** 309-325.

107. Ferguson, N., Capaldi, A. P., James, R., Kleanthous, C., & Radford, S. E. (1999) *J Mol Biol* **286,** 1597-1608.

108. Friel, C. T., Capaldi, A. P., & Radford, S. E. (2003) *J Mol Biol* **326,** 293-305.

109. Scott, K. A., Batey, S., Hooton, K. A., & Clarke, J. (2004) *J Mol Biol* **344,** 195-205.

110. Bohr, H. G. & Wolynes, P. G. (1992) *Phys Rev A* **46,** 5242-5248.

111. Plotkin, S. S., Wang, J., & Wolynes, P. G. (1997) *J Chem Phys* **106,** 2932-2948.

112. Plotkin, S. S. & Onuchic, J. N. (2000) *Proc Natl Acad Sci U S A* **97,** 6509-6514.

113. Plotkin, S. S. & Onuchic, J. N. (2002) *Q Rev Biophys* **35,** 205-286.

114. Villain, J. (1985) *J Phys-Paris* **46,** 1843-1852.

115. Suzuki, Y. & Onuchic, J. N. (2005) *The journal of physical chemistry* **109,** 16503-16510.

116. Miyazawa, S. & Jernigan, R. L. (1996) *J Mol Biol* **256,** 623-644.

117. Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., & Wolynes, P. G. (2004) *Proc Natl Acad Sci U S A* **101,** 3352-3357.

118. Cho, S. S., Levy, Y., & Wolynes, P. G. (2006) *P Natl Acad Sci USA* **103,** 586-591.

119. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L., & Oas, T. G. (1997) *Nat Struct Biol* **4,** 305-310.

120. Pearl, F. M. G., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., & Orengo, C. A. (2003) *Nucleic Acids Res* **31,** 452-455.

121. Jacobson, H. & Stockmayer, W. H. (1950) *J Chem Phys* **18,** 1600-1606.

122. Flory, P. J. (1956) *J Am Chem Soc* **78,** 5222-5234.

123. Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995) *Proteins* **21,** 167-195.