Designing Misfolded Proteins by Energy Landscaping

Arnab Bhattacherjee and Parbati Biswas*

Department of Chemistry, University of Delhi, Delhi-110007

Received: September 3, 2010; Revised Manuscript Received: November 6, 2010

Conformational fluctuations in the native state ensemble enhance the complexity in designing de novo protein sequences that may fold correctly into a desired target structure. In this work, the results of a self-consistent mean field theory are applied to a cubic lattice model of proteins and real nonhomologous proteins to assess the designability of folded, misfolded, and unfolded conformations. This theory, for the first time, accounts for the properties of misfolded sequences in terms of a generalized foldability criterion and characterizes the topography of the sequence energy landscape in terms of folded, misfolded, and unfolded ensemble of conformations. For a given foldability criterion, the folded, misfolded, and unfolded conformations may be distinctly classified by tuning the energy variance of the native state ensemble. This implies a promising route to de novo protein design and provides useful insights into understanding the impact of conformational similarity/diversity on the folding−misfolding−unfolding transition.

I. Introduction

Most proteins adopt specific three-dimensional structures as a prerequisite for its biological function. Usually proteins achieve this native/folded structure on a biologically relevant time scale even though the conformational space is astronomically large. This conformational change of the polypeptide chain is believed to be a minimally frustrated[1] process implying that the configurational energy landscape[2] is funnel-shaped with the native state lying at the free energy minimum. Landscape ruggedness arises due to the fact that the native state is stabilized by an interplay of various noncovalent interactions, all of which may not be simultaneously satisfied during folding. The covalent connectivity of the protein backbone prevents the interactions among amino acid residues from being favorable. This ruggedness in the energy landscape ensures that in addition to the global minimum corresponding to the folded state there are many additional local minima which correspond to partial/total misfolded states.

For many small single-domain proteins, folding proceeds through two main states, native (N) and denatured/unfolded (U). The energy landscape for this cooperative two-state transition is relatively smooth, though such an ideal folding scenario is rare.[3] Experimental techniques like hydrogen exchange[4−9] and NMR relaxation[10−12] studies reveal significant conformational heterogeneity and suggest that under native conditions a protein has a remarkably rich conformational manifold. Folding or refolding[13,14] to conformations which are of comparable energies as the native fold may lead to change in functionality or large-scale aggregation[15] are often known as amyloid fibril formation.[16] The myriad of near-native conformations often results in off-pathway aggregation[17,18] and causes a range of biological diseases or loss of protein functions particularly responsible for allosteric regulations of enzymes, site-to-site communication, and signal transduction.[19−21]

Designing sequences with funnel-like landscapes for a predetermined target structure requires that the target structure should be energetically stabilized against an ensemble of misfolded conformations.[22] To obtain an optimal sequence−structure compatibility, it is necessary to explore both sequence and conformation space exhaustively.[23,24] However, little is known about the detailed structure and energy of the ensemble of the near-native states. This is partly due to the fact that under native conditions each of the individual conformational states contributes only minutely to the overall observed properties, and there is considerable difficulty in experimentally resolving the contributions from these different states.[25] Also, the size of the sequence space compatible with a given target structure scales exponentially with the size of the protein; even moderately sized proteins of 100 amino acid residues can typically encode an enormous number of possible sequences (20^{100}). Exploring this vast sequence space is feasible, in part, by the enhanced consistency[26] observed in folded proteins. This consistency is often complex and involves a multitude of noncovalent interactions which partially or fully stabilize the target structure. The relative magnitudes of such interactions are intricate and most difficult to determine precisely.

Protein design strategies aim to identify the sequences which optimize these interactions by folding to a desired target conformation. Existing design algorithms are mostly based on either a positive design, in which the target native state is stabilized by a suitable choice of potential, or negative design strategy, where the unfolded/misfolded conformations are destabilized relative to the energy of the native state to ensure a pronounced global free energy minimum. Recent studies highlight the compositional determinants of thermostability including both positive and negative components of design in the context of monomeric soluble proteins[27] and protein complexes.[28] While positive design[29] protocols are commonly used, applying negative design strategy is nontrivial due to the difficulties in modeling a huge number of possible misfolded/unfolded conformations.

A probabilistic or statistical approach based on the simplified model of proteins[30] and scaling laws from polymer physics[31] may be appropriate to characterize the full diversity of sequences and identify the suitable ones corresponding to the target conformation. In this article, we present a self-consistent mean-field based theory[32−34] to investigate the role of near-native states

* To whom correspondence should be addressed. E-mail: pbiswas@chemistry.du.ac.in.

10.1021/jp108416c
Published on Web 12/15/2010

in designing wild-type protein sequences consistent with a generalized foldability criterion. Foldability criteria are usually energy based, quantifying the tendency of a sequence to fold reversibly to a unique native structure. This foldability criterion incorporates negative design features like the difference between the average energies of native state ensemble and unfolded state ensemble denoted by the stability gap (Δ) and the variance of the energies of the native state ensemble and the unfolded state ensemble, denoted by Γ^2_{native} and Γ^2_{unfold}, respectively. Such diverse folding behavior is also studied through a weakly frustrated minimalist protein model[35] in which all native state contacts are not optimal. A new measure of the folding affinity is defined in terms of the stability gap and accessibility of nonnative structures that strongly correlates with the folding kinetics.

The current theory estimates the designability of a structure by using a coarse-grained energy function to evaluate the sequence−structure compatibility. The local/global sequence characteristics are modeled by appropriate constraint conditions. The theory identifies two types of sequences: sequences which fold into the specified target conformation and the partially folded ensemble of conformational states which are either unfolded or misfolded. The theory is applied to cubic lattice proteins and real proteins with two-letter amino acid alphabet, hydrophobic (H) and polar (P). The generalized foldability criterion delineates the sequence energy landscape in terms of both the folded and the unfolded/misfolded conformational state ensemble. Protein design experiments may yield misfolded proteins which represent largely ordered structures without folding to a unique native state. This theory provides a statistical approach to protein sequence design and offers a new perspective in understanding the folding−misfolding−unfolding phenomenon.

II. THEORY

The protein chain is modeled as a 27-residue cubic lattice polymer consisting of sequences with two types of amino acids H and P. Protein conformations are represented by self-avoiding walks[36] on a maximally compact three-dimensional lattice,[37−39] where each residue/structural unit occupies only one lattice site. Exhaustive enumeration yields a total of 103 346 compact conformations which are not related by rotational, reflectional, or translational symmetry. The target/native conformation represents the most designable structure identified by Li et al.[40] and is the lowest-energy conformation for the maximum number of sequences. Extended conformations are rejected as they are typically noncompact and indicate higher-energy states. The input for the theory is the choice of a suitable energy function for evaluating the sequence−structure compatibility and a set of constraints required to specify the local/global features of sequences.

The energy of a particular sequence in a specific target conformation is a function of both the identity and the location of a residue in a particular structure.[41] The typical form of a one-body energy function may be given by

$$E = \sum_{i=1}^{N} \gamma_i(\alpha_i) \tag{1}$$

where $\gamma_i(\alpha_i)$ is the energy contribution at the ith site due to the interaction of the amino acid type α_i and N is the number of amino acid residues present in the protein. This term is dependent on both position and type of the ith amino acid present in a specified structural context in the protein. A variety of such structural contexts are possible which indicates whether the ith site is buried in the structure or accessible to the solvent or the type of secondary structure associated with the ith site. For k different types of structural contexts, the structural information parameter σ_{ik} is defined by

$$\sigma_{ik}^{(1)} = \begin{cases} 1 & \text{if site } i \text{ is in structural context } k, \\ 0 & \text{if not} \end{cases} \tag{2}$$

In this work, the structural contexts are classified according to the burial of hydrophobic residues in the interior of the protein or exposure to its solvent-accessible outer surface. A simple choice is based on the coordination number z_i of the respective residues in a particular structure which is equal to the number of nonbonded nearest neighbors about each site. A simple coarse-grained model of the energy function is given by

$$\gamma_k^{(1)}(H) = -k\varepsilon \text{ and } \gamma_k^{(1)}(P) = 0 \tag{3}$$

where $k = z_i$. This energy function ensures that hydrophobic forces play a dominant role in protein folding, and the hydrophobic residues are favored at high coordination sites, which form the hydrophobic core for a given protein.

Negative design strategies destabilize the unfolded state ensemble due to the repulsive polar interactions. A suitable choice of the coarse-grained potential is given by

$$\gamma_k^{(1)}(\mathrm{H}) = -k\varepsilon,\ \gamma_k^{(1)}(\mathrm{P}) = 0\ (k = 1 \text{ and } 2) \text{ and } \gamma_k^{(1)}(\mathrm{P}) = 1\ (k = 3 \text{ and } 4) \tag{4}$$

For the 27-mer lattice protein, positive values of γ at $k = 3$ and $k = 4$ indicate the repulsive polar interactions in its core. While these energy functions are very simple, the energetics of the sequences rather than their precise folding properties are relevant in the present study.

For the simple binary model of lattice proteins, the similarity measure between the target structure (t) and any other compact conformation (c) may be expressed in terms of the parameter Q defined as

$$Q(k,k') = \frac{1}{N}\sum_{i=1}^{N} \sigma_{ik}(t)\sigma_{ik'}(c) \tag{5}$$

where N is the number of amino acid residues present in the protein and $\sigma_{ik}(t)$ and $\sigma_{ik'}(c)$ denote the different kinds of structural contexts for the target structure and the chosen compact conformation, respectively. The similarity parameter Q in eq 5 measures the similarity in terms of the structural context between two conformations, i.e., whether any residue in both conformations has the same coordination number z_i. The most designable native conformation has four residues with $z_i = 4$ and three residues with $z_i = 3$. These seven residues are predominantly hydrophobic. The remaining 103 345 conformations are equally compact with seven hydrophobic residues, but the coordination number of these residues at these hydrophobic sites varies.

The native state ensemble comprises of the lowest-energy target structure and a set of near-native conformations which may have similar energies. These conformations are identified with a desired degree of structural context similarity with the

target structure. The criterion of selecting the native state ensemble is given by a set of conformations with $Q \geq 0.85$ with a total of 3587 conformations. The unfolded state ensemble is comprised of 99 759 conformations.

The average energy of the folded state ensemble $\langle E_f \rangle$ may be expressed as a function of the site-specific monomer probabilities assuming small fluctuations in E_f about its mean value due to the variation of sequence

$$\langle E_f \rangle = \sum_{i=1}^{N} \sum_{\alpha_i=1}^{2} \sum_{k} \langle \sigma_{ik}^{(1)} \rangle_f \gamma_k^{(1)}(\alpha_i) \omega_i(\alpha_i) \tag{6}$$

This averaging is performed over a suitable set of sequences which satisfy a particular set of imposed constraints. The one-body energy term $\gamma_k^1(\alpha_i)$ denotes the propensity of the ith monomer to reside in the kth structural context. Such one-body propensities are one of the simplest means of quantifying the structural propensities, such as propensities for solvent or for a particular structure. The sequence-averaged energy of an ensemble of the unfolded conformations may be similarly expressed as

$$\langle E_u \rangle = \sum_{i=1}^{N} \sum_{\alpha_i=1}^{2} \sum_{k} \langle \sigma_{ik}^{(1)} \rangle_u \gamma_{ik}^{(1)}(\alpha_i) \omega_i(\alpha_i) \tag{7}$$

The difference in the average folded state energy ($\langle E_f \rangle$) and the ensemble-averaged energy of the unfolded conformations ($\langle E_u \rangle$), denoted as the stability gap Δ, may be expressed as

$$\begin{aligned} \Delta &= \langle E_f \rangle - \langle E_u \rangle \\ &= \sum_{i=1}^{N} \sum_{\alpha_i=1}^{2} \omega_i(\alpha_i) \sum_{k} (\langle \sigma_{ik}^{(1)} \rangle_f - \langle \sigma_{ik}^{(1)} \rangle_u) \gamma_{ik}^{(1)}(\alpha_i) \end{aligned} \tag{8}$$

The energy fluctuations in the folded and unfolded ensemble of states is measured by the respective variance in energy, Γ_f^2 and Γ_u^2, which may be derived as a second-order term in the cumulant expansion of the natural logarithm of the partition function of the folded and unfolded state ensemble, respectively.[32,42] In terms of site-specific monomer identities, the variance of the unfolded ensemble energies is given by

$$\begin{aligned} \Gamma_u^2 &= \langle E_u^2 \rangle - \langle E_u \rangle^2 \\ &= \sum_{i,j} \sum_{\alpha_i,\alpha_j} \sum_{k,k'} \gamma_k^{(1)}(\alpha_i) \gamma_{k'}^{(1)}(\alpha_j) (\langle \sigma_{ik}^{(1)} \sigma_{jk'}^{(1)} \rangle - \langle \sigma_{ik}^{(1)} \rangle \langle \sigma_{jk'}^{(1)} \rangle) \omega_{ij}(\alpha_i, \alpha_j) \end{aligned} \tag{9}$$

The variance of the energy of the folded ensemble of states is given by

$$\begin{aligned} \Gamma_f^2 &= \langle E_f^2 \rangle - \langle E_f \rangle^2 \\ &= \sum_{i,j} \sum_{\alpha_i,\alpha_j} \sum_{k,k'} \gamma_k^{(1)}(\alpha_i) \gamma_{k'}^{(1)}(\alpha_j) (\langle \sigma_{ik}^{(1)} \sigma_{jk'}^{(1)} \rangle - \langle \sigma_{ik}^{(1)} \rangle \langle \sigma_{jk'}^{(1)} \rangle) \omega_{ij}(\alpha_i, \alpha_j) \end{aligned} \tag{10}$$

where the pairwise monomer probability $\omega_{ij}(\alpha_i,\alpha_j)$ may be approximated as

$$\omega_{ij}(\alpha_i, \alpha_j) = \begin{cases} \omega_i(\alpha_i)\omega_j(\alpha_j) & \text{if } i \neq j, \\ \omega_i(\alpha_i)\delta_{\alpha_i,\alpha_j} & \text{if } i = j \end{cases} \tag{11}$$

and $\delta_{\alpha_i,\alpha_j}$ is the Kronecker delta function.

For reversible folding, the free energy ΔF is related to the folding equilibrium constant K_{eq}

$$K_{eq} = \exp(-\beta \Delta F) = Z_f/Z_u \tag{12}$$

where Z_f and Z_u are the partition functions for the folded and unfolded conformational states of the protein, respectively. The partition function for the folded state ensemble comprised of Ω_f states is given by

$$Z_f = \sum_{i=1}^{\Omega_f} \exp(-\beta (E_i)_f) \tag{13}$$

The partition function for the ensemble of unfolded states Ω_u may be written as

$$Z_u = \sum_{i=1}^{\Omega_u} \exp(-\beta (E_i)_u) \tag{14}$$

where $(E_i)_f$ and $(E_i)_u$ are the energies of the ith folded and unfolded conformations, respectively. Minimizing ΔF (or maximizing K_{eq}) with respect to sequence may yield proteins that are stable in a particular target structure. However, with the exception of simple enumerable models, the evaluation of Z_f and Z_u is difficult.

However, $\ln Z_f$ and $\ln Z_u$ may be approximated using a series expansion involving averages over the folded and unfolded states, respectively. The partition function of the ensemble of the folded conformations may be expressed as an average over the folded conformations

$$\sum_{i}^{\Omega_f} e^{-\beta (E_i)_f} = \Omega_f \left(\frac{1}{\Omega_f} \sum_{i}^{\Omega_f} e^{-\beta (E_i)_f} \right) = \Omega_f \langle e^{-\beta (E_i)_f} \rangle \tag{15}$$

Using a cumulant expansion for approximating ensemble-averaged partition functions[43−45] and retaining terms up to second order, K_{eq} may be expressed as

$$\begin{aligned} \ln K_{eq} &= \ln \Omega_f + \left(-\beta \langle E_f \rangle + \frac{1}{2} \beta^2 \Gamma_f^2 \right) \\ &\quad - \ln \Omega_u - \left(-\beta \langle E_u \rangle + \frac{1}{2} \beta^2 \Gamma_u^2 \right) \\ &= \ln \frac{\Omega_f}{\Omega_u} - \left(\beta \Delta + \frac{1}{2} \beta^2 (-\Gamma_f^2 + \Gamma_u^2) \right) \end{aligned} \tag{16}$$

where $\Delta = \langle E_f \rangle - \langle E_u \rangle$. Γ_f^2 is the variance in energy among the folded states, while Γ_u^2 denotes the variance in energy of the unfolded states. The truncation at second order is exact assuming the energy fluctuations among the ensemble of states are a Gaussian process. The term $\ln \Omega_f/\Omega_u$ is identical for all sequences and hence is a constant. A new foldability criterion for quantifying the compatibility of sequences with the target structure can now be defined as

$$\phi = \beta\Delta + \frac{1}{2}\beta^2(-\Gamma_f^2 + \Gamma_u^2) \quad (17)$$

where ϕ is a dimensionless quantity and each of $\beta\Delta$ and $(\beta\Gamma)^2$ are also dimensionless, scaled by appropriate units of thermal energy β. Hence, ϕ is a physical parameter that quantifies sequence−structure compatibility. Although Δ is a useful statistical quantity, it only takes into account the mean energy of the folded $\langle E_f \rangle$ and unfolded conformations $\langle E_u \rangle$. The fluctuations in the folded and unfolded state energies are ignored. Δ may only be a useful foldability criterion for very simple cases, for which there is no information about the low-energy unfolded states that may compete with the target structure. Γ^2 is the simplest quantitative measure of the width of distribution of the energy of the ensemble of states which emerges naturally from the truncated cumulant expansion in eq 17. ϕ exhibits similar features as that of Δ/Γ often known as Z-score. It may be easily proved that minimizing Z-score is equivalent to minimizing ϕ.

The most probable set of the site-specific monomer probabilities $\omega_i(\alpha_i)$ is evaluated by maximizing the sequence entropy S subject to appropriate constraints on sequence identity and energies

$$S = -\sum_{i=1}^{N}\sum_{\alpha=1}^{m} \omega_i(\alpha)\ln \omega_i(\alpha) \quad (18)$$

Normalization of the site identity probability implies that each site is occupied and is given by

$$\sum_{\alpha=1}^{m} \omega_i(\alpha) = 1, \forall i \quad (19)$$

and the energy constraints are defined by any of the eqs 9, 10, and 17. Solving the simultaneous equations that define the maximum of the variational functional of the set of monomer identity probabilities subject to eqs 17 and 19, a set of coupled transcendental equations are obtained.

$$\begin{aligned} \omega_i(\alpha_i) &= \frac{1}{q_i}[\exp(-\beta_\phi \phi_i)] \\ \phi &= \Delta + \frac{1}{2}[-\Gamma_f^2 + \Gamma_u^2] \end{aligned} \quad (20)$$

where $q_i = \Sigma_{\alpha=1}^{m} \exp(-\beta_\phi \phi_i)$; $\phi_i = \partial\phi/\partial\omega_i(\alpha)$; and β_ϕ is the Lagrange multiplier for eq 17. This set of equations is solved numerically to yield the set of site-specific monomer probabilities and the Lagrange multipliers for a given value of ϕ. The individual contribution of Δ and Γ^2 is determined self-consistently from eqs 8, 10, 9, and 20, respectively, and thus can not have any arbitrary values. However, modifying ϕ necessarily means a change in Δ and Γ^2 terms. For the given choice of potentials, most sequences fold at lower values of ϕ, where the value of Δ is negative, which stabilizes the native state ensemble compared to the unfolded state ensemble. This incorporates positive design components in the present problem. Simultaneously, with the increase in ϕ values, the energy variance of the folded and the unfolded state ensemble increases which destabilizes the unfolded states compared to the folded ones. The present algorithm includes information about both stabilizing the native state ensemble and destabilizing the unfolded state ensemble, i.e., negative design.

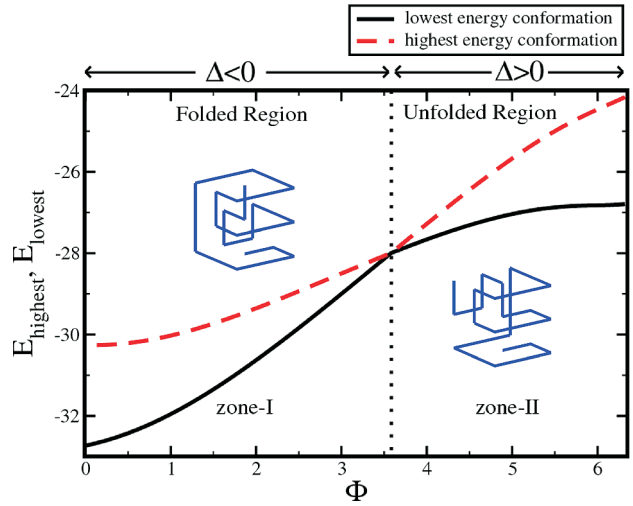

Figure 1. Plot of highest and lowest energy in the native state ensemble vs ϕ for the native state ensemble with 85% similarity to the target conformation.

This theory accounts for the possibility that some sequences may fold to other structures rather than the minimum energy target conformation. The extent to which the native state ensemble, which has a certain degree of similarity with the unique target state, dictates the conformational preferences of the designed protein sequences may be investigated by constraining eqs 10, 17, and 19 simultaneously. The energy fluctuations in the native state ensemble, Γ_f^2, may be suitably tuned to determine the optimal conformation of the protein, i.e., unfolded, misfolded, or folded for a given value of ϕ.

III. Results

For $\Delta < 0$, the target state ensemble is energetically stabilized relative to the competing nontarget structures, and the sequences resemble real proteins. The region with $\Delta > 0$ contains sequences whose ensemble-averaged energy of the unfolded conformations is less than that of the corresponding native state ensemble. These sequences do not fold to the chosen target structure and represent unfolded sequences.

Scanning the sequence space with ϕ values ranging from −0.01 to 6.32 yields 634 sequences constituting the native state ensemble with 85% similarity to the target structure. For each of the 634 sequences, comprising the native state ensemble having 85% similarity to the target structure, the energies of all conformations including the target conformation are calculated; the lowest (denoted by red line) and the highest energies (denoted by blue line) of the native state ensemble are plotted vs ϕ in Figure 1. Two regions observed in this figure correspond to two different Δ values, $\Delta < 0$ (folded region) and $\Delta > 0$ (unfolded region), respectively. In region I, $\Delta < 0$, all designed sequences fold correctly and choose the target structure as their unique native structure (shown in the figure) for values of ϕ ranging from −0.01 to 3.58. The figure clearly shows that the energy range of the native state ensemble decreases with increasing values of ϕ and tends to zero at $\phi = 3.6$. This implies that all conformations in the native state ensemble are of equal energies, representing a degenerate manifold. At $\phi = 3.6$, $\Delta = 0$, indicating that the average energy of the native state ensemble is equal to the average unfolded ensemble energy. Thus some of the unfolded conformations may compete for the degenerate native state. With further increase in ϕ values, the range of the

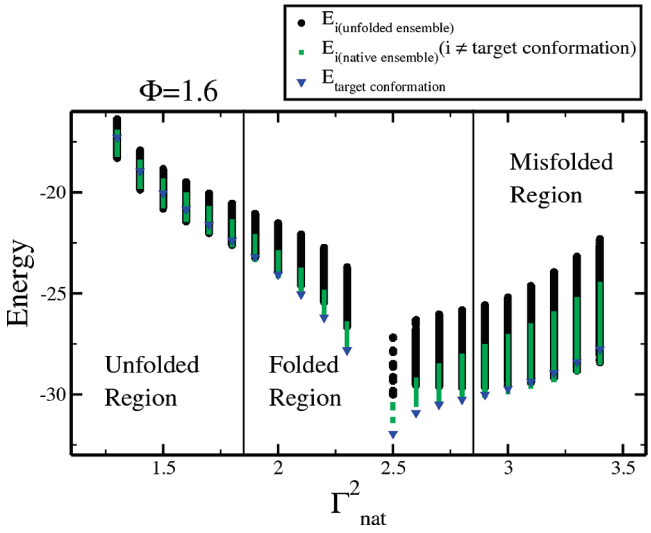

Figure 2. Plot of sequence energies for all conformations vs Γ_f^2 for the native state ensemble with 85% similarity to the target conformation and $\phi = 1.6$.

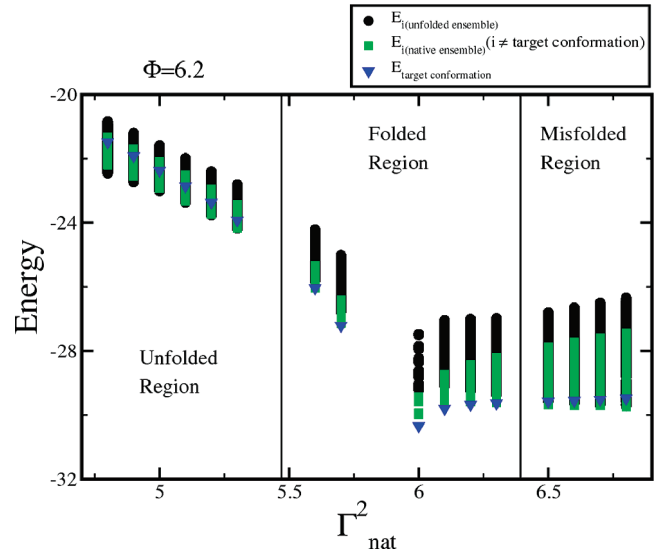

Figure 3. For different potential with repulsive polar interactions, the plot displays the sequence energies for all conformations vs Γ_f^2 for the native state ensemble with 85% similarity to the target conformation at $\phi = 6.2$.

native state ensemble energy increases. In region II, $\Delta > 0$, which implies that the average energy of the unfolded ensemble is lower than the average energy of the native fold ensemble. The protein sequences in this region select the unfolded conformations as the lowest energy state leading to unfolded structures.

To explore the effect of the conformational energy fluctuations of the native state ensemble, Γ_f^2 is varied for a given ϕ to study the *unfolding* → *folding* → *misfolding* transition. Theoretically, the energy of all conformations in the target state ensemble is evaluated self-consistently by solving the set of eq 20 for the site-specific monomer probabilities with an additional constraint of eq 10. For lattice proteins, the native state ensemble with 85% similarity to the target structure is comprised of 21 sequences generated at $\phi = 1.6$ with a range of Γ_f^2 values from 1.3 to 3.4. Figure 2 depicts the target state energy of different sequences (denoted by a blue downward triangle), the energy of all other states except the target state comprising the native state ensemble (denoted by green squares), and all conformations in the unfolded state ensemble (denoted by black circles) as a function of Γ_f^2.

In Figure 2 at $\phi = 1.6$, six sequences are obtained with Γ_f^2 values ranging from 1.3 to 1.8. These sequences possess minimum energy in one of the unfolded conformations as compared to the target state. These sequences do not fold to the target state and hence belong to the unfolded zone. Increasing Γ_f^2 values from 1.9 to 2.8, nine sequences having $\Delta < 0$ choose the specified target state as their unique lowest-energy state. Sequences in this region represent correctly folded sequences. With a further increase of Γ_f^2 in the range of 2.9 to 3.4, six protein sequences choose any conformation from the native state ensemble other than the target conformation. Though sequences in this region have $\Delta < 0$, they do not fold correctly to the specified target conformation and hence represent the misfolded zone.

Γ_f^2 plays a pivotal role in the *unfolding* → *folding* → *misfolding* transition, irrespective of chosen potential. For the potential given in eq 4, the modulation of Γ_f^2 exhibits a similar transition in Figure 3. At $\phi = 6.2$, six sequences are obtained with Γ_f^2 values ranging from 4.8 to 5.4. These sequences possess minimum energy in one of the unfolded conformations relative to the target state. These sequences belong to the unfolded zone. Increasing Γ_f^2 values from 5.5 to 6.3, six sequences having $\Delta < 0$ choose the specified target state as their unique lowest-energy state. Sequences in this region represent correctly folded sequences. With a further increase of Γ_f^2 values from 6.5 to 6.8, four protein sequences choose any conformation from the native state ensemble other than the target conformation. Though sequences in this region have $\Delta < 0$, they do not fold correctly to the specified target conformation and hence represent the misfolded zone.

A. Application to Real Proteins. The crystal structure of a small real protein, villin (PDB id 1WY3) is selected as an example to verify the applicability of this self-consistent mean field theory in the context of real proteins. 1WY3 is a small ultrafast folding protein which constitutes the subdomain of the headpiece of actin binding protein villin and is the smallest naturally occurring polypeptide that folds autonomously without any cofactors or disulfide bonds. The X-ray crystallographic structure of its target conformation is obtained from the protein data bank (http://www.rcsb.org) with a resolution of 0.95 Å and *R*-factor 0.145. Unlike other miniproteins, 1WY3 is a highly thermostable globular protein with a well-packed hydrophobic core. Hydrophobic interactions may be assumed to be the main driving force for folding of this type of well-packed globular proteins.[46] The target structure of 1WY3 is classified into three different coordination zones[47] depending upon the relative solvent accessibility of each site calculated from the ratio of DSSP[48] and free solvent accessibility values. Sites which have greater than 37% relative solvent accessibility are considered to be toward the surface of the protein and are the least hydrophobic. These sites are considered to be in the structural context $k = 1$. The sites with relative solvent accessibility between 7% and 37% are in the intermediate zone, and the corresponding structural context is given by $k = 2$. The innermost hydrophobic core consists of sites having less than 7% relative solvent accessibility and belongs to the structural context $k = 3$. The scoring function used to model the sequence−structure compatibility is based on the surface accessibility of the residues and is given by eq 3.

A set of 2351 nonhomologous protein conformations are obtained from the protein data bank. The selection criterion requires that all conformations should have equal or less number

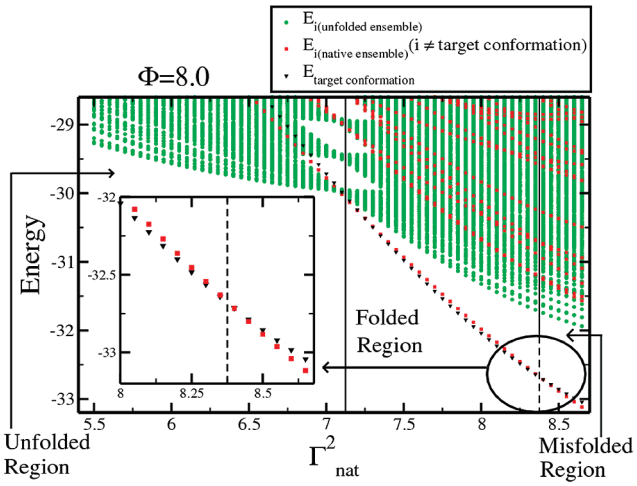

Figure 4. At $\phi = 8.0$ the plot shows the sequence energies for all conformations vs Γ_f^2 in real protein 1WY3 for the native state ensemble having 55% similarity in structural context with that of the target conformation. The energy of the target conformation is denoted by a black downward triangle, and other conformations in the native state ensemble are represented by red squares. The unfolded ensemble of states is represented by green circles. In the inset, the *folding* → *misfolding* transition is magnified to provide a clearer understanding.

of hydrophobic core sites as compared to the target conformation. This ensures optimal designability of protein sequences in the chosen target conformation. The similarity (Q) parameter of all conformations with the target state is determined by eq 5. A subset of 55 conformations is identified as the native ensemble for which $Q \geq 55\%$. The remaining 2296 conformations are considered as an ensemble of unfolded conformations. The structural information σ_{ik} for all the conformations is determined from eq 2. The real protein sequences are obtained from the theory by constraining both ϕ and Γ_f^2, which are calculated from eqs 8, 9, 10, and 17, respectively.

Figure 4 depicts the variation of the sequence energy as a function of Γ_f^2 for a fixed $\phi = 8.0$. For 1WY3, a total of 64 sequences are obtained at $\phi = 8.0$ with Γ_f^2 varying from 5.50 to 8.65. The chosen native state ensemble has 55% similarity to the target conformation of 1WY3. Similar *unfolding* → *folding* → *misfolding* transitions as lattice proteins are observed in the Γ_f^2 range of 5.50−7.10, 7.15−8.35, and 8.40−8.65, respectively. The unfolded region consists of 33 sequences; the folded zone contains 25 sequences; and 6 sequences are present in the misfolded zone. These plots exhibit three distinct regimes. Diverse sequences from unfolded, folded, and misfolded regions may be designed by tuning the variance in the target state energy, Γ_f^2, for a fixed value of ϕ. The inclusion of the native state ensemble topology is an important component in designing misfolded sequences.

Figures 2, 3, and 4 reveal that for a fixed value of ϕ, with an increase of Γ_f^2 values, the unfolded region appears first, followed by the folded and the misfolded region. For low Γ_f^2 values, $\Delta > 0$, the variance in the unfolded state ensemble increases consequently, and some conformations of the unfolded state ensemble may acquire lower energy as compared to the target state causing the respective sequences to unfold. For intermediate Γ_f^2 values, the designed sequences have $\Delta < 0$ and always choose the target state as the lowest-energy conformation. These are the stably folded sequences which prefer the target conformation compared to any other conformations. At the higher values of Γ_f^2 where the energy variance of the native state ensemble is large, some conformations from the native state ensemble may assume lower energy compared to the target conformation. These sequences always fold to near-native conformations rather than the chosen target conformation and represent the misfolded sequences. The results imply that the specific range of Γ_f^2 is an important tuning parameter for predicting the folding, unfolding, or misfolding patterns in designed protein sequences.

IV. Conclusions

The theory provides a quantitative framework to explore the differences in designabilities of the folded, misfolded, and unfolded conformations as it estimates the number of sequences and sequence composition as a function of a given foldability criterion. The theory uses a coarse grained energy function along with an ensemble of native and unfolded states as an input for evaluating the sequence−structure compatibility. Protein design experiments often yield misfolded conformations, which exhibit a large degree of ordering, but do not conform to the unique lowest-energy target structure. The theory complements for such experiments which accounts for the possibility to fold into other near-native structures. Foldable sequences may exhibit conformational diversity and under certain physiological conditions may misfold into structures other than the target one resulting in change or loss of its functional specificity. This theory, for the first time, accounts for the properties of misfolded sequences in terms of a generalized foldability criterion ϕ and the variance in energy of the native ensemble of states Γ_f^2. For a given ϕ, varying the degree of diversity of the native state ensemble results in the different conformations of the designed sequences. Γ_f^2 measures the ruggedness of the native state ensemble and characterizes the topography of the sequence energy landscape in terms of folded, misfolded, and unfolded ensembles of conformational states. For both lattice and real proteins, the specific range of Γ_f^2 may be suitably tuned to design folded, unfolded, and misfolded conformations for a fixed value of ϕ.

The study emphasizes that Δ alone may not solely determine the stability and foldability of sequences, but optimizing the energy fluctuations in the folded and unfolded ensemble of conformations plays a pivotal role in sampling sequences which select the correct target state conformation. More sophisticated and detailed potentials are necessary to design sequences with conformational specificity of real proteins, but the precise folding/misfolding properties of these scoring functions may not be relevant in examining the energetics of sequences as a function of ϕ and Γ_f^2. The theory may be used to evaluate design algorithms with a possibility of some sequences folding to structures other than the target one and provides a statistical foundation for understanding the compatibility between sequence and structure in folding/misfolding proteins.

Acknowledgment. This work was financially supported by DST (SR/S1/PC-07/06), India, and Delhi University Research Grant. A. Bhattacherjee acknowledges the support from CSIR, India, for providing a Senior Research Fellowship.

References and Notes

(1) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524.
(2) Onuchic, J. N.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.
(3) Kaya, H.; Chan, H. S. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 510.
(4) Englander, S. W. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 213.
(5) SwintKruse, L.; Robertson, A. D. *Biochemistry* **1996**, *35*, 171.

(6) Bai, Y. W.; Sosnick, T. R.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192.
(7) Chamberlain, A. K.; Handel, T. M.; Marqusee, S. *Nat. Struct. Biol.* **1996**, *3*, 782.
(8) Fuentes, E. J.; Wand, A. J. *Biochemistry* **1998**, *37*, 9877.
(9) Itzhaki, L. S.; Neira, J. L.; Fersht, A. R. *J. Mol. Biol.* **1997**, *270*, 89.
(10) Yang, D. W.; Kay, L. E. *J. Mol. Biol.* **1996**, *263*, 369.
(11) Li, Z. G.; Raychaudhuri, S.; Wand, A. J. *Protein Sci.* **1996**, *5*, 2647.
(12) Volkman, B. F.; Lipson, D.; Wemmer, D. E.; Kern, D. *Science* **2001**, *291*, 2429.
(13) Aguzzi, A.; Weissmann, C. *Nature* **1998**, *392*, 763.
(14) Ferreira, S. T.; De Felice, F. G. *FEBS Lett.* **2001**, *498*, 129.
(15) Shakhnovich, E. I. *Nat. Struct. Biol.* **1999**, *6*, 99.
(16) DeMarco, M. L.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2293.
(17) Wetzel, R. *Trends Biotechnol.* **1994**, *12*, 193.
(18) Powers, E. T.; Powers, D. L. *Biophys. J.* **2008**, *94*, 379.
(19) Benkovic, S. J.; Hammes-Schiffer, S. *Science* **2003**, *301*, 1196.
(20) Jardetzky, O. *Prog. Biophys. Mol. Biol.* **1996**, *65*, 171.
(21) Taylor, S. S.; Kim, C.; Vigil, D.; Haste, N. M.; Yang, J.; Wu, J.; Anand, G. S. *Biochem. Biophys. Acta* **2005**, *1754*, 25.
(22) Dobson, C. M. *Nature* **2003**, *426*, 884.
(23) Socci, N. D.; Onuchic, J. N. *J. Chem. Phys.* **1994**, *101*, 1519.
(24) Shakhnovich, E. I.; Gutin, A. M. *Nature (London)* **1993**, *346*, 773.
(25) Cremades, N.; Sancho, J.; Freire, E. *Trends Biochem. Sci.* **2006**, *31*, 494.
(26) Go, N. *Adv. Biophys.* **1984**, *18*, 149.
(27) Berezovsky, I. N.; Zeldovich, K. B.; Shakhnovich, E. I. *PLoS Comput. Biol.* **2007**, *3*, e52.
(28) Ma, B.-G.; Goncearenco, A.; Berezovsky, I. N. *Structure* **2010**, *18*, 819.
(29) Shifman, J. M.; Mayo, S. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13274.
(30) Go, N.; Taketomi, H. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 559.
(31) Degennes, P. G. *J. Phys. Lett.-Paris.* **1985**, *46*, L639.
(32) Biswas, P.; Zou, J.; Saven, J. G. *J. Chem. Phys.* **2005**, *123*, 154908.
(33) Bhattacherjee, A.; Biswas, P. *J. Phys. Chem. B* **2009**, *113*, 5520.
(34) Bhattacherjee, A.; Biswas, P. *J. Chem. Phys.* **2009**, *131*, 125101.
(35) Locker, C. R.; Hernandez, R. *J. Chem. Phys.* **2004**, *120*, 11292.
(36) Shaknovich, E.; Gutin, A. *J. Chem. Phys.* **1990**, *93*, 5967.
(37) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907.
(38) Socci, N. D.; Onuchic, J. N. *J. Chem. Phys.* **1995**, *103*, 4732.
(39) Hao, M. H.; Scherega, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4984.
(40) Li, H.; Helling, R.; Tang, C. *Science* **1996**, *273*, 666.
(41) Bowie, J. U.; Luthy, R.; Eisenberg, D. *Science* **1991**, *253*, 164.
(42) Saven, J. G. *J. Chem. Phys.* **2003**, *118*, 6133.
(43) Morrissey, M. P.; Shakhnovich, E. I. *Folding Des.* **1996**, *1*, 391.
(44) Boresch, S.; Karplus, M. *J. Mol. Biol.* **1995**, *254*, 801.
(45) Archontis, G.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 11246.
(46) Dill, K. A. *Biochemistry* **1990**, *29*, 7133.
(47) Guo, J.; Jaromczyk, J. W.; Xu, Y. *Proteins* **2007**, *67*, 548.
(48) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.

JP108416C