Statistical Theory of Protein Sequence Design by Random Mutation

Arnab Bhattacherjee and Parbati Biswas*

Department of Chemistry, University of Delhi, Delhi-110007

Received: September 3, 2008; Revised Manuscript Received: January 9, 2009

A self-consistent mean-field based theory is developed to evaluate the site-specific amino acid pair probabilities in a library of sequences to consider the effect of correlated mutations. This approach computes the entire residue−residue substitution pattern by completely characterizing all possible residue−residue combinations consistent with a given protein structure. Design involves screening a library of sequences with different monomer types to estimate the number and composition of sequences as a function of a generalized foldability criterion. The theory is applied to a simple lattice model of proteins. The theoretical results are respectively compared with real sequences obtained from both the lysozyme protein fold and 1789 nonhomologous globular proteins. The pairwise sequence probability profile of the real proteins show a reasonably good match with that of the lattice proteins with a simple coarse-grained potential. The theory may provide a framework for exploring site directed mutagenesis strategies in engineering known proteins and designing them de novo.

I. Introduction

The advent of directed evolution and DNA mutagenesis/in vitro recombination experiments[1,2] have paved the way to precisely tailor the structure of proteins corresponding to the desired functional changes. Attempts to design sequences which fold into a preassigned structure is the crux of protein design, which is often termed as inverse protein folding.[3,4] Designing novel proteins can yield molecules with new structure and properties which enables drug design and eventually leads to new functional materials. The problem of de novo protein design[5] involves the selection of the target structure and determining the sequences that fold into a given target structure from a huge ensemble of possible sequences. Even small proteins can potentially adopt an enormous number of conformations and can encode an astronomical number of sequences. This combinatorial complexity along with the subtle interplay of various noncovalent interactions[6] makes the folding energy landscape[7] dauntingly complex with virtually infinite possibilities. Most design methods involve energy functions which indicate the ability of a sequence to adopt a particular fold or a given structure. The rationale for choosing the potential function arises from the thermodynamic hypothesis that the native structure of the protein lies at the global minimum of the protein's free energy surface. Different potential energy functions predominate the protein design calculations and protein structure prediction. These range from detailed energy functions with explicit modeling of various interactions at the atomic level[8] to knowledge based potentials[9,10] derived from statistical analysis of a database of solved protein structures to empirical functions[11,12] based on experimental measurements to coarse grained potentials which attempt to model protein free energy.[13]

A statistical theory[14,15] with a coarse grained potential would be most befitting in probing nature's set of protein sequences and scanning the huge sequence space to estimate the number of sequences folding to a given structure. The input for the theory is a suitable target structure and a scoring function for measuring the sequence−structure compatibility. A foldability criterion provides a suitable measure of the compatibility between sequence and structure and is commonly used in both structure prediction and protein design. Such criteria require suitable energy functions; the simplest choice is the energy of a sequence in a particular folded structure. The constraints on the sequences can be tuned to specify the local/global features.

In this article, we present a second-order mean-field based formalism to determine the complete set of site specific monomer pair probabilities consistent with a generalized foldability criterion.[16] This foldability criterion includes negative design features like the mean energy of the unfolded ensemble of states denoted by Δ and the variance of the unfolded conformational energy characterized by Γ^2. The calculation of Γ^2 necessitates the evaluation of at least the pair correlation between the residue sites. The pairwise monomer probability profile explains the pattern of correlated mutations of amino acids between residue sites. This information can be used to identify the site-specific substitution pattern of different residues with respect to all possible residue−residue combinations for a given protein. This statistical formalism may be used to assess the designability of the structure as it yields a numerical estimate of the number of sequences as a function of the respective foldability criteria. The theory is applied to three-dimensional lattice proteins based on a simple binary patterning of hydrophobic and hydrophilic residues which encode the protein structure at the coarse grained level.[4,17] The results of the theory are compared to data sets of real proteins where the effects of side chain packing are neglected.

II. Theoretical Methods

Protein design motivates the choice of optimized energy functions which quantify the various interactions stabilizing the folded state. A common choice of a scoring function primarily includes pair interactions in the form of a contact potential. Although some have suggested that pair potentials alone are insufficient for protein structure prediction,[18] statistical pair potentials are useful in quantifying inter-residue contact propensities as well as excluded volume interactions. Another

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

10.1021/jp810515s CCC: $40.75
Published on Web 03/26/2009

work[19,20] along similar lines correlates protein design with folding stability based on a model of protein sequence evolution with mutations. Stability is assessed through an effective free energy function which results from a combination of hydrophobic effect with local interactions responsible for secondary structure formation. Here, the two-body contact interaction matrix is approximated through the main component of its spectral decomposition and expressed as the product of the hydrophobicities of two types of amino acids. The site specific amino acid distribution is determined analytically for this model. In this work, we self-consistently evaluate both the site specific monomer probability and the pairwise monomer probabilities with an optimized energy function consisting of two-body inter-residue contact potential.

For reversible folding, the target/native state represents a global free energy minimum. The target state energy E_f having different choices of amino acid residues for each pair of sequence positions may be expressed as a function of pairwise monomer probabilities. For a given set of sequences, the fluctuations in the target state energy about its mean value due to variation of sequences are assumed to be small. Thus, E_f can be written as

$$E_f \approx \overline{E_f} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{\alpha_i=1}^{m} \sum_{\alpha_j=1}^{m} \sigma_{ij}^{(2)} \gamma_{ij}^{(2)}(\alpha_i, \alpha_j) w_{ij}(\alpha_i, \alpha_j) \quad (1)$$

where the two-body interaction parameter $\gamma_{ij}^{(2)}(\alpha_i,\alpha_j)$ denotes the pair interaction between sites when their monomer types are α_i and α_j respectively. The structure information is contained in the parameter $\sigma_{ij}^{(2)}$, given by

$$\sigma_{ij}^{(2)} = \begin{cases} 1 \text{ if site } i \text{ and } j \text{ interact with one another} \\ 0 \text{ if not} \end{cases} \quad (2)$$

and $w_{ij}(\alpha_i,\alpha_j)$ is the monomer pair probability that the monomer type α_i occurs at position i and the monomer type α_j occurs at position j in a particular sequence.

Results of kinetic studies on multiple sequences reveal a large correlation of the folding rate with the folded state stability, especially with the Z-score value (Δ/Γ),[21,22] where Δ is the difference of the target state energy and the average energy of the ensemble of unfolded states and Γ is the width of the distribution of energy values of an ensemble of unfolded conformations.

$$\Delta \equiv E_f - \langle E \rangle_u \approx \overline{E_f} - \overline{\langle E \rangle_u} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{\alpha_i=1}^{m} \sum_{\alpha_j=1}^{m} (\sigma_{ij} - \langle \sigma_{ij} \rangle) \gamma_{ij}(\alpha_i, \alpha_j) w_{ij}(\alpha_i, \alpha_j) \quad (3)$$

where $\langle E_u \rangle$ is the average energy of the unfolded ensemble of states. Although Δ is a useful statistical quantity,[23] it does not characterize all sequences. It takes into account only the mean energy of the unfolded conformations, that is, $\langle E_u \rangle$. However, it contains no information about the low energy unfolded states which may compete with the target structure. The fluctuations in the unfolded ensemble of states are ignored. Γ^2 is the simplest quantitative measure of the variance of the energy of the unfolded states for each sequence, which emerges naturally as a second order term from the truncated cumulant expansion.[24] A suitable foldability criteria ϕ characterizing sequences compatible with a particular target structure can be thus obtained as a linear combination of both Δ and Γ^2.[15] The correlation between the different monomer identities can be directly obtained by considering the variance of the nonfolded states ensemble, Γ^2, which is given by

$$\Gamma^2 = \sum_{i<j} \sum_{k<l} \sum_{\alpha_i,\alpha_j,\alpha_k,\alpha_l} \gamma_{ij}(\alpha_i, \alpha_j) \gamma_{kl}(\alpha_k, \alpha_l) (\langle \sigma_{ij} \sigma_{kl} \rangle - \langle \sigma_{ij} \rangle \langle \sigma_{kl} \rangle) w_{ijkl} \quad (4)$$

Many-body correlation effects arising in the above equation, due to the multiple degrees of freedom, are typically complex and difficult to solve. A commonly used approximation to simplify the four-body probability $w_{ijkl}(\alpha_i,\alpha_j,\alpha_k,\alpha_l)$ in terms of respective two-body probabilities is given by

$$w_{ijkl}(\alpha_i, \alpha_j, \alpha_k, \alpha_l) = w_{ij}(\alpha_i, \alpha_j) w_{kl}(\alpha_k, \alpha_l) \quad (5)$$

The goal of the design problem is to identify a sequence that maximizes the sequence entropy subject to any constraints. The pairwise monomer identity probabilities are solved by maximizing[25] the entropy subject to the constraints on sequence identity and energies. The entropy of a set of sequences Ω_s is defined as

$$S = k_B \ln \Omega_s \quad (6)$$

where k_B is the Boltzmann constant. The cluster variation method (CVM)[26,27] is used to derive the sequence entropy when higher order correlations among residue sites are considered.

$$S^{(n)}(i_1, ..., i_n) = - \sum_{\alpha_{i_1},...,\alpha_{i_n}} w^{(n)}_{i_1,...,i_n}(\alpha_{i_1}, ..., \alpha_{i_n}) * \ln w^{(n)}_{i_1,...,i_n}(\alpha_{i_1}, ..., \alpha_{i_n}) \quad (7)$$

For a set of sequences satisfying a predetermined set of constraints, the Bethe approximation[28] is used to estimate the probability $W_N(\alpha_1,..., \alpha_N)$ of obtaining a particular sequence $(\alpha_1,..., \alpha_N)$ as a product of all pairwise monomer probabilities scaled appropriately so as to avoid double counting.

$$W_N(\alpha_1, ..., \alpha_N) = \prod_{i,j} \frac{w_{ij}(\alpha_i, \alpha_j)}{w_i(\alpha_i) * w_j(\alpha_j)} \quad (8)$$

Constraints on the structure and sequences couples the pair probabilities $w_{ij}(\alpha_i,\alpha_j)$. Within this approximation, considering only pair correlations among the residue sites, the total sequence entropy S can be recast into

$$S \approx \sum_i S^{(1)}(i) + 2 \sum_{i<j} (S^{(2)}(i,j) - S^{(1)}(i) - S^{(1)}(j)) \quad (9)$$

The pairwise monomer probabilities are equilibriated in the ensemble by maximizing a variational functional of the set of probabilities subject to the following constraints

$$\sum_{\alpha_i,\alpha_j} w_{ij}(\alpha_i,\alpha_j) = 1 \quad \forall i,j > i \tag{10}$$

$$w_i(\alpha_i) = \sum_{\alpha_j} w_{ij}(\alpha_i,\alpha_j) \quad \forall i,j \neq i \tag{11}$$

$$w_j(\alpha_j) = \sum_{\alpha_i} w_{ij}(\alpha_i,\alpha_j) \quad \forall j,j \neq i \tag{12}$$

and eqs 3 and 4.

Solving the simultaneous equations that define the maximum of the variational functional of probabilities and the constraint equations, the following set of coupled transcendental equations are obtained.

$$\begin{aligned} w_i(\alpha_i) &= \frac{1}{q_i}\left[\exp\left(\beta_\varphi \varphi_i(\alpha_i) + \sum_j \xi_{ij}\beta_{ij}(\alpha_i) + \sum_j \xi_{ji}\mu_{ij}(\alpha_i)\right)\frac{1}{N-2}\right] \\ w_{ij}(\alpha_i,\alpha_j) &= \exp(\beta_{ij}(\alpha_i) + \mu_{ij}(\alpha_j) - \beta_\phi \phi_{ij}(\alpha_i,\alpha_j)) - 1 \\ w_i(\alpha_i) &= \sum_{\alpha_j} w_{ij}(\alpha_i,\alpha_j) \\ w_j(\alpha_j) &= \sum_{\alpha_i} w_{ij}(\alpha_i,\alpha_j) \\ \phi &= \Delta + \frac{1}{2}\Gamma^2 \end{aligned} \tag{13}$$

where the single site partition sum q_i is given by

$$q_i = \sum_{\alpha_i}\left[\exp\left(\beta_\phi \phi_i(\alpha_i) + \sum_j \xi_{ij}\beta_{ij}(\alpha_i) + \sum_j \xi_{ji}\mu_{ij}(\alpha_i)\right)\frac{1}{N-2}\right]$$

The function ξ_{ij} arises due to possible constraints on the allowed types of residues at each sequence position and is given by

$$\xi_{ij} = \begin{cases} 1 & \text{if site } j > i \\ 0 & \text{if not} \end{cases} \tag{14}$$

The Lagrange multipliers β_{ij}, μ_{ij}, and β_ϕ arise due to the constraint conditions eqs 11, 12, and 14, respectively. Also,

$$\phi_i = \frac{\partial \phi}{\partial w_i(\alpha_i)} \tag{15}$$

and

$$\phi_{ij} = \frac{\partial \phi}{\partial w_{ij}(\alpha_i,\alpha_j)} \tag{16}$$

The pairwise probability w_{ij} has a form similar to that of Boltzmann statistics. This system of equations are nonlinear functions of the Lagrange multipliers and the probabilities. These coupled equations are solved numerically [http://www.netlib.org] for the respective Lagrange multipliers, monomer site-specific probability and the pairwise probability profile of amino acid residues consistent with a given value of the foldability criteria, ϕ. The fortran program finds a zero of a system of n nonlinear equations in n variables by a modification of the Powell hybrid method. An ordinary desktop computer takes approximately 20−30 min depending on the foldability criterion ϕ to solve this set of equations. For the types of constraints used to specify the sequences, the theory is equivalent to a heterogeneous mean field theory. The theory is characterized by the generalized foldability criteria ϕ which comprises the target state energy, E_f, the mean energy of the unfolded ensemble, $\langle E\rangle_u$ and the variance of the ensemble of unfolded states, Γ^2. E_f is usually fixed through the choice of the particular target state. The average local energy at each site and for each pair of sites along with the respective Lagrange multipliers are evaluated self-consistently. There are no other adjustable parameters in the theory.

The theory is applied to an exactly solvable system of a three-dimensional cubic lattice model of proteins[29] consisting of sequences made of two kinds of amino acids H and P configured as self-avoiding walks on a lattice. With the binary patterning of amino acid residues there are $2^{27} = 134\,217\,728$ possible sequences which is large, yet computationally enumerable. The optimized energy function contains only two-body terms characterized by the following energy parameters:[30]

$$\gamma^{(2)}(H,H) = -3\epsilon, \gamma^{(2)}(H,P) = -1\epsilon \quad \text{and} \quad \gamma^{(2)}(P,P) = 0 \tag{17}$$

For the target structure, the conformation discussed by Li et al. is considered.[30] The choice of the energy function also ensures that the target structure is the most designable structure and represents the lowest energy conformation for the largest number (3794) of sequences.

The value of Δ and Γ^2 are specified by considering the ensemble of unfolded states. Exhaustive enumeration of all conformations are performed by the first depth algorithm,[29] which yields 103 346 compact cubic conformations unrelated by rotation or reflection symmetry. The lowest energy state is the target/native state and the ensemble of unfolded states comprise the remaining 103 345 conformations. Other extended conformations are rejected as they are typically noncompact and higher in energy. From this ensemble of unfolded conformations, the average value of the structure parameter $\langle \sigma_{ij}^2\rangle$ for each possible $N(N-1)/2 = 351$ pairs can be calculated. However, all pairs are not possible for the cubic lattice model; residues which are not in contact with one another due to the topological constraint of compact conformations on the lattice, never contribute and consequently $\sigma_{ij}^2 = 0$. Δ and Γ^2 are calculated from eq 3 and eq 4, respectively.

III. Results and Discussions

A. Numerical Study. The ability of the present model to design protein sequences can also be verified numerically by determining the probability that a randomly drawn sequence from the analytic distribution can choose the target structure as its unique native state structure as a function of any foldability criterion. Here, ϕ is the required foldability criterion which takes into account both mean and variance of the energy of the ensemble of unfolded conformations. A set of 21 sequences are chosen randomly corresponding to 21 unique ϕ values ranging from −10 to 10 with the chosen target structure as its native state. The number of possible conformations at different ϕ values which can compete with the chosen target state are evaluated by scanning the entire

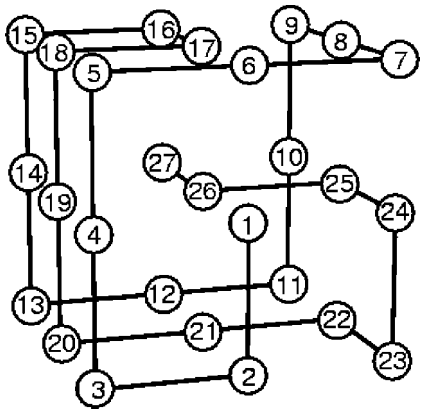

Figure 1. Target structure in cubic lattice.

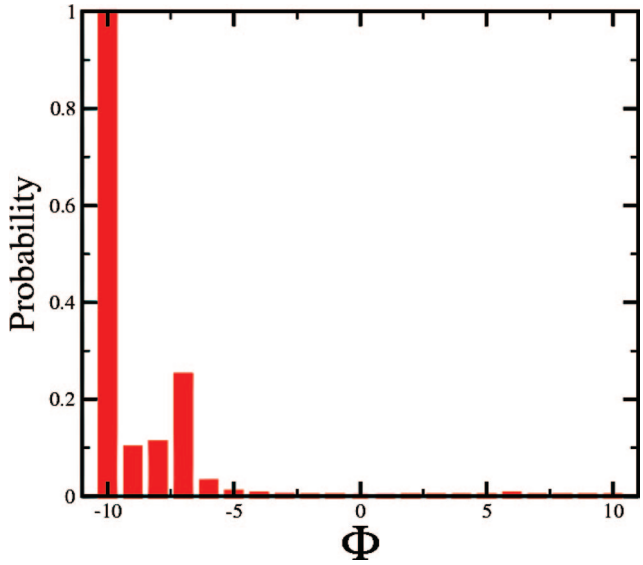

Figure 2. Variation of the probability of finding the target structure as the native one by random sequence vs ϕ.

conformation space of 103 346 conformations for each of the 21 sequences. Results are plotted in Figure 2which depicts the probability that randomly selected sequences from the sequence pool can choose the target structure as their unique native conformation as a function of ϕ. For a given sequence, this probability is found to be the inverse of the number of conformations consistent with a given value of ϕ. The probability peaks to a maximum at the extreme negative value of ϕ (Figure 2). Negative values of ϕ are correlated with more negative values of Δ which suggest that the energy of the target structure is much lower than that of the average energy of the unfolded ensemble of states. The target structure is energetically stabilized in this regime relative to the competing nontarget structures and is favorable for the folding of protein sequences. Nevertheless, the choice of the target conformation as the unique native conformation by a set of sequences can not be solely assesssed on the basis of the probabilities found for a single randomly chosen sequence.

To test the applicability of the theory in designing sequences, the theoretically designed sequences for a particular value of ϕ are selected with respect to their choice of the specified target structure as the unique native state. For each value of ϕ ranging from -8 to 8, the present theory yields 17 sequences probabilistically, which assumes this target structure as their unique native structure. For these 17 sequences, the energy (E_f) of all possible compact conformations are calculated and plotted with respect to corresponding ϕ values. Figure 3 demonstrates that a designed

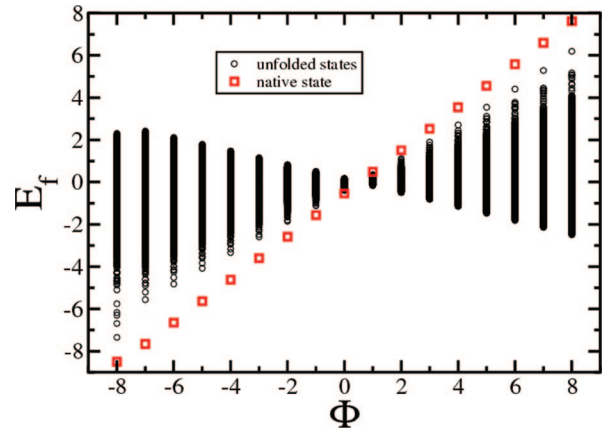

Figure 3. Energy of all 103 346 conformations are calculated for the theoretically predicted sequences at a given ϕ. The sequences choose the target structure as their native structure (red square) as a pronounced energy minimum.

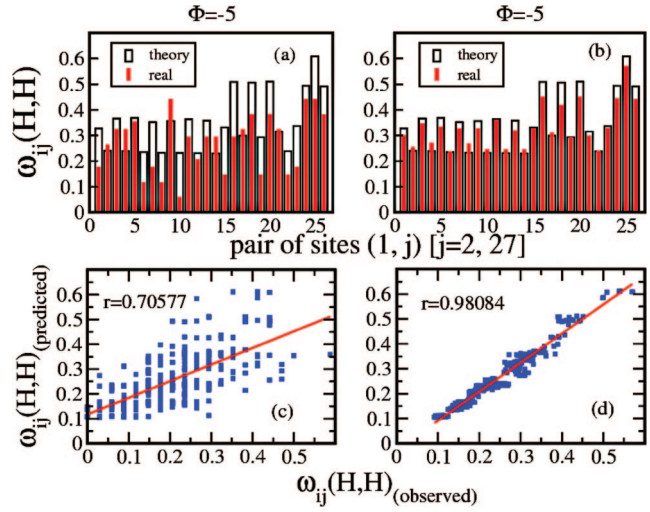

Figure 4. (a,b) Pairwise hydrophobic−hydrophobic probability profile for all possible pairs of site 1. (a) Comparison of results between theory (black bars) and that of lysozyme fold (red solid bars). (b) Comparison of results between theory (black bars) and all globular protein segments (red solid bars). (c,d) Correlation between results corresponding to a and b, respectively.

sequence, corresponding to any given negative ϕ value, can identify the target conformation as the distinct native conformation characterized by minimum energy which may compensate the loss of conformational entropy upon folding. For positive ϕ values, the trend reverses completely, and the mean energy of the ensemble of unfolded conformations is much lower compared to that of the target structure. The target structure is destabilized in this region and here the sequences do not fold to the chosen target structure. This folding pattern is dependent on the choice of the energy function.

B. Application to Real Proteins. The results obtained from the self-consistent mean field theory are compared with two different sets of real proteins. The first data set comprises of the lysozyme fold which contains 978 globular proteins, with a total of 1270 protein chains. Their X-ray crystallographic structures are obtained from the protein databank (http://www.rcsb.org). The maximum sequence identity between any two sequences is 100%. The second data set consists of a database of 1787 globular proteins having a total of 4428 chains. Their X-ray crystallographic structures are obtained from the protein database PDBselect (http://bioinfo.tg.fh-giessen.de/pdbselect/recent.pdb_select25). The criteria

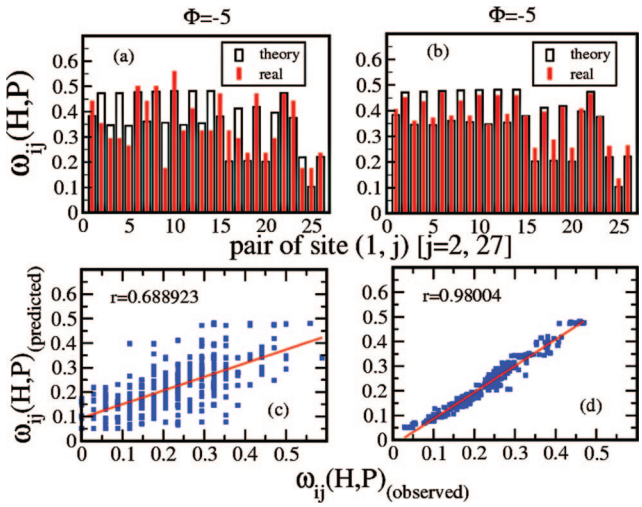

Figure 5. (a,b) Pairwise hydrophobic−polar probability profile for all possible pairs of site 1. (a) Comparison of results between theory (black bars) and that of lysozyme fold (red solid bars). (b) Comparison of results between theory (black bars) and all globular protein segments (red solid bars). (c,d) Correlation between results corresponding to a and b, respectively.

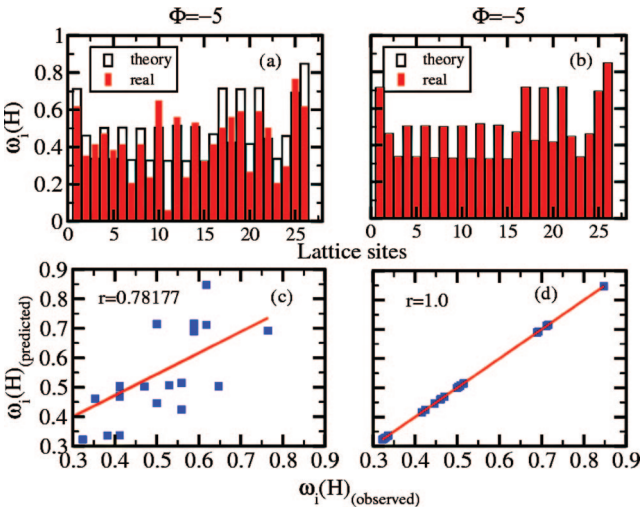

Figure 6. (a,b) Monomer hydrophobic probability profile for all lattice sites. (a) Comparison of results between theory (black bars) and that of lysozyme fold (red solid bars). (b) Comparison of results between theory (black bars) and all globular protein segments (red solid bars). (c,d) Correlation between results corresponding to a and b, respectively.

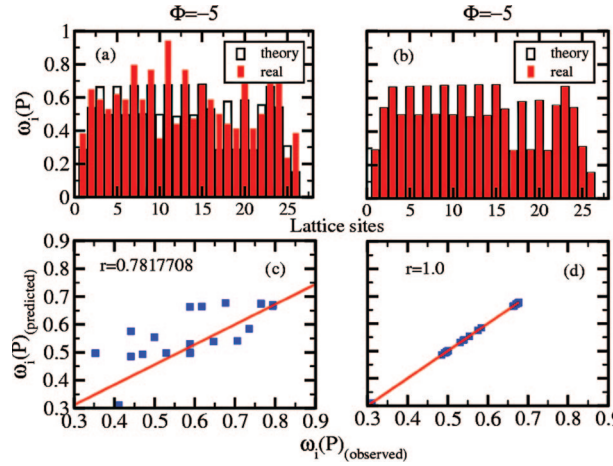

Figure 7. (a,b) Monomer polar probability profile for all lattice sites. (a) Comparison of results between theory (black bars) and that of lysozyme fold (red solid bars). (b) Comparison of results between theory (black bars) and all globular protein segments (red solid bars). (c,d) Correlation between results corresponding to a and b, respectively.

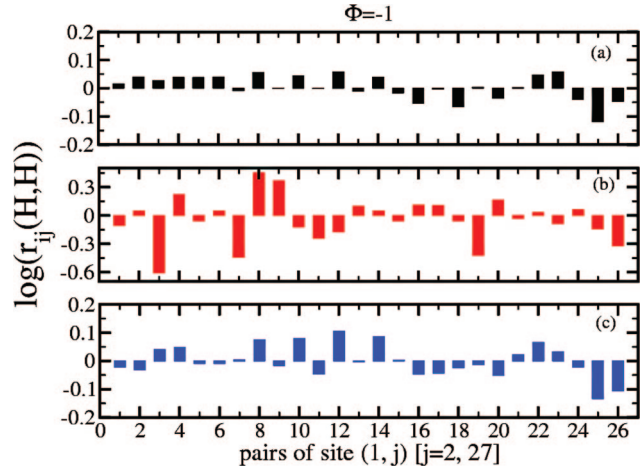

Figure 8. Semilog plot of tolerance r_{ij} (H−H) vs pairwise sequence positions for all possible pairs of site 1. Comparisons of results among (a) theory, (b) lysozyme fold, and (c) globular proteins.

of selection of the structures are resolution <2.0 Å and R-factor <20%, and maximum sequence identity between any two sequences is <25%. Depending on the respective physicochemical properties of the amino acid residues, a binary patterning of these sequences yield {V, L, I, F, W, M, A, G, P} as hydrophobic (H) and {R, N, D, C, Q, E, H, K, S, T, Y} as polar (P) groups.[33] In accordance to this scheme, all sequences from both of the data sets are converted to binary patterned sequences. Sampling sequences from the different proteins of lysozyme fold and the database of globular proteins yields a subset of 546 and 26 798 unique binary sequences respectively. The choice of sequences from the two data sets provides a means to explore the range of local sequence variability in real proteins. Such empirical measures are useful by themselves even if their full implications are not fully realized in terms of the respective structural features. For comparing the results of the mean field theory with that of the real proteins, it must be noted that the theory scans the entire sequence space comprising of 2^{27} = 134 217 728 sequences while the real protein sequences in both cases constitute a very limited set.

For both sets, all binary sequences are mapped into the 3 × 3× 3 lattice to select a set of sequences which fold into the specified target structure as their unique native structure. This set of binary sequences are classified according to their respective ϕ values which are calculated using eq 3 and eq 4. For each subset of binary sequences, the site specific monomer probability and the pairwise monomer probability profile are calculated from the frequency of occurrence of a particular residue/residue-pair for a definite value of ϕ. The results obtained from the theory for a specific ϕ value are compared with that obtained from the data set of lysozyme fold and the data set of globular proteins. The correlation coefficient between the observed and the predicted probabilities are also calculated for each case.

Figure 4 and Figure 5 compare the pairwise monomer probabilities of $H-H$ and $H-P$ pairs obtained from the theory and from the selected data sets. In Figure 4a the theoretical $H-H$ pair probabilities are compared with that of lysozyme fold at $\phi = -5$. The correlation between the observed and predicted probabilities is shown in Figure 4c and the calculated correlation coefficient (r) is 0.70577. Figure 4b compares the $H-H$ monomer pair probabilities between the theory and the data set of all globular proteins

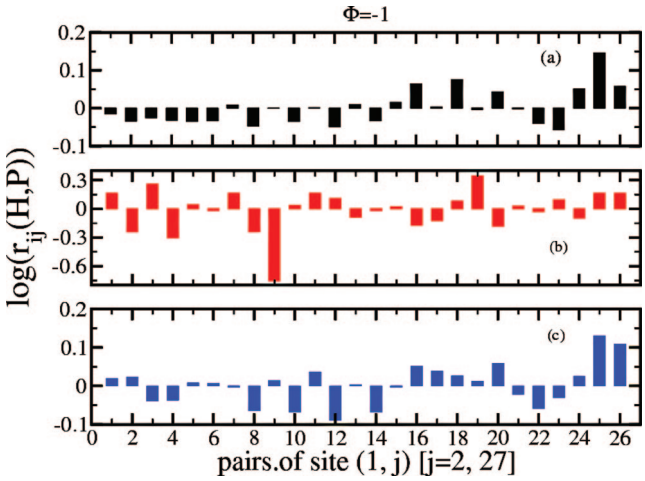

Figure 9. Semilog plot of tolerance r_{ij} (H−P) vs pairwise sequence positions for all possible pairs of site 1. Comparisons of results among (a) theory, (b) lysozyme fold, and (c) globular proteins.

at the same value of ϕ. The result shows a very high correlation between the observed and the predicted probabilities (Figure 4d) as the calculated correlation coefficient is $r = 0.98084$. These results suggest that for a broad range of sequence variability the theory provides a more realistic representation of protein sequences and predicts the pairwise monomer probabilities very accurately. A similar trend is reflected in the $H-P$ pairwise monomer probabilities.

Figure 6 and Figure 7 compare the site-specific monomer probabilities from the theory and the selected data sets of real proteins. Results depict an excellent correlation between the observed and the predicted probabilities.

Tolerance of a protein[34] to random mutations can be quantitatively characterized by the probability of random amino acid substitutions in different sequence positions. This is defined as the tolerance factor

$$r_{ij} = \frac{\omega_{ij}(\alpha_i, \alpha_j)}{\omega_i(\alpha_i)\omega_j(\alpha_j)} \tag{18}$$

if

$$r_{ij}(\alpha_i, \alpha_j) = \begin{cases} >1 & \alpha_i \text{ and } \alpha_j \text{ residues are favored at sites } i \text{ and } j \\ <1 & \alpha_i \text{ and } \alpha_j \text{ residues are unfavored at sites } i \text{ and } j \\ =1 & \text{no preference} \end{cases}$$

Favorable or unfavorable residue pairs can be identified by calculating the probablity ratio r_{ij} that quantifies the departure of the pairwise probability from the site identity probablities. Figure 8 and Figure 9 compare the semilog plots of hydrophobic−hydrophobic ($r_{ij}(H,H)$) and hydrophobic−polar ($r_{ij}(H,P)$) tolerance results for the 27-mer lattice protein and two data sets of real proteins by analyzing the monomer pair probability profile respectively. Tolerance values are calculated for all possible pairs,

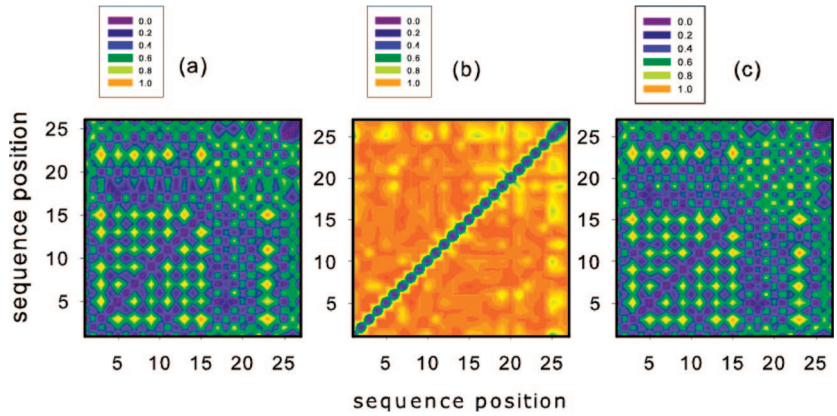

Figure 10. Hydrophobic−hydrophobic dependency plots are shown for (a) theoretical results, (b) results obtained for lysozyme folds, and for (c) set of all globular proteins.

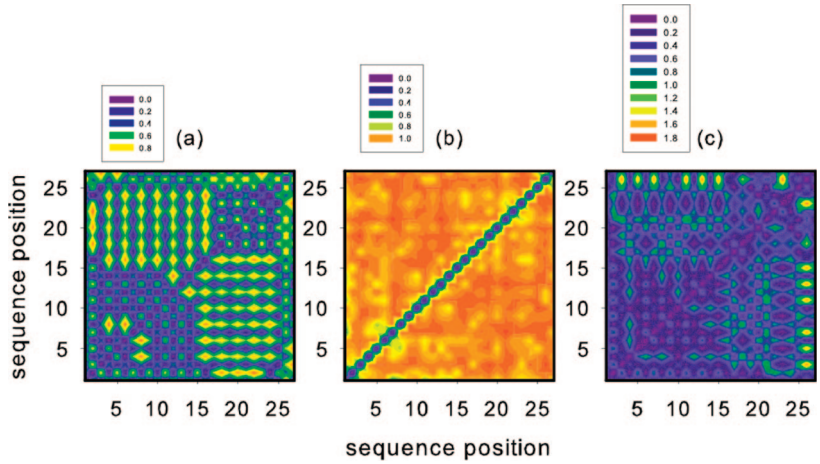

Figure 11. Hydrophobic−polar dependency plots are shown for (a) theoretical results, (b) results obtained for lysozyme folds, and for (c) set of all globular proteins.

but the results are displayed graphically for some representative pairs in which one of the inner hydrophobic sites is fixed. A value greater than zero along the *y* axis indicates that the respective residues at the given pair of sites are preferred whereas a negative value along the *y* axis suggests that the respective residues at the given pair of sites are unfavored. A zero value implies no preference. The hydrophobic core of the protein favorably tolerates replacements with hydrophobic residues but is less tolerant of any substitutions with polar residues as compared with the exterior surface. The quantitative agreement between the theory and the real protein results are highlighted in the corresponding correlation plots. For the lysozyme fold, the correlation coefficient (r) is significantly lower compared with the data set of globular proteins due to the lack of sequence variability among the limited number of available protein sequences. The theory, on the other hand, considers the whole sequence space consisting of $2^{27} = 134\,217\,728$ sequences. In both cases, real protein data display a reasonable agreement with the results of the lattice protein.

The standard deviation of the tolerance values over all residue combinations provides a quantitative measure of the substitution dependency.[34]

$$D_{ij} = \left[\sum_{\alpha_i=1}^{2} \sum_{\alpha_j=1}^{2} \omega_{ij}(\alpha_i, \alpha_j)(\log_2 r_{ij} - \mu_{ij})^2 \right]^{1/2} \qquad (19)$$

where,

$$\mu_{ij} = \sum_{\alpha_i=1}^{2} \sum_{\alpha_j=1}^{2} \omega_{ij}(\alpha_i, \alpha_j) \log_2 r_{ij} \qquad (20)$$

Nonzero values imply a correlation between the amino acid residues at sites i and j respectively whereas a zero value indicates that the substitution patterns of residue positions i and j are mutually independent. A zero value of D_{ij} is only possible when $r_{ij}(\alpha_i,\alpha_j) = 1$, which implies no preference in residue substitution in sites i and j. A zero value of D_{ij} necessarily indicates mutually independent substitution patterns between the sites i and j.

Figure 10 and Figure 11 illustrate the $H-H$ and $H-P$ dependency plots for the theory and the sequences of two sets of real proteins, respectively. For both plots, violet to blue color implies less correlation, orange to red denotes strongest correlation, and yellow to green indicates moderate correlation. For both Figure 10 and Figure 11, the theoretical results closely resemble the results obtained from the data set of all globular proteins. This is due to the fact that this data set encompasses diverse sequences and explores a broader range of sequence variability. Both plots depict a pronounced substitution dependency for all residues. The plots obtained from the lysozyme fold display a limited match due to the restricted number of similar sequences. However, a precise agreement with the substitution patterns of a real protein are not expected to be captured with such simplified coarse-grained potentials. Rather the complete characterization of all possible pairwise substitution patterns for this energy function are investigated.

IV. Conclusions

To summarize, a second-order mean-field theory based formalism is developed to completely characterize the sequence landscape for combinatorial designing of protein libraries that provide a complete sampling of all possible sequences. This statistical theory of sequences addresses both the number and the composition of sequences which are compatible with a particular folded state. The theory also computes the entire set of site specific pairwise monomer probabilities and identifies the residue−residue substitution patterns consistent with a generalized foldability criterion by incorporating the correlations between residue mutations. For large molecules like proteins, this method is especially useful as exact enumeration of all possible sequences is feasible only with a small number of amino acids or with restricted sequence variability. The theory evaluates the sequence−structure compatibility in terms of two-body inter-residue interactions in the form of a contact potential.

The theory is applied to two different data sets of protein sequences: (i) the lysozyme protein fold and (ii) a database of nonhomologous globular proteins. The theory scans the entire sequence space and not just the limited fraction accessible for a real protein. The calculated monomer pair probability profile displays a good match with the corresponding theoretical values obtained from a simple lattice model with a simple coarse-grained energy function. A precise match is however not expected as the theory scans the entire sequence space and not just the limited fraction of sequences accessible to a specific protein family/fold. The highly simplified energy function used in this model also contributes to the quantitative discrepancy. Rather, this energy function is used to assess the feasibility of designing sequences and examining correlations among residue−residue mutations.

The design of particular sequences of a protein family/fold using this potential is difficult since the search for particular sequences is very sensitive to the choice of the energy function. More sophisticated and detailed potentials are undoutedly necessary for the design of specific sequences, since specific bonding between the residues or the complementary packing of side chains may be the likely prerequisites for designing structures with the conformational specificity of natural proteins. However, this theory in its present form may be used to evaluate design strategies for target protein structures to engineer existing proteins and crafting new ones de novo by incorporating the effect of correlated mutations.

Acknowledgment. We gratefully acknowledge the financial assistance from DST (Project No. SR/S1/PC-07/06), India, and Delhi University research grant.

References and Notes

(1) Reidhaar-Olson, J. F.; Sauer, R. T. *Science* **1988**, *241*, 53.
(2) Jones, D. H.; Sakamoto, K.; Vorce, R. L.; Howard, B. H. *Nature (London)* **1990**, *344*, 793.
(3) Hecht, M. H.; Richardson, J. S.; Richardson, D. C.; Ogden, R. C. *Science* **1990**, *249*, 884.
(4) Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. *Science* **1993**, *262*, 1680.
(5) DeGrado, W. F.; Wasserman, Z. R.; Lear, J. D. *Science* **1989**, *243*, 622.
(6) Dill, K. A. *Biochemistry* **1990**, *29*, 7133.
(7) Onuchic, J. N.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.
(8) Gordon, D. B.; Marshall, S. A.; Mayo, S. L. *Curr. Opin. Struct. Biol.* **1999**, *9*, 509.
(9) Miyazawa, S; Jernigan, R. L. *Macromolecules* **1985**, *218*, 534.
(10) Russ, W. P.; Ranganathan, R. *Curr. Opin. Struct. Biol.* **2002**, *12*, 447.
(11) Vajda, S.; Sippl, M.; Novotny, J. *Curr. Opin. Struct. Biol.* **1997**, *7*, 222.
(12) Lu, H.; Skolnick, J. *Proteins* **2001**, *44*, 223.
(13) Archontis, G.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 11246.
(14) Zou, J.; Saven, J. G. *J. Mol. Biol.* **2000**, *296*, 281.
(15) Biswas, P.; Zou, J.; Saven, J. G. *J. Chem. Phys.* **2005**, *123*, 154908.
(16) Saven, J. G. *Chem. Rev. (Washington, D.C.)* **2001**, *101*, 3113.
(17) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986.

(18) Vendruscolo, M.; Domany, E. *J. Chem. Phys.* **1998**, *109*, 11101.
(19) Bastolla, U.; Porto, M; Ortiz, A. R. *Proteins.* **2008**, *71*, 278.
(20) Bastolla, U.; Porto, M; Roman, H. E.; Vendruscolo, M. BMC. *Evol. Biol.* **2006**, *6*, 1.
(21) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3908.
(22) Ejtehadi, M. R.; Hamedani, N.; Shahrezaei, V. *Phys. Rev. Lett.* **1999**, *82*, 4723.
(23) Deutsch, J. M.; Kurosky, T. *Phys. Rev. Lett.* **1996**, *76*, 323.
(24) Saven, J. G. *J. Chem. Phys.* **2003**, *118*, 6133.
(25) McQuarrie, D. A. *Statistical Mechanics. 2nd ed.*; Harper and Row: New York, 1976.
(26) Morita, T.; Tanaka, T. *Phys. Rev.* **1966**, *145*, 288.
(27) Sciretti, D.; Bruscolini, P; Pelizzola, A; Pretti, M; Jaramillo, A. *Proteins: Struct., Funct., Bioinf.* **2008**, *74*, 176.
(28) Bethe, H. A. *Proc. R. Soc. London, Ser. A* **1935**, *150*, 552.
(29) Shakhnovich, E. I.; Gutin, A. M. *J. Chem. Phys.* **1990**, *93*, 5967.
(30) Li, H.; Helling, R.; Tang, C.; Wingreen, N. *Science* **1996**, *273*, 666.
(31) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59.
(32) Zhou, H; Zhou, Y. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 483.
(33) Madel-Gutfreund, Y.; Gregoret, L. M. *J. Mol. Biol.* **2002**, *323*, 453.
(34) Moore, G. L.; Maranas, C. D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5091.

JP810515S