J. Chem. Sci., Vol. 121, No. 5, September 2009, pp. 887–896.

Statistical theory of neutral protein evolution by random site mutations†

ARNAB BHATTACHERJEE and PARBATI BISWAS*
Department of Chemistry, University of Delhi, Delhi 110 007
e-mail: pbiswas@chemistry.du.ac.in

Abstract. Understanding the features of the protein conformational space represents a key component to characterize protein structural evolution at the molecular level. This problem is approached in a two-fold manner; simple lattice models are used to represent protein structures with the ability of a protein sequence to fold into the lowest energy native conformation, quantified as the foldability, which measures the fitness of the sequence. Alternatively, a self-consistent mean-field based theory is developed to evaluate the protein neutrality through random single-point and multiple-point mutations by calculating the pair-wise probability profile of the amino acid residues in a library of sequences, consistent with a particular foldability criterion. The theory predicts the change in sequence plasticity with the foldability criterion and also correlates the effect of hydrophobic residues on the variation of the free energy surface of the protein as a function of the number of cumulative mutations. The results obtained from the theory are compared with the exact enumeration results of $3 \times 3 \times 3$ lattice protein and also with some small real proteins chosen from the protein databank. An excellent match of the results obtained from theory and exact enumeration with those of real proteins validates the range of applicability of the theory. The theory may provide a new perspective in *de novo* protein design, *in-vivo*/*in-vitro* protein evolution and site-directed mutagenesis experiments.

Keywords. Neutral evolution; protein design; mutations; foldability criteria.

1. Introduction

Understanding the relationship between protein sequences and structures can be partially achieved by predicting the folded structure from the amino acid sequence, and by predicting the compatible sequences of a known structure from physical principles. Since proteins are the building blocks of life and a direct by-product of evolution, it is important to comprehend how evolution shapes the global relationship between protein sequences and structures by directly simulating the process of evolution. Molecular evolution can be studied not only in the context of population genetics but also by considering the thermodynamic and kinetic stability of biomolecules involved in evolution.[1] This structural approach of molecular evolution has been applied to study neutral networks[2–5] of RNA,[6] to probe the sequence-secondary structure relationship.[7,8]

Motivated by various theories of polymer physics simple tractable models, such as lattice[9,10] and spin-glass[11] models, combined with simulation of evolution are often used to investigate the protein sequence-structure relationship. Computational studies can also guide experiments on *in vitro* evolution.[12] For a broader perspective, of molecular evolution, an analysis of the vast evolutionary landscape is required which represents the fitness of a protein as a function of the protein sequence parameters such as structural integrity or characteristics of the folding process. Evidences show that most accepted mutations have little effect on the fitness of the organism.[13–17] This constitutes the basis of the neutral theory of evolution[18,19] which proposes that many mutations are neutral, conserving the native fold and the biochemical function of the protein. This random mutation imparts extra thermodynamic stability[20] to the protein's native conformation and makes it more robust by increasing the fraction of mutants which possesses the minimal stability required to fold. This also implies some tolerance in the sequence optimization of proteins in keeping with Go's principle of minimal frustration[21] while allowing for sufficient native state stability and efficient folding dynamics.

†Dedicated to the memory of the late Professor S K Rangarajan
*For correspondence

Constructing models of appropriate fitness of proteins are complicated by the number of parameters involving stability, function and its survivability in different physiological conditions. Minimalist protein models with reduced amino acid alphabets are extremely handy in this regard and they are frequently used for studying various aspects of protein evolution folding and design. Foldability is one of the factors which determines the fitness function and plays a significant role in the selection process. Understanding the evolution of proteins is based on some optimized foldability criteria and adjusting this parameter we can unravel how the evolutionary process is influenced by the selective pressure. A quantitative measure of foldability characterizes all intra-molecular interactions in the protein which may not be individually optimal.

In an alternative approach, several studies in protein folding and evolution assume biological proteins to exist in three-dimensional lattices with a pre-determined energy function.[22,23] Encoding evolutionary selection in the input energy function it is possible to model the folding behaviour of these optimized proteins. Confining the proteins to a lattice makes their conformational space amenable to exact enumeration and motivates various theoretical models to calculate simple measures of foldability and how foldability affects the course of evolution. However, the choice of energy function is crucial in the sense that it should be realistic but sufficiently coarse-grained to capture the essential qualitative features.

In this work, we have used a simple three-dimensional lattice model of proteins[9,24–29] with binary patterning of amino acid residues to investigate the detailed effects of mutation on the evolution of protein stability and neutrality. The properties of these proteins are optimized for maximum fitness where the fitness represents the foldability of the protein. The model computes the probability $P_f(m)$ with which a protein retains its native fold after m substitutions as a function of the free energy change caused by each substitution. A tacit assumption which is in-built in the model is the concept that the free energy change due to the respective amino acid substitutions are independent of each other and does not influence the correct folding pattern of the protein as long as its free energy of folding remains below some critical threshold value. It also depicts the variation in the free energy surface with the number of mutations and the role of hydrophobicity in the mutation process. Moreover, the theory depicts the neutral–network pattern and points out the effect of cumulative mutations on protein sequences. The results obtained from this theory are compared with those of exact enumeration and also to a set of seven small real proteins, each of 27 residues, chosen from the PDB database. The predicted results of the self-consistent theory and exact enumeration show excellent agreement between each other and with that of the real proteins of widely different functional classes.

2. Theory

We present here a second-order mean field based formalism[30,31] to evaluate the pair-wise sequence probability profile for a given foldability criterion. A foldability criterion provides a suitable measure for the sequence–structure compatibility in form of a pre-determined energy function. This is the main input for the theory along with a given set of constraints on the sequences which can be modulated to specify local/global features of the protein structure. Choice of optimized energy functions requires accurate quantification of the various interactions stabilizing the native state. Statistical pair potentials prove to be a useful option in this regard as they measure both the inter-residue contact propensities[32–34] and excluded volume interactions. The energy of a sequence in a particular target state conformation E_f may be expressed as a function of pair-wise probabilities of the amino acid residues at each pair of sequence positions. Assuming small fluctuations in E_f about its mean value due to the variation of sequences, the target state energy can be expressed as

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

where the two-body interaction parameter $\gamma_{ij}^{(2)}(\alpha_i, \alpha_j)$ is a measure of the inter-residue contact propensity when the monomer types at sites i and j are α_i and α_j respectively. The structural information is contained in the parameter $\sigma_{ij}^{(2)}$ which is given by

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

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.

The sequence averaged energy of the unfolded ensemble of conformations can be written similarly as

$$\overline{\langle E_u \rangle} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{\alpha_i=1}^{m} \sum_{\alpha_j=1}^{m} \langle \sigma_{ij}^{(2)} \rangle \gamma_{ij}^{(2)} w_{ij}(\alpha_i, \alpha_j). \tag{3}$$

The stability gap Δ is a measure of the energy difference between the target structure and the ensemble-averaged energy of the unfolded conformations and is therefore expressed as

$$\begin{aligned} \Delta &\equiv E_f - \langle E_u \rangle \approx \overline{E_f} - \overline{\langle E_u \rangle} \\ &= \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sum_{\alpha_i=1}^{m} \sum_{\alpha_j=1}^{m} (\sigma_{ij} - \langle \sigma_{ij} \rangle) \gamma_{ij}(\alpha_i, \alpha_j) w_{ij}(\alpha_i, \alpha_j). \end{aligned} \tag{4}$$

Although Δ is a potential foldability criterion[35] which explicitly includes information about the unfolded ensemble of states, it does not take into account the low-energy unfolded structures which may compete with the target structure. It ignores the fluctuations in the energy of the unfolded ensemble of states. Therefore it would be rather accurate to include more information about the distribution of unfolded state energies for each sequence. Γ^2 is the simplest quantitative measure of the width of distribution of the energy of the unfolded states which emerges naturally as a second order term from the truncated cumulant expansion.[36] The variance of the unfolded state ensemble provides the correlation between the amino acid residues occupying different sites of the lattice protein and can be written as

$$\begin{aligned} \Gamma^2 \approx \overline{\Gamma^2} = \sum_{i,j,k,l} \sum_{\alpha_i,\alpha_j,\alpha_k,\alpha_l} & \gamma_{ij}(\alpha_i, \alpha_j) \gamma_{kl}(\alpha_k, \alpha_l) \\ & (\langle \sigma_{ij} \sigma_{kl} \rangle - \langle \sigma_{ij} \rangle \langle \sigma_{kl} \rangle) \\ & w_{ijkl}(\alpha_i, \alpha_j, \alpha_k, \alpha_l) \end{aligned} \tag{5}$$

A commonly used approximation to simplify the four-body probability $w_{ijkl}(\alpha_i, \alpha_j, \alpha_k, \alpha_l)$ in terms of the respective two-body probabilities is given by

$$w_{ijkl}(\alpha_i, \alpha_j, \alpha_k, \alpha_l) = w_{ij}(\alpha_i, \alpha_j) w_{kl}(\alpha_k, \alpha_l). \tag{6}$$

A generalized foldability criterion for measuring the sequence-structure compatibility can thus be obtained by a linear combination[31] of Δ and Γ^2,

$$\phi = \beta\Delta + \frac{1}{2}\beta^2\Gamma^2, \tag{7}$$

where ϕ is a dimensionless quantity and each of Δ and Γ^2 are also dimensionless, scaled by appropriate units of thermal energy β.

The pair wise monomer probabilities $w_{ij}(\alpha_i, \alpha_j)$ are directly solved by maximizing the sequence entropy S subject to appropriate constraints on sequence identity and energies. The entropy of a set of sequences Ω_s is defined as

$$S = k_B \ln \Omega_s \tag{8}$$

where k_B is the Boltzmann constant.

The Cluster Variation Method[37] (CVM) is used to derive the sequence entropy when higher order correlations among residue sites are considered.

$$\begin{aligned} S^{(n)}(i_1, ..., i_n) = - \sum_{\alpha_{i_1}, ..., \alpha_{i_n}} & w^{(n)}_{i_1, ... i_n}(\alpha_{i_1}, .., \alpha_{i_n}) \\ & \ln w^{(n)}_{i_1, ... i_n}(\alpha_{i_1}, .., \alpha_{i_n}). \end{aligned} \tag{9}$$

For a set of sequences satisfying a predetermined set of constraints the Bethe approximation[38] 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 pair wise monomer probabilities scaled appropriately so as to avoid double counting.

$$W_N(\alpha_1, ..., \alpha_N) = \prod_{ij} \frac{w_{ij}(\alpha_i, \alpha_j)}{w_i(\alpha_i) w_j(\alpha_j)}, \tag{10}$$

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

$$S \approx \sum_i S^{(1)}(i) + 2 \sum_{i<j} (S^{(2)}(i, j) - S^{(1)}(i) - S^{(1)}(j)). \tag{11}$$

The pair-wise monomer probabilities are equilibrated in the ensemble by maximizing a variational functional of the set of probabilities $w_i(\alpha_i)$ and $w_{ij}(\alpha_i, \alpha_j)$ subject to the following constraints,

$$\sum_{i<j} w_{ij}(\alpha_i,\alpha_j) = 1 \ \forall\ i, j > i, \tag{12}$$

$$w_i(\alpha_i) = \sum_{\alpha_j} w_{ij}(\alpha_i,\alpha_j) \quad \forall\ i, j \neq i, \tag{13}$$

$$w_j(\alpha_j) = \sum_{\alpha_i} w_{ij}(\alpha_i,\alpha_j) \quad \forall\ j, j \neq i, \tag{14}$$

and (4) and (5).

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

$$w_{ij}(\alpha_i,\alpha_j) = \exp(\beta_{ij}(\alpha_i) + \mu_{ij}(\alpha_j) - \beta_\phi \phi_{ij}(\alpha_i,\alpha_j) - 1)$$

$$w_i(\alpha_i) = \frac{1}{q_i}\left[\begin{array}{l}\exp(\beta_\phi \phi_i(\alpha_i) + \sum_j \xi_{ij}\beta_{ij}(\alpha_i) \\ + \sum_j \xi_{ji}\mu_{ij}(\alpha_j))\dfrac{1}{(N-2)}\end{array}\right]$$

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

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

$$\phi = \Delta + \frac{1}{2}\Gamma^2, \tag{15}$$

where the monomer partition function q_i is given by,

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

Also,

$$\phi_i = \frac{\partial \phi}{\partial w_i(\alpha_i)} \text{ and } \phi_{ij} = \frac{\partial \phi}{\partial w_{ij}(\alpha_i,\alpha_j)}.$$

These set of equations are then solved numerically (http://www.netlib.org) to yield the probabilities $w_i(\alpha_i)$, $w_{ij}(\alpha_i,\alpha_j)$ and the respective Lagrange multipliers for a given value of ϕ.

3. Model of neutral mutations

3.1 *Lattice protein simulation*

The protein is modelled as a cubic lattice polymer consisting of 27 residues on a maximally compact three-dimensional lattice. Protein conformations are represented by self-avoiding walks that occupy all lattice vertices, and a total of 103346 compact conformations are possible[39] which are not related by symmetry. The non-compact states will have less relevance on the qualitative nature of the foldability landscape. These models crudely represent larger proteins where each lattice point corresponds to the position of a structural unit stabilized by local cooperative interaction.[27] Lattice protein models have been proved to be extremely useful[40,41] in the context of protein folding and evolution. With a binary patterning of monomers, the number of possible sequences for a 27-mer is large, $2^{27} = 134217728$, but computationally enumerable. Pair-wise contact energy function is used here to characterize sequence-structure compatibility[42]

$$\gamma^{(2)}(H,H) = -3\varepsilon,\ \gamma^{(2)}(H,P) = \gamma^{(2)}(P,H) = -\varepsilon$$
$$\text{and } \gamma^{(2)}(P,P) = 0. \tag{16}$$

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.[22]

All possible sequences are enumerated exhaustively and are classified according to the different ϕ values which lie within a small interval $\partial\phi\partial\phi$ centered about a definite value of ϕ. The site-specific monomer probabilities and the pair-wise monomer probabilities are calculated[43] from the corresponding frequencies of occurrence of the polar and hydrophobic residues for a specific value of ϕ. Random single-point mutations are performed sequentially at different sites till all the sites of the lattice protein are mutated once. In each generation, the parent sequence differs from the mutated sequence by a single residue. The mutation in one generation is not retained in the successive generations.

Mutations in all possible sites of the protein result in a set of 27 singly mutated sequences. The theory yields different sequences for different values of ϕ, differing in terms of the site specific monomer probability and pair-wise monomer probability. Three

distinct sequences for three different ϕ values (–8·5, 6·5, 7·5) are chosen for performing the site-specific mutations by the method described above. These sequences are unique and closely resemble some of the natural protein sequences.

3.2 *Real proteins*

Seven small real proteins, (1SMZ, 1TTL, 2CCO, 2HFR, 2HGO, 2HTG and 1AV3), each consisting of 27 residues, are selected from the protein databank (PDB) (http://www.rcsb.org). Six of these proteins (1TTL, 2CCO, 2HFR, 2HGO, 2HTG and 1AV3) are from the conotoxin family and 1SMZ belongs to the family of peptide hormone. The energies of these proteins are calculated from their sequence composition without knowing its native conformation. The energy per amino acid is calculated by,[44]

$$\frac{E_{\text{fold}}}{L} = \sum_{ij}^{20} n_i P_{ij} n_j, \tag{17}$$

where L is the length of the protein sequence and P_{ij} is the energy predictor matrix,[44] which contains information about how the energy of ith amino acid depends on the jth element of the amino acid composition vector.

The amino acid sequences of the real proteins are converted into the binary code hydrophobic (H) and polar (P) based on their respective physico-chemical properties. The hydrophobic amino acids are {V, L, I, F, W, M, Y, C, P} and the polar ones are {A, R, N, D, Q, E, G, H, K, S, T}.[45–47] This classification of the amino acid residues is also experimentally confirmed by Fauchere and Pliska.[48] All real proteins are randomly mutated at single points according to the above mentioned procedure. These mutations yield a set of 27 mutated sequences for each real protein sequence. All single-point mutations are non-Poissonian and occur only once. The thermodynamic changes due to random site mutations in the prototype sequences are reflected by the change in their free energies ($\Delta G_{\text{folding}}$) which is given by,[49]

$$\Delta G_{\text{folding}} = E_f + k_B T \ln\left(Z - \exp - \left(\frac{E_f}{k_B T}\right)\right), \tag{18}$$

where Z is the partition function is ($k_B T$ is chosen to be 1). Mutant sequences are considered viable if the native conformation of the protein remains unchanged and sufficiently stable, which implies that the free energy of folding $\Delta G_{\text{folding}}$ should be less than some fixed parameterΔG_{crit}. The robustness of these sequences is monitored by measuring the change in free energy $\Delta\Delta G_{\text{stability}}$ as a function of the stability of the protein prior to mutation.

$$\Delta\Delta G_{\text{stability}} = \Delta G_{\text{folding}}^{\text{mut}} - \Delta G_{\text{folding}}^{\text{wild-type}}. \tag{19}$$

4. Results and discussions

Mutant sequences for which the $\Delta\Delta G_{\text{stability}}$ values lie below certain critical value ΔG_{crit} are termed as viable sequences and can fold back to the native conformation. The corresponding mutations are known as 'good' mutations.[50] For each of the 27 mutated sequences corresponding to a given parent protein sequence, the $\Delta\Delta G_{\text{stability}}$ is calculated and compared with ΔG_{crit}. The protein sequences for which $\Delta\Delta G_{\text{stability}} < \Delta G_{\text{crit}}$ are considered stably folded. After each generation of mutations the fraction of such prototype sequences ($P_f(m)$) are calculated. For three arbitrary cut-off free energy values, $\Delta G_{\text{crit}} = -0{\cdot}1$, $-0{\cdot}5$, $-1{\cdot}0$, $P_f(m)$'s are calculated for the real protein sequences as well as the sequences generated from exhaustive enumeration and self-consistent field theory. Results are depicted for the variation of $P_f(m)$ with the generation of each random site mutations in figure 1.

Theoretically calculated values of $P_f(m)$'s show an excellent match with that of exact enumeration and real proteins in some cases but for others (e.g. in figure 1(d), (f) and (i)) the matches are not perfect. A quantitative assessment of the performance of the theory is done by calculating the root mean square (RMS) deviation of the logarithm of $P_f(m)$ between the real proteins and the SCF theory results.

The logarithmic RMS deviation is given by,[51]

$$\rho = \left(\sum_{i=1}^{n=27} [\ln P_f^{\text{real}}(m) - \ln P_f^{\text{theory}}(m)]^2\right)^{1/2}. \tag{20}$$

For the data presented here, figure 2 clearly points out cases where the theory matches extremely well, the measured logarithmic RMS deviation is less than 1, mostly it lies within the range 0 to 0·1.

Very few cases are observed where ρ values exceed 1. The sequence plasticity of proteins makes them more robust to random mutations and imparts extra stability to their respective native conforma-

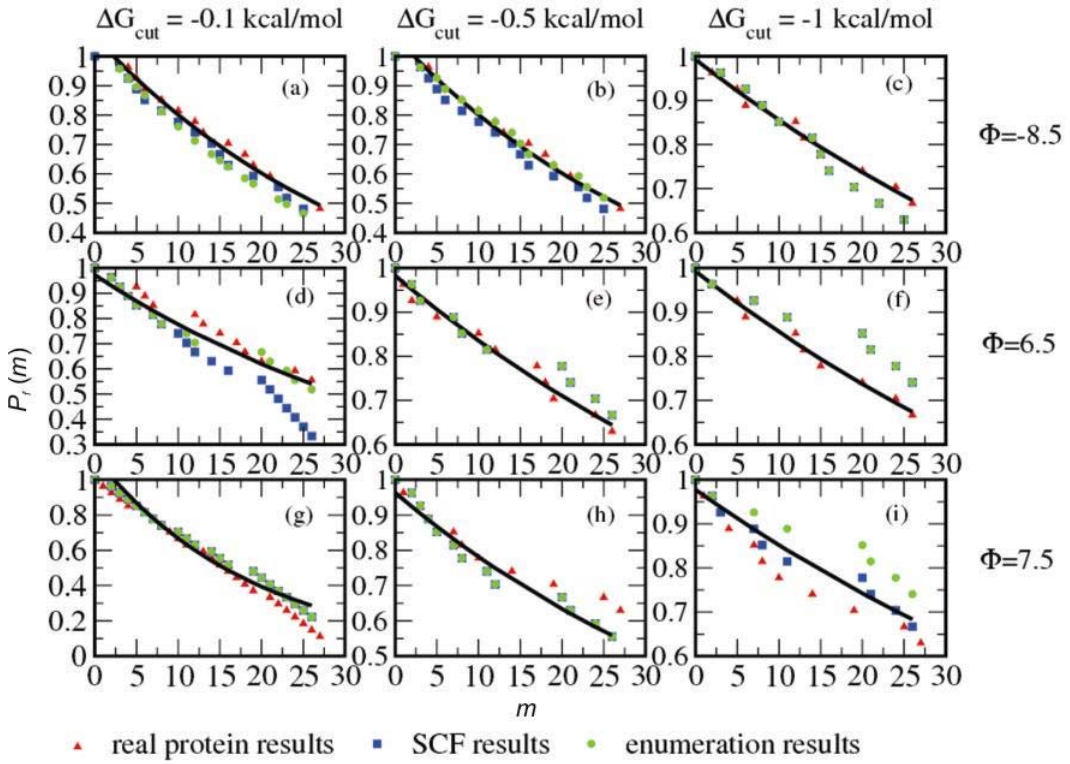

Figure 1. (**a**)–(**c**). Φ chosen for SCF and enumeration is –8·5. Three plots are corresponding to different ΔG_{crit} values taken in decreasing order as marked above. For ΔG_{crit} = –0·1, –0·5 the real protein considered is 2HTG, for ΔG_{crit} = –1·0 the considered real protein is 2CCO. (**d**)–(**f**) Φ chosen for SCF and enumeration is 6·5. For ΔG_{crit} = –0·1, –1·0 the real protein considered is 1TTL and 1SMZ is for ΔG_{crit} = –0·5. (**g**)–(**i**) Φ chosen for SCF and enumeration is 7·5. Protein 2HFR is chosen for ΔG_{crit} = –0·5, –1·0 whereas protein 2HGO is for ΔG_{crit} = –0·1.

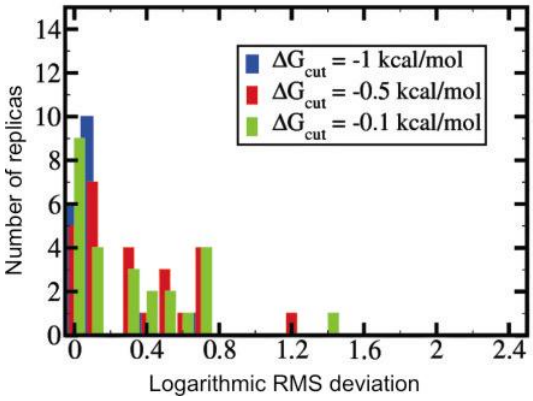

Figure 2. Histogram of logarithmic RMS deviation for randomly mutated 27 sequences at ϕ = –0·5 for each cut-off level.

tions. However, multiple point mutations especially in the active site of the protein can cause a change in its biochemical function. It is observed that the replacement of amino acid residues in the internal region can alter the entire properties of tropomysin molecule.[52] These mutations considerably affect the sequence composition and the free energy surface of the protein molecule. Thus, it is important to know the variation of composition in sequence with increasing generations of cumulative mutations for designing *de novo* proteins.

The contour diagram of two real proteins' free energy surfaces are studied as a function of generations of mutations and fraction of hydrophobic residues. The two real proteins considered are 1 AV3, a potassium channel blocker and 2 HTG, a membrane protein. The free energy surface for the exactly enumerated sequences show an excellent match with those of the self-consistent theory and real proteins. Comparison of results, shown in figures 3–8, suggest that protein sequences are stable up to a certain degree of random and cumulative site mutations after which they collapse. The plasticity of protein sequences is solely dependent on the stability of native fold. More stable the native fold, more will be its tolerance towards cumulative mutations. Another

important aspect is the effect of hydrophobic residues on the inherent sequence plasticity of proteins. It is observed that for each protein there exists an upper and lower cut-off bound for the number of hydrophobic residues within which the protein sequence is robust to mutations. This cut-off range may vary[53] with the length of a specific protein and its native fold stability. The present study demonstrates that on an average it lies within a range of 40% to 50%. The result is bit surprising as

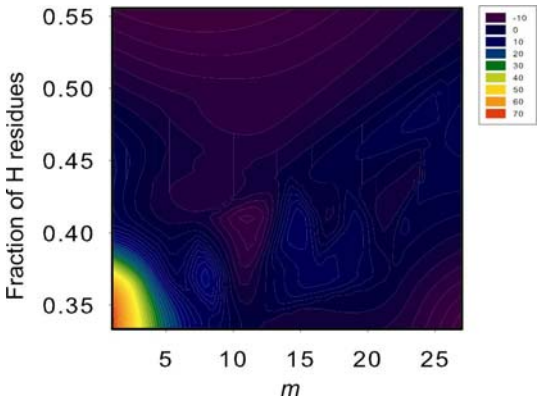

Figure 3. Free energy surface $\Delta G(n_H, m)$ as function of mutations and fraction of hydrophobic residues. The real protein considered here is 1AV3.

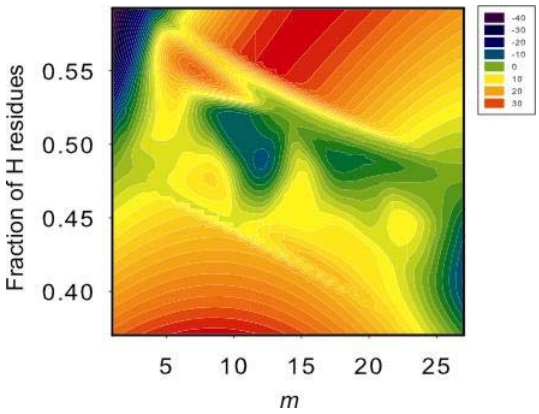

Figure 6. Free energy surface $\Delta G(n_H, m)$ as a function of number of mutations and fraction of hydrophobic residues for 2 HTG.

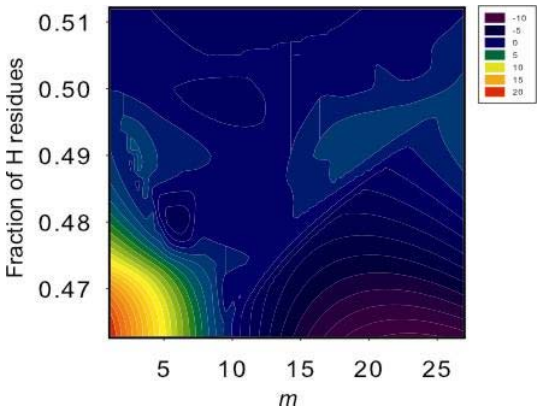

Figure 4. Enumeration result at $\phi = -4$.

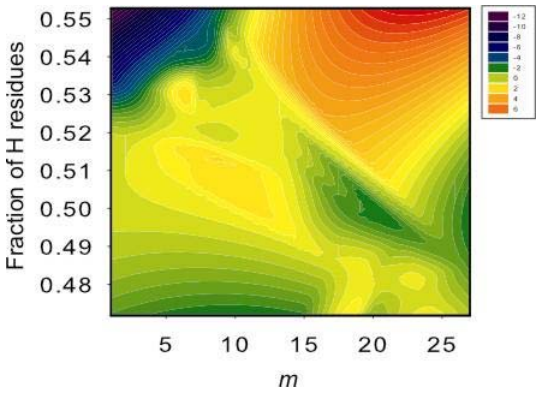

Figure 7. Enumeration result at $\phi = 5{\cdot}5$.

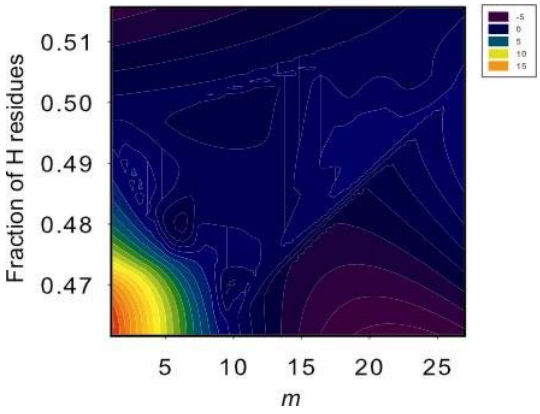

Figure 5. SCF results at $\phi = -4$.

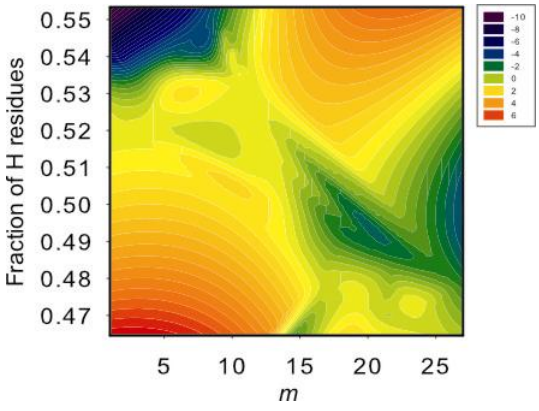

Figure 8. SCF result at $\phi = 5{\cdot}5$.

the ratio of hydrophobic residues to polar is less than half, though hydrophobic interaction is believed to be the major driving force to fold protein sequences. This is probably due to the increasing repulsions within the hydrophobic core with the increase in the number of hydrophobic residues in the protein.

A neutrally evolving protein can be looked upon as a node in the neutral network. Hence it is important to know the topology of neutral-network and to verify whether it can affect the plasticity of a protein sequence and the manner in which it does. For the membrane protein 2HTG, the sequence is mutated cumulatively up to four generations at first four sites chosen in all possible (i.e. 4! = 24) ways. The mutation is done by the following scheme: $H \rightarrow P$ or vice versa.[54]

The free energies ΔG of all the mutated as well as the wild type sequences are calculated. This protein is robust for the four cumulative mutations. Hence all the 96 prototype sequences are interconnected in neutral network. The nodes are said to be neutral-neighbours in the network on the basis of the most natural metric between different points of a sequence space, the Hamming distance.[55,56] Hamming distance represents the number of amino acid changes necessary to go from one amino acid sequence k to another sequence l which is exactly same as the number of mutated sites. Figure 9 depicts the neutral network observed for the real protein 2HTG and the sequence generated from the theory for $\phi = 5{\cdot}5$. Except the unique native fold, there exist degenerate states at each Hamming distance. More the number of degenerate states more would be the interconnections within the network and thus the protein is more amenable to evolve neutrally at a faster rate. The protein is expected to be more robust towards mutations and hence more 'designable'.

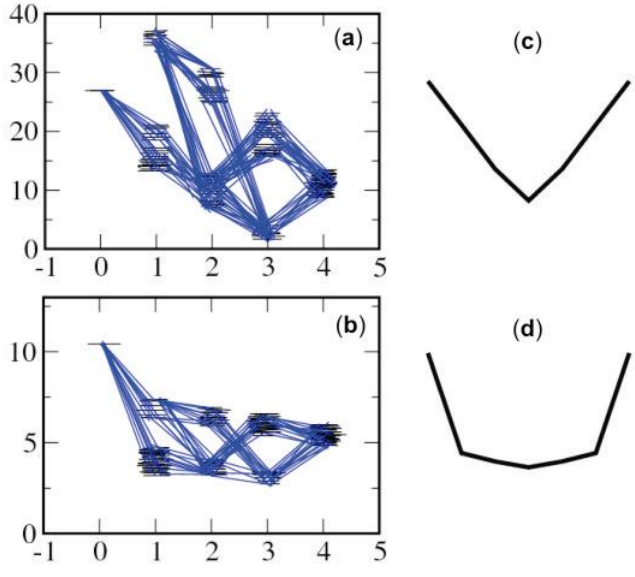

Figure 9. (**a**) and (**b**) represent the native stabilities of protein sequences of real protein 2HTG and theory respectively in neutral-net. Neutral mutations are indicated by lines connecting the horizontal levels. (**c**) and (**d**) represent the heuristic view of neutral-net topology for real protein and theory respectively.

The neutral-network pattern obtained from the self-consistent field theory bears a close resemblance to the real proteins but differ in the energy scale. The heuristic views of the funnel-like topology (figure 9c–d) of the neutral-network is presented by plotting the average stability of the sequences as a function of the Hamming distances. Theory presents a comparatively wider funnel compared to the shallow one obtained from the real protein. This difference in the depth of the funnel may be due to the simple coarse-grained potential used in the theory as compared to the complex interplay of diverse interactions in a real protein.

Successive mutations in the wild type sequences of some real proteins enhance its function and stability.[57] This observation is quite contradictory from the evolutionary perspective.[58] Though the present study suggest the possibility of existence of some sequences after successive mutations (at Hamming distance 3), which are even more stable and hence may be of more pronounced functionality compare to the native fold. In the self-consistent field theory, the choices of sequences are motivated partly by the foldability criterion (ϕ). Hence, the variation of the sequence plasticity with the change in ϕ values can provide a better understanding regarding protein's

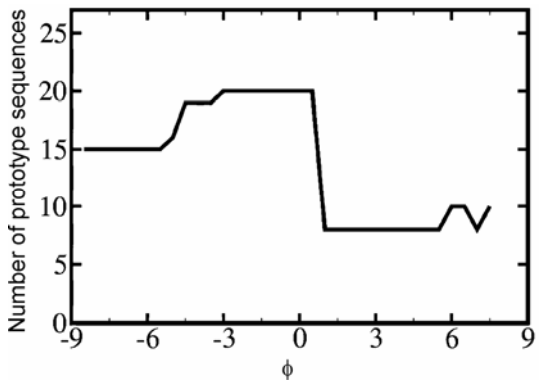

Figure 10. Variation of protein plasticity with the foldability criterion.

robustness. Figure 10 shows that the protein sequences having negative ϕ values are more robust to site mutations, especially the sequences lying within the foldability range $-3{\cdot}0$ to $0{\cdot}5$. The mutational robustness of a sequence is identified by the number of prototype sequences which decrease at both extremes of the ϕ values. This is because the sequences are either predominantly hydrophobic or hydrophilic at both extremes of ϕ values.

This is because the sequences are either predominantly hydrophobic or hydrophilic at both extreme of ϕ values.

5. Conclusions

To summarize, a self-consistent mean-field theory based formalism is developed to predict the protein neutrality and its robustness for single and multiple point random mutations. The theory is applied to a 27-mer lattice protein on a simple three-dimensional cubic lattice. All maximally compact conformations of the lattice protein are considered. Non-compact lattice proteins rarely fold into maximally compact conformations.[59,60] Using a simple coarse-grained scoring function and the theory can be used to identify native fold sequences for pre-assigned conformations consistent with a generalized foldability criterion ϕ. Sequences with different ϕ values closely resemble real proteins. The theory yields the fraction of mutant proteins which fold stably to the native conformation, $P_f(m)$ as a function of single-point mutations.

Results from the theory and exact enumeration are compared with small real proteins obtained from the PDB. Theoretically predicted protein neutrality matches extremely well with those of real proteins and with the results of exact enumeration. The theory also provides an insight about the sequence plasticity and correlates between the fraction of hydrophobic residues and mutational robustness. Moreover, the theory depicts the neutral-network pattern in terms of the Hamming distance and points out the effect of cumulative mutations on protein sequences. An important aspect of the theory is its applicability to real proteins with different functionalities. Hence, this theory provides a suitable framework for designing *de novo* protein sequences with improved functionalities by site-directed mutagenesis experiments and a rationale to the inherently different designability of different protein sequences based on the differences in their neutrality.

Acknowledgements

We gratefully acknowledge the financial assistance from Department of Science and Technology (DST) (project no. SR/S1/PC-07/06), India and Delhi University Research Grant.

References

1. Schuster P, Fontana W, Stadler P F and Hofacker I L 1994 *Proc. Roy. Soc. London* **B255** 279
2. Govindarajan S and Goldstein R A 1997 *Biopolymers* **42** 427
3. Bornberg-Bauer E and Chan H S 1999 *Proc. Natl. Acad. Sci. USA* **96** 10689
4. van Nimwegen E, Crutchfiels J P and Huynen M 1999 *Proc. Natl. Acad. Sci. USA* **96** 9716
5. Bastolla U, Porto M, Roman H E and Vendruscolo M 2002 *Phys. Rev. Lett.* **89** 208101
6. Thomas J, Martin O C and Wagner A 2008 *BMC Bioinformatics* **9** 464
7. Reidys C, Stadler P F and Schuster P 1997 *Bull. Math. Biol.* **59** 339
8. Schuster P 1995 *J. Biotechnol.* **41** 239
9. Sali A, Shaknovich E and Karplus M 1994 *J. Mol. Biol.* **235** 1614
10. Shakhnovich E I and Gutin A M 1990 *J. Chem. Phys.* **93** 5967
11. Bryngelson J D and Wolynes P G 1987 *Proc. Natl. Acad. Sci. USA* **84** 7524
12. Martinez M A, Pezo V, Marliere P and Wain-Hobson S 1996 *The EMBO J.* **15** 1203
13. Shortle D and Lin B 1985 *Genetics* **110** 539
14. Pakula A A, Young V B and Sauer R T 1986 *Proc. Natl. Acad. Sci. USA* **83** 8829
15. Guo H H, Choe J and Loeb L A 2004 *Proc. Natl. Acad. Sci. USA* **101** 9205
16. Bloom J D, Labthavikul S T, Otey C R and Arnold F H 2006 *Proc. Natl. Acad. Sci. USA* **103** 5869
17. Serrano L, Day A G and Fersht A R 1993 *J. Mol. Biol.* **233** 305
18. King J L and Jukes T H 1969 *Science* **164** 788
19. Kimura M 1983 *The neutral theory of molecular evolution* (Cambridge University Press)
20. Bloom J D, Silberg J J, Wilke C O, Drummond D A, Adami C and Arnold F H 2005 *Proc. Natl. Acad. Sci. USA* **606** 102
21. Go N 1983 *Annu. Rev. Biophys. Bioeng.* **12** 183
22. Li H, Helling R, Tang C and Wingreen N 1996 *Science* **273** 666
23. Russ W P and Ranganathan R 2002 *Curr. Opin. Struct. Biol.* **12** 447
24. Go N and Taketomi H 1978 *Proc. Natl. Acad. Sci. USA* **75** 559
25. Chan H S and Dill K A 1991 *Annu. Rev. Biophys. Biophys. Chem.* **20** 447
26. Shaknovich E I 1994 *Phys. Rev. Lett.* **72** 3907
27. Socci N D and Onuchic J N 1995 *J. Chem. Phys.* **103** 4732

28. Abkevich V I, Gutin A M and Shakhnovich E I 1995 *Protein. Sci.* **4** 1167
29. Hao M-H and Scheraga H A 1996 *Proc. Natl. Acad. Sci. USA* **93** 4984
30. Zou J and Saven J G 2000 *J. Mol. Biol.* **296** 281
31. Biswas P, Zou J and Saven J G 2005 *J. Chem. Phys.* **123** 154908
32. Miyazawa S and Jernigan R L 1985 *Macromolecules* **218** 534
33. Sippl M J 1990 *J. Mol. Biol.* **213** 859
34. Goldstein R, Luthey-Schulten Z Z and Wolynes P G 1992 *Proc. Natl. Acad. Sci. USA* **89** 9029
35. Deutsch J M and Kurosky T 1996 *Phys. Rev. Lett.* **76** 323
36. Saven J G 2003 *J. Chem. Phys.* **118** 6133
37. Morita T and Tanaka T 1966 *Phys. Rev.* **145** 288
38. Pathria R K 1972 *Statistical mechanics* (Pergamon Press)
39. Shaknovich E I and Gutin A 1990 *J. Chem. Phys.* **93** 5967
40. Go N 1975 *Int. J. Pept. Protein Res.* **7** 313
41. Dill K A, Bromberg S, Yue K S, Fiebig K M, Yee D P, Thomas P D and Chan H S 1995 *Protein Sci.* **4** 561
42. Saven J G 2001 *Chem. Rev.* (Washington DC) **101** 3113
43. Bhattacherjee A and Biswas P 2009 *J. Phys. Chem. B* (ASAP), DOI 10.1021/jp810515s
44. Dosztanyi Z, Csizmok V, Tompa P and Simon I 2005 *J. Mol. Biol.* **347** 827
45. Sharp K A, Nicholls A, Friedmann R and Honig B 1991 *Biochemistry* **30** 9686
46. Levitt M 1976 *J. Mol. Biol.* **104** 59
47. Zhou H and Zhou Y 2002 *Proteins: Struct. Funct. Genet.* **49** 483
48. Fauchere J-L and Pliska V 1983 *Eur. J. Med. Chem.* (*Chim. Ther.*) **18** 369
49. Taverna D M and Goldstein R A 2002 *J. Mol. Biol.* **315** 479
50. Oliveira L C, Silva R T H, Leite V B P and Chahine J 2006 *J. Chem. Phys.* **125** 084904
51. Wilke C O, Bloom J D, Drummond D A and Raval A 2005 *Biophys. J.* **89** 3714
52. Sano K-I, Maeda K, Taniguchi H and Maeda Y 2000 *Eur. J. Biochem.* **267** 4870
53. Miao J, Klein-Seetharaman J and Meirovitch H 2004 *J. Mol. Biol.* **344** 797
54. Chan K A and Dill H S 1994 *J. Chem. Phys.* **100** 9238
55. Hamming R W 1950 *Bell. Syst. Tech. J.* **29** 147
56. Hamming R W 1986 *Coding and information theory* (Englewood Cliffs: Prentice Hall) 2nd edn
57. Hecht M H, Hehir K M, Nelson H C M, Sturtevant J M and Sauer R T 1985 *J. Cell. Biochem.* **29** 217
58. Hecht M H, Sturtevant J M and Sauer R T 1986 *Proteins. Struct. Funct. Genet.* **1** 43
59. Irback A and Troein C 2002 *J. Biol. Phys.* **28** 1
60. Chan H S and Dill K A K 1996 *Proteins: Struct. Funct. Genet.* **24** 335