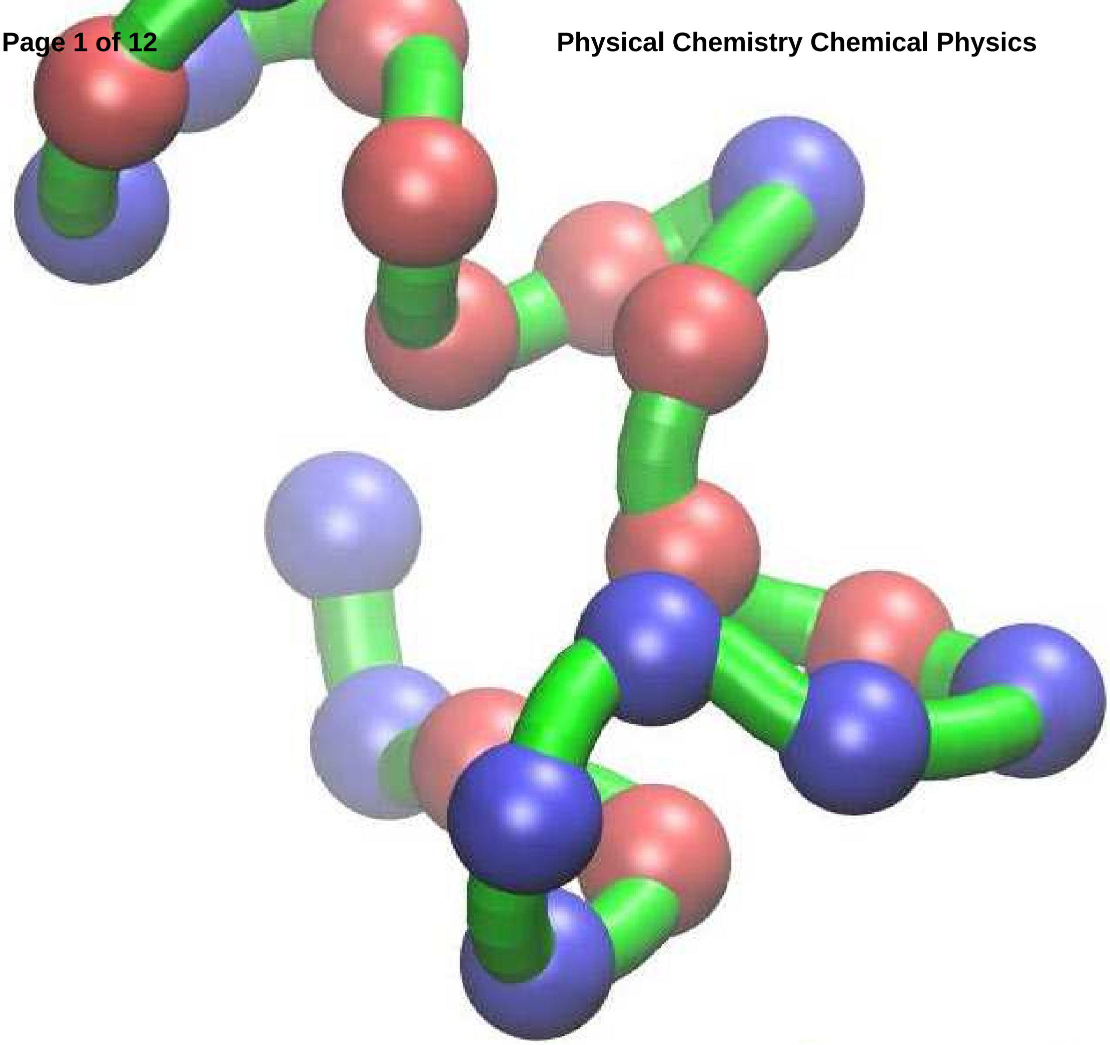
PCCP

Accepted Manuscript

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
PCCP
Physical Chemistry Chemical Physics
www.rsc.org/pccp
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp



Maintained critical

Intact critica

Increased matchi

clashing r

Role of Site-Directed Point Mutations in Protein Misfolding[†]

Anupaul Baruah, and Parbati Biswas[*]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

First published on the web Xth XXXXXXXXXX 200X

DOI: 10.1039/b000000x

A self-consistent mean-field based model is presented to explore the effect of site-directed point mutations in designing folded and/or misfolded sequences with a reduced hydrophobic-polar(HP) patterning of amino acids. This site-directed point mutation procedure is developed and applied on both real and lattice proteins to generate a diverse set of sequences. The respective roles of core and surface residues are analyzed with respect to the optimum hydrophobicity required for the structural stability of the protein. The core sites are found to have a critical number of hydrophobic residues, below which a protein may misfold, while the surface sites show a clear preference for the polar residues with an ability to tolerate some hydrophobic residues. Although core sites play an important role in the structural stability of proteins, some specific surface sites are also found to be equally important. A clash and match calculation procedure is proposed, which may be used to predict the number of residue pairs in a sequence with unfavorable and favorable interactions respectively due to site-directed point mutations. The number of clashing and matching residue pairs may indicate whether the mutated sequence would be folded or misfolded. The results are independent of the secondary structure topology of the protein. This model may provide new insights to the effect of point mutations on protein stability and may introduce a new method to predict the outcome of a mutation in terms of its probability to fold or misfold.

1 Introduction

Proteins may be distinguished from other biological and synthetic polymers by the presence of a rugged funnel-shaped conformational energy landscape with an overall energy gradient towards the native state located at the free energy minimum. Designing sequences with such landscapes involves stabilizing the ‘minimally frustrated’[1] native (target) structure against the ensemble of unfolded/misfolded conformations. Mutations of such optimized sequences almost always increases the frustration even though in most cases, it does not substantially change the native structure. Mutational robustness is the characteristic property of naturally evolved proteins, which increases the number of sequences within the constraints of a given structure facilitating viability and sequence diversity. Incorporating point mutations by site-directed mutagenesis in natural and designed protein sequences may be used to explore the stability, flexibility and functional features of proteins.[2–5] Numerous studies have investigated the effects of point mutations on protein stability and function. The alignment of the homologous amino acid sequences has lead to a ‘consensus’ approach, which assumes that the conserved amino acid residues play a dominant role in implementing protein stability.[6–11] However, sequences generated by ‘consensus’ approach are not necessarily the most stable ones.[6]

Random point mutation studies have revealed that some of these mutations dramatically affect the stability of a protein, which cannot be explained from the molecular principles of structural stability.[12,13] Rational design of point mutations are often complicated by the choice of the type and location of appropriate residues, which may cause any desired change in the protein function. Thus, site directed mutagenesis on the protein surface may have less impact in terms of its stability and function as compared to the core comprising of hydrophobic residues. It is normally accepted that core forming hydrophobic residues are sensitive to mutations, but several experimental studies have shown that optimization of surface residues also have important role in protein stability.[14–16] Optimization of surface electrostatics by mutations in the surface residues is used to design thermostable proteins.[15] Other studies have reported effects on local/global flexibility of proteins induced by point mutations.[4,5,17] In this context, a model that can (i) design and optimize sequences through site-directed point mutations and (ii) identify the role of core/surface residues on the protein stability may help in identifying the site specific mutation patterns of different residues for a given protein. Thus this model may provide the foundation for predicting the outcome of site-directed mutations in engineering the stability/foldability and evolvability of known proteins and designing *de novo* ones.

In this article, a self-consistent mean field based model is presented to investigate the effect of site-directed point muta-

[*] *Department of Chemistry, University of Delhi, Delhi-110007, India ; E-mail: pbiswas@chemistry.du.ac.in*
[†] Electronic Supplementary Information (ESI) available. See DOI:

tions in designing folded/misfolded sequences with two letter HP (Hydrophobic-Polar) amino acid alphabet. A sequence, with marginal stability in the native/target conformation, is randomly selected for site directed point mutations. These point mutations are site directed for highly correlated sites and is incorporated by changing the residue either from hydrophobic to polar or from polar to hydrophobic such that the two-body residue propensity of that specific highly correlated site pair is opposed by the mutation. This directed mutation is applied to investigate whether such mutations lead to a directed outcome in terms of stability or not i.e. whether these mutations are always destabilizing or always stabilizing? The site-directed mutations does not yield a directed outcome; in fact, the outcome of the mutation is quite random with respect to the stability of the mutated sequence in the native/target conformation. The diverse set of sequences generated by cumulative point mutations are analyzed to assess the role of hydrophobicity in the core and the surface sites on the stability of the sequence in the native/target conformation. The mutation sensitive sites and their spatial positions in the native/target conformation i.e. core or surface is also determined. A method is proposed to calculate the number of clashing (unfavorable interaction) and matching (favorable interaction) residue pairs. The number of clashing/matching residue pairs is found to be strongly correlated with the stability of the mutated sequence in the native state.

2 Theory

The polypeptide chain is configured as a self-avoiding walk (SAW) on a $3 \times 3 \times 3$ cubic lattice, where each lattice point represents a residue/structural unit. Exact enumeration by the first depth algorithm yields 103346 unique, compact conformations unrelated by rotational, reflectional or translational symmetry.[18] The total number of possible sequences for the 27-mer cubic lattice protein is $2^{27} = 134217728$ with a reduced 2 letter amino acid representation i.e. hydrophobic and polar residues. The target/native conformation represents the most designable conformation as it corresponds to the lowest energy conformation for maximum number of sequences.[18–20] This HP lattice model is widely used to have important physical insights to many complex phenomena of proteins.[19,21–24] Even though progress in computational techniques have made atomistic modeling and simulation of proteins feasible, yet minimalist models are still relevant to achieve qualitative physical insights of complex and computation intensive problems.[25–29]

A suitable energy function is required to characterize the sequence-structure compatibility by quantifying the stability of a sequence in any conformation. The energy of a sequence in any conformation, E may be expressed as a function of the site-specific and pairwise monomer interactions.

$$E = \sum_{i=1}^{N} \gamma_i(\alpha_i) + \sum_{i<j} \gamma_{i,j}(\alpha_i, \alpha_j) \tag{1}$$

where N is the total number of amino acid residues present in the protein. The one-body term $\gamma_i(\alpha_i)$ quantifies the propensity of the monomer type α_i to reside in a particular structural context/environment.[30,31] The two-body term $\gamma_{i,j}(\alpha_i, \alpha_j)$ represents the inter-residue contact interactions of the monomer types α_i and α_j located at i-th and the j-th sites respectively.[31–33] Both one-body and two-body interaction terms may be expressed as

$$\begin{aligned} \gamma_i(\alpha_i) &= \textstyle\sum_k \sigma_{ik}^{(1)} \gamma_k^{(1)}(\alpha_i) \\ \gamma_{i,j}(\alpha_i, \alpha_j) &= \sigma_{i,j}^{(2)} \gamma^{(2)}(\alpha_i, \alpha_j) \end{aligned} \tag{2}$$

where k is the number of structural contexts and $\sigma_{ik}^{(1)}$ is the one-body structural parameter which indicates whether the i-th site is in the k-th structural context. Such structural contexts carries the information whether site i is buried in the interior of the protein or accessible to solvent.[34] The two-body structural parameter $\sigma_{ij}^{(2)}$ denotes the inter-residue contact interaction between the site pair i and j with monomer types α_i and α_j respectively.[35,36] Non-bonded nearest neighbors are assumed to be in contact.

$$\sigma_{ik}^{(1)} = \begin{cases} 1 & \text{if site } i \text{ is in structural context } k, \\ 0 & \text{if not.} \end{cases} \tag{3}$$

$$\sigma_{i,j}^{(2)} = \begin{cases} 1 & \text{if sites } i \text{ and } j \text{ are non-bonded neighbors,} \\ 0 & \text{if not.} \end{cases} \tag{4}$$

The one-body energy parameter, $\gamma_k^{(1)}(\alpha_i)$, denotes the energy contribution of the amino acid type α_i in k-th structural context. Here, solvent accessibility is chosen as an appropriate structural context, which is quantified in terms of the coordination number. Higher co-ordinated sites have lower solvent accessibility and reside in the core of the protein. Structural context $k = 1$ is chosen for sites with coordination number 1 or 2 while $k = 2$ is chosen for sites with coordination number 3 or 4. In this work, one-body energy parameters for different residue types in different structural contexts are chosen as[34]

$$\gamma_1^{(1)}(H) = 0 \quad \gamma_1^{(1)}(P) = 0 \quad \gamma_2^{(1)}(H) = -1\varepsilon \quad \gamma_2^{(1)}(P) = 0 \tag{5}$$

where, ε denotes the scaled energy unit, which measures the one-body interaction of the residue in a particular structural context.

The two-body energy parameter $\gamma^{(2)}(\alpha_i, \alpha_j)$ quantifies the inter-residue contact propensities[32] between a residue-pair α_i and α_j.

$$\gamma^{(2)}(H,H) = -3\varepsilon' \quad \gamma^{(2)}(H,P) = -1\varepsilon' \quad \gamma^{(2)}(P,H) = -1\varepsilon' \\ \gamma^{(2)}(P,P) = 0 \tag{6}$$

where, ε' is the dimensionless inter-residue interaction energy parameter .

Assuming small fluctuations due to sequence variations, the energy of any sequence in a particular native/target conformation may be expressed as a sum of energies of the residues due to local site-specific interactions and pair interactions with its neighbours

$$\overline{E_{nat}} = \sum_i \sum_{\alpha_i} \sum_k \sigma_{ik} \gamma_k^{(1)}(\alpha_i)\omega_i(\alpha_i) + \\ \sum_{i,j} \sum_{\alpha_i,\alpha_j} \sigma_{i,j} \gamma^{(2)}(\alpha_i,\alpha_j)\omega_{i,j}(\alpha_i,\alpha_j) \tag{7}$$

where $\omega_i(\alpha_i)$ is the site-specific monomer (one-body) probability of finding α_i type of residue at i-th site, while $\omega_{i,j}(\alpha_i,\alpha_j)$ is the pairwise monomer (two-body) probability such that α_i and α_j residue types are at i-th and j-th sites in the sequence respectively. Similarly, the ensemble averaged energy of the unfolded conformations may be expressed as

$$\langle\overline{E_{unf}}\rangle = \sum_i \sum_{\alpha_i} \sum_k \langle\sigma_{ik}\rangle \gamma_k^{(1)}(\alpha_i)\omega_i(\alpha_i) + \\ \sum_{i,j} \sum_{\alpha_i,\alpha_j} \langle\sigma_{i,j}\rangle \gamma^{(2)}(\alpha_i,\alpha_j)\omega_{i,j}(\alpha_i,\alpha_j) \tag{8}$$

The stability gap, Δ, represents the energy difference between the native state and the ensemble-averaged unfolded state energy, which may be expressed as

$$\Delta = \overline{E_{nat}} - \langle\overline{E_{unf}}\rangle \tag{9}$$

and Γ^2_{unf} quantifies the fluctuations in the energies of the unfolded state ensemble.

$$\Gamma^2_{unf} \quad = \langle\overline{E^2_{unf}}\rangle - \langle\overline{E_{unf}}\rangle^2 \tag{10}$$

In our earlier works, a foldability criterion, ϕ was derived using the cumulant expansion of the free energy of folding[34,36] which provides a measure of the sequence-structure compatibility.[37]

$$\phi = \Delta + \frac{1}{2}\Gamma^2_{unf} \tag{11}$$

where ϕ is a dimensionless quantity appropriately scaled with respect to k_BT. Eq. 11 clearly indicates that the more negative the value of ϕ the larger is the stability gap Δ representing a highly stable sequence in the target/native conformation. Hence, a sequence with lower negative ϕ value represents a marginally stable sequence. This foldability criterion is derived from a cumulant expansion approximating the free energy of folding and hence is a thermodynamic quantity. The folding kinetics may also be a major determinant of the foldability of a protein sequence[38] but results of the kinetic studies on model protein sequences indicate that there is a large correlation of the folding rate with the thermodynamic foldability criteria like Δ/Γ, T_f/T_g.[34,39–41] These studies confirm that such foldability criterion, which considers the stability gap and the fluctuation in the energies of the unfolded ensemble, may also be a good indicator of the kinetic foldability.

In this work, we assume that for a specific site-pair, the pairwise monomer probability is not explicitly coupled to the respective site-specific monomer probabilities. The one-body residue propensities are dependent on each site's overall structural context, while the pairwise monomer probability of a specific site pair is dependent on their contact interactions. However, the one-body and two-body probabilities are coupled to each other and among themselves through the set of constraints, which specify local/global features of the structure and sequence. Within this approximation the sequence entropy may be expressed as the sum of the contributions from one-body and two-body residue probabilities[42–45]

$$S = -\sum_i \sum_{\alpha_i} \omega_i(\alpha_i) ln(\omega_i(\alpha_i)) - \\ \sum_{i,j} \sum_{\alpha_i,\alpha_j} \omega_{i,j}(\alpha_i,\alpha_j) ln(\omega_{i,j}(\alpha_i,\alpha_j)) \tag{12}$$

The most probable set of one-body $(\omega_i(\alpha_i))$ and two-body probabilities $(\omega_{i,j}(\alpha_i,\alpha_j))$ may be determined by maximizing the entropy subject to the relevant constraints. These constraints are:

i) The normalization of site-specific probabilities at each site,

$$\sum_{\alpha_i=1}^{m} \omega_i(\alpha_i) = 1 \tag{13}$$

ii) The normalization of pairwise monomer probabilities for each pair of sites,

$$\sum_{\alpha_i,\alpha_j} \omega_{i,j}(\alpha_i,\alpha_j) = 1 \tag{14}$$

iii) The foldability criteria, ϕ, defined by Eq. 11

The variational functional, V, of the set of one-body and two-body probabilities may be expressed as

$$V = S - \sum_{i=1}^{N} (\beta_{norm1})_i(-1 + \sum_{\alpha_i=1}^{m} \omega_i(\alpha_i)) - \\ \sum_{i,j} (\beta_{norm2})_{i,j}(-1 + \sum_{\alpha_i,\alpha_j} \omega_{i,j}(\alpha_i,\alpha_j)) - \beta_\phi \phi \tag{15}$$

where, $(\beta_{norm1})_i$,$(\beta_{norm2})_{i,j}$ and β_ϕ are the Lagrange multipliers for the constraint Eqns. 13, 14 and 11 respectively. The inclusion of ϕ as a constraint in the variational functional makes it possible to design sequences with varied foldability. Solving the simultaneous equations that define the maximum of the variational functional subject to the appropriate

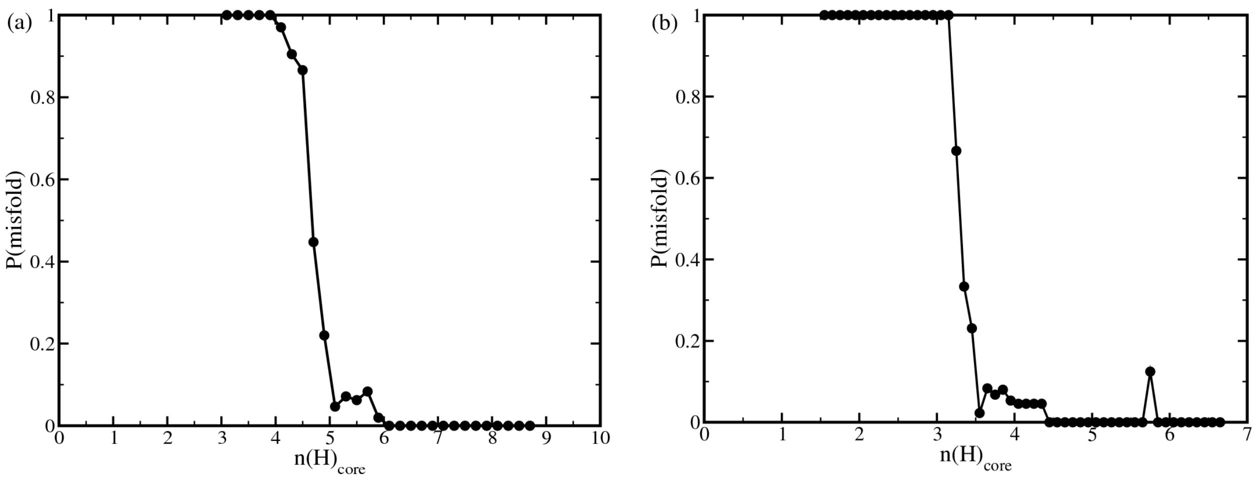


Fig. 1 Probability of misfolding (P(misfold)) of the mutated sequences is plotted against $n(H)_{core}$ for (a) real and (b) lattice protein.

constraint equations, a set of coupled non-linear equations are obtained.

$$\begin{aligned} &\omega_i(\alpha_i) = \tfrac{1}{q_i}(\exp(-\beta_\phi \phi_i)) \\ &\omega_{i,j}(\alpha_i,\alpha_j) = \tfrac{1}{q_{i,j}}(\exp(-\beta_\phi \phi_{i,j})) \\ &\phi = \Delta + \tfrac{1}{2}\Gamma^2 \end{aligned} \tag{16}$$

where,

$$\begin{aligned} &q_i = \textstyle\sum_{\alpha_i} \exp(-\beta_\phi \phi_i) \\ &q_{i,j} = \textstyle\sum_{\alpha_i,\alpha_j} \exp(-\beta_\phi \phi_{i,j}) \\ &\phi_i = \tfrac{\partial \phi}{\partial \omega_i(\alpha_i)} \\ &\phi_{i,j} = \tfrac{\partial \phi}{\partial \omega_{i,j}(\alpha_i,\alpha_j)} \end{aligned} \tag{17}$$

This set of equations are solved numerically to yield the one-body $(\omega_i(\alpha_i))$ and two-body probabilities $(\omega_{i,j}(\alpha_i,\alpha_j))$ and the Lagrange mutipliers consistent with a particular value of the foldability criterion, ϕ.

Among these designed sequences, a marginally stable sequence (lower negative ϕ), which represents a model real protein sequence, is selected for site-directed point mutations. A pair of sites (i, j) with highest correlation between the respective residue pairs is identified. The correlation between the i-th and j-th sites may be calculated as

$$C_{i,j}(\alpha_i,\alpha_j) = \frac{\omega_{i,j}(\alpha_i,\alpha_j)}{\omega_i(\alpha_i) \times \omega_j(\alpha_j)} \tag{18}$$

This term $(C_{i,j}(\alpha_i,\alpha_j))$ is a measure of the two-body interaction correlation between two specific sites. The denominator predicts the individual two-body probability of α_i and α_j residues at i-th and j-th sites simultaneously when the two sites do not interact. The numerator provides the two-body probability in presence of the specific two-body interaction. Thus the ratio is a measure of the correlation between the two specific sites via two-body interactions. $C_{i,j}(\alpha_i,\alpha_j) > 1.0$ implies a higher two-body propensity for α_i and α_j residues at i-th and j-th sites respectively while $C_{i,j}(\alpha_i,\alpha_j) < 1.0$ indicates lower two-body propensity for the same. Now the point mutations are incorporated such that the two-body propensity of a highly correlated residue pair is opposed by the mutation. For example, if the i-th site is occupied by a H (hydrophobic) residue then $C_{i,j}(H,\alpha_j)$ for each j and α_j are calculated. Depending on the value of the correlation, the j-th site with highest $C_{i,j}(H,\alpha_j)$ value is selected for mutation. Thus, the mutation process is site-directed. The selected site j is mutated such that it opposes the two-body propensity $(\omega_{i,j}(\alpha_i,\alpha_j))$ of that site pair for the specific pair of residues H and α_j. Suppose, if $C_{i,j}(H,P)$ has the maximum value for the i-th site, implying a high pairwise monomer probability $(\omega_{i,j}(H,P))$ for H and P residues at i-th and j-th sites respectively, then the j-th site is mutated with a hydrophobic residue such that their two-body propensity is opposed by the directed mutation. This mutation will change the foldability (ϕ) of the sequence. From the new foldability value of the sequence, a set of optimized pairwise monomer probabilities are self-consistently calculated by constraining Eq. 14 and Eq. 11, while keeping all the one-body monomer probabilities constant. This procedure is then repeated till a foldable sequence is obtained self-consistently such that a diverse set of sequences is generated from cumulative point mutations. The procedure is repeated with the same initial wild-type sequence for all possible sites i. This site-directed mutation procedure provides a platform to examine the outcome of such cumulative point mutations.

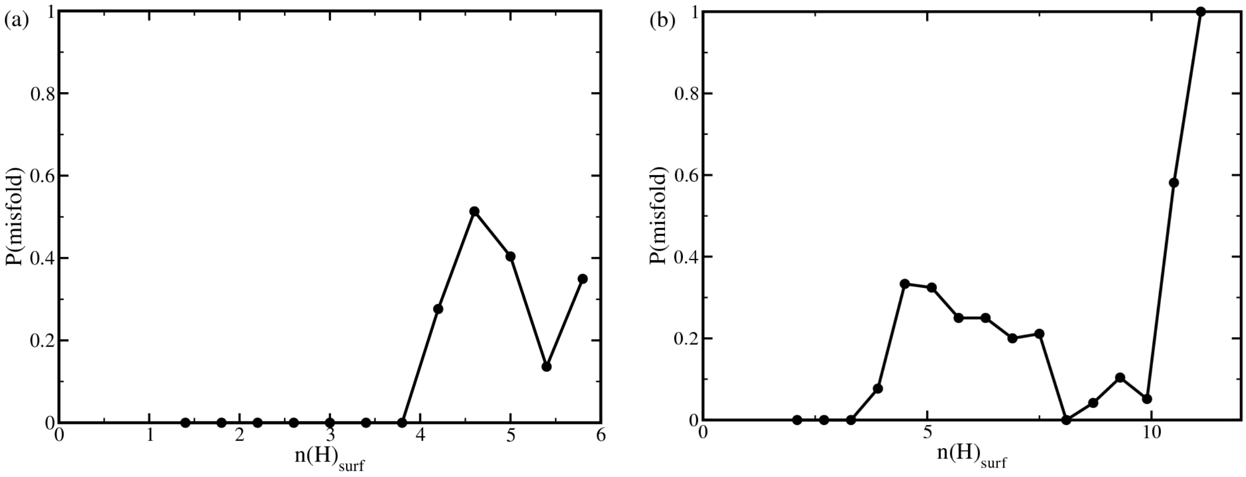


Fig. 2 Probability of misfolding (P(misfold)) of the mutated sequences is plotted against $n(H)_{surf}$ for (a) real and (b) lattice protein.

2.1 Real Protein

The mean-field theory is applied to a 21-mer real protein with pdb id 1EDN (x-ray resolution = 2.18 Å, R-value = 0.19)(http://www.rcsb.org/pdb/).[46] The co-ordinates of the C-α chain backbone in the crystal structure are chosen as the initial template for a Monte Carlo (MC) simulation to generate an ensemble of unfolded conformations. A simple $6-12$ LJ (Lennard-Jones) potential is applied and the pseudo C-α-C-α bond length is restricted to be 3.8 ± 0.15 Å.[47] The surface accessibility of the sites are measured using DSSP[48] and the corresponding relative surface accessibility are divided into 3 coordination zones.[49,50] Sites in the coordination zone 1 with relative surface accessibility $\geq 37\%$ are considered as surface sites in the structural context $k = 1$, while sites in the coordination zone 2 and 3 with relative surface accessibility $< 37\%$ represent buried sites in the structural context $k = 2$. A subset of 35539 generated conformations are selected such that these conformations have equal or lower number of buried sites and equal or lower number of two-body contacts as compared to the crystal (native) structure of the protein. Higher number of buried sites form a well defined core and higher number of two-body contacts ensures the compactness of the native/target structure. Thus, this selection criterion ensures highest designability of the native/target structure of the protein as mentioned in the “Theory” section. Sequences of varied foldability are designed by minimizing the energetic frustration and using the reduced HP alphabets, which are compatible to the native structure of 1EDN through the self consistent mean field theory as explained in the “Theory” section. This protein (1EDN) primarily consists of helical and loop structure. To study the geometric effects of the protein conformations, we have repeated the entire calculation for another real protein with a predominant β-sheet structure (PDB id 2PM1) with a resolution of 1.60 Åand crystallographic R-factor = 0.154. This protein is a derivative of human alpha-defensin 1 with 90% sequence identity. The high sequence identity implies similar energy landscape properties as that of a naturally occuring real protein sequence. The selection of these two proteins (1EDN and 2PM1) is based on the following criteria i) the structures are of reasonably high resolution, ii) they are monomeric without the presence of any ligand, DNA and RNA, ensuring autonomous folding of the sequence to the structure. This indicates that the target structures are not highly topologically frustrated and designing a foldable sequence is feasible by minimization of energetic frustration only, since this model does not explicitly account for topological frustration[51,52].

3 Results and Discussions

A marginally stable sequence with $\phi = -0.4$ is randomly selected from the designed real protein sequences for site directed mutations. For each site, the corresponding highest correlated site is identified and mutated such that it opposes the two-body probability of that specific site pair as explained in the “Theory” section. This point mutation procedure is repeated to accumulate mutations till a mutated sequence may be designed self-consistently, i.e., till the set of coupled transcendental equations may be solved to yield specific values for site-specific and pairwise monomer probabilities (beyond this the equations fail to converge). Thus a diverse range of mutated sequences are obtained. Similarly, a lattice protein sequence with $\phi = -1.0$ is selected and mutated repeatedly. The mutated sequences are analyzed to determine whether they are stabilized or destabilized against the most stable unfolded conformation. This can be calculated as

$$\Delta St = \left(\left((E_{lowest})_{unf} - E_{nat}\right)_{mut}\right)_i - \left(\left((E_{lowest})_{unf} - E_{nat}\right)_{mut}\right)_{i-1} \qquad (19)$$

The first term of RHS in Eq. 19 measures the energy difference between most stable conformation of the unfolded ensemble and the native/target state after *i*-th mutation while the second term measures the same after $i-1$-th mutation. The range of ΔSt corresponding to -0.0005 to $+0.0005$ is considered as neutral mutation i.e. mutations having negligible effect on the stability of a protein. $\Delta St < -0.0005$ represents a destabilizing mutation; the more negative the value, the more destabilizing is the mutation. The overall probabilities of destabilizing, stabilizing and neutral mutations are 0.15, 0.17 and 0.68 respectively. These values confirm that although mutations are performed in a directed way but the outcome is random, while stabilizing and destabilizing mutations are almost equally probable but most mutations are neutral.[53–55] Thus a directed mutation may not necessarily lead to a directed outcome i.e evenif the mutations are incorporated such that the two-body propensity of a specific highly correlated site pair is opposed by the mutation yet the outcome of the mutation in terms of the stability of the mutated sequence in the target/native structure is not always destabilizing or stabilizing.

Often, destabilizing mutations lead to misfolded sequences in which a conformation from the unfolded state ensemble becomes more energetically stable compared to the native/target conformation. How does the sequence composition dictate folding/misfolding? The composition of the mutated sequences are analyzed to explore the roles of hydrophobicity of the core and surface sites in protein misfolding. In Figure 1(a) and (b) the probability of misfolding(P(misfold)) is plotted against the average number of hydrophobic residues in the core sites of real and lattice proteins respectively. All the mutated sequences are mapped into different bins with respect to their hydrophobicity in the core (buried) sites. Thus the probability of misfolding of the *i*-th bin may be expressed as

$$P(misfold)_i = \frac{(N_m)_i}{(N_t)_i} \qquad (20)$$

$(N_m)_i$ represents the number of sequences in the *i*-th bin having lowest energy in a conformation other than the native conformation. $(N_t)_i$ is the total numebr of sequences in the *i*th bin. The plot suggests that the number of buried hydrophobic residues in the core play a crucial role on the stability of the protein. When the buried sites are predominantly occupied by the hydrophobic residues, the probability of misfolding is low and the sequence stably folds to the target/native conformation. The inverted sigmoidal-shaped plot indicates that the probability of misfolding is not linearly proportional to the number of hydrophobic residues in the core. A critical number of hydrophobic residues must be present in the core to impart stability to the native/target conformation of the protein. The importance of hydrophobicity in the core sites is shown by many experimental studies.[54,56–60] The native conformation of the real protein 1*EDN* has 10 buried residues, while the target conformation of the lattice protein has 7 buried sites. For the real protein conformations, the transition from misfolded to folded sequence occurs at around 5 buried hydrophobic residues as depicted in Figure 1(a). The probability of misfolding is minimum when the number of hydrophobic residues is greater than 5 (among 10 buried sites), leading to a stable native state. Similarly, four or more hydrophobic residues out of 7 buried sites of the lattice protein ensures higher stability of the sequence in the target/native conformation. Thus, a critical number of hydrophobic residues of at least 50% in the buried sites may be a requirement to stabilize the sequence in the native conformation. This critical number may vary for different globular proteins (See Figure S1 for corresponding result of protein 2PM1).

Varying hydrophobicity in the surface sites of the protein may profoundly affect the stability of a sequence in the native conformation. Figure 2(a) and (b) depicts the probability of misfolding as a function of the number of hydrophobic residues in the surface sites of real and lattice proteins respectively. From the plot, it is evident that lower number of hydrophobic residues in the surface sites stabilize the native state, hence the probability of misfolding is lower. But higher number of hydrophobic residues in the surface sites do not always lead to misfolding. This implies that if the hydrophobicity of the buried sites are above the critical value of hydrophobicity then the surface sites may tolerate some hydrophobic residues without misfolding. This observation is in accord with experiments. In one α-helical region of bacteriophage P22 Arc repressor, five polar to hydrophobic mutations at the surface are accomodated to retain its biological activity.[61] Schwehm *et. al.* have also demonstrated that in staphylococcal nuclease protein the surface sites are tolerant to many polar to hydrophobic mutations.[62] For both real and lattice protein conformations this trend is observed for the hydrophobicity of the surface sites (See Figure S2 for the corresponding result of the protein 2PM1). This study considers a simplified HP model of proteins. Inclusion of more detailed amino acid alphabets[63,64] in such models may be of interest. Such models can provide important information on the effect of aromaticity, charge or size of amino acids in the core or the surface of proteins. Detailed amino acid alphabets may be incorporated in this generalized model by accounting for the appropriate energetic contribution terms for such amino acids in the interaction potential. Inclusion of such amino acid alphabets and energetic terms may accurately describe the ge-

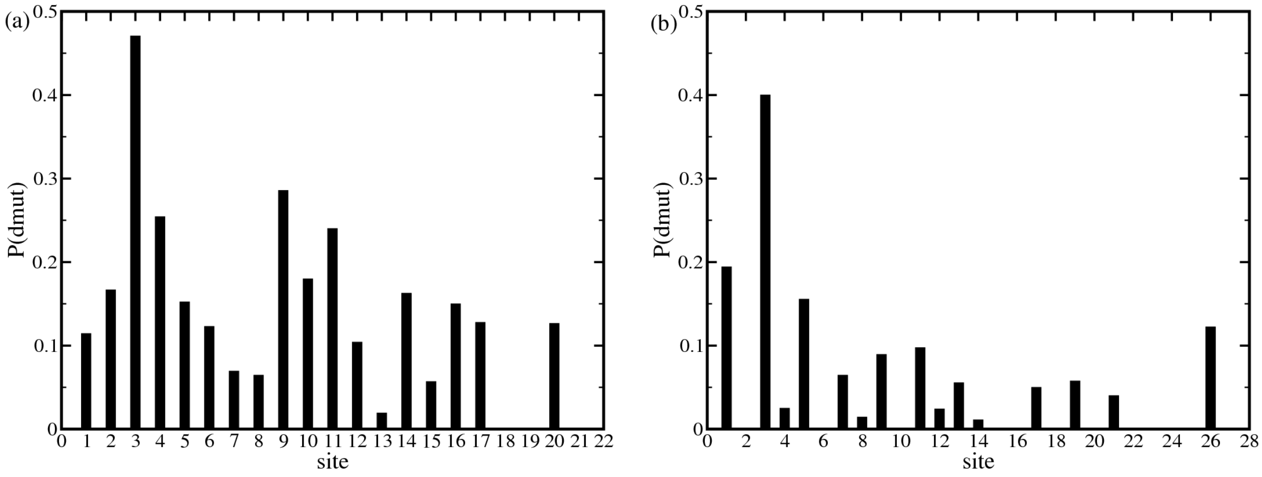


Fig. 3 Probability of destabilizing mutations (P(dmut)) is plotted against the corresponding mutated site for (a) real and (b) lattice protein.

ometric effects of protein secondary structure conformations on mutational stability at the expense of an increased cost of computation.

To investigate the mutation sensitivity of specific sites, the probability of destabilizing mutation is plotted against the corresponding mutated site in Figure 3. The probability of the destabilizing mutation of any i-th site may be calculated as

$$P(dmut)_i = \frac{(N_{\Delta St<-0.0005})_i}{(N_{total})_i} \tag{21}$$

where $(N_{\Delta St<-0.0005})_i$ represents the number of times the mutation is destabilizing when the i-th site is mutated and $(N_{total})_i$ is the total number of mutations of the i-th site. In Figure 3(a) and (b) the P(dmut) is plotted against corresponding sites for real and lattice proteins respectively. The plot clearly shows that all sites are not equivalent, while some mutated sites have a pronounced destabilizing effect others hardly perturb the stability of the native conformation. For real proteins, (Figure 3(a)) the sites 3, 9, 4 and 11 depict a markedly destabilizing effect upon mutation while for the lattice proteins, (Figure 3(b)) the sites 1, 3, 5 and 26 are highly sensitive to mutation with a high destabilizing probability as compared to other sites. Interestingly, although the coarse-grained potential of this work accounts for the hydrophobic collapse and hydrophobic-hydrophobic interactions only, the sites which show pronounced destabilization upon mutation (with high value of P(dmut)) are not just located in the core. For the real protein, the surface site 9, show destabilization upon mutation, while 3, 4and 11 form the core. Similarly,for lattice proteins the sites, 3 and 5, are from the surface while the 1st and the 26th sites are from the core. Figure 3(a) shows that 36% (4 out of 11)surface residues (13, 18, 19 and 21) have negligible destabilizing effect on mutation, while almost all core sites have some destabilizing effect for real proteins (See Figure S3 for the corresponding result for the protein 2PM1). Similarly, 50% (10 out of 20) surface sites (2, 6, 10, 15, 16, 18, 20, 22, 23 and 24) have negligible destabilizing effect, while 71% core sites (1, 17, 19, 21 and 26) show marked destabilization for lattice proteins (See Figure 4 for the position of the above mentioned sites in both 1EDN native structure and lattice target struture and Figure 5 for some of the misfolded structures of 1EDN and lattice protein). Thus, mutating surface and core sites exert non-equivalent effect in determining the stability of a a protein. In addition to the core sites, some specific surface sites may play a key role in dictating the stability of the protein in its native state. This finding is supported by the experimental studies[14,15,62] which show the importance of specific surface residues in determining the stability of a sequence. One study[62] shows that some specific surface sites have equivalent destabilizing effect upon mutation as that of the core sites even though most of polar to hydrophobic mutations on the surface hardly affects the stability of the protein. Another study[15] suggests the importance of the optimization of surface residues for designing thermostable proteins.

The interaction of a pair of residues present at i-th and j-th sites may be favorable or unfavorable depending on the residue types present at those sites. The two-body probabilities of any specific residue pair at i-th and j-th sites determine the preference of that residue pair for the given sites. The two-body probability of a residue pair, $\omega_{i,j}(\alpha_i, \alpha_j)$=0.25, denotes completely random preference for that residue pair for a given site pair i, j. Any value above 0.25 represents a favorable interaction and below 0.25 represents an unfavorable interaction. Thus, the clashing (unfavorable) or matching (fa-

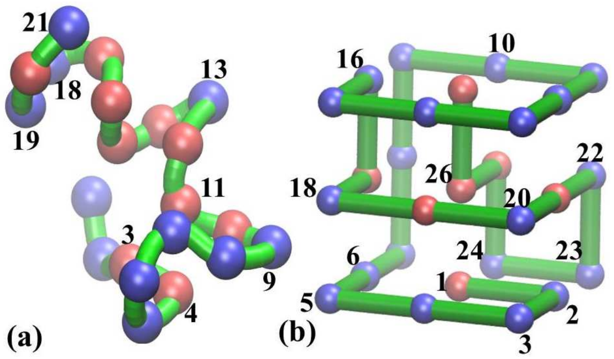


Fig. 4 (a) 1EDN native structure (b) lattice target structure. Blue beads represents polar residues and Red beads represents hydrophobic residues.

vorable) interactions between a given pair of residues present at i-th and j-th sites may be calculated as

$$I_{i,j}(\alpha_i,\alpha_j) = \frac{(\omega_i(\alpha_i)\times\omega_j(\alpha_j)) - 0.25}{(\omega_{i,j}(\alpha_i,\alpha_j) - 0.25)} \qquad (22)$$

The denominator is positive when there is a favorable interaction between the residue of type α_i at site i with the residue of type α_j at site j. The numerator is positive when the product of the site-specific probabilities of the residue α_i at site i and residue α_j at site j is greater than 0.25. i.e. both residues have a definite preference of co-existing at the respective sites. The numerator is always positive for the residues which have a high probability of co-existence, while the denominator is positive for residue pairs with favorable interactions. Thus positive value of $I_{i,j}(\alpha_i,\alpha_j)$ implies that the residues that have high probability of co-existence also interact favorably and form a matching residue pair. A negative value of $I_{i,j}(\alpha_i,\alpha_j)$ implies that residues with high co-existence probability have an unfavorable interaction constituting a clashing residue pair.

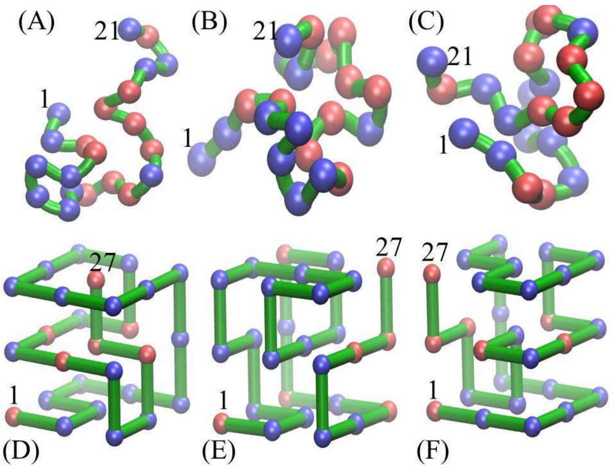


Fig. 5 (A),(B) and (C) are some of the misfolded structures of 1EDN and (D), (E) and (F) are some of the misfolded structures of lattice protein. Blue beads represents polar residues and Red beads represents hydrophobic residues.

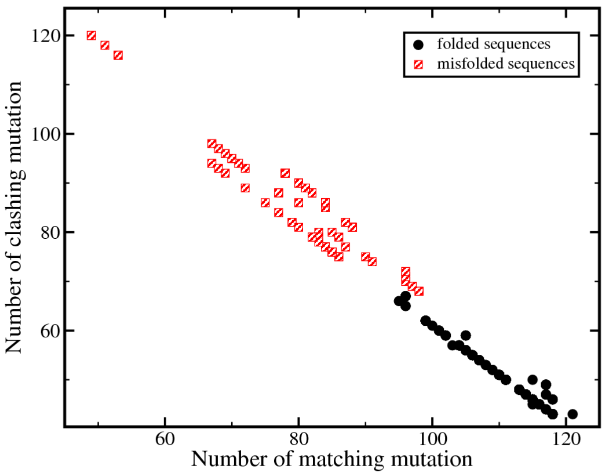


Fig. 6 Number of clashing mutation is plotted against the number matching mutation for folded and misfolded sequences.

The effect of a mutation may be estimated by evaluating either clash or match for both wild-type and mutated sequences. The gain or loss in stability may be calculated as

$$\Delta I_{i,j} = (I_{i,j})_{mut} - (I_{i,j})_{wild} \qquad (23)$$

A negative value of $\Delta I_{i,j}$ represents a clash inducing mutation and a positive value implies a match inducing mutation. Thus for a given mutated sequence, $\Delta I_{i,j}$ values for all possible residue pairs can be calculated relative to corresponding residue pairs of the wild type sequence; total number of negative $\Delta I_{i,j}$ values represents the number of clash inducing mutations in that specific mutated sequence. Similarly, total number of positive $\Delta I_{i,j}$ values represents the number of match inducing mutations. For all mutated sequences the number of clash and match inducing mutations are calculated. In Figure 6 the number of clash inducing mutations are plotted against the number match inducing mutations for all mutated sequences (folded and misfolded). The misfolded (red) and folded (black) sequences occupy two distinct regimes in this plot. Misfolded sequences occupy high clash and low match region, while low clash and high match zone is populated by the folded sequences. Thus there exists a cut-off value for the total number of clashing residue pairs that a sequence can accommodate without being misfolded. This indicates that the number of interactions of clashing residue pairs, rather than the specificity of clashing residue pairs, are important. Thus, it may be concluded from Figure 3 and Figure 6 that site-directed mutations, which lead to higher number of clashing residue pairs, results in misfolded sequences (See Figure S4 for corresponding result of protein 2PM1). Specific sites may be highly sensitive to mutations due to higher number clashing interactions induced by point mutations at those sites. Thus this evaluation of the clashing and matching residue pairs may be used to predict the outcome of a site-directed point mutation.

4 Conclusions

In the present article, a self-consistent mean-field based model is presented to investigate the effect of site-directed point mutations in designing folded/misfolded sequences with reduced amino acid alphabets. This site-directed point mutation method is developed and applied on designed protein sequences of both lattice and real proteins. The input to the model comprises of a coarse-grained potential which takes into account only the hydrophobic collapse and the hydrophobic-hydrophobic interactions. These point mutations target highly correlated residue-pairs such that it opposes the pairwise monomer probability of that specific site pair. Even though the point mutations are site-directed, but the outcome turns out to be random; stabilizing and destabilizing mutations are found to be equally probable. Analysis of the mutated sequences suggest the presence of a critical number of hydrophobic residues in the core to ensure that the mutated sequences fold to the native/target conformation. The misfolding probability of the mutated sequences increases sharply below this critical number. The surface sites clearly prefer polar residues, but increasing number of hydrophobic residues in the surface sites does not necessarily lead to misfolding, indicating the ability of the surface sites to accommodate hydrophobic residues, while the optimum hydrophobicity of the core sites is maintained. This result is also supported by experimental findings.[61]

Mutations at some sites are found to exert higher destabilizing effect, reaffirming the fact that all sites in the protein are not equivalent. Although mutations at the core sites are expected to destabilize the protein but specific surface sites are equally important and may be identified with such a coarse-grained potential. Experimental studies have already confirmed the importance of surface residues.[14,15] A method of identifying the clashing and matching residue pairs in the mutated and wild type sequences is proposed. Mutated sequences in the higher clash and lower match region are invariably misfolded, while those in the higher match and lower clash region are correctly folded to the native/target conformation. Thus mutations at sites which results in higher number of residue-residue clashes are sensitive to mutation and lead to misfolding. This clash-match method evaluates the comparative mutability of different residues resulting in a folded/misfolded sequence, which may provide the necessary framework to complement the outcome of site-directed mutagenesis experiments.

It should be noted that the selected real proteins are small and topologically simple. The results of the two real proteins are in good agreement to each other indicating that the findings may be independent of the secondary structure content and should be applicable to larger proteins without high topological frustration, at the expense of an exponentially increased computational cost. Future works, with sophisticated potential and consideration of topological frustration[51,52,65] may be required to further generalize the findings of this study for highly complex topologies like knotted proteins.

Acknowledgment

A. Baruah acknowledges University of Delhi for providing University Teaching Assistantship and DST, India for providing Senior Research Fellowship. The authors acknowledge Bioinformatics Resources and Applications Facility (BRAF) of the Centre for Development of Advanced Computing (CDAC), India for providing adequate computational facility in the Biogene cluster. The authors acknowledge financial assistance from DST, India (Project No. SB/S1/PC-023/2013) and Delhi University Research Grant.

References

1 J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci.*, 1987, **84**, 7524–7528.
2 J. D. Hermes, S. C. Blacklow and J. R. Knowles, *Proc. Natl. Acad. Sci.*, 1990, **87**, 696–700.
3 J. U. Bowie, *Curr. Opin. Struct. Biol.*, 2001, **11**, 397–402.
4 J. L. Battiste, R. Li and C. Woodward, *Biochemistry*, 2002, **41**, 2237–2245.
5 I. N. Smirnova and H. R. Kaback, *Biochemistry*, 2003, **42**, 3025–3031.
6 B. van den Burg and V. G. Eijsink, *Curr. Opin. Biotechnol.*, 2002, **13**, 333–337.
7 M. Lehmann, D. Kostrewa, M. Wyss, R. Brugger, A. D'Arcy, L. Pasamontes and A. P. van Loon, *Protein Eng.*, 2000, **13**, 49–57.
8 A. Suemori, *Protein Eng. Des. Sel.*, 2013, **26**, 479–488.
9 B. Synstad, S. Gåseidnes, D. M. van Aalten, G. Vriend, J. E. Nielsen and V. G. Eijsink, *Eur. J. Biochem.*, 2004, **271**, 253–262.
10 M. Guharoy and P. Chakrabarti, *BMC Bioinf.*, 2010, **11**, 286.
11 S. Z. Stepanovic, F. Potet, C. I. Petersen, J. A. Smith, J. Meiler, J. R. Balser and S. Kupershmidt, *J. Physiol.*, 2009, **587**, 2555–2566.
12 P. L. Wintrode, K. Miyazaki and F. H. Arnold, *Biochim. Biophys. Acta, Protein Struct. Mol. Enzymol.*, 2001, **1549**, 1–8.
13 B. Spiller, A. Gershenson, F. H. Arnold and R. C. Stevens, *Proc. Natl. Acad. Sci.*, 1999, **96**, 12305–12310.
14 A. A. Pakula and R. T. Sauer, *Nature*, 1990, **344**, 363–364.
15 A. Martin, V. Sieber and F. X. Schmid, *J. Mol. Biol.*, 2001, **309**, 717–726.
16 A. J. Lawson, E. A. Walker, S. A. White, T. R. Dafforn, P. M. Stewart and J. P. Ride, *Protein Sci.*, 2009, **18**, 1552–1563.
17 D. Verma, D. J. Jacobs and D. R. Livesay, *PLoS Comput. Biol.*, 2012, **8**, e1002409.

18 E. Shakhnovich and A. Gutin, *J. Chem. Phys.*, 1990, **93**, 5967.

19 H. Li, R. Helling, C. Tang and N. Wingreen, *Science*, 1996, **273**, 666–669.

20 H. S. Chan and K. A. Dill, *J. Chem. Phys.*, 1991, **95**, 3775–3787.

21 H. Taketomi, Y. Ueda and N. Gō, *Int. J. Pept. Protein Res.*, 1975, **7**, 445–459.

22 H. S. Chan and K. A. Dill, *J. Chem. Phys.*, 1989, **90**, 492.

23 E. I. Shakhnovich, *Phys. Rev. Lett.*, 1994, **72**, 3907.

24 C. J. Camacho and D. Thirumalai, *Proc. Natl. Acad. Sci.*, 1993, **90**, 6369–6372.

25 S. Moreno-Hernández and M. Levitt, *Proteins: Struct. Funct. Bioinf.*, 2012, **80**, 1683–1693.

26 T. Wüst and D. P. Landau, *J. Chem. Phys.*, 2012, **137**, 064903.

27 Z. Wang, L. Wang and X. He, *Soft Matter*, 2013, **9**, 3106–3116.

28 C. Holzgräfe, A. Irbäck and C. Troein, *J. Chem. Phys.*, 2011, **135**, 195101.

29 N. C. Shirai and M. Kikuchi, *J. Chem. Phys.*, 2013, **139**, 225103.

30 J. U. Bowie, R. Luthy and D. Eisenberg, *Science*, 1991, **253**, 164–170.

31 R. A. Goldstein, Z. A. Luthey-Schulten and P. G. Wolynes, *Proc. Natl. Acad. Sci.*, 1992, **89**, 9029–9033.

32 J. G. Saven and P. G. Wolynes, *J. Phys. Chem. B*, 1997, **101**, 8375–8389.

33 S. Miyazawa and R. L. Jernigan, *Macromolecules*, 1985, **18**, 534–552.

34 P. Biswas, J. Zou and J. G. Saven, *J. Chem. Phys.*, 2005, **123**, 154908.

35 J. Zou and J. G. Saven, *J. Mol. Biol.*, 2000, **296**, 281–294.

36 A. Bhattacherjee and P. Biswas, *J. Phys. Chem. B*, 2010, **115**, 113–119.

37 J. G. Saven, *Chem. Rev.*, 2001, **101**, 3113–3130.

38 P. E. Leopold, M. Montal and J. N. Onuchic, *Proc. Natl. Acad. Sci.*, 1992, **89**, 8721–8725.

39 D. Klimov and D. Thirumalai, *J. Chem. Phys.*, 1998, **109**, 4119–4125.

40 A. R. Dinner, V. Abkevich, E. Shakhnovich and M. Karplus, *Proteins: Struct. Funct. Bioinf.*, 1999, **35**, 34–40.

41 R. Mélin, H. Li, N. S. Wingreen and C. Tang, *J. Chem. Phys.*, 1999, **110**, 1252–1262.

42 E. T. Jaynes, *Phys. Rev.*, 1957, **106**, 620.

43 E. T. Jaynes, *Phys. Rev.*, 1957, **108**, 171.

44 R. Nettleton and M. Green, *J. Chem. Phys.*, 1958, **29**, 1365–1370.

45 H. J. Raveché, *J. Chem. Phys.*, 1971, **55**, 2242–2250.

46 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

47 A. Baruah, A. Bhattacherjee and P. Biswas, *Soft Matter*, 2012, **8**, 4432–4440.

48 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.

49 D. Kim, D. Xu, J.-t. Guo, K. Ellrott and Y. Xu, *Protein Eng.*, 2003, **16**, 641–650.

50 J.-T. Guo, J. W. Jaromczyk and Y. Xu, *Proteins: Struct. Funct. Bioinf.*, 2007, **67**, 548–558.

51 C. Clementi, P. A. Jennings and J. N. Onuchic, *Proceedings of the National Academy of Sciences*, 2000, **97**, 5871–5876.

52 J.-E. Shea, J. N. Onuchic and C. L. Brooks, *Proceedings of the National Academy of Sciences*, 1999, **96**, 12512–12517.

53 M. Kimura *et al.*, *Nature*, 1968, **217**, 624–626.

54 N. Tokuriki, F. Stricher, J. Schymkowitz, L. Serrano and D. S. Tawfik, *J. Mol. Biol.*, 2007, **369**, 1318–1332.

55 N. Tokuriki and D. S. Tawfik, *Curr. Opin. Struct. Biol.*, 2009, **19**, 596–604.

56 D. Shortle, W. E. Stites and A. K. Meeker, *Biochemistry*, 1990, **29**, 8033–8041.

57 B. W. Matthews, *Annu. Rev. Biochem.*, 1993, **62**, 139–160.

58 R. Liu, W. A. Baase and B. W. Matthews, *J. Mol. Biol.*, 2000, **295**, 127–145.

59 A. Hernández-Santoyo, L. Domínguez-Ramírez, C. A. Reyes-López, E. González-Mondragón, A. Hernández-Arana and A. Rodríguez-Romero, *Int. J. Mol. Sci.*, 2012, **13**, 10010–10021.

60 H. Chen and H.-X. Zhou, *Nucleic Acids Res.*, 2005, **33**, 3193–3199.

61 M. H. Cordes and R. T. Sauer, *Protein Sci.*, 1999, **8**, 318–325.

62 J. M. Schwehm, E. S. Kristyanne, C. C. Biggers and W. E. Stites, *Biochemistry*, 1998, **37**, 6939–6948.

63 K.-C. Chou, *Proteins: Struct. Funct. Bioinf.*, 2001, **43**, 246–255.

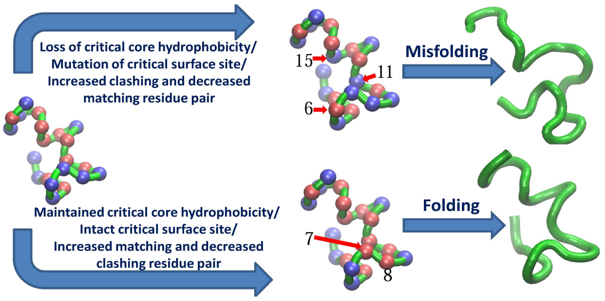
64 S. Pape, F. Hoffgaard and K. Hamacher, *Proteins: Struct. Funct. Bioinf.*, 2010, **78**, 2322–2328.

65 S. Gosavi, L. L. Chavez, P. A. Jennings and J. N. Onuchic, *Journal of molecular biology*, 2006, **357**, 986–996.

“For Table of Content Use Only”

Title : Role of Site-Directed Point Mutations in Protein Misfolding

Authors : Anupaul Baruah and Parbati Biswas*


Loss of critical core hydrophobicity/
Mutation of critical surface site/
Increased clashing and decreased
matching residue pair
15
11
6
Misfolding
Maintained critical core hydrophobicity/
Intact critical surface site/
Increased matching and decreased
clashing residue pair
7
8
Folding

Mutations inducing higher clashing and lower matching residue pairs lead to misfolding.