Stability Effects of Protein Mutations: The Role of Long-Range Contacts

Lavi S. Bigman and Yaakov Levy*

Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT: Predicting the effect of a single point mutation on protein thermodynamic stability (ΔΔG) is an ongoing challenge with high relevance for both fundamental and applicable aspects of protein science. Drawbacks that limit the predictive power of stability prediction tools include the lack of representations for the explicit energetic terms of the unfolded state. Using coarse-grained simulations and analytical modeling analysis, we found that a mutation that involves the breaking of long-range contacts may lead to an increase in the unfolded state entropy, which can lead to an overall destabilization of the protein. A bioinformatics analysis indicates that the effect of mutation on the unfolded state is greater for hydrophobic or charged (compared with polar) residues that participate in long-range contacts through a loop length longer than 18 amino acids and whose formation probabilities are relatively high.

INTRODUCTION

Protein structure and stability are strongly linked to the protein amino acid sequence. The stability of a protein is often particularly sensitive to mutations. Similarly, protein function can also be affected by mutations that may even lead to undesirable diseases via various mechanisms, among them reducing stability, protein misfolding and aggregation, change in allosteric flexibility, or a change in the network of interactions with other biomolecules. Protein mutagenesis is a common means to probe the role of a specific site in protein function, structure, folding kinetics, and stability. Understanding the effect of substituting various amino acids at a single site is valuable for the ongoing effort to engineer and design proteins with improved or even novel function.

Change in the thermodynamic stability of a protein upon mutation, ΔΔG = ΔGmutant − ΔGwild-type, where ΔG is the difference between the free energies of the folded and unfolded states, was measured experimentally for various proteins, at various sites, and with different substitutions. A point mutation with ΔΔG > 0 indicates destabilization. However, determining the thermodynamic stability of protein mutants through experimentation is very time-consuming; therefore, computational approaches to estimate the effect of mutations on stability are essential and of high practical value. Many methods have been developed to estimate ΔΔG, but this remains challenging because of the complexity of the physical interactions in folded proteins and the accuracy needed to predict thermodynamic parameters, as discussed below.

The heart of a prediction algorithm is the energy function, which is often simplified by a force-field that encompasses different terms that together should be able to predict correctly the stability of a given protein mutant. Many prediction algorithms for protein stability have been developed, with most relying on a 3D structure as their starting point. Thermodynamic stability prediction algorithms can be classified into three main groups: physical-based algorithms, knowledge-based models, and training-based models. The physical-based models, which aim to calculate ΔΔG from simulations involving detailed atomistic models that capture all of the physical interactions in proteins, are computationally too intense to be applied to a large number of mutations. The knowledge-based predictors often ignore protein dynamical structures have been used to estimate ΔΔG with reasonable accuracy. The existing methods vary in the structural information used to estimate ΔΔG. In training-based models, the parameters of the scoring functions are compared with a small database of experimental values that is used to train the parameters. For example, the energy function of FoldX includes terms for van der Waals (VDW) interactions, solvation, water bridges, and intramolecular hydrogen bonds, all based on empirical data. Additional terms that account for entropy and electrostatic interactions are included. All components of the Fold-X energy function have weights, which were tuned against a training set of >300 mutants. Many of the stability prediction tools exhibit correlations of 0.6 to 0.8 with experimentally measured ΔΔG. However, despite the success of the various algorithms in predicting the effect of single-point mutations on protein stability, many of them have limited performance and suffer from caveats.

The knowledge-based predictors often ignore protein dynamics and flexibility, although some mutations are expected to introduce strains in proteins’ backbone. Furthermore, training the energy function using a small experimental ΔΔG database may limit the transferability of the method and its accuracy. A common drawback in many of the prediction algorithms is an overly simplified representation of the unfolded state. These methods thus assume that the unfolded state can be

Special Issue: William A. Eaton Festschrift

Received: July 31, 2018
Revised: September 6, 2018
Published: September 10, 2018

DOI: 10.1021/acs.jpcb.8b07379
protein stability. Various global measures for the degree of long-range interactions in the native states (e.g., contact order) were shown to be correlated with the folding rates of proteins having different folds. The loop length of each of the native contacts was used to present a free-energy functional for the calculation of protein folding pathways and kinetics. Whereas the contact order of different proteins was found to be uncorrelated with their ΔG, it was shown that the mean loop length (⟨L⟩) of a mutated site in a studied protein is correlated with the relative stability of the mutants. However, the effect that deletion of long-range contacts will have on the stability of the unfolded state of a protein remains elusive.

In this study, we explored the effect of long-range contacts on protein stability. Our hypothesis was that mutating a site that is involved in long-range contacts may result in greater destabilization than mutating a site that participates in shorter-range contacts because of the larger entropic effect of the former on the unfolded state. The effect on thermodynamic stability of a mutation affecting short- and long-range contacts to different degrees was studied using coarse-grained molecular dynamics simulations (CG-MD). The effect of eliminating contacts on the enthalpy and entropy of the unfolded state was further quantified using a free-energy functional model for the protein folding energy landscape. Finally, the effect of long-range contacts on ΔΔG was examined for hundreds of experimentally measured mutants. Not only did ΔΔG correlate with the number of long-range contacts, but also the stability predictor struggled to capture this additional destabilization from the entropic changes of the unfolded state.

Figure 1. Entropy–enthalpy compensation in protein unfolded state. Unfolded proteins often have some residual structure, represented here by contacts that are far apart in sequence (a long-range contact, red beads) and contacts that are close in sequence (cyan beads). The deletion of a long-range contact leads to an increase in the protein’s conformational flexibility (left side, blurred beads) and the entropy of its unfolded state (ΔSU > 0, left side), whereas the deletion of a short-range contact leads to a minor change in the conformational flexibility and entropy of the unfolded state (ΔSF ≈ 0, right side). The enthalpic change due to the deletion of contacts in both cases is similar, and thus the free energy of the unfolded state is expected to be lower when removing contacts that are long range (i.e., with larger L).

determination of Protein Stability Using Coarse-Grained Molecular Dynamics Simulations. The effect of mutation on ΔΔG was studied computationally for the SH3 and CI2 domains (PDB IDs: 1SRL and 2CI2). The proteins were represented using a coarse-grained model in which each residue was represented by a single bead at the position of its Cα atom. The force field applied in our simulations used a native-topology-based potential. The potential in this model rewarded conformations that resemble the native fold and ensured a funnel-like energy landscape by excluding nonnative interactions.

The potential of a particular conformation V(Γ, Γ₀), where Γ denotes a particular conformation and Γ₀ denotes the native conformation along the coarse-grained simulation trajectory, consists of the following terms

\[
V(\Gamma, \Gamma_0) = \sum_{bonds} K_{bonds}(b_j - b_j^0)^2 + \sum_{angles} K_{angles}(\theta_{ijk} - \theta_{ijk}^0)^2 + \sum_{dihedrals} K_{dihedrals}(1 - \cos(\phi_{ijk} - \phi_{ijk}^0)) + \frac{1}{2}[1 - \cos(3(\phi_{ijk} - \phi_{ijk}^0))] + \sum_{\neq j} K_{repulsion} \left( \frac{C_{ij}}{r_{ij}} \right)^{12}
\]

where \(K_{bonds} = 100 \text{ kcal mol}^{-1} \text{ Å}^{-2}, K_{angles} = 20 \text{ kcal mol}^{-1}, \text{ and } K_{dihedrals, contact, repulsion} \) are each valued at 1 kcal mol⁻¹. The term \(b_j\) is the distance (in Å) between bonded beads \(i\) to \(j\).
and \( \theta_{ijk}^0 \) is the distance (in Å) between bonded beads \( i-j \) in the native conformation. The term \( \theta_{ijk} \) is the angle (in radians) between sequentially bonded beads \( i-j-k \), and \( \theta_{ijk}^0 \) is the angle between subsequently bonded beads \( i-j-k \) in the native conformation. The term \( \varphi_{ijk} \) is the dihedral angle (in radians) between subsequently bonded backbone beads \( i-j-k-\ell \), and \( \varphi_{ijk}^0 \) is the dihedral angle between subsequently bonded backbone beads \( i-j-k-\ell \) in the native conformation. The native contact interactions are modeled using the Lennard-Jones potential. \( A_q \) is the native distance (in Å) between beads \( i-j \) that are in contact with each other, and \( r_q \) is the distance (in Å) between beads \( i-j \) in a given conformation along the trajectory. Values of the native conformation parameters were calculated from the atomic coordinates of the X-ray structures. \( C_q \) is the sum of radii for any two beads not forming a native contact; the repulsion radius of the backbone bead was 2.0 Å. Electrostatic interactions between charged residues of the proteins were not included in this study.

Mutations were introduced by the removal of native contacts. We constructed five variants of both SH3 and CI2, each with four native contacts deleted per variant. The deleted contacts in a specific variant all had the same sequence separation between contacting residues. The variants of SH3 had sequence separations of 5, 10, 15, 18, and 22 amino acids, and the CI2 variants had sequence separations of 5, 8, 12, 17, and 20 amino acids. Eliminating the same number of native contacts that differ in their loop length minimizes enthalpic effects and enabled us to focus on the entropic consequences. We note that because of the simplicity of the model used in this study, the calculated enthalpy is effective and does not refer quantitatively to the experimentally measured values. Similarly, the estimate of entropy lacks the contribution of the solvent entropy, and therefore it refers to configurational entropy.

Similar native-topology-based models have been successfully used previously to capture the essential details of the folding of various proteins, including modified proteins. Further details can be found in previous studies. The folding of each protein mutant was studied at a temperature range that covers transitions from unfolded to folded states. The thermodynamic properties of each mutant were obtained by the weighted histogram analysis method (WHAM). In particular, the specific heat capacity, \( C_V \), as a function of temperature was used to identify the folding temperature, \( T_g \) (the peak of the \( C_V \) curve), as a measure for the relative stability of each protein variant. The effect of mutations (i.e., eliminating contacts) on stability is defined as

\[
\Delta T_g(\%) = \frac{T_g^{\text{WT}} - T_g^{\text{mut}}}{T_g^{\text{WT}}} \times 100%
\]

**Analytical Model for the Thermodynamic Characterization of the Unfolded State.** To obtain a quantitative measure of the competition between the changes in entropy and enthalpy of the unfolded state of the proteins, we used an analytical model based on the geometrical properties of the protein. This model was applied to the coarse-grained simulations of SH3 and CI2 and successfully reproduced many of their folding characteristics. The entropy of a specific state of the protein, defined by the fraction of contacts in that state, \( Q \) (0 < \( Q < 1 \)), is given by the following expression

\[
S_{\text{ref}}(Q) = N S_0 + S_{\text{bond}}(Q) + S_{\text{route}}(Q)
\]

where \( N \) is the number of residues in the protein (57 and 64 for SH3 and CI2, respectively), \( S_0 \) is the entropy of a residue when no contacts are formed, and \( S_{\text{bond}} \) is the entropic cost due to the formation of native contacts along the folding pathway. \( S_{\text{bond}} \) is given by

\[
S_{\text{bond}} = S_{\text{MF}} - \frac{3}{2} k_B M \log(\log(L))
\]

where \( S_{\text{MF}} \) is the entropy calculated using a mean-field approximation, and the second term in eq 2 reflects a decrease in entropy due to loop–loop fluctuations (\( L \) is the loop length between two residues that form a native contact). The exact term for \( S_{\text{MF}} \) is

\[
S_{\text{MF}} = -Q N S_0 - \frac{3k_B}{2} M Q \log(L) + \frac{3k_B}{2} M \log[1 + (1-Q)]
\]

\[
= \log[1 + (1-Q)]
\]

\( M \) is the total number of contacts in the protein (\( M = 137 \) and 142 for SH3 and CI2, respectively), \( L \) is the mean loop length, and \( k_B \) is the Boltzmann constant. The second term in eq 2 is given by

\[
M \log(\log(L)) = \sum_{i=1}^{M} \log(L_i - \log(L))
\]

where \( Q_i \) is the probability of formation of a specific contact and \( L_i \) is the loop length of that contact. \( S_{\text{route}} \) is the entropy gained from all of the different ways to arrange a specific set of MQ contacts in a specific state and is given by

\[
S_{\text{route}} = k_B \lambda(Q) \sum_{i=1}^{M} [-Q_i \log(Q_i) - (1 - Q_i) \log(1 - Q_i)]
\]

The function \( \lambda(Q) \) accounts for the decrease in entropy due to the connectivity of the polypeptide chain (further details can be found elsewhere). The effective enthalpy of a specific state is governed by the fraction of contacts in that state and is given by

\[
H(Q) = -\sum_{i=1}^{M} \epsilon_i Q_i
\]

The exact values of \( S_0 \) and \( \epsilon_i \) were tailored by Suzuki et al., specifically to be compatible with the native-topology-based simulations of SH3 and CI2; therefore, the same values are applied here. For SH3, \( \epsilon_i = 1.11 \text{ kcal mol}^{-1} \) and \( S_0 = 2.49 \text{ kcal mol}^{-1} \), and for CI2, \( \epsilon_i = 1.13 \text{ kcal mol}^{-1} \) and \( S_0 = 2.34 \text{ kcal mol}^{-1} \). Given that this study focuses on the unfolded state, a three-body term that accounts for the cooperativity was excluded.

**Structural Characterization of Protein Mutants.** In this study, a set of 607 mutants with experimentally measured \( \Delta G \), which were collected and summarized by Gueriois et al., was used to examine the effect of long-range contacts on protein thermodynamic stability. This set includes mutants of 33 proteins in which mutations are to either Ala or Gly. For each mutant, \( \langle L \rangle \) was calculated based on the 3D structure of the WT protein. For that purpose, the mutated residue was considered to be in contact with another residue if the distance, \( d \), between any side-chain heavy atom of the mutated...
amino acid and any other heavy atom in the protein was shorter than 5 Å. (In mutations to Ala, the Cα atoms were not included for counting contacts with neighboring residues.) The value of \( \langle L \rangle \) for each mutant was calculated by averaging the loop length, \( L \) (i.e., the sequence separation between the pair of residues that constitute each native contact), of its \( N \) contacts per mutated amino acid \( L = \frac{\sum_{i=1}^{N} L_i}{N} \). Note that only residues that had two or more interactions with the mutated amino acid were included in the contact list. The same definition of \( \langle L \rangle \) was used for the analysis of the mutations of the ribosomal protein S6.44

Each mutant was also characterized by the number of long-range contacts with neighboring residues. Using the same definition for contacts described above, a contact was considered “long” if it satisfies \( 15 < L_i < 35 \). When counting long-range contacts, only a single contact was considered between any pair of amino acids, even if their interaction was stabilized by multiple contacts. We note that there are other alternative ways to estimate the number of long-range contacts in the unfolded state. We classified the amino acids into three groups: hydrophobic (Ala, Val, Leu, Ile, Phe, Trp, Pro), charged (Lys, Arg, Glu, Asp), and polar (Gly, Ser, Thr, Cys, Tyr, Asn, Gln, Met).

\section*{RESULTS AND DISCUSSION}

**Contact Deletion Leads to Loop Length-Dependent Protein Destabilization.** Because various energetic and structural features may affect protein thermodynamic stability, isolating the contribution of a single feature to protein stability is a complex task. Here CG-MD simulations were used to highlight the effect of the contact loop length, \( L \), on protein stability. Although CG-MD simulations are too simplified to predict experimental \( \Delta G \) values, because they represent the protein at low resolution, both molecularly and energetically, they are nevertheless a powerful tool to dissect the contributions of topological parameters to the overall protein stability. To capture the effect of different loop lengths while minimizing changes to other aspects of protein stability, we constructed five variants of the src homology domain SH3 and the chymotrypsin inhibitor CI2, and in each variant, we deleted four native contacts. The deleted contacts in every variant all had the same loop length.

Contact deletion led to protein destabilization in a loop-length-dependent manner for both the SH3 and CI2 domains (Figure 2). Deletions of contacts with larger \( L \) resulted in greater destabilization. Because these variants differ only in the loop length of the deleted contacts, it is plausible that the driving force for the loop-length-dependent destabilization is entropic rather than enthalpic. In addition, because all variants had the same number of contacts, it is unlikely that the stability of the folded state will be significantly different between the variants. A similar effect of modification resulting in entropically driven protein destabilization was reported for other systems.46,49,55

\section*{Deletion of Long-Range Contacts Leads to Entropy-Driven Stabilization of the Protein Unfolded State.}

To directly probe the effect of loop length on the free energy of the unfolded state of proteins, we used an analytical model with an energy functional that explicitly takes into account the loop length of each native contact29–31,34 (see the Methods for details). This model was tailored such that it could be compared to the results of CG-MD simulations of SH3 and CI2.31 We focused on the unfolded state of the SH3 and CI2 proteins, defined here as the ensemble of conformations in which 20% of the native contacts were formed (i.e., \( Q = 0.2 \)). To mimic the effect of mutations on the unfolded state, we deleted each native contact one at a time and recalculated the free energy of the unfolded state (i.e., \( \Delta G_U = G_{U_{\text{WT}}} - G_{U_{\text{Mut}}} \)). We found that \( \Delta G_U \) decreases as loop length increases, and this trend is independent of the probability of contact formation. Interestingly, when the probability of contact formation in the unfolded state is \( Q > 0.5 \) (Figure 3A,C, yellow circles) and \( L_i \geq 18 \), the unfolded state experiences overall stabilization, as reflected by \( \Delta G_U < 0 \) (Figure 3A,C, shaded area). We then sought to determine whether this stabilization of the unfolded state is entropic or enthalpic in origin.

Decomposition of \( \Delta G_U \) into its enthalpic (\( \Delta H_U \)) and entropic (\( T \Delta S_U \)) components reveals that the stabilization is, indeed, entropic (Figure 3B,D). The diagonal lines in Figure 3B,D represent full compensation between entropy and enthalpy (\( \Delta H_U = T \Delta S_U \)). For clarity, we shaded the area that corresponds to \( T \Delta S_U > \Delta H_U \) to highlight the parameters for which \( \Delta G_U < 0 \) (shaded gray areas in Figure 3A,C). Interestingly, the data points above the diagonal, which represent entropic stabilization, correspond to the deletion of long-range contacts with \( L_i \geq 18 \). By contrast, the points below the diagonal, which represent enthalpic destabilization of the unfolded state, represent shorter range contacts with \( L_i < 18 \). Hence, we conclude that the deletion of long-range contacts leads to entropic stabilization of the unfolded state, which can explain the overall destabilization of the proteins, as was observed in the CG-MD simulations (Figure 2).

It is noteworthy that loop length alone is insufficient to explain the stabilization of the unfolded state upon the deletion of nonlocal interactions. The main reason for the increase in entropy when deleting a long-range contact is the removal of a configurational constraint. However, if the probability of the formation of a long-range contact is too low, then removing a long-range contact will not contribute much to the entropy of the unfolded state. This argument is reflected in eq 4, in which the contribution of long-range interactions (\( L \)) is coupled to the probability of contact formation, \( Q \), between them. In Figure 4, we demonstrate that when \( Q \) is low (Figure 4 left
Deletion of long-range contacts leads to entropy-driven stabilization of the unfolded state of the protein. The thermodynamic properties of the unfolded state of SH3 and CI2 were calculated based on an analytical model that explicitly takes into account the loop length formed by a pair of interacting residues (see main text and the Methods for details). (A,C) Change in free energy of the unfolded state ($ΔG_u$) upon contact deletion is shown as a function of the loop length ($L_i$) of the deleted contact for SH3 (A) and CI2 (C). The deleted contacts are grouped into three groups based on the probability of their contact formation, $Q_i$, and colored differently, as indicated in the Figure legend. Only the unfolded state, which was defined as the state in which 20% of the native contacts were formed (i.e., conformations with $Q = 0.2$), was included in this analysis. The area in which the unfolded state is stabilized is shaded gray. (B,D) The black diagonal line represents full compensation between enthalpy ($ΔH_u$, x axis) and entropy ($ΔS_u$, y axis) of the unfolded states of SH3 (B) and CI2 (D). Note that entropic stabilization of the unfolded state occurs for variants that are characterized by long loop length ($L_i ≥ 18$, red circles), whereas enthalpic destabilization occurs for variants with short loop length ($L_i < 18$, blue circles).

To further characterize the structural properties of mutations that result in larger $ΔΔG$, we calculated the number of long-range contacts in which each of the mutated amino acids participates (see the Methods for details). We found that mutants with a greater number of long-range contacts exhibit greater destabilization (i.e., have higher values of $ΔΔG$) than mutants with fewer long-range contacts (Figures S5C,D). Interestingly, the slope of the plot of $ΔΔG$ versus the number of long-range contacts is slightly higher for the S6 mutants than the slope for the larger data set of 607 mutants. This difference may be due to the fact that the S6 mutants are almost exclusively mutations of hydrophobic amino acids to Ala. Therefore, this may indicate that the correlation between the degree of lost long-range contacts and the magnitude of protein destabilization is more profound for hydrophobic amino acids. To further examine the linkage between the loss of long-range contacts and protein stability, we classified the 607 mutants into three groups: hydrophobic, charged, and polar amino acids. It appears that for all three groups of amino acids there is a positive correlation between $ΔΔG$ and the number of long-range contacts (Figure 6, top panels).

Taken together, the analysis of experimental data supports our conclusions that the deletion of long-range contacts enhances thermodynamic destabilization, which originates from an increased unfolded state entropy.

Destabilization by Mutations Might Be Underestimated by Current Predictors. One question remains open: To what extent can the removal of long-range interactions due to a mutation be applied to improve the performance of protein stability predictions? For that purpose, we calculated the difference between the calculated and experimental $ΔΔG$, entropic stabilization of the unfolded state and overall destabilization of CI2 and SH3 (Figures 2 and 3) may suggest that this is a general feature in protein mutants. To explore this aspect more thoroughly, we analyzed the relationship between the loop length and the thermodynamic stability of 607 mutants (from 33 proteins) with single-point mutations to either Ala or Gly (data extracted from ref 15). We defined a critical loop length, $L_C$, such that mutants are classified as “long” if $L > L_C$ and as “short” otherwise. Using this definition, it appears that long mutants are less stable than short mutants (Figure S5A).

Figure 4. Effect of the probability of contact formation on the entropy–enthalpy compensation due to deletion of a contact of loop length $L_i$. The entropic term that is mostly affected by a single mutation is $M(ΔQ_i(\log L_i)) = \sum_{i=1}^{\text{mut}} (Q_i - Q) (\log L_i - \log \bar{L}_i)$. The entropic gain ($TS$, solid line) for deleting a single contact with varying loop length, $L_i$ (x axis), depends on the probability of the formation of that specific contact. In this Figure, we used the value of ($\bar{L}_i$) = 10. The enthalpy ($H$, open circles) associated with breaking a contact depends only on its $Q_i$ and not on its $L_i$. 

Figure 3. Deletion of long-range contacts leads to entropy-driven stabilization of the unfolded state of the protein. The thermodynamic properties of the unfolded state of SH3 and CI2 were calculated based on an analytical model that explicitly takes into account the loop length formed by a pair of interacting residues (see main text and the Methods for details). (A,C) Change in free energy of the unfolded state ($ΔG_u$) upon contact deletion is shown as a function of the loop length ($L_i$) of the deleted contact for SH3 (A) and CI2 (C). The deleted contacts are grouped into three groups based on the probability of their contact formation, $Q_i$, and colored differently, as indicated in the Figure legend. Only the unfolded state, which was defined as the state in which 20% of the native contacts were formed (i.e., conformations with $Q = 0.2$), was included in this analysis. The area in which the unfolded state is stabilized is shaded gray. (B,D) The black diagonal line represents full compensation between enthalpy ($ΔH_u$, x axis) and entropy ($ΔS_u$, y axis) of the unfolded states of SH3 (B) and CI2 (D).
sufficiently high. It is possible that the unfolded state energy value is high enough for hydrophobic and charged amino acids but not for polar amino acids.

We note that the correlations presented in Figure 6 are relatively weak for several reasons that are related to the simplification of characterizing the entropy of the unfolded state by a crude structural characterization based on the number of native long-range contacts. There are various ways to count long-range contacts that may contribute to the residual structure of the unfolded state. As argued above, it is not trivial to estimate the formation probabilities of these contacts. Furthermore, additional non-native long-range interactions can also affect the unfolded state entropy. Thus the bioinformatics analysis can serve solely as a support for the reported biophysical effect of long-range contacts on stability, and further quantitative analysis demands a more refined description of the unfolded state.

The effect of long-range contacts on protein stability can be nicely demonstrated for the case of the villin headpiece, a small helical protein. Because three Phe residues are closely packed in its hydrophobic core and interact with each other with different loop lengths \(L_c\) between them, the villin headpiece can serve as a good model system to demonstrate our main hypothesis. The mutation of different Phe residues leads to destabilization of the villin headpiece in a way that depends on \(L_c\) of the deleted contacts (Figure 7, left y axis, circles). However, Fold-X does not capture the experimentally observed trend (Figure 7, right y axis, triangles). Hence, the villin headpiece serves as another demonstration that deletion of long-range contacts can lead to larger destabilization of a protein and that this effect is difficult to predict using the existing prediction algorithm for the effect of mutation on protein stability.

## CONCLUSIONS

Predicting the effect of a single point mutation on protein thermodynamic stability is an ongoing challenge. Although current prediction algorithms exhibit good performance, with a correlation between predicted and measured \(\Delta G\) of \(-0.8\), there are still several biophysical aspects that are not understood and are not explicitly represented in the energy functions for the calculation of thermodynamic stability. The success of the current knowledge-based stability prediction tools is most likely linked to training their scoring functions, which indirectly leads to their capturing complex biophysical effects. Drawbacks limiting the predictive power of \(\Delta G\) predictor tools include that they lack representation of explicit energetic terms for the unfolded state. The free energy of the unfolded state can, in principle, be affected by mutations that modify its configurational entropy, primarily by disruption of long-range interactions. Although the role of long-range contacts was studied extensively in protein folding and the effect of such contacts on the mutants remains elusive.

In this study, we used a combination of CG-MD simulations together with analytical and bioinformatic analysis to study the effect of long-range contacts on the stability of protein mutants. CG-MD simulations showed that contact deletion leads to loop length-dependent destabilization for both the SH3 and CI2 domains. The deletion of long-range contacts results in a larger destabilization than when shorter range contacts are deleted. An analytical model illustrates that this effect originates from the increased entropy of the unfolded state. Similarly to the CG-MD simulations, the analytical
The principal finding of this study is that a mutation that breaks long-range contacts may lead to an increase in the unfolded state entropy, which can lead to an overall destabilization of the protein. Introducing this finding into prediction algorithms could be valuable because current algorithms need improvement in modeling the effect of the unfolded state on the stability of mutant proteins. Whereas modeling the unfolded state can be simplified by assuming that its entropy is mostly hydrophobic and electrostatic interactions, it may involve long-range native as well as non-native interactions whose frequencies are difficult to predict computationally. Accordingly, incorporating the free energy of the unfolded state into prediction algorithms is not trivial. Some NMR techniques, such as paramagnetic resonance enhancements (PREs), residual dipolar coupling (RDCs), and long-range nuclear Overhauser effects (NOEs), can provide such information on inter-residue distances in unfolded states, which can be translated into probabilities of contact formation. An accurate estimation of
the probabilities of long-rang contacts can be used to improve the performance of existing prediction algorithms.

**AUTHOR INFORMATION**

**Corresponding Author**
*E-mail: Koby.Levy@weizmann.ac.il. Tel: 972-8-9344587.*

**ORCID**
Yaakov Levy: 0000-0002-9929-973X

**Notes**
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**
This work was supported by the Benoziyo Fund for the Advancement of Science and by the Kimmelman Center for Macromolecular Assemblies. Y.L. is The Morton and Gladys Pickman professional chair in Structural Biology.

**REFERENCES**


(35) Giri Rao, V. V.; Gosavi, S. The Multi-Domain Protein Adenylate Kinase, Domain Insertion Facilitates Cooperative Folding
Protein Folding? An Investigation For Small Globular Proteins.

Protein Folding: The Energy Landscape Perspective.


Nabuurs, S. M.; Westphal, A. H.; Van Mierlo, C. P. M. Extensive Formation Of Off-Pathway Species During Folding Of An


(82) Dyson, H. J.; Wright, P. E. Unfolded Proteins and Protein Folding Studied by NMR. Chem. Rev. 2004, 104 (8), 3607−3622.
