Chapter III

Analysis of rate-limiting long-range contacts in the folding rates of three-state and two-state proteins

3.1 Introduction

In the past decade, the number of models developed for predicting the folding rates of three-state proteins was found to be lower when compared to models developed for two-state proteins. This is due to the reason that the folding pathway of three-state proteins has an intermediate state which makes the folding of three-state proteins a complicated process, whereas no intermediate state was observed in the folding pathway of two-state proteins. Also, the three-state proteins have higher chain length and 3D structures mostly organized with more than a single domain, whereas two-state proteins have single domain with chain length around 100 residues. Because of these complexities present in three-state proteins, theoretical models developed to predict the folding rates of three-state proteins remains limited. However overcoming these complexities, considerable number of models has been developed to predict the folding rates of three-state proteins. The models developed for predicting the folding rates of three-state proteins have been discussed under section 1.9.5 in chapter-I.

Contact formation between two residues along the sequence in an unfolded polypeptide chain is one of the most elementary processes of protein folding. Experimental support for the importance of long-range contacts has been observed in the unfolded non-native state of several proteins using the NMR method. Ittah and Haas (1995) analyzed the folding intermediates of bovine pancreatic trypsin inhibitor using NMR and revealed that the loops
formed in early, non-native states are stabilized by nonlocal interactions. Seetharaman et al. (2002) showed that the NMR study on the early formation of hydrophobic clusters in the non-native state of lysozyme was linked together by long-range interactions. Lietzow et al. (2002) studied the acid-unfolded state of apomyoglobin using paramagnetic spin labeling and observed that significant interactions are observed between N- and C- terminal regions. Mizuguchi et al. (2003) studied apoplastocyanin using heteronuclear single quantum correlation spectroscopy and observed that local and long-range interactions in native apoplastocyanin are formed simultaneously, consistent with the highly cooperative formation of native structure. Dedmon et al. (2005) used paramagnetic relaxation enhancement NMR and molecular dynamics simulations studies on the protein α-synuclein with 140-residues and showed that the aggregation of α-synuclein is inhibited by the existence of long-range interactions within the native structure and suggested that these interactions may affect the role of α-Syn in the pathogenesis of Parkinson’s disease. Mok et al. (2007) observed residual structure due to hydrophobic collapse in the unfolded state of a small protein TC5b with strong inter-residue contacts between side chains that are relatively distant from one another in the native state. Felitsky et al. (2008) showed that hydrophobic clusters with transient long-range contacts observed in the acid-unfolded state of apomyoglobin played key role in initiating chain collapse and folding. Meier et al. (2008) suggested that by using novel NMR techniques, it has become possible to obtain information about local conformations as well as long-range interactions present under denaturing conditions. Zarrine-Afsar et al. (2008) analyzed the energetics and position specificity of nonnative hydrophobic interactions in the Fyn SH3 domain by a continuum coarse-grained chain model and predicted that energetically significant nonnative interactions led to
acceleration/deceleration of the folding rate and concluded that specific nonnative interactions significantly influence folding energetics. Nabuurs et al. (2010) observed long-range mediated nonnative contacts between transiently ordered structures of unfolded apoflavodoxin. The above experimental observations using NMR technique in the unfolded, non-native states of several proteins showed that long-range contact formation play a key role in initiating native state formation and in stabilization of native structure.

3.2 Importance of long-range interactions in protein folding

Gromiha and Selvaraj (1999) grouped long-range contacts (Residues that are within a sequence separation distance of >± 4 residues from the central residue) into several bin intervals with a step of 10 as 4-10, 11-20, 21-30, 31-40, 41-50 and >50 for all the residues in a protein molecule for a sphere of 8Å radius and various numbers of long-range contacts occurring at various bins for all the residues in a protein molecule were computed. The importance of long-range interactions from these various bins have been analyzed in four different structural classes of 150 globular proteins and an opposite trend in long-range contact preference between all-α and all-β proteins was observed. The all-α proteins had more number of long-range contacts in the 4-10 range, whereas all-β proteins posses more number of long-range contacts in 11-20 range. This was suggested to be due to the specific hydrogen bonding pattern of α-helices and β-sheets in these structural classes. In mixed-class proteins, the range 4-10 was favored by α+β class proteins, while the α/β class proteins prefer the 21-30 range. These results showed that long-range contacts from various sequence separation bins are crucial in the folding of proteins belonging to the four major structural classes and different structural classes.
prefer long-range contacts from different sequence separation bins to reach their 3D structure.

Since, it was observed that long-range interactions plays a crucial role in protein folding, we have used the parameter long-range order (LRO) derived from the long-range interactions to understand the folding rates of three-state proteins. A correlation of -0.70 was observed between folding rates of 35 three-state proteins and LRO obtained at a sequence separation distance of 7 residues at spatial distance cut-off of 10Å.

Hence in the present work, following the methodology of Gromiha and Selvaraj (1999) long-range contacts from different bin intervals (4-10; 11-20; 21-30; 31-40; 41-50; 51-60; >60) were computed and their roles were analyzed in predicting the folding rates of three-state proteins. Results observed from the present work showed a significant correlation between folding rates and long-range contacts from various sequence separation bins. Linear regression equations were developed using long-range contacts for predicting folding rates of 35 three-state proteins and for comparison, analysis of long-range contacts from various sequence separation bins in 45 two-state proteins was also carried out and their folding rates were predicted with better accuracy.

3.3 Materials and methods

In the present work, we have taken a data set of 35 three-state and 45 two-state proteins belonging to three major structural classes (all-α, all-β and mixed-class). Atomic co-ordinates for the proteins were taken from Protein Data Bank (Berman et al. 2000). The experimental folding rates ln(kf) for this proteins were collected and reported in Ouyang and Liang (2008). The computation of
long-range contacts from various bin intervals has been detailed in the following section.

3.3.1 Computation of long-range contacts from various bin intervals

Following the methodology of Gromiha and Selvaraj (1999) various numbers of long-range contacts occurring at various bins (4-10, 11-20, 21-30, 31-40, 41-50, 51-60 and >60) for all the residues in a protein molecule was computed. Each residue in a protein molecule has different number of long-range contacts at different sequence separation bins. The number of long-range contacts occurring in each bin for all the residues in 35 three-state and 45 two-state proteins were computed using an in-house FORTRAN program. Correlation between long-range contacts from various sequence separation bins (4-10; 11-20; 21-30; 31-40; 41-50; 51-60; >60) and experimental folding rates for both the 35 three-state and 45 two-state proteins was computed.

3.3.2 Prediction of folding rates based on long-range contacts using multiple regression technique

We have predicted the folding rates for the 35 three-state and 45 two-state proteins by relating long-range contacts obtained from various sequence separation bins and experimental folding rates using multiple regression technique. In the multiple regression case, when there is more than one independent variable, the regression line cannot be visualized in the two-dimensional space. Hence in the multiple regression technique the relationship between several independent or predictor variables and a dependent or criterion variable is analyzed. In our present work, experimental folding rates are considered as the dependent variables and long-range contacts from various bin intervals were considered as the independent variables. In general,
multiple regression procedures was estimated with a linear equation of the form
\[ Y = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p. \]
‘Y’ is the experimental folding rate, ‘a’ is the constant and ‘b’ is the slope and the ‘X’ variables are the long-range contacts obtained from various sequence separation bins.

We have predicted the folding rates of 35 three-state and 45 two-state proteins using back-check and jack-knife methods. The methodology of back-check and jack-knife prediction methods have been explained under section 2.2.2 in chapter-II.

3.4 Results and Discussion

3.4.1 Relationship between long-range contacts from various sequence separation bins and folding rates for three-state proteins

In Table 3.1, correlation coefficient obtained between experimental folding rates and long-range contacts for combinations of bins ranging from 4-10, 11-20, 21-30, 31-40, 41-50, 51-60 and >60 for the 35 three-state proteins are presented. For the 35 three-state proteins, a correlation of 0.88 was observed between experimental folding rates and long-range contacts from all the sequence separation bins ranging from 4-10, 11-20, 21-30, 31-40, 41-50, 51-60 and > 60 were included in the multiple regression equation. A correlation of \( r = 0.69 \) is observed between experimental folding rates and long-range contacts for the combination of first two bins, whereas including contacts from the corresponding bins one by one in the multiple regression equation improved the correlation. While correlation was computed between experimental folding rates and long-range contacts from individual sequence separation bins, long-range contacts from the first three bins showed a correlation of -0.66, -0.60 and -0.66 with experimental folding rates. Including
only these three bins in the regression equation showed a correlation of 0.81 with experimental folding rates. From this result, it is clear that most of the rate-limiting long-range contacts necessary for structure formation of 35 three-state proteins are present in the first three bins themselves.

### 3.4.2 Folding rate prediction of three-state proteins based on long-range contacts

In Table 3.2, experimental folding rates, along with the computed long-range contacts from various sequence separation bins, results of back-check, jack-knife predicted folding rates and their deviations from experimental folding rates for the 35 three-state proteins were given. For the full set of 35 three-state proteins, the back-check method predicted the folding rates of 15 proteins with a deviation less than 1 from experimental folding rates. Particularly the protein 1QOP-b with 388 residues and folding rates value of -6.9 has been excellently predicted as -7.1. The average deviation between experimental and predicted folding rates for the 35 three-state proteins through back-check method is 1.82. The regression equation for the 35 three-state proteins developed using long-range contacts from the above seven bins is

\[
\ln(k_f) = (-0.0019) (x1) + (-0.0106) (x2) + (-0.0104) (x3) + (0.0115) (x4) + (-0.0353) (x5) + (-0.0145) (x6) + (-0.0034) (x7) + 7.5375
\]

where \( x1, x2, x3, x4, x5, x6, x7 \) = Number of long-range contacts in the 1\(^{\text{st}}\), 2\(^{\text{nd}}\), 3\(^{\text{rd}}\), 4\(^{\text{th}}\), 5\(^{\text{th}}\), 6\(^{\text{th}}\) and 7\(^{\text{th}}\) bins.

In the stringent jack-knife prediction, experimental folding rates of 13 proteins were predicted with a deviation less than 1 when compared with experimental folding rates. The average deviation between experimental and
predicted folding rates for the 35 three-state proteins through jack-knife method is 2.12.

Figure 3.1 represents the scatter plot between experimental and back-check, jack-knife predicted folding rates for the 35 three-state proteins.

3.4.3 Relationship between long-range contacts from various sequence separation bins and folding rates for 45 two-state proteins

In Table 3.3, correlation obtained between experimental folding rates and long-range contacts for various combinations of bins ranging from 4-10, 11-20, 21-30, 31-40, 41-50, 51-60 and >60 for the 45 two-state proteins is presented. While computing correlation between folding rates and individual sequence separation bins for the full set of 45 two-state proteins, the first four bins showed a considerable correlation of -0.51, -0.75, -0.73 and -0.49 respectively, whereas the remaining bins showed a poor correlation. Also, including long-range contacts from these four bins in the regression equation showed a correlation of 0.84 with experimental folding rates. From this result, it is clear that most of the rate-limiting long-range contacts necessary for predicting the folding rates are present in these bins itself. Hence long-range contacts from the first 4 bins were used in developing the regression equation. Some interesting results were observed in the case of slow folding two-state proteins. It was observed that proteins with very low folding rates (slow folding proteins) seems to have considerable number of long-range contacts through out all the sequence separation bins and especially in bins from higher sequence separation distances (31-40; 41-50; 51-60; >60). Inversely, in proteins with very high folding rates (fast folding proteins), no long-range contacts are found in the sequence separation bins of 31-40; 41-50; 51-60; >60. These results clearly
showed that rate-limiting long-range contacts of proteins with lower folding rates are present at distant sequence separation bins. In other words, as there are more number of long-range contacts from distant sequence separation bins (higher sequence separation distance), the time taken by the proteins to fold to their native state increases. Inversely, proteins that fold with higher folding rate have no long-range contacts are at distant sequence separation bins.

3.4.4 Folding rate prediction of two-state proteins based on long-range contacts from various sequence separation bins

In Table 3.4, experimental folding rates, along with the computed long-range contacts from various sequence separation bins, results of back-check, jack-knife predicted folding rates and their deviations from experimental folding rates for the 45 two-state proteins were given. The 45 two-state proteins based on their folding rate values have been arranged in ascending order, in which the folding rates ranges from -1.47 to 12.9 respectively. For the full set of 45 two-state proteins, the back-check method predicted the folding rates of 18 proteins with a deviation less than 1 from experimental folding rates. The regression equation for the 35 three-state proteins developed using long-range contacts from the first four bins is

\[
\ln(k_i) = (-0.01342) \times (x_1) + (-0.05217) \times (x_2) + (-0.07355) \times (x_3) + (0.01351) \times (x_4) + 11.7312.
\]

\[x_1, x_2, x_3, x_4 = \text{Number of long-range contacts in the } 1^{st}, 2^{nd}, 3^{rd} \text{ and } 4^{th} \text{ bins.}\]

From the 45 two-state proteins, more than 50% of proteins have no or very low number of long-range contacts in the last three bins. Hence long-range
contacts only from the first four bins are included in the regression equations. In the stringent jack-knife prediction, experimental folding rates of 17 two-state proteins were predicted with a deviation less than 1 in comparison with experimental folding rates. The average deviation for the 45 two-state proteins through back-check and jack-knife method is 2.31 and 2.63 respectively.

Figure 3.2 represents the scatter plot between experimental and back-check, jack-knife predicted folding rates for the 45 two-state proteins.

3.5 Implication of long-range contacts in protein folding rates

A number of experimental studies have been carried out to understand the fundamental process of contact formation between residues that are far away in sequence. Hagen et al. (1996) carried out the first experimental determination for the time scale of contact formation through nanosecond resolved spectroscopy and showed that the regions of unfolded Cytochrome C separated by ~50 residues diffuse together in 35-40μs and was suggested that the approximate upper limit on the rate of collapse of a random coil to a compact structure should be $\sim 10^6 \text{ S}^{-1}$. Interactions between side chains on opposite strands of anti-parallel $\beta$-sheets are required for their stability and initiation of a parallel $\beta$-sheet requires contact between residues distant in sequence (Gellman, 1998). Contact formation rates over longer distances decrease with increasing chain length, indicating different rate-limiting steps for motions over short- and long-chain segments (Krieger et al., 2003).

Apart from experimental observations, importance of long-range interactions is also observed in theoretical and simulation works. Mirny et al. (1998) developed a lattice model of simulation with 48 residues on a cubic
lattice that showed acceleration of folding was accompanied by strengthening of interactions observed in the folding nucleus residues. Dokholyan et al. (2002) constructed contact network topology for two proteins, chymotrypsin inhibitor 2 (CI2) and C-Src SH3 domain using average graph connectivity and revealed the importance of long-range contacts in their structures. Fuxreiter and Simon (2002) identified residue clusters stabilized by cooperative long-range interactions which serve as anchoring points for arranging secondary structural elements and play an important role in preventing decay of 3D structures of proteins. Mukherjee and Bagchi (2003) studied the folding of a model protein by simulation techniques and observed that slow late stage of folding is due to long-range contact formation. Papoian et al. (2004) concluded that long-range water-mediated potentials guide the folding of proteins and further showed that long-range pairing of hydrophilic groups were found to serve as an integral part of protein architecture.

3.6 Conclusion

In the present work, we have elucidated how long-range contacts play a crucial role in the folding mechanism of three-state and two-state proteins belonging to all-α, all-β and mixed-structural classes. Our present method makes use of long-range contacts observed in the 3D structure of proteins for describing the folding mechanism of proteins that fold through three-state and two-state kinetics. Without including any other information such as secondary structure content, amino acid composition and chain length our method predict the folding rates of both three-state and two-state proteins by using only long-range contacts observed in the 3D structures. Folding rates of 35 three-state and 45 two-state proteins with known 3D structures have been easily predicted with reasonable accuracy. This study reveals the importance of long-range
contacts from various sequence separation bins in predicting the folding rates of three-state and two-state proteins. Results observed from this present work strongly evidence that long-range contacts present in the 3D structure at various sequence separation bins were found to be an important descriptor in predicting the folding rates of both two-state and three-state proteins.