Open Access

Intrinsic Relationship of Amino Acid Composition/Occurrence with Topological Parameters and Protein Folding Rates

M. Michael Gromiha*

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Abstract: Understanding the relationship between amino acid sequences and folding rates of proteins is an important task in computational and molecular biology. It has been shown that topological parameters, contact order, long-range order and total contact distance relate well with protein folding rates. In this work, we have systematically analyzed the relationship between amino acid composition/occurrence and protein folding rates along with topological parameters derived from protein three-dimensional structures. We found that the classification of proteins based on their structural classes and folding types (two and three-state proteins) could explain the relationship very well. The amino acid composition showed good correlation with protein folding rates for two-state proteins whereas the correlation is high with amino acid occurrence for three-state proteins. The composition of polar amino acids, Asn, Gln and Ser directly correlated with protein folding rates and a reverse trend was observed between the occurrence of hydrophobic amino acids, Ile and Gly and protein folding rates. The amino acid occurrence showed a positive correlation with folding rates in two-state proteins and a negative correlation in three-state proteins, which reveals that the presence of more number of amino acids in three-state proteins slows down the folding process. The analysis on slow and fast folding proteins showed that the slow folding proteins have appreciable number of residues that form multiple contacts with other residues. Further, we have combined different amino acids based on their chemical properties and analyzed the relationship with protein folding rates, and set up multiple regression equations for predicting protein folding rates.

INTRODUCTION

Folding rate is a measure of slow/fast folding of a protein from its unfolded state to the native three-dimensional structure. Studies on protein folding rates enhance our understanding on the variations in protein folding kinetics, which may lead to several pathologies such as prion and Alzheimer diseases. Fulton *et al.* [1] collected the experimental data on protein folding rates and developed a protein folding database for understanding protein folding and stability.

As an advancement to understand/predict protein folding rates, Plaxco *et al.* [2] proposed the concept of contact order (CO) using the information about the average sequence separation of all contacting residues in the native state of two-state proteins and found a significant correlation between CO and folding rates of two-state proteins. Gromiha and Selvaraj [3] defined a novel parameter, long-range order (LRO) from the knowledge of long-range contacts (contact between two residues that are close in space and far in the sequence) in protein structure and established a simple statistical model for predicting the protein folding rates. Recently it has been reported that LRO is the only parameter that shows excellent correlation with the folding rates of all structural classes of proteins [4]. These two parameters, CO and LRO are incorporated into a new parameter, total contact distance (TCD), which shows a good relationship with protein folding rates [5]. All these parameters are derived from the knowledge of inter-residue interactions in protein three-dimensional structures [6].

In the past several years, investigations have been carried out to understand/predict the folding rates of proteins from protein three-dimensional structures, secondary structure information and amino acid sequences, and the details are extensively reviewed [7]. The prediction of protein folding rates from amino acid sequence includes the relationship between protein folding rates and amino acid properties [8-11], predicted secondary structures [12], predicted contacts between amino acid residues [13], amino acid composition [14,15] and secondary structure length [16].

The main aim of the present study is to explore the relationship between amino acid composition/occurrence in different structural classes of proteins as well as in different folding types of proteins. We found that the occurrence of polar residues Asn, Gln and Ser have direct correlation with protein folding rates whereas an opposite trend was observed for the hydrophobic residues. In all-α proteins the polar residues showed the highest positive correlation with folding rates whereas in all-β proteins the hydrophobic residues the highest negative correlation with folding rates. The two and three-state proteins showed different trends that the folding rates of former ones are positively correlated whereas the latter ones are negatively correlated with amino acid occurrence. Further, the relationship between amino acid composition/occurrence with structure based parameters, CO, LRO and TCD will be discussed. The multiple contacts established by amino acid residues in protein structures are

*Address correspondence to this author at the Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan; Tel: +81-3-3599-8046; Fax: +81-3-3599-8081; E-mail: michael-gromiha@aist.go.jp

1874-1991/09 2009 Bentham Open

found to be important for understanding protein folding rates.

MATERIALS AND METHODS

Experimental Folding Rates

The experimental folding rates of 75 two and three-state proteins used in related works [10,12] form the basis for the present study. The Protein Data Bank codes [17] and experimental ln(k_f) values are available in our earlier article [10,15]. The structural classification of these proteins yielded 16 all-α (dominated by α-helices; $\alpha > 40\%$ and $\beta < 5\%$), 26 all-β (dominated by β-strands; $\beta > 40\%$ and $\alpha < 5\%$) and 33 mixed class proteins (contain both α-helices and β-strands; $\alpha > 15\%$ and $\beta > 10\%$). The dataset has 50 two-state and 25 three-state proteins.

Computation of Amino Acid Composition and Occurrence

The amino acid composition for a protein has been computed using the number of amino acids of each type and the total number of residues. It is defined as:

$$\mathrm{Comp(i)} = \Sigma\, \mathrm{n_i/N} \qquad (1)$$

where i stands for the 20 amino acid residues. n_i is the number of residues of each type and N is the total number of residues. The summation is through all the residues in the considered protein.

The amino acid occurrence is the actual number of amino acid residues of each type present in a protein without normalizing with chain length.

Computation of Contact Order, Long-Range Order and Total Contact Distance

The parameter, CO reflects the relative importance of local and non-local contacts to the native structure of a protein [2]. It is defined as: $\mathrm{CO} = \Sigma\Delta S_{ij} /L.N$, where, ΔS_{ij}, is the sequence separation between contacting residues i and j, L is the total number of residues in the protein and N is the total number of contacts. In this definition, two residues are considered to be in contact with each other if the distance between any non-hydrogen atoms in these residues is within the distance of 6Å.

Gromiha and Selvaraj [3] defined a parameter, LRO for a protein from the knowledge of long-range contacts (contacts between two residues that are close in space and far in the sequence) in protein structure [3]. It is defined as, $\mathrm{LRO} = \Sigma n_{ij}/N$; $n_{ij}=1$ if $|i-j| > 12$; 0 otherwise, where i and j are two contacting residues in space within a distance of 8Å and N is the total number of residues in a protein.

Total contact distance is defined as:

$$\mathrm{TCD} = 1/\mathrm{nr}^2 \sum_{k=1}^{\mathrm{nc}} |\,i - j\,|,$$

where, i and j are contacting residues, nr and nc are, respectively, number of residues and number of contacts in a protein [4]. The summation is done for any cutoff residue separation (l_{cut}) and if $|i\text{-}j| > l_{cut}$. TCD is related to CO and LRO by a simple multiplication (TCD = CO × LRO) if LRO is calculated with the same l_{cut} value as CO.

Estimation of Medium and Long-Range Contacts

Each residue in a protein molecule is represented by its α-carbon atom. The center is fixed at the α-carbon atom of the first (N-terminal) residue and the distances between this atom and the rest of the α carbon atoms in the protein molecule are computed. The composition of the surrounding residues associated with this residue is calculated for a sphere of radius 8Å. It has been shown that the influence of each residue over the surrounding medium extends effectively only up to 8Å [18] and this limit is sufficient to characterize the hydrophobic behavior of amino acid residues [19] and to accommodate both the local and non-local interactions [6]. Further, 8Å limit has been used in several studies, such as, to understand the folding rate of proteins [3,20], protein stability upon mutations [21], thermal stability of proteins [22], transition state structures of two-state protein mutants [23] and to understand the relationship between hydrophobic clusters and long-range contacts in proteins [24].

For a given residue, the composition of surrounding residues is analyzed in terms of their location at the sequence level. The residues that are within a distance of two residues from the central residue are considered to contribute to short-range interactions, those within a distance of ±3 or ±4 residues to medium range and those more than four residues away to long-range interactions [6].

Preference of Residue Pairs Influenced by Medium and Long-Range Contacts

For each medium and long-range interaction, we have computed the average preference of surrounding residues for all the 20 amino acid residues. It is defined as $\langle N\rangle_{ij} = \Sigma N_{ij}/(\Sigma N_i + \Sigma N_j)$, where N_{ij} is the number of surrounding residues (contacts) of type j around residue i (400 combinations), and the summation is over all the residues in the considered proteins. ΣN_i and ΣN_j are respectively the total number of residues of type i and j [25]. We have derived sets of 20x20 matrices for medium and long-range interactions for all structural classes and folding types of proteins.

Classification Based on Amino Acid Properties

We have classified the amino acids into 18 groups based on the procedure used in Caurgo [26] and the details are summarized below: the first seven groups were proposed by Chakrabarti and Pal [27] for describing the conformational similarity between the 20 amino acids, by monitoring the ψ, φ, and $\chi 1$ torsions: [CMQLEKRA], [P], [ND], [G], [HWFY], [S], and [TIV]. The next eight groups are proposed by Murphy *et al.* [28] on the basis of the possibility to identify foldable structures by simplifying the BLOSUM50 matrix: [P], [KR], [EDNQ], [ST], [AG], [H], [CILMV], and [YWF]. Other five groups are proposed by Rose *et al.* [29] based on hydrophobicity: [CFILMVW], [AG], [PH], [EDRK], [NQSTY]. These classifications yielded 20 types of residue groups and some of them {[P] and [AG]} are present in more than one groups and are considered only once. Hence, we

Table 1. Representative Amino Acids that Show High Correlation with Folding Rate, CO, LRO and TCD Based on Composition

Parameter	Overall (75)	all-α			all-β			Mixed		All		
		2 (12)	3 (4)	All (16)	2 (17)	3 (9)	All (26)	2 (21)	3 (12)	All (33)	2-all (50)	3-all (25)
ln(k_f)	0.33 (Gln)	0.67 (Leu)	0.93 (Asn)	0.72 (Asn)	0.69 (Gln)	0.55 (Ser)	-0.53 (Leu)	0.50 (Leu)	0.68 (Arg)	-0.39 (Gly)	0.39 (Asn)	-0.32 (Ala)
		0.64 (Asn)			-0.65 (Leu)				-0.56 (Gly)			
		0.60 (Gln)							-0.53 (Ala)			
CO	-0.47 (Ala)	0.58 (Trp)	-0.97 (Gly)	-0.62 (Ala)	-0.70 (Met)	-0.87 (Met)	-0.63 (Met)	0.53 (Val)	0.64 (Cys)	0.54 (Val)	-0.43 (Ala)	0.63 (Val)
	0.45 (Val)	-0.52 (Ala)	-0.95 (Ala)						0.60 (Gln)			
			0.94 (Ile)						-0.58 (Ala)			
LRO	0.39 (Val)	-0.78 (Leu)	-0.99 (Ile)	-0.67 (Leu)	0.50 (Pro)	0.35 (Phe)	-0.29 (Gln)	0.70 (Cys)	-0.69 (Tyr)	0.43 (Cys)	0.44 (Val)	0.42 (Gly)
	-0.31 (Leu)	-0.68 (Gln)	0.98 (Ala)	0.61 (Ala)				-0.57 (Asn)	0.65 (Gly)			
			0.91 (Gly)	-0.61 (Gln)								
			-0.91 (Phe)									
TCD	0.43 (Val)	0.54 (His)	0.97 (Ile)	0.60 (Lys)	-0.63 (Met)	-0.79 (Met)	-0.55 (Met)	0.57 (Val)	0.67 (Cys)	0.60 (Val)	0.44 (Val)	0.59 (Val)
		-0.50 (Leu)	-0.95 (Gly)			0.63 (Ala)			0.60 (Val)			

2: two-state; 3: three-state
Number of proteins in each group is given in parenthesis

used the total of 18 variables and are listed in Table **1**. The distribution of correlation coefficients computed for each pair of variables showed that in majority of the cases, there is no correlation between the pairs [26].

Single and Multiple Correlations

The single correlation between the amino acid occurrence/composition and topological parameters/folding rates has been calculated using the familiar expression:

$$r = [\mathrm{N}\ \Sigma\mathrm{XY} - (\Sigma\mathrm{X}\ \Sigma\mathrm{Y})]/\{[\mathrm{N}\ \Sigma\mathrm{X}^2 - (\Sigma\mathrm{X})^2][\mathrm{N}\ \Sigma\mathrm{Y}^2 - (\Sigma\mathrm{Y})^2]\}^{1/2} \qquad (2)$$

where, *r* is the correlation coefficient, N, X, and Y are the number of data, pair of variables (E.g. amino acid occurrence and folding rates) respectively.

We have combined the amino acid compositions/occurrences using multiple regression technique: multiple correlation coefficients and regression equations were determined using standard procedures [30].

RESULTS AND DISCUSSION

Relationship between Amino Acid Composition and Protein Folding Rates in Different Structural Classes and Folding Types

We have computed the correlation between amino acid composition and protein folding rates in a set of 50 two and 25 three-state proteins. We observed that the amino acid residues showed different trends of positive and negative correlation with protein folding rates. The highest correlation obtained for each group of proteins is presented in Table **1**. We observed that Asn and Ala are the most correlated residues with the folding rates of two-and three-state proteins, respectively, which agrees with the results of Ma *et al.* [14]. In all-α proteins, Asn shows the strongest correlation of 0.72 whereas Leu shows the highest negative correlation of -0.53 with the folding rates of all-β proteins. The classification did not improve the correlation in mixed class proteins. This result reveals that the distinct folding patterns of all-α and all-β classes of proteins enhanced the correlation between amino acid composition and protein folding rates [6]. Further, it is noteworthy that a polar residue has the highest positive correlation in all-α class proteins whereas a hydrophobic residue has the highest negative correlation in all-β class proteins. The correlation has been improved further by classifying the proteins based on both structural classes and folding types. However, the number of data are not significant and hence the results are not analyzed in detail.

Relationship between Amino Acid Occurrence and Protein Folding Rates in Different Structural Classes and Folding Types

The influence of chain length has been extensively analyzed by considering the relationship between amino acid occurrence and protein folding rates. The results are presented in Table **2**. We found a striking correlation between amino acid occurrence and folding rates of three-state proteins whereas there is no significant improvement in two-state proteins. The highest correlation between amino acid occurrence and folding rates of three-state proteins is -0.69 whereas amino acid composition showed the highest correlation of -0.32. This results agrees with the observation that chain-length is a major factor to determine the folding rates of three-state proteins [31]. In addition, most of the amino acid residues show a positive correlation with protein folding rates whereas an opposite trend was observed in three-state proteins. Especially high correlation was observed for the hydrophobic residues, Gly, Ile, Ala and Met. This result suggests that the two state proteins are small proteins and the amino acid residues tend to interact with each other so quickly to attain their native three-dimensional structures. The three-state proteins have long chains and the residues take time to make the necessary long-range contacts and attain the stable structure. It has been reported that the small proteins generally follow a two-state folding mechanism where as the three-state proteins are mainly have the long amino acid sequence: the limit of chain length is estimated to be about 110 residues [32].

Considering different structural classes, amino acid occurrence shows a strong correlation in mixed class proteins. The correlation rose from -0.39 to -0.58 due to the removal of normalization with chain length. In three-state mixed class proteins, several residues have the absolute correlation of more than 0.80. It is worth to mention that the correlation is less compared with amino acid composition in all-α and all-β structural classes of proteins (Table **2**).

Influence of Amino Acid Occurrence and Composition in Determining Contact Order

The results obtained for the correlation between amino acid composition and contact order are included in Table **1**. The amino acid occurrence (Table **2**) showed a negative correlation with contact order and the performance with amino acid occurrence is better than that obtained with composition. There is a significant improvement in the correlation between amino acid occurrence and contact order in all the three structural classes of proteins as well as in the whole dataset. As the contact order includes the contact information about the nearby and far apart residues the occurrence is able to explain the contact order of proteins. In two-state proteins occurrence performs better than composition whereas there is no difference of correlation in three-state proteins. This result suggests that the number of amino acid residues present in small proteins is able to explain about their contact order. The correlation between amino acid occurrence and contact order lies in the range of -

Table 2. Representative Amino Acids that Show High Correlation with Folding Rate, CO, LRO and TCD Based on Occurrence

Parameter	Overall (75)	all-α			all-β			Mixed		All		
		2 (12)	3 (4)	All (16)	2 (17)	3 (9)	All (26)	2 (21)	3 (12)	All (33)	2-all (50)	3-all (25)
ln(k_f)	-0.28 (Asp)	-0.63 (Lys)	-0.82 (His)	-0.45 (Trp)	0.68 (Gln)	0.47 (Ser)	0.35 (Gln)	0.49 (Leu)	-0.85 (Gly)	-0.58 (Met)	0.34 (Met)	-0.69 (Ile)
		-0.60 (Ser)	-0.80 (Leu)		0.54 (Phe)		-0.34 (Leu)		-0.85 (Ile)	-0.50 (Ala)		-0.68 (Gly)
									-0.85 (Asp)			
CO	-0.68 (Leu)	-0.85 (Leu)	-0.99 (Ala)	-0.83 (Leu)	-0.82 (Gln)	-0.82 (Met)	-0.78 (Met)	-0.83 (Lys)	-0.68 (Lys)	-0.75 (Asn)	-0.69 (Lys)	-0.62 (Ala)
	-0.64 (Ala)	-0.85 (Gln)	-0.97 (Gly)	-0.81 (Pro)	-0.76 (Met)		-0.70 (Gln)	-0.82 (Ala)	-0.65 (Asn)	-0.74 (Leu)	-0.65 (Leu)	-0.62 (Leu)
			-0.95 (Pro)									
LRO	0.49 (Val)	0.95 (Lys)	0.98 (Arg)	0.83 (Gly)	0.51 (Thr)	0.86 (Val)	0.60 (Glu)	0.54 (Val)	0.66 (Val)	0.55 (Val)	0.41 (Val)	0.72 (Val)
		0.85 (Asp)	0.96 (Ala)	0.82 (Ala)		0.84 (Gly)	0.60 (Val)	0.53 (Cys)				
			0.95 (Pro)	0.81 (Asp)		0.80 (Ile)	0.59 (Ile)					
TCD	-0.50 (Leu)	-0.64 (Gln)	-0.99 (Pro)	-0.62 (Arg)	-0.54 (Gln)	-0.74 (Met)	-0.60 (Met)	-0.65 (Lys)	-0.52 (Tyr)	-0.59 (Asn)	-0.50 (Lys)	-0.40 (Met)
			-0.96 (Arg)			-0.71 (Arg)	-0.57 (Trp)	-0.64 (Ala)		-0.55 (Leu)		
			-0.92 (Ala)			-0.71 (Trp)						

2: two-state; 3: three-state
Number of proteins in each group is given in parenthesis

0.68 to -0.99 in different groups of proteins based on structural classes and folding types.

Amino Acid Composition/Occurrence and Long-Range Order

Long-range order is reported to be a parameter that explains well the folding rates in all structural classes of proteins [4]. The amino acid occurrence improved the correlation in all-α class proteins whereas there is no appreciable change in the correlation of other two structural classes as well as the classification based on folding types. On the other hand, amino acid occurrence improved the correlation in all structural classes of proteins and the folding rates of three-state proteins. It is noteworthy that in contrast to contact order, amino acid occurrence shows a positive correlation with long-range order. It shows that the presence of more number of amino acids increases the long-range order, especially in three-state proteins. The formation of disulfide bonds from distant residues and hydrophobic contacts between the residues that are far away in the sequence increased the long-range order. On the other hand, there is no significant correlation between amino acid occurrence and long-range order in two-state proteins. Furthermore we observed a significant improvement in the correlation between amino acid occurrence and long-range order in all-β class proteins. The absolute correlation increased from 0.29 to 0.60. This might be due to the folding behavior of such class of proteins.

Amino Acid Composition/Occurrence and Total Contact Distance

The amino acid composition shows a moderate correlation (less than 0.60) with total contact distance in all structural classes of proteins and different folding types. Similar trend is also observed with the amino acid occurrence. This might be due to the fact that total contact distance is the combination of contact order and long-range order, and these two parameters shows inverse relationships with amino acid occurrence.

Relationship between Groups of Amino Acids and Protein Folding Rates

We have analyzed the relationship between several groups of amino acid residues classified by their properties and folding rates of two and three-state proteins.

We observed that most of the groups of amino acids did not improve the correlation, which suggests that the folding behavior of amino acids belonging to the same group (for example, hydrophobic or polar or charged) could be explained with a single amino acid residue. The group of amino acids [T, I, V] based on the conformational preference of amino acids improved the correlation from 0.39 to 0.51 in two-state proteins. This property also increased the correlation from 0.44 to 0.56 with LRO of two-state proteins. Another property that increased the correlation in three-state proteins and LRO of all-α proteins is based on hydrophobicity/simplified BLOSSUM matrix, [A, G]. In all-α proteins, it improved the correlation from 0.67 to 0.72.

The amino acid occurrence of specific groups of amino acids marginally improved the correlation in several categories of proteins based on structural classes and folding types. The main groups involved in the enhancement are based on hydrophobicity/conformational preferences, [T, I, V], [C, M, Q, L, E, K, R, A] and [K, R].

Table 3. Difference in Preference of Residue Pairs between Two- and Three-State Proteins Influenced with Long-Range Contacts

	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	3.9	-0.2	-0.4	0.8	2.5	-2.0	-0.2	-0.4	3.3	-3.5	-1.2	1.1	-1.5	1.7	-2.8	-0.9	0.3	0.2	0.0	-0.8
Asp	-0.6	-3.0	-0.6	0.7	2.7	-0.4	-0.5	1.6	2.9	0.7	-1.2	-1.0	-1.6	2.7	-3.6	-0.2	-0.1	1.1	0.5	-0.1
Cys	-0.1	-0.8	3.5	-3.3	-1.9	1.9	-0.8	-3.2	4.5	2.0	3.0	-1.3	2.0	-1.7	-1.3	3.9	**-6.2**	1.8	-0.3	-1.5
Glu	0.7	0.4	-1.5	-0.9	3.0	2.0	-0.8	-1.6	-0.1	-2.8	0.8	-0.6	-0.1	1.6	-1.8	0.5	0.0	1.2	0.1	-0.2
Phe	1.7	1.0	-1.6	1.9	2.4	0.5	0.1	-1.5	0.1	-3.7	1.0	-0.5	-0.5	0.5	-0.2	-0.3	-0.4	-0.3	0.6	-0.5
Gly	-1.7	0.2	-0.3	2.2	2.2	-5.1	0.7	-0.9	3.4	-3.2	-2.5	1.2	-1.9	1.4	-0.3	0.3	-0.1	1.6	1.0	1.7
His	-1.2	-1.3	-1.2	-2.7	1.5	1.6	1.0	0.5	1.1	-1.4	-0.6	-0.3	1.9	1.7	-0.5	-2.9	-0.2	0.0	-1.4	4.1
Ile	-0.7	0.8	-0.9	-1.0	0.8	-1.4	0.2	-3.4	2.3	-1.2	-1.5	0.5	0.2	1.9	0.5	0.9	0.6	2.3	0.8	-1.4
Lys	2.7	0.4	0.0	-2.3	0.7	1.6	0.1	1.0	1.2	-3.5	-0.1	-0.2	-1.9	1.4	-1.5	1.1	0.5	0.3	-1.2	-0.2
Leu	-1.6	0.8	-0.3	-0.5	1.1	-1.9	0.0	0.7	0.4	-2.4	-0.6	0.7	-0.5	2.9	-1.7	1.9	0.4	2.0	-1.1	-0.1
Met	-2.1	-1.0	1.3	5.5	6.0	**-7.6**	0.0	-2.6	2.6	-1.5	-0.9	0.5	-0.2	0.1	-0.6	2.6	-2.1	-0.4	-0.4	0.9
Asn	1.9	-1.7	-1.3	-1.5	0.4	1.2	-0.2	0.9	0.5	-0.2	-0.6	-2.3	0.7	0.5	-2.4	-0.1	1.2	2.0	1.4	-0.3
Pro	-2.9	-1.3	-0.2	0.5	1.5	-4.0	1.3	1.1	-1.5	-2.2	-0.8	1.1	1.5	1.7	-3.2	0.1	3.3	1.4	3.0	-0.2
Gln	1.9	1.8	-1.6	-0.4	-0.6	-1.2	0.1	0.5	1.2	1.9	-1.4	-1.5	-1.0	-1.6	-0.6	-0.2	-3.2	6.2	-0.1	-0.6
Arg	-4.7	-2.2	-0.7	-0.3	2.0	0.8	0.2	2.8	-0.2	-3.3	-0.6	-0.7	-2.2	2.4	2.1	1.3	-1.9	3.9	0.0	1.2
Ser	-3.2	-1.0	0.0	-0.1	0.7	-1.4	-1.3	0.5	2.1	0.9	-0.1	-0.3	-0.8	1.4	-0.6	-0.8	1.8	0.6	1.1	0.3
Thr	0.0	-0.2	-1.8	0.0	1.4	-1.2	-0.1	0.5	2.0	-1.2	-1.2	0.8	1.4	0.6	-2.3	1.9	1.2	0.2	-1.6	0.1
Val	-1.6	-0.1	-0.5	0.0	0.9	-0.4	-0.2	0.3	0.7	-2.1	-1.1	0.4	-0.3	3.1	-0.1	0.3	-0.6	1.1	-0.4	0.4
Trp	-0.5	0.9	-1.2	0.1	3.2	3.9	-1.7	2.9	-3.5	**-8.8**	-1.2	2.8	5.5	1.1	-1.4	3.3	**-6.7**	-2.0	-0.9	4.0
Tyr	-2.9	-0.3	-1.1	-0.6	0.8	2.6	1.9	-3.7	0.8	-3.0	-0.3	-0.2	-0.9	0.9	-0.3	0.6	-0.2	1.8	1.8	2.6

Residue Contacts in Two and Three-State Proteins

We have analyzed the average medium and long-range contacts for each amino acids and the preference of residues to form the contacts in two and three-state proteins. We observed that generally there is no appreciable difference and the three-state proteins have minor increase in total number of contacts. The residue-wise analysis shows that Cys, Gln and Val prefer to have more number of long-range contacts in two-state proteins whereas the three-state proteins have the prefernce for Met and Ser.

The analysis on residue pair preference shows that the polar or nonpolar residues have similar preference to form medium-range contacts. Specifically, the pair of residues DK, QE, YA and CC are dominant in two-state proteins and CN, CI and HA have high preference in three-state proteins. In long-range contacts, hydrophobic residue pairs have higher preference than pairs of polar residues. The residue pairs MF, QV and WP are dominant in two-state proteins whereas three-state proteins prefer the contacting pairs of CT, MG, WL and WT (Table **3**). While Trp has high preference to have long-range contacts with other amino acids there is no such preference for the other aromatic residues, Phe and Tyr. This analysis reveals that there is no specific preference between two-state and three-state proteins in terms of the chemical properties of amino acid residues.

Influence of Residues with Multiple Contacts in Folding Process

We have analyzed the number of residues that have multiple contacts in different structural classes, folding types, and slow and fast folding proteins. The results obtained with the cutoff of more than five and 10 contacts are presented in Table **4**. We found that the number of residues with more than five contacts is less in all-α proteins compared with other structural classes of proteins. On the other hand, the difference is not significant between the two-state and three-state proteins. Interestingly, only 13% of residues have more than five contacts in fast folding proteins whereas 23% of residues have such contacts in slow folding proteins. In order to understand the importance of multiple contacts we have selected the proteins with logarithmic folding rates of less than one and more than 9 and repeated the analysis. Remarkably, 25% of residues form multiple contacts in slow folding proteins (ln(kf) < 1.0) whereas only 6% of residues have multiple contacts in fast folding proteins (ln(kf) > 9.0). This result reveals that the multiple contacts established between amino acid residues play an important role to determine the folding rates.

Multiple Regression Analysis

We have combined the compositions/occurrences of different amino acid residues to relate with protein folding rates as well as with the structural parameters. The results obtained with multiple regression technique are presented in Table **5**. We noticed that the classification based on structural classes remarkably improved the correlation. Considering all proteins together, amino acid composition/occurrence correlation well with CO compared to other structural parameters. As the main objective is not prediction we did not carry out the validation procedure extensively.

Table 4. Number and Percentage of Residues that have Multiple Contacts in Different Structural Classes and Folding Types

Classification	N_p	N_{res}	$N_{cont} > 5$	%cont	$N_{cont} > 10$	%cont
All-α	16	1532	147	9.6	6	0.4
All-β	26	2345	442	18.8	12	0.5
Mixed	33	3814	769	20.2	51	1.3
2-state	50	4190	715	17.1	30	0.7
3-state	25	3501	629	18.0	39	1.1
Fast	47	3925	494	12.6	16	0.4
Slow	28	3766	850	22.6	53	1.4
Fastest (ln(kf) > 9.0)	8	596	36	6.0	2	0.3
Slowest [ln(kf) < 0.0]	7	1329	333	25.1	23	1.7

N_p: Number of proteins; N_{res}: Number of residues; N_{cont}: Number of contacts

Table 5. Multiple Correlation Coefficients Obtained with the Combination of Amino Acid Composition/Occurrence*

Dataset	Composition				Occurrence			
	ln(kf)	CO	LRO	TCD	ln(kf)	CO	LRO	TCD
All	0.53 (0.58)	0.70 (0.74)	0.55 (0.60)	0.66 (0.69)	0.59 (0.60)	0.82 (0.83)	0.69 (0.69)	0.72 (0.72)
All-a	0.94 (0.99)	0.93 (0.99)	0.99 (0.99)	0.95 (0.99)	0.95 (0.99)	0.98 (0.99)	0.99 (0.99)	0.97 (0.97)
All-b	0.88 (0.92)	0.84 (0.89)	0.60 (0.90)	0.84 (0.93)	0.86 (0.94)	0.92 (0.94)	0.86 (0.92)	0.86 (0.91)
Mixed	0.78 (0.88)	0.87 (0.93)	0.87 (0.93)	0.89 (0.94)	0.89 (0.94)	0.93 (0.94)	0.90 (0.94)	0.90 (0.95)
Two-state	0.74 (0.83)	0.72 (0.78)	0.77 (0.81)	0.70 (0.77)	0.72 (0.86)	0.85 (0.88)	0.74 (0.80)	0.76 (0.80)
Three-state	0.78 (0.95)	0.94 (0.98)	0.66 (0.83)	0.89 (0.97)	0.86 (0.99)	0.93 (0.99)	0.90 (0.95)	0.87 (0.96)

*correlation coefficients were obtained with a combination of selected seven amino acid residues; the correlation coefficients obtained with all the 20 amino acid residues are given in parentheses.

CONCLUSIONS

We have systematically analyzed the relationship between amino acid occurrence/composition with protein folding rates as well as with structural parameters in different structural classes, folding types, and slow and fast folding proteins. We observed that the polar residues have positive correlation with protein folding rates whereas an opposite trend was observed for hydrophobic residues. The amino acid composition/occurrence relates well with contact order in different structural classes and folding types, which reveals that the local contacts are well accounted in terms of amino acid composition/occurrence. Further, the residues with multiple contacts are found to be important for understanding protein folding rates.

REFERENCES

[1] Fulton KF, Bate MA, Faux NG, Mahmood K, Betts C, Buckle AM. Protein Folding Database (PFD 2.0): an online environment for the International Foldeomics Consortium. Nucleic Acids Res 2007; 35: D304-7.

[2] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998; 277: 985-94.

[3] Gromiha MM, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol 2001; 310: 27-32.

[4] Istomin AY, Jacobs DJ, Livesay DR. On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. Protein Sci 2007; 16:2564-9.

[5] Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J 2002; 82: 458-63.

[6] Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004; 86: 235-77.

[7] Gromiha MM, Selvaraj S. Bioinformatics approaches for understanding and predicting protein folding rates. Curr Bioinformatics 2008; 3: 1-8.

[8] Gromiha MM. Importance of native state topology for determining the folding rate of two-state proteins. J Chem Inf Comp Sci 2003; 43:1481-85.

[9] Gromiha MM. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model 2005; 45: 494-501.

[10] Gromiha MM, Thangakani AM, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res 2006; 34: W70-W74.

[11] Huang JT, Tian J. Amino acid sequence predicts folding rate for middle-size two-state proteins. Proteins 2006; 63: 551-4.

[12] Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 2004; 101: 8942-44.

[13] Punta M, Rost B. Protein folding rates estimated from contact predictions. J Mol Biol 2005; 348: 507-12.

[14] Ma BG, Guo JX, Zhang HY. Direct correlation between proteins' folding rates and their amino acid compositions: an *ab initio* folding rate prediction. Proteins 2006; 65: 362-72

[15] Huang L-T, Gromiha MM. Prediction of protein folding rates using quadratic response surface models. J Comp Chem 2008; 29: 1675-83.

[16] Huang JT, Cheng JP, Chen H. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins 2007; 67(1): 12-17.

[17] Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 2007; 35: D301-3.

[18] Manavalan P, Ponnuswamy PK. A study of the preferred environment of amino acid residues in globular proteins. Arch Biochem Biophys 1977; 184: 476-87.

[19] Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular protein. Nature 1978; 275: 673-74.

[20] Debe DA, Goddard WA. First principles prediction of protein folding rates. J Mol Biol 1999; 294: 619-25.

[21] Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. Protein Eng 1999; 12: 549-55.

[22] Gromiha MM. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. Biophys Chem 2001; 91: 71-77.

[23] Gromiha MM, Selvaraj S. Important amino acid properties for determining the transition state structures of two-state protein mutants. FEBS Lett 2002; 526: 129-34.

[24] Selvaraj S, Gromiha MM. Importance of hydrophobic cluster formation through long-range contacts in the folding transition state of two-state proteins. Proteins 2004; 55: 1023-35.

[25] Gromiha MM, Selvaraj S. Importance of long-range interactions in protein folding. Biophys Chem 1999; 77: 49-68.

[26] Carugo O. Metalloproteins: metal binding predicted on the basis of the amino acid sequence. J Appl Crystallogr 2008; 41: 104-109.

[27] Chakrabarti P, Pal, D. The interrelationships of side-chain and main-chain conformations in proteins. Prog Biophys Mol Biol 2001; 76: 1-102.

[28] Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000; 13: 149-152.

[29] Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985; 229: 834-838.

[30] Grewal PS. Numerical methods of statistical analysis, Sterling publishers, New Delhi 1987.

[31] Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins 2003; 51: 162-66.

[32] Huang JT, Cheng JP. Differentiation between two-state and multi-state folding proteins based on sequence. Proteins 2008; 72: 44-9.

Received: December 26, 2008 Revised: February 24, 2009 Accepted: March 02, 2009