Progress in Biophysics & Molecular Biology 86 (2004) 235–277

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

Review

Inter-residue interactions in protein folding and stability☆

M. Michael Gromiha[a,*], S. Selvaraj[b]

[a] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Aomi Frontier Building 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan

[b] Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, Tamil Nadu, India

Abstract

During the process of protein folding, the amino acid residues along the polypeptide chain interact with each other in a cooperative manner to form the stable native structure. The knowledge about inter-residue interactions in protein structures is very helpful to understand the mechanism of protein folding and stability. In this review, we introduce the classification of inter-residue interactions into short, medium and long range based on a simple geometric approach. The features of these interactions in different structural classes of globular and membrane proteins, and in various folds have been delineated. The development of contact potentials and the application of inter-residue contacts for predicting the structural class and secondary structures of globular proteins, solvent accessibility, fold recognition and ab initio tertiary structure prediction have been evaluated. Further, the relationship between inter-residue contacts and protein-folding rates has been highlighted. Moreover, the importance of inter-residue interactions in protein-folding kinetics and for understanding the stability of proteins has been discussed. In essence, the information gained from the studies on inter-residue interactions provides valuable insights for understanding protein folding and de novo protein design.

Keywords: Inter-residue interactions; Protein folding; Stability; Contact potential; Structural class; Folding rate; Kinetics; Φ value analysis

Contents

Abbreviations: HC, hydrophobic cluster; TMH, transmembrane helical; TMS, transmembrane strand; H bond, hydrogen bond; CO, contact order; LRO, long-range order; CI2, chymotrypsin inhibitor; ACBP, acyl coenzyme binding protein; FKBP12, FK506 binding protein; srcSH3, SH3 domain of src

☆ Dedicated to Professor P.K. Ponnuswamy, our mentor.

*Corresponding author. Tel.: +81-3-3599-8046; fax: +81-3-3599-8081.

E-mail address: michael-gromiha@aist.go.jp (M.M. Gromiha).

0079-6107/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pbiomolbio.2003.09.003

1. Introduction

The formation of stable secondary structures and a unique tertiary structure of proteins are dictated by intra-molecular interactions between the constituent amino acid residues along the polypeptide chain as well as their interaction with the surrounding medium. During the process of protein folding, the hydrophobic force drives the polypeptide chain to the folded state overcoming the entropic factors, while hydrogen bonds, ion-pairs, disulfide bonds and van der Waals interactions, define the shape and keep it from falling apart (Ponnuswamy and Gromiha, 1994). These non-covalent forces that arise from inter-residue interactions among sequential neighbors as well as those far away in the sequence, cooperatively form and stabilize the native structure.

Tanaka and Scheraga (1975) categorized the inter-residue interactions into short, medium and long range and proposed a hypothesis for protein folding by a three-step mechanism based on these interactions. For the past three decades, the concept of inter-residue interactions has been the main focus to understand the mechanism of protein folding and stability (Gromiha and Selvaraj, 2002a). Consequently, a number of investigations based on the atomic coordinates obtained from X-ray crystallography/NMR spectroscopy have been carried out to understand the role of various interactions in the folding and stability of proteins (Bahar et al., 1997; Russell et al., 1998; Selbig and Argos, 1998; Gromiha and Selvaraj, 1999a; Drablos, 1999). Further, the short-, medium- and long-range interactions have been classified according to the distance of separation between the residues along the polypeptide chain. This classification has been used successfully to address the problem of protein folding and sequence recognition (Gromiha and Selvaraj, 1997a; Miyazawa and Jernigan, 1999a).

Inter-residue interactions in proteins have been viewed from different perspectives, such as, development of empirical potentials (Reva et al., 1997; Zhang and Skolnick, 1998; Seno et al., 1998), partition of energetic components to the folding and stability of globular proteins (Ponnuswamy and Gromiha, 1994; Pace et al., 1996) and identification of amino acid side-chain clusters (Heringa and Argos, 1991; Karlin and Zhu, 1996; Dosztanyi et al., 1997; Selvaraj and Gromiha, 1998a; Kannan and Vishveshwara, 1999). Recently, inter-residue interactions have been characterized by the parameters, contact order (CO) and long-range order (LRO), which show a very strong correlation with the folding rate of small proteins (Plaxco et al., 1998; Baker, 2000; Gromiha and Selvaraj, 2001a).

In this review, we provide a comprehensive account of the role of inter-residue interactions in the structure, folding and stability of proteins.

2. Definition of short-, medium- and long-range interactions

The definition of short-, medium- and long-range interactions in a protein is based on (i) the amino acid residues, which are in contact with each other in the native structure, and (ii) their respective locations in the sequence.

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 8 Å. It has been shown that the influence of each residue over the surrounding medium extends effectively only up to 8 Å (Manavalan and Ponnuswamy, 1977) and this limit is sufficient to characterize the hydrophobic behavior of amino acid residues (Manavalan and Ponnuswamy, 1978; Ponnuswamy, 1993) and to accommodate both the local and non-local interactions (Gromiha and Selvaraj, 2000; Jiang et al., 2002). Further, 8 Å limit has been used in several studies, such as, to understand the folding rate of proteins (Debe and Goddard, 1999; Gromiha and Selvaraj, 2001a), protein stability upon mutations (Gromiha et al., 1999a), thermal stability of proteins (Gromiha, 2001; Gromiha and Thangakani, 2001), transition state structures of two-state protein mutants (Gromiha and Selvaraj, 2002b) and to understand the relationship between hydrophobic clusters (HCs) and long-range contacts in $(\alpha/\beta)_8$ barrel proteins (Selvaraj and Gromiha, 2003). On the other hand, various contacting distances in space and different sets of atoms in protein structure have also been used in the literature for protein-folding studies. Tudos et al. (1994) used the C_α atoms and the distance of 7 Å for defining long-range contacts. Kocher et al. (1994) used the centroid of each residue and the distance of 8 Å for deriving contact potentials. Jernigan and colleagues (Miyazawa and Jernigan, 1996; Bahar and Jernigan, 1997) represented the hydrophilic and hydrophobic contacts with the cutoff distance of 4 and 7 Å, respectively. Plaxco et al. (1998) proposed the concept of CO, using all the atoms in a protein within the distance of 6 Å. Fariselli et al. (2001) used the cutoff distance of 8 Å between C_β atoms for representing inter-residue contacts. On the other hand, the information about the van der Waals radii of the atoms has also been used to define inter-residue contacts (Selbig, 1995; Dosztanyi et al., 1997).

For a given residue, the composition of surrounding residues (say, within a sphere of 8 Å radius) 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 (Ponnuswamy et al., 1973, 1980; Gromiha and Selvaraj, 1997a, 1999b, 2000). Miller et al. (2002) used the same cutoff of four residues to represent long-range contacts. However, the limit of residue separation has been varied for several studies. Gilis and Rooman (1997) considered the residues separated by more than 15 residues along the sequence as long-range contacts while Tudos et al. (1994) proposed the limit of 20 residues. Dosztanyi et al. (1997) made the cutoff of 10 residues for defining stabilization centers in proteins. Gugolya et al. (1997) considered the interactions between residues that are separated by no more than four residues in the primary structure as short-range interactions, 5–20 residues as medium range and more than 20 residues as long-range interactions. The limit of 12 residues shows the highest correlation between LRO and protein-folding rates in a set of 23 two-state

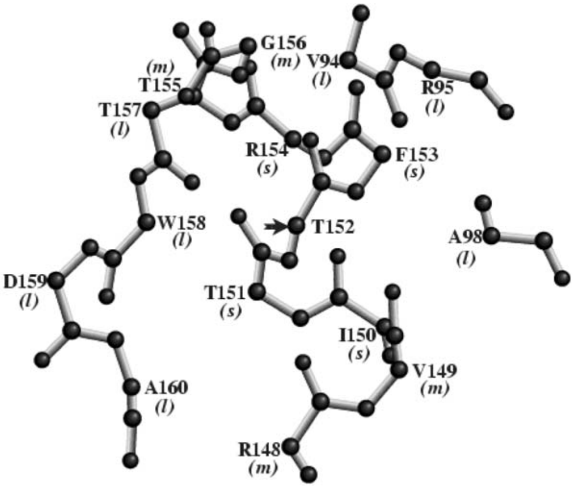

Fig. 1. Representation of short-, medium- and long-range contacts in protein structures. A typical example of surrounding residues around T152 of T4 lysozyme within 8 Å is shown (Gromiha and Selvaraj, 2000); s: short-range contacts; m: medium-range contacts; and l: long-range contacts.

proteins and this limit varies for the proteins in different structural classes (Gromiha and Selvaraj, 2001a).

The definitions of the short-, medium- and long-range interactions in protein structures are illustrated in Fig. 1. In this figure, we show the surrounding residues of T152 (indicated by an arrow) in T4 Lysozyme (2LZM) within the sphere of radius 8 Å. The residues I150, T151, F153 and R154 (marked as *s*) are close to the residue T152 (central residue) along N- and C-directions and these residues contribute to short-range interactions; the residues R148, V149, T155 and G156 (marked as *m*) are separated by three or four residues from the central residue (T152) in the amino acid sequence and these residues contribute to medium-range interactions; other residues, T157, W158, D159, A160, are separated by more than four residues and V94, R95 and A98 are far in the sequence level from T152. Accordingly, these residues (marked as *l*) contribute to long-range interactions.

This classification and representation enables one to take into consideration of both local and non-local interactions involved in the three-dimensional (3D) structures of proteins.

3. Inter-residue interactions in protein structures

The nature of inter-residue interactions varies for each class of proteins and specifically in secondary structures. In this section, we discuss the role of these interactions in proteins belonging to different structural classes and folds of globular and membrane proteins.

3.1. Different structural classes of globular proteins

Proteins are categorized into four structural classes, namely, all-α, all-β, $\alpha+\beta$ and α/β (Levitt and Chothia, 1976). The ribbon diagrams illustrating the structures in each class are shown in

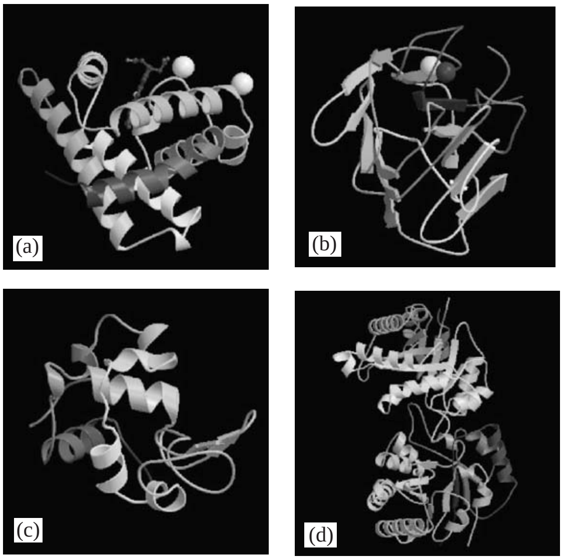

Fig. 2. Ribbon diagram for four typical protein structures in different structural classes: (a) all-α (4MBN), (b) all-β (3CNA), (c) $\alpha+\beta$ (4LYZ) and (d) α/β (1TIM). Figure was adapted from Gromiha (2003b).

Fig. 2. The all-α and all-β classes are dominated by α-helices ($\alpha > 40\%$ and $\beta < 5\%$) and β-strands ($\beta > 40\%$ and $\alpha < 5\%$), respectively (Figs. 2a and b). The $\alpha+\beta$ class contains both α-helices ($>15\%$) and anti-parallel β-strands ($>10\%$) that do not mix, but tend to segregate along the polypeptide chain (Fig. 2c). The α/β class proteins (Fig. 2d) have mixed or approximately alternating segments of α-helical ($>15\%$) and parallel β-strands ($>10\%$).

The average short-, medium- and long-range contacts for the residues in each of the four structural classes (all-α, all-β, $\alpha+\beta$ and α/β) of globular proteins are presented in Table 1. The short-range contacts are similar in all classes and the role of medium- and long-range contacts are distinct in each class. The average medium-range contacts are higher for all-α class proteins than all-β proteins, indicating the vital influence of medium-range contacts in the formation of α helices. Conversely, long-range contacts are remarkably higher in the all-β proteins than all-α proteins, implying the importance of long-range effects in β-strands. In the $\alpha+\beta$ and α/β classes the medium- and long-range contacts lie in the range between all-α and all-β class proteins. Further analysis showed that each residue in all structural classes of proteins except all-α has an average of 10 contacts (sum of short, medium and long range) whereas the all-α class residues have nine contacts (Gromiha and Selvaraj, 1997a, 1999b). It has also been reported that the number of interactions is almost proportional to the number of residues in a protein (Gromiha and Selvaraj, 1997b; Gugolya et al., 1997) and the relationship between number of residues and total number of inter-residue interactions in a set of 150 globular proteins is shown in Fig. 3. The correlation between total number of interactions and number of residues is 0.996.

3.1.1. Proteins of different size

The dominance of inter-residue interactions in proteins differs with respect to size. The computed short-, medium- and long-range contacts in small, medium and large size of proteins

Table 1
Average short-, medium- and long-range contacts in different structural classes of globular proteins

Class	Size	Average residue contacts			
		Short	Medium	Long	Total
All-α	Small	3.8	2.9	1.8	8.5
	Medium	4.0	3.1	2.0	9.1
	Large	4.0	2.8	2.7	9.5
	Combined	**3.9**	**3.0**	**2.1**	**9.1**
All-β	Small	3.9	0.8	4.5	9.2
	Medium	4.0	0.7	5.1	9.8
	Large	4.0	0.9	5.3	10.2
	Combined	**4.0**	**0.8**	**5.1**	**9.9**
$\alpha + \beta$	Small	3.9	1.6	3.3	8.8
	Medium	3.9	1.7	3.6	9.2
	Large	4.0	1.8	4.2	10.0
	Combined	**4.0**	**1.7**	**3.9**	**9.6**
α/β	Small	3.9	1.6	3.4	8.9
	Medium	4.0	2.0	3.8	9.8
	Large	4.0	1.9	4.3	10.2
	Combined	**4.0**	**1.9**	**4.3**	**10.2**
Chaperones	Papd	4.0	0.7	4.2	8.9
	GroES	3.9	0.8	4.4	9.1
	HSC	4.0	2.0	4.2	10.2
	GroEL	4.0	2.3	4.1	10.4
TMH	Small	3.9	3.0	0.3	7.2
	Medium	4.0	2.8	1.1	7.9
	Large	4.0	3.0	2.6	9.6
TMS	Large	3.9	0.7	5.9	10.5

Note: Bold numerals show the average short-, medium- and long-range contacts in four structural classes (Gromiha and Selvaraj, 2001a, c; Kumarevel et al., 1998).

are included in Table 1. We observed that in all structural classes, the average number of long-range contacts increases with increase in size. The total contacts/residue is increased by one residue in larger proteins compared to the smaller ones. On the other hand, average medium-range contact is similar in all structural classes of proteins, irrespective of the size.

3.1.2. Preference of amino acid residues in medium- and long-range contacts

The preference of each amino acid to form medium- and long-range contact has been analyzed and the results are presented in Table 2. The residue Met has the highest medium-range contact followed by Leu, Ala, Glu and Gln (Gromiha and Selvaraj, 1997a) and all of them are found to be

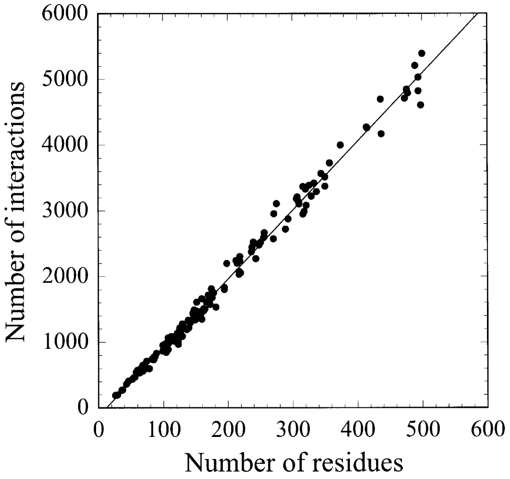

Fig. 3. Relationship between number of residues and total number of inter-residue interactions (within 8 Å) in a set of 150 protein structures.

Table 2
Average medium- and long-range contacts for the 20 amino acid residues in globular proteins

Residue	Medium	Long
Ala	2.11	3.92
Asp	1.80	**2.85**
Cys	1.88	5.55
Glu	2.09	**2.72**
Phe	1.98	4.53
Gly	1.53	4.31
His	1.98	3.77
Ile	1.77	5.58
Lys	1.96	**2.79**
Leu	2.19	4.59
Met	2.27	4.14
Asn	1.84	3.64
Pro	**1.32**	3.57
Gln	2.03	3.06
Arg	1.94	3.78
Ser	1.57	3.75
Thr	1.57	4.09
Val	1.63	5.43
Trp	1.90	4.83
Tyr	1.67	4.93

Note: The lowest medium-range contact and bottommost three long-range contacts are bold (Gromiha and Selvaraj, 1997a).

helix forming residues (Fasman, 1989; Gromiha and Ponnuswamy, 1995). This implies that these residues, at the level of medium range, influence the formation, propagation and stabilization of helices. On the other hand, Pro has the lowest medium-range contact, indicating the fact that it is not a favored residue in α-helical conformation (Fasman, 1989; MacArthur and Thornton, 1991; Gromiha and Ponnuswamy, 1995).

The residue Ile has the highest long-range contact followed by Cys, Val, Tyr, Trp, Phe and Leu. This shows that hydrophobic residues mainly influence the long-range contacts, and these residues can serve as nucleation centers during the process of protein folding and get buried in the interior. The lowest value is observed for Glu and other residues having lowest long-range contacts are Lys and Asp. This result indicates that the like-charged and oppositely charged interactions may be formed between neighboring residues (Barlow and Thornton, 1983; Matthews, 1993). From the analysis on ion-pairs (oppositely charged groups $\leqslant 4$ Å apart) and like-charged interactions ($\leqslant 4$ Å apart) in 38 protein structures, Barlow and Thornton (1983) reported that the sequence separation between 54% of like-charged groups and 34% of ion pairs is less than seven residues.

3.1.3. Long-range contacts in different residue intervals

In order to evaluate the effective inter-residue separation for the formation of long-range contacts, a bin size of 10 residues (4–10, 11–20, 21–30, 31–40, 41–50 and >50) was used. The observed average long-range contacts in four structural classes for these intervals are presented in Fig. 4 (Gromiha and Selvaraj, 1999a; Selvaraj and Gromiha, 2000). This figure reveals the opposite trends in the formation of long-range contacts between all-α and all-β proteins. The all-α class proteins have more long-range contacts in the 4–10 range and the all-β class proteins have more long-range contacts in the 11–20 range. This may be due to the specific hydrogen-bonding pattern of α-helices and β-strands in these classes of proteins.

The behavior of proteins in $\alpha+\beta$ and α/β classes is interesting. The $\alpha+\beta$ class of proteins favors the range 4–10 while the α/β class of proteins prefers the 21–30 range. The helical and strand segments are segregated into separate domains in $\alpha+\beta$ proteins and the proteins in this class behave like either all-α or all-β type. We observed that the features of $\alpha+\beta$ proteins are similar to that of all-α proteins. In α/β class, the α-helices and β-strands occur alternatively and some residue distances are necessary to form β-strand, which leads to having higher contacts in 21–30 range. These results indicate that the long-range contacts from different intervals play a considerable role in the folding of proteins belonging to different structural classes.

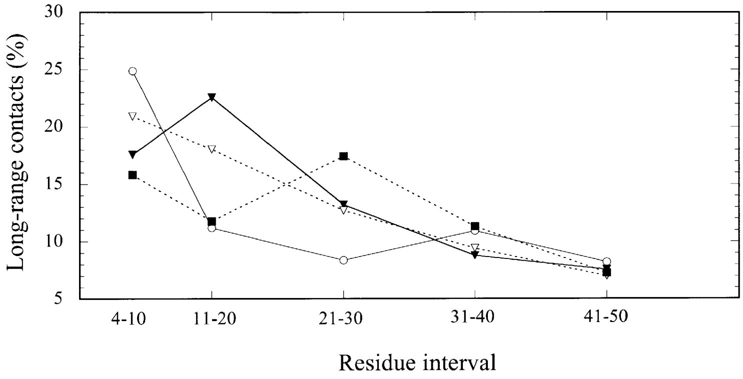

Fig. 4. Average percentage of long-range contacts in different intervals for the four structural classes of globular proteins (Gromiha and Selvaraj, 1999a). ○: all-α; ▼: all-β; ▽: $\alpha+\beta$; ■: α/β.

Further, this figure indicates that the limit of residual distances to form long-range contacts is 21–30. This is consistent with the analysis on the number of interactions per residue as a function of sequential distance between the interacting residues, which showed a significant margin after 25th neighbor (Gugolya et al., 1997).

3.2. Proteins of different fold

The average inter-residue contacts for various folds of globular proteins (Orengo et al., 1997) showed that in all-α proteins, the globin and four-helical bundle folds have more medium-range contacts than DNA/RNA binding three-helical bundle fold. Immunoglobulin fold in all-β proteins has the highest number of long-range contact among all folding types. Ribonuclease fold prefers the range 4–10 while the residues in all other folds of α/β class proteins interact with distant residues in the range of 21–30 to form long-range contacts (Kumarevel et al., 2002). In the following section, we discuss the results for $(\alpha/\beta)_8$ barrel fold and molecular chaperones.

3.2.1. $(\alpha/\beta)_8$ barrel fold

$(\alpha/\beta)_8$ barrel fold is one of the most frequent and regular domain structures of globular proteins. The overall behavior of medium- and long-range contacts in $(\alpha/\beta)_8$ barrel fold is similar to α/β class of globular proteins, reflecting the common intra-molecular interaction pattern in these proteins (Gromiha and Selvaraj, 1997b, 1999c). In different intervals of long-range contacts, most of the residues prefer the range, 21–30 followed by 4–10. We note that in this type of protein the range 11–20 is less often found.

The sequential distances between the N-terminal residues of successive β-strand segments in $(\alpha/\beta)_8$ barrel proteins showed that a significant number of successive β-strands are formed at a sequential distance of 21–30 residues. This result is consistent with the range having more long-range contacts in the native 3D structures of $(\alpha/\beta)_8$ barrel proteins (Selvaraj and Gromiha, 1998b).

Recently, we have observed a pattern of long-range contact network in $(\alpha/\beta)_8$ barrel proteins (Selvaraj and Gromiha, 2003) and that for a typical $(\alpha/\beta)_8$ barrel protein, Taka amylase (PDB code: 2TAA) is shown in Fig. 5. Consider the residue I60 in the second β-strand (S2), which has the maximum local surrounding hydrophobicity, termed as HC, we observed long-range contacts with the residues, I11, Y12, F13, L14, M112, Y113, L114 and M115. At the N-terminal side, the farthermost long-range contact was observed at I11, belonging to the first strand (S1) and part of the hydrophobic cluster (HC1). Further examination at I11 indicated no long-range contact in the

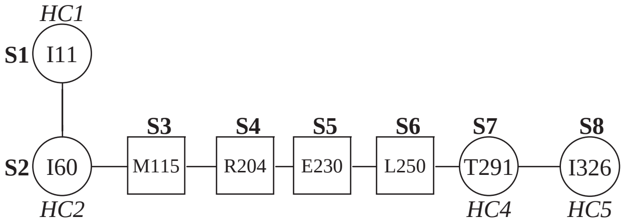

Fig. 5. Long-range contact network in Taka amylase. The residue inside the circles and squares represents the farthest long-range contact (Selvaraj and Gromiha, 2003). HC: hydrophobic clusters and S: strand.

same direction. In the C-terminal side of residue I60, we observed the farthest long-range contact at M115 (S3). Further search resulted in residues at R204, E230, L250, T291 and I326. Interestingly, all these residues are part of the β-strands, S4, S5, S6, S7 and S8, respectively, in which the residues T291 and I326 are themselves HCs of 2TAA. Thus, we observed a good pattern of long-range contact network with links to all β-strands constituting the $(\alpha/\beta)_8$ barrel domain and also among the HCs.

3.3. Molecular chaperones

The average short-, medium- and long-range contacts in four typical chaperone proteins (PapD, GroES, HSC and GroEL) are included in Table 1. We observed that the average medium-range contacts in PapD and GroES are lower than that of all-β class of proteins (Kumarevel et al., 1998). In the case of HSC and GroEL the contacts are similar to α/β class proteins. On the other hand, all the four proteins have similar level of average long-range contacts (4.2 ± 0.1 contacts/residue). Residue-wise analysis showed that Cys and all the aromatic residues have lower medium-range contacts than globular proteins. Similar trend is also observed in all-β class proteins. Further, we noticed that these residues in chaperones have lower medium-range contacts than that in all structural classes of globular proteins except in all-β class proteins. This reveals the difference in the medium-range contacts between globular proteins and chaperones and this information might be helpful to distinguish chaperones from globular proteins.

3.4. Membrane proteins

Membrane proteins, requiring to be embedded into the lipid bilayers, have amino acid sequences that will fold with a hydrophobic surface in contact with the alkane chains of the lipids and polar surface in contact with the aqueous phases on both sides of the membrane and the polar head groups of the lipids. The membrane assemblies of transmembrane helical (TMH) and transmembrane strand (TMS) proteins are different between each other as well as from globular proteins. In the following sections, we discuss the distinct role of inter-residue interactions in membrane proteins and compare the results with all-α and all-β proteins.

3.4.1. Transmembrane helical proteins

The average medium-range contact in TMH proteins is similar to all-α globular proteins, indicating a common behavior of helix packing in these two groups of proteins. Residue-wise analysis showed that Cys, Leu, Ile, Met, Val, Phe, Trp and Ala are the topmost residues with high medium-range contacts in TMH proteins (Gromiha and Selvaraj, 2001b). Strikingly, these residues are highly preferred in the membrane environment (Gromiha, 1999). All the charged residues (Asp, Glu, His, Lys and Arg) in TMH proteins have less medium-range contacts than all-α proteins. In coincidence with globular proteins, Pro has the lowest medium-range contacts due to the fact that it is a helix breaker (Fasman, 1989) and Pro is not a favored residue in membrane environment (Deber et al., 1990; Ponnuswamy and Gromiha, 1993).

Further, all the residues except Met and Pro in all-α proteins have higher medium-range contacts than long-range contacts, whereas in TMH proteins, residues Cys, Gly, Asn, Pro, Gln, Ser and Tyr have higher long-range contacts than medium-range contacts. It is noteworthy that the ratio between long and medium-range contacts for all charged residues is higher in TMH proteins than all-α proteins. This may be due to the presence of these residues near the termini of membrane spanning helices (Andersson et al., 1992).

The number of medium- and long-range contacts for a typical TMH protein (M-chain of 1PRC) is displayed in Fig. 6a. We observed that most of the residues in the membrane spanning helical segments have four medium-range contacts. Interestingly, few of the N-terminal residues in helices H1, H2 and H4 have higher long-range contacts than medium-range contacts. A similar tendency was also observed for the C-terminal residues in helices H3 and H5. Further analysis showed the presence of long-range contacts between residues in helices H1 and H2, H3 and H5, and H4 and H5. This result indicates the importance of long-range interactions in the stabilization of helix–helix interactions within the membrane. Considering the total contacts (both medium and long range) the local minima always occur in the non-helical regions of the protein.

The residues having maximum long-range contacts Gly92, Pro95 and Trp169 are also indicated in Fig. 6a. These residues have 9–11 long-range contacts and occur in the long loop regions connecting H1 and H2, and H3 and H4. Most of the contacts for the residues Gly92 and Pro95 are with the residues in the loop connecting H3 and H4, and that of Trp165 is with the residues in the loop connecting H1 and H2. This shows the favorable interactions between the residues outside the membrane. Also, these residues have few long-range contacts with the termini residues of the nearest neighboring transmembrane helix.

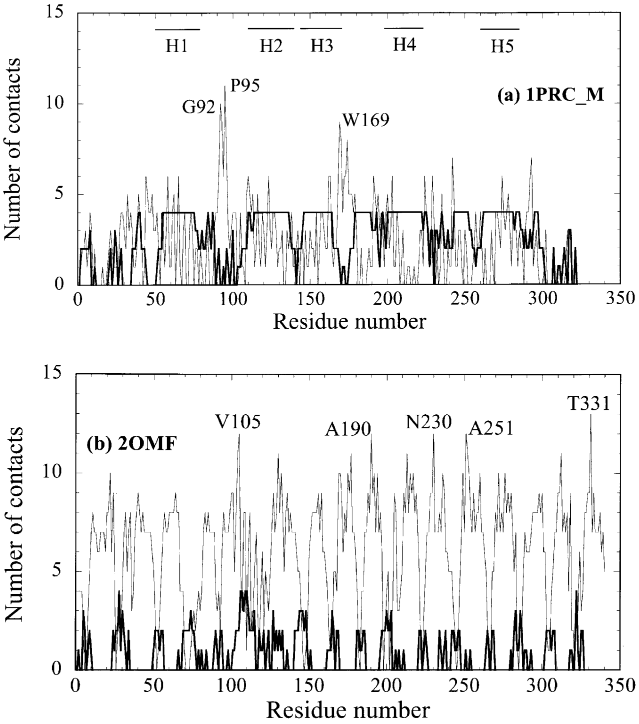

Fig. 6. Number of medium- and long-range contacts in two typical membrane proteins (Gromiha and Selvaraj, 2001b). (a) M-chain of 1PRC and (b) 2OMF. The thick and thin lines indicate, respectively, the medium- and long-range contacts.

3.4.2. Transmembrane strand proteins

The TMS proteins have appreciably higher long-range contacts and lower medium-range contacts than all-β proteins (Gromiha and Selvaraj, 1997a, 2001b). Considering the total contacts, one additional contact (11 contacts per residue) is observed for all residues in TMS proteins compared to all-β proteins, indicating the closer packing of β-strands in TMS proteins.

Residue-wise analysis showed that the residue Cys has the highest long-range contact followed by Met and Pro has the lowest long-range contact (Gromiha and Selvaraj, 2001b). Indeed, it has been shown that Met and Pro are, respectively, high and low preferred residues in membrane spanning β-strands (Gromiha and Ponnuswamy, 1993, 1996; Gromiha et al., 1997). The ratio between the number of long- and medium-range contacts for the positively charged residues is approximately 8–12 in TMS proteins, while it is 5–7 in all-β proteins. In the membrane spanning β-strand segments, the positive charged residues, Lys and Arg make significant contacts with other residues in the membrane (Gromiha and Suwa, 2003). This might be due to the formation of ion-pairs and/or cation–π interactions (Gromiha, 2003a).

The number of medium- and long-range contacts for all the residues in a typical TMS protein (2OMF) is shown in Fig. 6b. We observed that about 75% of the residues in membrane spanning β-strands have no medium-range contact. The maximum long-range contacts are observed for the residues V105, A190, N230, A251 and T331. Among these residues, A190, N230 and T331 are in membrane spanning strands and V105 and A251 are very close to β-strands. Hence, the long-range contacts arise mainly from the residues in β-strands.

The comparison of total medium- and long-range contacts in TMS proteins revealed that on an average, all the residues have more long-range contacts that are remarkably 18 times higher than medium-range contacts. This might be due to the complex architecture of TMS proteins interspersed with polar residues in membrane, which make hydrogen-bonding networks and electrostatic contacts with residues in neighboring β-strands that are far in sequence level.

3.5. Mesophilic and thermophilic proteins

The average medium- and long-range contacts are similar in mesophilic and thermophilic proteins due to their structural similarities. However, the residues Cys, Arg, Asn and Trp have different behaviors in mesophilic and thermophilic proteins (Gromiha, 2001). Arg has more number of medium-range contacts in thermophiles than mesophiles and it might be due to the presence of more number of ion pairs and salt bridges in thermophilic proteins, which are reported to be important factors for enhancing protein thermostability (Kumar et al., 2000).

3.6. Formation of amino acid clusters

The amino acid residues in protein structures interact with each other and form clusters. Heringa and Argos (1991) used a cutoff radius of 4.5 Å between side-chain atoms to delineate amino acid clusters and showed that most of the clusters are comprised of three to four residues and are localized near the protein surface. On the other hand, Zehfus (1995) reported that an average of 65% of hydrophobic residues are involved in residue clusters and each HC contains at least seven residues. Karlin and colleagues (Karlin et al., 1994; Karlin and Zhu, 1996) proposed several distance measures between residues in protein structures based on average, minimum and maximum distances of both all atom (backbone and side chain) and side-chain atom coordinates to extract the clusters. The applications of this method have been demonstrated with the delineation of different types of clusters, namely, acidic clusters, cysteine clusters, iron–sulfur proteins and charge clusters, and clusters of multiple histidine residues. Chirgadze and Larionova (1999) suggested a different criterion for identifying charged clusters: charged groups were included in a cluster if their charged N and O atoms were located at distances between 2.4 and 7 Å. This approach identified charged clusters in 86% of the considered 275 protein structures. Selvaraj and Gromiha (1998a) identified the residue clusters in $(\alpha/\beta)_8$ barrel proteins based on the distance criteria of 8 Å limit between the C_α atoms of two residues in protein structure. This procedure showed the presence of 14 identical amino acid clusters and a large number of physicochemically similar clusters based on surrounding hydrophobicity, turn preference, bulkiness, refractive index and anti-parallel β-strand preference in a set of 36 $(\alpha/\beta)_8$ barrel proteins. Further, graph theoretical approach has been used to identify residue clusters within the distance of 6.5 Å in protein structures (Kannan and Vishveshwara, 1999; Kannan et al., 2001). These results show the influence of inter-residue interactions to the formation of residue clusters, which are important for the folding and stability of protein structures.

4. Contact potentials based on inter-residue interactions

The information about the preference of residue pairs to form medium- and long-range contacts has been used to understand the residue–residue cooperativity in protein folding.

4.1. Development of effective potentials

The most widely used approach for deriving effective potentials from an ensemble of experimentally determined protein structures consists of computing frequencies of sequence and structure features, and converting these frequencies into free energies (Sippl, 1990, 1995). Rooman and Wodak (1995) reviewed the two main approaches for developing database-derived potentials and their applications to protein structure prediction.

The first approach is in the context of standard statistical models for deriving scores based on log(frequencies) (Bowie et al., 1991; Ouzounis et al., 1993; Wilmanns and Eisenberg, 1993). The second one is in the context of statistical mechanics for deriving potential terms expressed as $-kT$ log(frequencies), where k is Boltzmann's constant and T is the temperature (Sippl, 1990; Jones et al., 1992; Kocher et al., 1994). Zhang and Skolnick (1998) constructed an artificial protein structural database using contact and secondary structure propensity potentials (called as "true" potentials) and then derived new sets of potentials to see how they are related to the true potentials. They found that using the Boltzmann distribution method, when the stability of the structures in the database lies within a certain range, both contact potentials and secondary structure propensities could be derived separately with remarkable accuracy.

Further, several potentials have been derived based on the interactions between amino acid residues in a protein and are used to understand the protein-folding problem, to predict protein structure, stability and fold recognition, and for designing novel proteins (Flockner et al., 1995; Mirny and Shakhnovich, 1996; Reva et al., 1997; Gilis and Rooman, 1997; Furuichi and Koehl, 1998; Chiu and Goldstein, 1998; Miyazawa and Jernigan, 1999b; Tobi et al., 2000; Russ and Ranganathan, 2002).

4.1.1. Backbone torsion potentials

The influence of amino acids on the backbone conformation of neighboring residues along the chain of a protein has been considered for developing backbone torsion potentials (Rooman et al., 1992; Kocher et al., 1994). They have reported two types of torsion potentials, such as, residue-to-torsion potential and torsion-to-residue potential. The residue-to-torsion potential has been computed from the probabilities $P^{ai,ai}_{i-k,\,i-k}(t_k)$ and $P^{ai,aj}_{i-k,\,j-k}(t_k)$ that amino acid a at position i along the sequence and pairs of amino acids a at positions i and j, respectively, are associated with the torsion domain t at position k; i and k can be anywhere within a window of 17 sequence positions $[k-8, k+8]$ centered around k. The torsion-to-residue potential is computed from the probabilities $P^{ti,ti}_{i-k,\,i-k}(a_k)$ and $P^{ti,tj}_{i-k,\,j-k}(a_k)$ that torsion domains t at position i along the sequence and pairs of torsion domains t at positions i and j, respectively, are associated with the amino acid a at position k. Both the residue-to-torsion and torsion-to-residue potentials have been subdivided into two parts: a short-range part that includes only contributions from residues and torsion domains in the interval $[k-1, k+1]$, and a middle-range part that considers all the remaining contributions in the $[k-8, k+8]$ window. These torsion potentials represent best the interactions in protein surface (Gilis and Rooman, 1997) and perform well for predicting the stability of surface mutations.

4.1.2. Residue–residue interaction potentials

Residue–residue potentials describe both local interactions (short and medium range) along the chain and interactions between residues that are far apart along the sequence but close in space (long range). They are computed from the propensities $P^{ai,aj}_{|i-j|}(d_{ij})$ of two residues a_i and a_j at positions i and j along the sequence, to be separated by a spatial distance d_{ij} (Kocher et al., 1994), and the pairs separated by 1–7 sequence positions have been considered as middle-range potentials. Pairs separated by more than eight positions represent non-local (long-range) interactions along the chain. This distance potential is dominated by hydrophobic interactions and represents the main interactions that stabilize the protein core. Similar approach has also been proposed for devising a residue–residue potential function to calculate the conformational energy of proteins (Oobatake and Crippen, 1981).

4.1.3. Inter-residue contact potentials

Miyazawa and Jernigan (1985, 1996) estimated the effective inter-residue contact energies from the number of residue–residue contacts (within 6.5 Å) observed in crystal structures of globular proteins by means of a quasi-chemical approximation. This empirical energy function includes solvent effects and provides an estimate of the long-range component of conformational energies. The interaction energies e_{ij} (contact energy between the residues i and j) and e'_{ij} (the energy difference accompanying the formation of a contact pair i–j from contact pairs i–i and j–j) are given in Table 3 (Miyazawa and Jernigan, 1985, 1996). They have observed the following results: (i) the formation of Cys–X contacts from Cys–Cys and X–X contacts represents a relatively large energy loss, because Cys–Cys often form disulfide bonds; (ii) the contact formation between negatively charged (Asp and Glu) and positively charged (Lys and Arg) residues are preferable

Table 3
Contact energies derived from protein crystal structures

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro
Cys	−5.44	−4.99	−5.80	−5.50	−5.83	−4.96	−4.95	−4.16	−3.57	−3.16	−3.11	−2.86	−2.59	−2.85	−2.41	−2.27	−3.60	−2.57	−1.95	−3.07
Met	0.46	−5.46	−6.56	−6.02	−6.41	−5.32	−5.55	−4.91	−3.94	−3.39	−3.51	−3.03	−2.95	−3.30	−2.57	−2.89	−3.98	−3.12	−2.48	−3.45
Phe	0.54	−0.20	−7.26	−6.84	−7.28	−6.29	−6.16	−5.66	−4.81	−4.13	−4.28	−4.02	−3.75	−4.10	−3.48	−3.56	−4.77	−3.98	−3.36	−4.25
Ile	0.49	−0.01	0.06	−6.54	−7.04	−6.05	−5.78	−5.25	−4.58	−3.78	−4.03	−3.52	−3.24	−3.67	−3.17	−3.27	−4.14	−3.63	−3.01	−3.76
Leu	0.57	0.01	0.03	−0.08	−7.37	−6.48	−6.14	−5.67	−4.91	−4.16	−4.34	−3.92	−3.74	−4.04	−3.40	−3.59	−4.54	−4.03	−3.37	−4.20
Val	0.52	0.18	0.10	−0.01	−0.04	−5.52	−5.18	−4.62	−4.04	−3.38	−3.46	−3.05	−2.83	−3.07	−2.48	−2.67	−3.58	−3.07	−2.49	−3.32
Trp	0.30	−0.29	0.00	0.02	0.08	0.11	−5.06	−4.66	−3.82	−3.42	−3.22	−2.99	−3.07	−3.11	−2.84	−2.99	−3.98	−3.41	−2.69	−3.73
Tyr	0.64	−0.10	0.05	0.11	0.10	0.23	−0.04	−4.17	−3.36	−3.01	−3.01	−2.78	−2.76	−2.97	−2.76	−2.79	−3.52	−3.16	−2.60	−3.19
Ala	0.51	0.15	0.17	0.05	0.13	0.08	0.07	0.09	−2.72	−2.31	−2.32	−2.01	−1.84	−1.89	−1.70	−1.51	−2.41	−1.83	−1.31	−2.03
Gly	0.68	0.46	0.62	0.62	0.65	0.51	0.24	0.20	0.18	−2.24	−2.08	−1.82	−1.74	−1.66	−1.59	−1.22	−2.15	−1.72	−1.15	−1.87
Thr	0.67	0.28	0.41	0.30	0.40	0.36	0.37	0.13	0.10	0.10	−2.12	−1.96	−1.88	−1.90	−1.80	−1.74	−2.42	−1.90	−1.31	−1.90
Ser	0.69	0.53	0.44	0.59	0.60	0.55	0.38	0.14	0.18	0.14	−0.06	−1.67	−1.58	−1.49	−1.63	−1.48	−2.11	−1.62	−1.05	−1.57
Asn	0.97	0.62	0.72	0.87	0.79	0.77	0.30	0.17	0.36	0.22	0.02	0.10	−1.68	−1.71	−1.68	−1.51	−2.08	−1.64	−1.21	−1.53
Gln	0.64	0.20	0.30	0.37	0.42	0.46	0.19	−0.12	0.24	0.24	−0.08	0.11	−0.10	−1.54	−1.46	−1.42	−1.98	−1.80	−1.29	−1.73
Asp	0.91	0.77	0.75	0.71	0.89	0.89	0.30	−0.07	0.26	0.13	−0.14	−0.19	−0.24	−0.09	−1.21	−1.02	−2.32	−2.29	−1.68	−1.33
Glu	0.91	0.30	0.52	0.46	0.55	0.55	0.00	−0.25	0.30	0.36	−0.22	−0.19	−0.21	−0.19	0.05	−0.91	−2.15	−2.27	−1.80	−1.26
His	0.65	0.28	0.39	0.66	0.67	0.70	0.08	0.09	0.47	0.50	0.16	0.26	0.29	0.31	−0.19	−0.16	−3.05	−2.16	−1.35	−2.25
Arg	0.93	0.38	0.42	0.41	0.43	0.47	−0.11	−0.30	0.30	0.18	−0.07	−0.01	−0.02	−0.26	−0.91	−1.04	0.14	−1.55	−0.59	−1.70
Lys	0.83	0.31	0.33	0.32	0.37	0.33	−0.10	−0.46	0.11	0.03	−0.19	−0.15	−0.30	−0.46	−1.01	−1.28	0.23	0.24	−0.12	−0.97
Pro	0.53	0.16	0.25	0.39	0.35	0.31	−0.33	−0.23	0.20	0.13	0.04	0.14	0.18	−0.08	0.14	0.07	0.15	−0.05	−0.04	−1.75

Note: The upper half and diagonal elements represent for e_{ij} and lower half indicates e'_{ij}. The energies are in RT units. Data from Miyazawa and Jernigan (1996).

due to electrostatic interactions; the magnitudes of the interaction energies of Asp and Glu with His are smaller than that with Lys and Arg because of its less average charge; (iii) Tyr and Trp (to some extent) prefer contacts with polar residues because of the presence of polar atoms in their side chains, although they have hydrophobic characteristics as indicated by large negative values of e_{ij}; and (iv) the segregation of hydrophobic and hydrophilic residues can be directly seen from the values of e'_{ij}; e'_{ij} among hydrophobic residues (Met, Phe, Ile, Leu and Val) takes small positive or negative values, indicating that these residues do not have strong specific preferences but are almost randomly mixed in protein structures. Hydrophilic residues (Thr, Ser, Asn, Gln, His, Arg, Lys and Pro) for the most part prefer contacts with each other to those between the same types of residues; in the case of charged residues, the subtracted unfavorable electrostatic interactions would in part be responsible for this.

Zhang and Kim (2000) proposed that the residue contact energies strongly depend on the secondary structural environment and derived contact potentials in the context of secondary structural environment in proteins. These potentials have been used in threading and to predict the contacts in 3D structures of proteins.

4.2. Potentials based on distance criteria

The distance criterion has also been used to describe the residue pairs influenced by local (medium-range) and non-local (long-range) interactions. The inter-residue contacts in protein structures have been pictorially represented with the aid of contact maps. The contact map for a typical protein within the distance of 8 Å is shown in Fig. 7. In this figure, the diagonal residues show the short-range contacts, the residues close to the diagonal represent the medium-range contacts and the residues far away from the diagonal indicate the 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

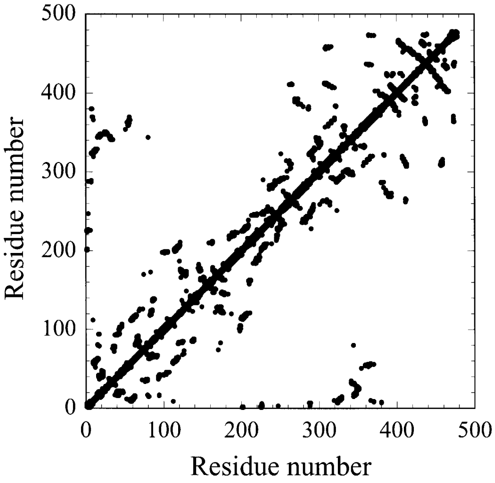

Fig. 7. The 8 Å contact map for Taka amylase.

are, respectively, the total number of residues of type i and j (Gromiha and Selvaraj, 1999a). We have derived sets of 20×20 matrices for medium- and long-range interactions for each structural class of globular proteins and membrane proteins (Gromiha and Selvaraj, 1999a, 2001b). In principle, it should be possible to convert the number of contacts into free energies using the standard procedures (Sippl, 1990, 1995; Miyazawa and Jernigan, 1996).

4.2.1. Globular proteins

The topmost 10 residue pairs that are influenced by medium- and long-range contacts in all structural classes of globular proteins are presented in Table 4. In medium-range contacts, we found that the effect of C–C is very less; hydrophobic and polar residues have equal role in forming medium-range contacts. The charged residues, Asp, Glu and Lys, have more medium-range contacts than other residues. Interestingly, D–K and E–K are one of the topmost three preferred residue pairs for the classes, $\alpha+\beta$ and α/β, respectively. This may be due to the formation of ion pairs (Barlow and Thornton, 1983).

In long-range contacts, about 50% of residue pairs are observed with the same residue (C–C; V–V; G–G; L–L, A–A and I–I). The highest preference is observed for C–C, which is due to the formation of disulfide bridges (Thornton, 1981). The hydrophobic residues Ala, Val and Leu

Table 4
Topmost 10 residue pairs influenced by medium- and long-range contacts in globular (all structural classes) and membrane (TMH and TMS) proteins

No.	All-α	TMH	All-β	TMS	$\alpha+\beta$	α/β	Globular	Membrane
Medium-range contacts								
1	A–A	L–L	A–A	D–G	L–V	A–A	A–A	L–L
2	L–L	G–L	S–S	G–S	A–A	E–K	L–L	G–L
3	A–K	L–T	C–C	E–G	D–K	L–L	E–K	A–A
4	E–K	A–L	D–G	A–A	G–G	A–G	A–K	A–L
5	D–K	L–S	G–G	G–T	A–L	G–G	D–K	L–T
6	L–V	A–I	I–V	A–S	E–K	A–L	A–L	L–S
7	F–F	F–F	L–V	Q–Q	C–C	A–D	L–V	I–V
8	I–L	I–V	D–R	S–T	A–D	I–L	G–G	A–F
9	A–E	A–F	G–S	I–I	A–S	A–K	A–G	A–V
10	G–L	L–V	G–V	G–V	W–W	L–V	I–L	A–I
Long-range contacts								
1	A–L	C–C	C–C	C–C	C–C	V–V	C–C	C–C
2	L–L	L–L	V–V	G–G	V–V	G–G	V–V	L–L
3	C–C	I–L	L–V	A–A	L–L	A–V	G–G	G–G
4	A–A	A–L	G–G	A–G	A–V	I-V	L–V	A–V
5	L–V	L–V	L–L	A-L	A–L	L–V	A–V	A–L
6	A–V	F–G	S–S	V–Y	L–V	I–L	L–L	A–G
7	V–V	G–L	A–V	G–Y	I–L	A–L	A–L	I–L
8	E–K	P–P	T–V	G–V	I–I	A–A	A–A	A–V
9	I–L	G–P	G–T	N–N	I–V	I–I	I–V	G–V
10	A–I	A–I	A–A	D–G	A–I	L–L	I–L	V–V

Note: Data from Gromiha and Selvaraj (1999a, 2001b).

contribute more for long-range contacts, due to the formation of HCs (Table 4). The contribution of Ile is less compared to other hydrophobic residues.

The comparison of medium- and long-range contacts shows that the charged and polar residues play a main role in forming medium-range contacts although hydrophobic residues are also making significant contribution. In long-range contacts, hydrophobic residues dominate and the role of polar residues is minimal.

4.2.2. Membrane proteins

In Table 4, we also included the topmost 10 residue pairs that are influenced by medium- and long-range contacts in membrane (TMH and TMS) proteins. In TMH proteins, hydrophobic–hydrophobic residue pairs are predominant to form medium-range contacts, owing to the fact that the membrane spanning helical segments are highly accommodated with a stretch of hydrophobic residues. Among the hydrophobic residues, Leu is the most preferred one. Interestingly, Leu has a preference for small residues (Ala, Gly, Ser and Thr) to form residue pairs. In TMS proteins, the polar residues Gln, Thr and Ser and the negatively charged residues Asp and Glu have higher preference to form medium-range contacts.

In TMH proteins, as expected, the hydrophobic residue pairs mainly contribute for long-range contacts. Interestingly, the pair P–P has also higher influence for long-range contacts. This may be due to the occurrence of Pro at or near the termini of TMH segments and form higher long-range contacts. In TMS proteins, Gly has the highest preference. Also, Tyr has higher long-range contacts with other residues and Asn has higher contacts with the same residue.

4.2.3. Mesophilic and thermophilic proteins

The thermophilic proteins have a significant number of inter-residue contacts between residues of H-bond forming capability other than specific interactions, such as HCs that are observed in both mesophilic and thermophilic proteins (Gromiha, 2001). Recent structural analysis on a set of mesophilic and thermophilic proteins showed that the increase in number of hydrogen bonds increases the stability (Vogt et al., 1997). Further, we observed that apart from hydrophobic residue pairs, thermophilic proteins prefer to form contacts between hydrophobic and polar residues. The residue, Tyr prefers to have contacts with other residues, Asp, Ile, Lys, Ser and Val in thermophiles while its contacts in mesophiles are less.

These results reveal the distinct residue pair-preference between (i) globular and membrane proteins and (ii) mesophilic and thermophilic proteins. These preferred residue pairs interact in a cooperative manner to direct and stabilize the process of protein folding.

5. Application of inter-residue interactions in protein structure prediction

5.1. Secondary structure of proteins

The information about the inter-residue interactions in different structural classes showed that secondary structure formation is mainly influenced by medium-range interactions in all-α proteins whereas it is influenced by long-range interactions in all-β proteins. As most of the secondary structure prediction algorithms take into account only the effect of neighboring residues along the sequence which include short- and medium-range interactions, we hypothesized that secondary structure prediction of all-α proteins would be better than the other classes of proteins (Gromiha and Selvaraj, 1998).

The average predicted accuracy levels of all the structural classes by different methods have been computed and the results are presented in Table 5 (Gromiha and Selvaraj, 1998). We found that all the methods predict the secondary structures of all-α class better than other classes. The enhanced neural network method (Kneller et al., 1990) and the PHD algorithm (Rost and Sandor, 1994a) predict with a difference of 9%. Also other methods (Chou and Fasman, 1974; Garnier et al., 1978; Qian and Sejnowski, 1988; Gromiha and Ponnuswamy, 1995; Chandonia and Karplus, 1996; Ito et al., 1997) predict about 5% more accurately in all-α class than all other classes of globular proteins.

A plausible reason for this tendency can be ascribed to the predominant role of short- and medium-range interactions in all-α proteins. Similarly, lower accuracy in the prediction of secondary structures in other classes of proteins implies the dominance of long-range interactions in them. Further, inter-residue interactions have been successfully used to predict the structural class of globular proteins (Kumarevel et al., 2000). Hence, developing secondary structure prediction techniques that are specific for each structural class incorporating the influence of short-, medium- and long-range interactions may pave a way for improving the secondary structure prediction of proteins.

5.2. *Solvent accessibility*

Solvent accessibility is one of the important factors for the structure and function of proteins. Recently, Ahmad et al. (2003a, b) proposed a method based on neural networks for predicting the real value solvent accessibility of amino acid residues. The basic design of a neural network is shown in Fig. 8 and it contains input, hidden and output layers. The details about the construction of neural network have been available in the literature (Rost and Sander, 1994b; Ahmad and Gromiha, 2002, 2003; Pollastri et al., 2002; Ahmad et al., 2003a, b). In the network designed for predicting the real value solvent accessibility, the neighboring residue information has been taken into account for training the network and hence the predictive accuracy mainly depends on short-range interactions. It has been reported that the inclusion of residues

Table 5
Accuracy of prediction in different structural classes by eight methods

Method	All-α	All-β	Mixed	Reference
Statistical analysis	**62.2**	49.5	55.3	Chou and Fasman (1974)
Information theory	**61.9**	58.2	58.1	Garnier et al. (1978)
Neural network	**67.0**	64.0	64.9	Qian and Sejnowski (1988)
Enhanced neural network	**79.0**	70.0	64.0	Kneller et al. (1990)
Neural network, PHD	**80.8**	68.8	72.4	Rost and Sandor (1994a, b)
Hydrophobicity profile	**84.1**	83.6	79.8	Gromiha and Ponnuswamy (1993)
Neural network	**71.1**	66.3	66.9	Chandonia and Karplus (1996)
3D-1D compatibility	**74.5**	70.6	67.3	Ito et al. (1997)

Note: The highest accuracy among the three structural classes is indicated in bold (Gromiha and Selvaraj, 1998).

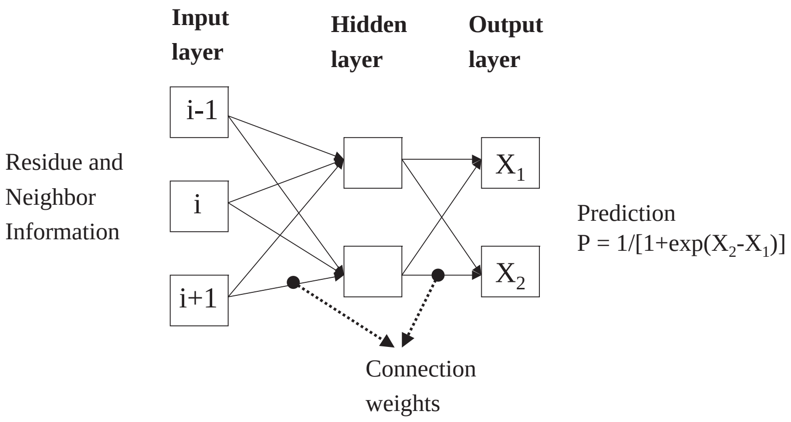

Fig. 8. Design of a neural network for predicting the solvent accessibility of proteins (Ahmad et al., 2003a).

contributing for short- and medium-range interactions improved the accuracy by 5.7% and the information from more than four residue neighbors did not show any significant improvement. However, they mentioned that the mean absolute error of 18% has been determined by long-range contacts between residues in a protein.

5.3. Identification of domains

Islam et al. (1995) proposed an automatic algorithm for identifying domains in proteins based on inter-residue contacts. This approach is based on dividing the chain segment to minimize the density of inter-domain contacts in a distance plot. The distance plot has been constructed from a distance matrix, D_{ij}, where i and j are two residues in a protein. The value of D_{ij} is taken as 1 if the C_α–C_α distance is less than a cutoff distance d or 0 otherwise. They have correctly identified the domains in 78% of the 284 non-redundant protein structures and most of the proteins have one or two domains.

5.4. Inter-residue contact prediction

Gobel et al. (1994) proposed a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment. They have then used these correlations to predict contact maps of protein structures. Further, a neural network based prediction has been proposed for predicting the inter-residue contacts in proteins (Fariselli et al., 2001). In this procedure, two residues in a protein are in contact with each other if the distance between the C_β atoms is within 8 Å, which are separated by more than seven residues. The program CORNET has been developed including the evolutionary information in the form of sequence profile, sequence conservation, correlated mutations and predicted secondary structures. The predictor was trained and cross-validated on a data set of 173 non-homologous proteins. It has been reported that the performance of this method is poor for all-α proteins and the accuracy is high for all-β and mixed proteins.

5.5. Protein fold and sequence recognition

Miyazawa and Jernigan (1999a) developed a tertiary-structure potential consisting of a long-range interaction, pairwise contact potential and a repulsive packing potential for protein fold and sequence recognition. This potential is devised to evaluate together the total conformational energy of a protein at the coarse-grained residue level. The stability of native structures is assumed as a primary requirement for proteins to fold into their native structures. In order to remove the protein size dependence and to represent protein stabilities for monomeric and multimeric states, a collapse energy is subtracted from the contact energies. The free energy of the whole ensemble of protein conformations that is subtracted from the conformational energy to represent protein stability as approximated as the average energy expected for a typical native structure with the same amino acid composition. This term may be constant in fold recognition but essentially varies in sequence recognition. They have employed a simple test of threading sequences into structures without gaps and showed that the new potential may be used for both fold and sequence recognition.

5.6. Discriminating correct from incorrect folds

The information about inter-residue interactions has been used to distinguish correct fold from incorrect folds of protein structure. Park et al. (1997) tested a set of 18 energy functions, which include 13 pairwise potentials, four hydrophobic potentials and one environment-based potential for detecting the correct folds in three different sets of data. First set contains eight small proteins, all of which have correct native secondary structures and are reasonably compact, second is the set of all sub-conformations in a database of known 3D structure and the third is a set of ensembles of 1000 conformations each for seven small proteins obtained from molecular dynamics simulations at 298 and 498 K. They found that the potentials obtained from inter-residue contacts are performing reasonably well to discriminate the correct folds from incorrect folds.

5.7. Ab initio protein structure prediction

Skolnick and colleagues (Skolnick and Kihara, 2001; Skolnick et al., 2001) have developed a program, PROSPECTOR for identifying templates of protein structures, side-group contact prediction and for the prediction of short- and intermediate-range distances between the side groups along the polypeptide chain. The accuracy of contact and distance predictions is high if any homologous protein exists in the structural database. In such situation, the remaining protocols follow the comparative modeling and this methodology is not used for the novel folds or for different cases of fold recognition. This problem has been overcome by culling the predicted contacts from templates that can have a different global fold, with corresponding low accuracy. They have also built the starting lattice models from very fragmentary templates, which may contain small elements of super-secondary structures resembling structural motifs in a target structure. This ab initio method has been fully automated and it predicts the side-group contacts with reasonable accuracy even if the proteins do not have significantly high scoring in the threading stage.

6. Inter-residue interactions and protein-folding rates

Understanding the protein-folding pathways and kinetics is a challenging task. It has been recently established that the topology of proteins plays an important role in kinetics of protein folding (Makarov et al., 2002; Makarov and Plaxco, 2003; Jewett et al., 2003; Gromiha, 2003c). Recently, two parameters, CO and LRO have been proposed to relate the total number of contacts and long-range contacts, respectively, with protein-folding rates (Plaxco et al., 1998; Gromiha and Selvaraj, 2001a).

6.1. Concept of contact order

The parameter, CO reflects the relative importance of local and non-local contacts to the native structure of a protein (Plaxco et al., 1998). It is defined as $\Sigma\Delta S_{ij}/LN$, where N is the total number of contacts, ΔS_{ij} is the sequence separation between contacting residues i and j, and L is the total number of residues in the protein. In a protein with low CO, the residues that interact with others are close in sequence. A high CO implies that there are a large number of long-range interactions. Plaxco et al. (1998) reported that the CO shows a significant correlation with folding rate of small, two-state proteins.

6.2. Definition of long-range order

We 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 (Gromiha and Selvaraj, 2001a). It is defined as

$$\mathrm{LRO} = \Sigma n_{ij}/N, \quad n_{ij} = \begin{cases} 1 & \text{if } |i-j| > 12, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where i and j are two contacting residues in space within a distance of 8 Å and N is the total number of residues in a protein.

6.2.1. Relationship between long-range order and folding rate of proteins

The computed LRO using Eq. (1) for a set of 23 small, two-state proteins are presented in Table 6 along with the CO and folding rate of each protein. We found a good inverse relationship ($r = -0.78$) between LRO and experimental folding rate, ln(k) for all the considered proteins. We have varied the minimum distance of separation between two interacting residues from 1 to 50 residues and examined the correlation between LRO and folding rates. The results presented in Fig. 9 indicate that the minimum distance of 12 residues in defining LRO has the best correlation between LRO and folding rates and significant correlation is obtained for the minimum residue separation of 10–15 residues (Gromiha and Selvaraj, 2001a). This observation suggests the presence of key residues, responsible for the formation of folding nucleus at an interval of approximately 25 residues, in agreement with the experimentally observed distance of separation between residues forming folding nuclei in *SH3 domain* and *chymotrypsin inhibitor* (Itzhaki et al., 1995; Grantcharova et al., 1998). Further, we have set up separate limits for the proteins in different structural classes.

Table 6
Contact order, long-range order and folding rate of 23 two-state proteins

PDB code	CO	LRO1	LRO2	ln(k)
All-α proteins				
1LMB	9.40	1.126	0.851	8.19
1HRC	11.20	2.212	1.173	8.76
2ABD	14.00	2.302	1.814	6.55
1YCC	11.60	2.214	1.126	9.62
All-β proteins				
1CSP	16.40	3.045	0.269	6.98
1TEN	17.40	3.888	1.101	1.06
1SHF	18.30	2.847	0.678	4.55
2AIT	21.60	4.135	0.703	4.20
3MEF	17.70	2.957	0.348	5.30
1MJC	16.00	2.986	0.377	5.24
1AEY	19.90	3.000	0.759	2.09
1SHG	19.10	3.018	0.737	1.41
1SRL	19.60	3.107	0.714	4.04
1PKS	20.00	3.842	1.184	−1.05
Mixed class proteins				
1UBQ	15.10	2.368	2.500	7.33
1CIS	16.40	3.333	3.485	3.87
1PCA	17.00	2.553	2.660	6.80
2PTL	17.60	2.231	2.538	4.10
1HDN	18.40	3.459	3.647	2.70
1APS	21.20	4.184	4.306	−1.48
1URN	16.90	2.917	3.063	5.76
1FKB	17.80	3.963	4.131	1.46
1VIK	12.30	2.970	3.212	6.80

Note: LRO1, LRO computed with Eq. (1); LRO2, the minimum distance of separation for computing LRO is 27 in all-α proteins, 44 for all-β proteins and 10 for mixed class proteins. Data from Gromiha and Selvaraj (2001a).

6.2.2. Contact order versus long-range order

The relationship between CO and folding rate of protein showed that the CO has a strong negative correlation for mixed-class proteins ($r = -0.82$) as shown in Fig. 10e and the correlation is much weaker for all-α ($r = -0.56$) and all-β ($r = -0.46$) proteins (Fig. 10a and c). This is probably due to the same treatment of all structural classes in defining the parameter. On the other hand, LRO performs extremely well ($r = -0.72$ for all-α, $r = -0.92$ for all-β and $r = -0.86$ for mixed-class proteins) in determining the folding rate of two-state proteins belonging to all structural classes (Fig. 10b, d and f).

6.2.3. Prediction of folding rate based on long-range order

We set up regression equations for all-α, all-β and mixed-class proteins by relating the folding rate and LRO. A back-check test was carried out to verify the self-consistency of the analysis. We found an excellent agreement between the predicted folding rates and experimental observations.

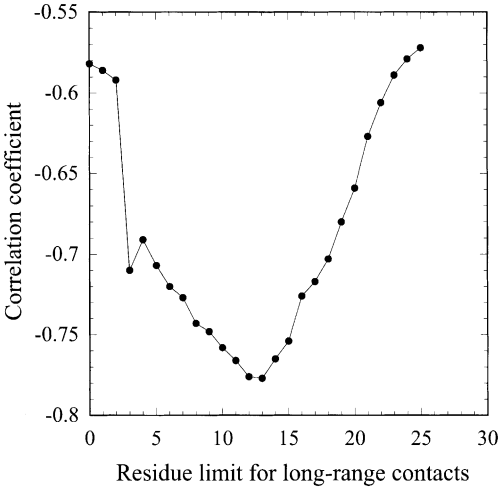

Fig. 9. Plot connecting the correlation coefficient obtained between long-range contacts and folding rate of proteins, and the minimum limit to define long-range contacts (Gromiha and Selvaraj, 2001a).

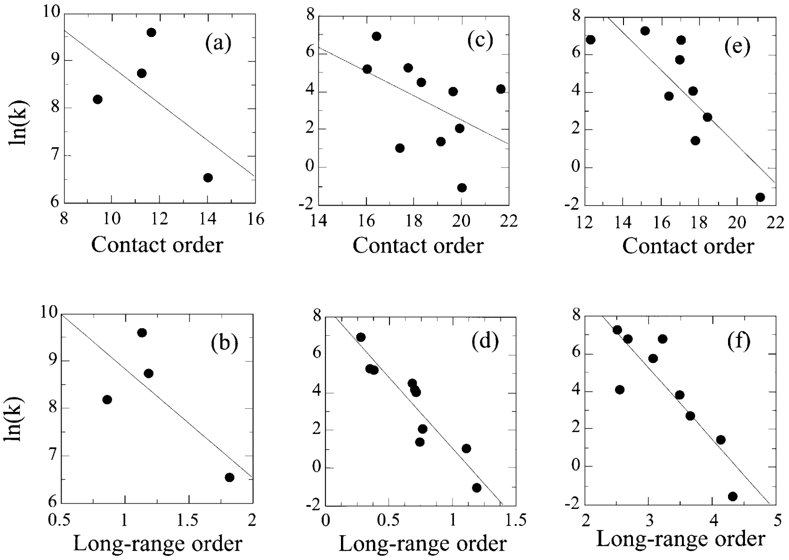

Fig. 10. Comparison between CO and LRO for determining the folding rate in different structural classes (a) CO in all-α ($r = -0.56$), (b) LRO in all-α ($r = -0.72$), (c) CO in all-β ($r = -0.46$), (d) LRO in all-β ($r = -0.92$), (e) CO in mixed ($r = -0.82$) and (f) LRO in mixed-class proteins ($r = -0.86$).

Further, we have performed the jack-knife test to examine the validity of the present method. We found that more than 50% of the considered proteins agreed very well with the experiment and the deviation is less than one unit and the average deviation for all the 23 proteins is 1.782 (Gromiha and Selvaraj, 2001a).

6.3. Other models for predicting protein-folding rates

Debe and Goddard (1999) proposed a first-principles approach for predicting the experimentally determined folding rates. This approach is based on a nucleation–condensation folding mechanism, where the rate-limiting step is a random, diffusive search for the native tertiary topology. To estimate the rates of folding for various proteins via this mechanism, they have first determined the probability of randomly sampling a conformation with the native fold topology. Next, these probabilities have been converted into folding rates by estimating the rate that a protein samples different topologies during diffusive folding.

Munoz and Eaton (1999) used an elementary statistical mechanical model to calculate the protein-folding rates for 22 proteins from their known 3D structures. In this model, residues come into contact only after all of the intervening chain is in the native conformation. An additional simplifying assumption is that native structure grows from localized regions and then fuses to form the complete native molecule. The free energy function for this model contains two contributions, conformational entropy of the backbone and the energy of the inter-residue contacts.

Dinner and Karplus (2001) performed a statistical analysis to predict the protein-folding rates and reported that both CO and stability play important roles in determining the folding rate. Zhou and Zhou (2002) combined the two parameters, CO and LRO and proposed total contact distance for predicting the protein-folding rates. Further, neural networks based models have been suggested to relate folding rates of proteins from the topological parameters, CO and LRO, and the combination of these terms, total contact distance (Zhang et al., 2003).

Recently, Miller et al. (2002) experimentally evaluated the role of structural topology to determine the protein-folding rates and pathways. They have measured the folding rates for a set of circular permutants of the ribosomal protein S6 from *Thermus thermophilus* and estimated the correlation between folding rates and other topological parameters, CO (Plaxco et al., 1998), LRO (Gromiha and Selvaraj, 2001a, b) and fraction of short-range contacts (Mirny and Shakhnovich, 2001). They observed that despite a wide range of relative CO, the permuted proteins all fold with similar rates. On the other hand, LRO and fraction of short-range contacts correlate very well with protein-refolding rates including circular permutations of the ribosomal protein S6 from *Thermus thermophilus* (Miller et al., 2002).

7. Inter-residue interactions in protein-folding kinetics and Φ value analysis

Many small proteins are known to fold rapidly by simple two-state kinetics (Jackson, 1998). The information about the transition state structures at the level of individual residues has been obtained through the protein engineering method and Φ value analysis (Matouschek et al., 1989; Itzhaki et al., 1995).

The Φ value analysis of a protein that folds via a two-state mechanism is illustrated in Fig. 11 (Nolting, 1999a). A mutation causes a change of stability, $\Delta\Delta G_{\mathrm{F-U}}$, between the folded (F) and unfolded states (U). In the transition state, #, the energy difference between mutant and wild type is $\Delta\Delta G_{\#\text{-}\mathrm{U}}$. In this case, the fraction of energy difference, $\Phi_{\#} = \Delta\Delta G_{\#\text{-}\mathrm{U}}/\Delta\Delta G_{\mathrm{F-U}}$, depends on the amount of structure that has built up in the transition state, #, at the position of the mutation.

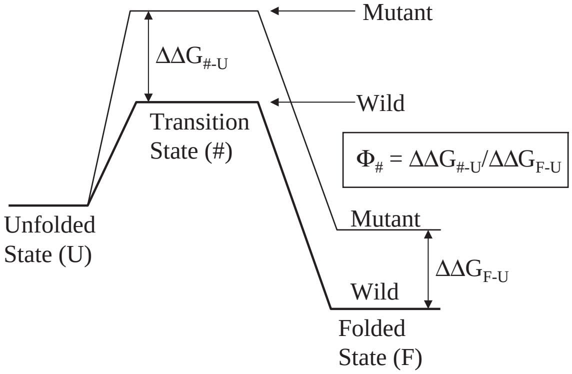
Mutant
$\Delta\Delta G_{\#\text{-}U}$
Wild
Transition
State (#)
$\Phi_{\#} = \Delta\Delta G_{\#\text{-}U}/\Delta\Delta G_{F\text{-}U}$
Mutant
Unfolded
State (U)
$\Delta\Delta G_{F\text{-}U}$
Wild
Folded
State (F)

Fig. 11. Φ value analysis of a protein that folds via a two-state mechanism (Nolting, 1999a).

$\Phi = 0$ corresponds to no structure formation and for the complete formation of structure at the position of the mutation, $\Phi = 1$ (Itzhaki et al., 1995). A set of single mutants, strategically distributed over the molecule is used to map out the structure of the transition state at the resolution of single amino acid residues. This approach may analogously be applied on proteins with more complicated energy landscapes which contain intermediates on the reaction pathway, and on proteins which have residual structure in their unfolded states.

7.1. Experimental studies on Φ value analysis

Several two-state proteins from different folds, such as, α-helical, all-β, β-sandwich and α/β fold have been characterized by Φ value analysis (Jackson, 1998). Acyl coenzyme binding protein (ACBP) belongs to the α-helical fold and in ACBP, eight conserved hydrophobic residues have been identified as critical for fast folding. These residues are located in the interface between N- and C-terminal helices (Kragelund et al., 1999). The SH3 domains of Src, α-spectrin and Fyn belong to β-proteins. The Sso7d protein also has a similar topology to that of SH3 domains (Guerois and Serrano, 2000).

The transition state of srcSH3 domain appears to be quite polarized, with one portion of the molecule highly ordered than the rest; a hydrogen bond network involving two β-turns and an adjacent HC appear to be formed (Grantcharova et al., 1998). There is good qualitative similarity among the transition state structures of srcSH3, α-spectrin and Fyn SH3 (Martinez et al., 1998; Northey et al., 2002). In Fyn SH3 it has further been shown that residues at three positions are tightly packed in the transition state than the rest and these residues are designated as the folding nucleus (Northey et al., 2002).

The transition state structures of TNfn3 (third fibronectin type III domain of human tenascin), Fnfn10 (III domain of human fibronectin) and TI127 (human cardiac titin) belong to the immunoglobulin fold, have been characterized in detail (Hamill et al., 2000; Cota et al., 2001; Fowler and Clarke, 2001). The Φ values indicate that the formation of folding nuclei in all these proteins is due to tertiary interactions between residues in four strands namely B, C, E and F. The residues with high Φ values are clustered together in space in the folded structure.

The 64-residue chymotrypsin inhibitor 2 (CI2) belonging to the α/β fold is one of the well-characterized proteins through Φ value analysis by more than 100 mutations under a variety of conditions (Otzen et al., 1994; Itzhaki et al., 1995). This analysis has revealed that CI2 folds around an extended nucleus that is composed of a contiguous region of structure (α-helix) and long-range interactions with groups distant in sequence (Fersht, 2000).

These experimental studies on Φ value analysis show the presence of several key residues, known as folding nucleus, important for initiating the folding and maintaining the stability. Table 7 provides the list of folding nuclei identified in a set of 17 proteins by protein engineering experiments. Further, Gromiha and Selvaraj (2002b) listed the experimental Φ values obtained for the mutants in CI2, FK506 binding protein (FKBP12) and SH3 domain of src (srcSH3). This information can be used to develop a theoretical model to predict the folding nuclei and for understanding the transition state structures in two-state proteins (Poupon and Mornon, 1999; Galzitskaya and Finkelstein, 1999; Gromiha and Selvaraj, 2002b).

7.2. Relationship between inter-residue contacts and Φ values

The relationship between inter-residue interactions and Φ values upon mutations has been explored by correlation coefficient approach (Gromiha and Selvaraj, 2002b). The amino acid properties representing long-range interactions and experimental Φ values from FK506 binding protein (FKBP12), chymotrypsin inhibitor (CI2) and scrSH3 have been used for the study. The brief descriptions about the amino acid properties along with the numerical values for the 20 amino acid residues have been available in our previous articles (Gromiha et al., 1999b, 2000a).

Table 7
Identified folding nuclei in 17 proteins

Protein name	PDB code	Folding nuclei	Reference
CI2	2ci2	A35, L68, I76	Itzhaki et al. (1995)
Tenascin	1ten	I821, Y837, I860, V871	Hamill et al. (2000)
CD2.d1	1hnf	L19, I21, I33, A45, V83, L96, W35	Lorch et al. (1999)
CheY	3chy	D12, D13, D57, V10, V11, V33, A36, D38, A42, V54	Lopez-Hernandez and Serrano (1996)
ADA2h	1aye	I15, L26, F67, V54, I23	Villegas et al. (1998)
Acp	1aps	Y11, P54, F84, Y25, A30, G45, V47	Chiti et al. (1999)
U1A	1urn	I43, V45, L30, F34, I40, I14, L17, L26	Ternstrom et al. (1999)
ACBP	2abd	F5, A9, V12, L15, Y73, I74, V77, L80	Kragelund et al. (1999)
FKBP12	1fkj	V2, V4, V24, V63, I76, I101, L50	Fulton et al. (1999)
Fyn SH3	1fyn	I28, A39, I50	Northey et al. (2002)
srcSH3	2src	E30, A45, I56	Grantcharova et al. (1998)
α-spectrin	1shg	V23, V44, V53	Martinez et al. (1998)
Sso7d	1bf4	V23, I30, A45, L59	Guerois and Serrano (2000)
WW	1pin	R14, S19, Y23, A31	Jäger et al. (2001)
Titin	1tit	I2, I23, L36, I49, L58, F73	Fowler and Clarke (2001)
Fnfn	1ttf	L8, I20, I34, V50, A74, Y92	Cota et al. (2001)
Villin	2vik	V7, I18, I23, M28, C44, I61, A77, T81, M84	Choe et al. (2000)

Note: The data for the first nine proteins were taken from Mirny and Shakhnovich (2001).

We found that the inter-residue interactions are important for determining the transition state structures of partially buried and exposed mutants of two-state proteins. Further, Nolting (1998, 1999b) mapped the Φ values of barnase and chymotrypsin inhibitor to inter-residue contact plots and observed a good correlation between inter-residue contacts and Φ values.

7.2.1. Partially buried mutations

The changes in each of the amino acid property values upon partially buried mutations have been computed using the equation, $P_{\mathrm{seq}}(i) = [\sum_{j=i-k}^{j=i+k} P_j(i)] - P_{\mathrm{mut}}(i)$, where, $P_{\mathrm{mut}}(i)$ is the property value of the ith mutant residue and $\Sigma P_j(i)$ is the total property value of the segment of $(2k+1)$ residues ranging from $i-k$ to $i+k$ about the ith residue of wild type. We found that the inclusion of nearest neighboring residue information improves the correlation between $P_{\mathrm{seq}}(i)$ and Φ values with an increase of 13% in FKBP12. In Table 8, we present the correlation coefficient obtained for seven properties reflecting short-, medium- and long-range interactions with experimental Φ values in FKBP12 (Gromiha and Selvaraj, 2002b). We found that the short- and medium-range energy (E_{sm}) strongly correlated with Φ values, expressing the influence of medium-range interactions in the transition state structures of partially buried mutants. In Fig. 12, we show the direct relationship between E_{sm} and Φ and the correlation coefficient is 0.72.

Table 8
Single property correlation with Φ values in the mutants of two-state proteins

Property	Correlation coefficient, r	
	Partially buried	Exposed
FK506 binding protein		
H_p	−0.30	0.69
E_{sm}	**0.72**	0.05
E_l	−0.36	0.67
P_β	−0.15	0.76
N_s	−0.21	0.76
N_l	−0.13	**0.79**
H_{gm}	−0.15	0.64
Chymotrypsin inhibitor		
H_p	0.26	0.80
E_{sm}	**0.72**	−0.11
N_s	0.15	0.79
N_l	0.09	0.76
H_{gm}	0.28	**0.81**
srcSH3 domain		
H_p	0.73	**0.76**
E_{sm}	0.66	−0.05
P_β	**0.87**	0.63
N_s	0.77	0.74
N_l	0.85	0.66
H_{gm}	0.72	0.74

Note: The highest correlation coefficient in each set of data is shown in bold (Gromiha and Selvaraj, 2002b).

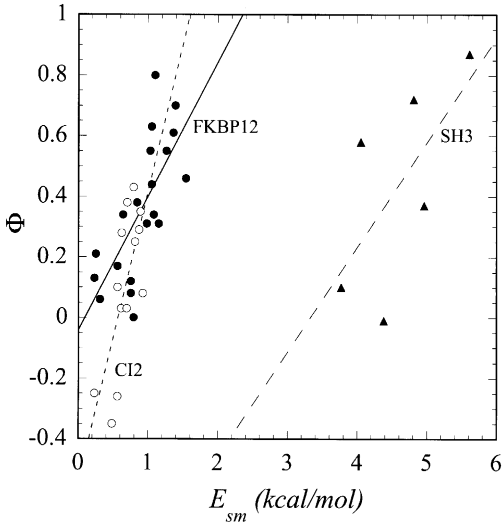

Fig. 12. Relationship between experimental Φ values and E_{sm} in partially buried mutations of two-state proteins. The symbols ●, ○ and ▲ represent FKBP12, CI2 and SH3 mutants and their respective r-values are 0.72, 0.72 and 0.66 (Gromiha and Selvaraj, 2002b).

Our further analysis on CI2 and srcSH3 showed that the inclusion of nearest neighboring residue information improves the correlation of property values with Φ and the correlation coefficients for selected properties are also included in Table 8. As observed in FKBP12, the property E_{sm} has the highest correlation ($r = 0.72$) with Φ in the helical mutants of CI2, which reveals the importance of medium-range interactions in the formation of transition state structures. Further, an appreciable correlation ($r = 0.66$) was observed between E_{sm} and Φ in srcSH3. The relationship between E_{sm} and Φ in CI2 and srcSH3 is shown in Fig. 12.

7.2.2. Exposed mutations

In exposed mutations, we found that the inclusion of sequence information using nine-residue window length remarkably improved the correlation (20%) with Φ values in FKBP12. The correlation coefficient obtained for the selected amino acid properties by the inclusion of information from nine-residue window in the sequence with Φ values of FKBP12 is presented in Table 8. We found that the average long-range contact (N_{l}) has the strongest correlation with Φ ($r = 0.79$). Other properties reflecting long-range interactions such as long-range non-bonded energy, average number of surrounding residues and β-strand tendency show significant correlation with Φ values. It is noteworthy that the hydrophobicity scales including the effect of surrounding residues have a good correlation, ranging from 0.65 to 0.75.

Similar analysis has been carried out for CI2 and srcSH3 and the r-values are included in Table 8. We found that the properties reflecting long-range interactions have a strong correlation with Φ ($r > 0.75$) in both CI2 and srcSH3. This result is consistent with experimental observations that the long-range contacts play an important role in the formation of folding nucleus (Itzhaki et al., 1995; Grantcharova et al., 1998).

8. Influence of inter-residue interactions to protein stability

The importance of inter-residue interactions to protein stability has been viewed from several perspectives, such as (i) relating amino acid properties representing inter-residue interactions to protein mutant stability, (ii) predicting protein mutant stability from effective potentials, (iii) estimating the contribution of non-covalent interactions to protein stability and (iv) delineating the importance of inter-residue interactions to the extreme stability of thermophilic proteins. In this section, we discuss these approaches in detail.

8.1. *Amino acid properties and protein mutant stability*

We have analyzed the relationship between several amino acid properties and protein stability upon mutations. The stability data have been obtained from our thermodynamic database for proteins and mutants, ProTherm available at the web, http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html (Gromiha et al., 1999c, 2000a, 2002a). We found that the properties representing inter-residue interactions have good correlation with the stability of protein mutants (Gromiha et al., 1999a, b, 2000b, 2002b; Gromiha, 2003d).

8.1.1. *Buried, partially buried and exposed mutations*

The properties reflecting inter-residue interactions, long-range non-bonded energy (E_l) and number of surrounding residues (N_s) have significant correlation with stability in buried and partially buried mutants, respectively. Interestingly, N_s has the highest correlation with $\Delta\Delta G$ for the subset of hydrophobic mutations (Gromiha et al., 1999b). In the coil region of surface mutations, N_l (number of long-range contacts) has the second topmost negative correlation ($r = -0.67$) with $\Delta\Delta G$ (Gromiha et al., 2002b). Hence, the properties reflecting inter-residue interactions play an essential role to the stability of buried, partially buried and exposed protein mutants. Further, we observed the inclusion of sequence (neighboring residue information) and structural information (residues which are close to each other within the limit of 8 Å) improved the correlation between amino acid properties and protein stability upon partially buried and exposed mutations. This result emphasizes the importance of inter-residue interactions for the stability of protein structures.

8.1.2. *Secondary structural regions*

The relationship between amino acid properties and protein stability at different secondary structural regions showed that the properties representing inter-residue interactions, E_l, E_t (total non-bonded energy) and N_s have good correlation in helical mutants and E_l, N_s and N_l have significant correlation with stability for mutations in strand segments of protein core (Gromiha et al., 1999a, b). It is noteworthy that E_l has the highest correlation with $\Delta\Delta G$ and all the three properties have good correlation with stability are influenced by long-range interactions. The properties N_s and N_l have the highest negative correlation with stability in the coil regions of protein surface (Gromiha et al., 2002b).

8.1.3. Effect of sequence information

We examined the effect of sequence window length from 1 to 12 residues on each side of the mutant residue. The highest correlation coefficients obtained using sequence information at different window lengths in helical, strand and coil mutations are presented in Table 9 (Gromiha et al., 2000b). We observed that each secondary structure preferred a specific window length for obtaining the highest correlation with stability. The preferred window lengths for the secondary structures, helix, strand and coil, are, respectively, 0, 9 and 4 residues on both sides of the mutant residue. The preference of nine residues on both sides of the mutants in strand segments includes the information from neighboring strands, which are close to each other in 3D structures. The information from four residues on both sides of the mutants in coil regions includes the short- and medium-range interactions.

8.1.4. Influence of structural information

We have analyzed the influence of structural information by computing the correlation coefficient at various contacting distances, from 4 to 20 Å around the mutated residue and the results for helical, strand and coil mutations are presented in Fig. 13 (Gromiha et al., 2000b). The observed highest correlation coefficient for each set of mutations showed that all the mutations prefer the distance of 6–8 Å and specifically, helical mutations prefer 8 Å; strand mutations, 6 Å and coil mutations, 7 Å. The distance of 6–8 Å is sufficient to accommodate the nearest neighboring residues and several residues that are far in sequence level.

8.2. Effective potentials and protein mutant stability

The information about inter-residue interactions in protein structures has been successfully used to predict the stability of protein mutants. Gilis and Rooman (1996) analyzed a set of 106

Table 9
Correlation coefficients for different sequence window lengths in partially buried helical, strand and coil mutations

Window length	Correlation coefficient		
	Helix	Strand	Coil
0	0.61	0.56	0.41
1	0.50	0.64	0.56
2	0.29	0.50	0.47
3	0.31	0.66	0.51
4	0.30	0.61	0.56
5	0.26	0.66	0.41
6	0.28	0.65	0.34
7	0.28	0.60	0.40
8	0.23	0.66	0.47
9	0.30	0.68	0.44
10	0.32	0.60	0.37
11	0.35	0.56	0.43
12	0.31	0.48	0.38

Note: Data from Gromiha et al. (2000b).

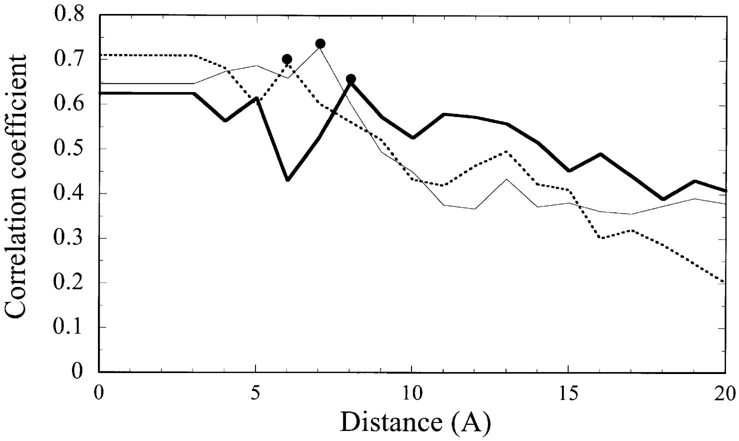

Fig. 13. Single correlation coefficients obtained for different distances from the mutant residue: thick line, helix; dash line, strand; and thin line, coil (Gromiha et al., 2000b).

surface mutations and reported that the correlation coefficient between experimental and computed free energy change using backbone torsion potentials (see Section 4.1.1) is 0.67 for all the mutations and it raised up to 0.87 for a subset of 96 mutations. This torsion potential is mainly based on the neighboring residues in a sequence and hence the local interactions along the chain are more dominant than hydrophobic interactions at the protein surface (Gilis and Rooman, 1996). Further, the buried and partially buried mutations have been systematically analyzed with distance potentials (see Section 4.4.2; Kocher et al., 1994; Gilis and Rooman, 1997). They reported that for a set of completely buried mutations, the combination of distance potential and torsion potential weighted by a factor of 0.4 yielded the correlation coefficient of 0.80 between the computed and measured changes in folding free energy. For mutations of partially buried residues, the best potential is a combination of torsion potential and a distance potential weighted by a factor of 0.7 and the correlation coefficient is 0.82. These results indicate that the distance potentials, dominated by hydrophobic interactions, represent best the main interactions stabilizing the protein core.

Topham et al. (1997) proposed an approach based on the knowledge of amino acid replacements that are tolerated within the families of homologous proteins of known 3D structure. Amino acid variations in families of homologous proteins are converted into propensity and substitution tables. These tables provide quantitative information about the existence of an amino acid in a structural environment and the probability of replacement by any amino acids and this approach is similar to the development of knowledge-based potentials. Hence, the information about local and long-range contacts is included in these environment dependent propensity and substitution tables. It has been reported that the stability score obtained with these tables to a set of 159 mutations in T4 lysozyme showed a correlation of 0.83 with differences in melting temperatures. Further, 86% of the mutants were correctly classified as stabilizing or destabilizing.

Our analysis showed that the information about local protein sequence and structural effects were more important for the prediction of stability changes caused by partially buried and surface mutations than for buried mutations. Further, we observed a direct correlation between solvent accessibility and raise in correlation due to the inclusion of sequence and structural information (Gromiha et al., 1999a, b, 2000b, 2002b).

8.3. *Non-covalent interactions and protein stability*

The inter-residue interactions in protein structures can be viewed in terms of major non-covalent interactions. We have computed the stability factors, hydrophobic, hydrogen bonds, ion-pairs, disulfide bonds and van der Waals interactions from the 3D structures of proteins and translated them into free energy contributions to the stability of the native state (Ponnuswamy and Gromiha, 1994). We found that the hydrophobic factor is a dominant force and hydrogen bonds do contribute significantly to the stability of the native state. Interestingly, hydrophobic interactions are mainly dominated by long-range interactions and hydrogen bonds are dictated by both medium- and long-range interactions due to the formation of α-helices and β-strands, respectively. This result shows that the inter-residue interactions contain the necessary information about all physical interactions responsible for protein folding and to maintain the stability of folded proteins.

8.4. *Extreme stability of thermophilic proteins*

Comprehending the mechanisms responsible for enhancing the thermostability of proteins is an interesting topic of research. The inter-residue interactions play an important role to the stability via several physical and non-covalent interactions. Querol et al. (1996) analyzed the influence of protein conformational characteristics to thermal stability. Vogt et al. (1997) reported that the increase in number of hydrogen bonds and number of ion pairs increases the stability. Facchiano et al. (1998) suggested the importance of intrinsic helical propensities, residue–residue interactions and helix–dipole interactions for the stability. Our studies showed that the number of long-range contacts and long-range interaction energy increases the stability apart from negative free energy of hydration and shape (Gromiha et al., 1999d). Further, several investigations have been carried out to stress the importance of ion-pairs, salt bridges and electrostatic interactions (Karshikoff and Ladenstein, 1998; Xiao and Honig, 1999; Kumar et al., 2000; Szilagyi and Zavodszky, 2000). Recently, we have analyzed the effect of inter-residue contacts and cation–π interactions to the stability of thermophilic proteins (Gromiha, 2001, 2002; Gromiha et al., 2002c). The details about the importance of different interactions to the thermal stability of thermophilic proteins have been extensively reviewed in the literature (Jaenicke and Bohm, 1998; Ladenstein and Antranikian, 1998; Kumar and Nussinov, 2001). We infer from these studies that the stability of proteins has been dictated with inter-residue interactions.

9. Future perspectives

The foregoing studies have revealed the vital role of inter-residue interactions in determining the structure, folding and stability of proteins. The recent advances achieved by various workers in this field have opened the way for further studies. In view of the significant differences in the contribution of inter-residue interactions in (i) different structural classes of globular proteins, (ii) mesophilic and thermophilic proteins, (iii) globular and membrane proteins and (iv) proteins of different folds, it may be possible to develop promising methods to discriminate proteins of different kinds. These approaches can be successfully used to extract specific proteins from genome sequences and hence understanding the function.

The folding architecture of proteins depends on the interactions between amino acid residues. The favorable amino acid residue pairs and amino acid clusters in protein structures provide deep insight to our understanding about the initiating elements upon protein folding and the important contacts for stabilizing the structure. Hence, the potentials derived from the knowledge of inter-residue contacts can be effectively used for protein structure prediction and fold recognition. In secondary structure prediction, the inclusion of long-range interactions may improve the accuracy of identifying β-strand segments. On the other hand, the prediction of solvent accessibility, inter-residue contacts, discrimination of correct folds from incorrect folds based on the information from inter-residue interactions can be utilized for predicting the 3D structure of proteins.

Recent experimental studies show the strong relationship between folding rate and topology of the native state. The topology of native state protein structure has been demonstrated with inter-residue contacts in terms of CO and LRO and hence it will be possible to predict the protein-folding rates using residue contacts at high accuracy. The folding rate of proteins reveals the mechanism of proteins folding in terms of the prior formation α-helices/β-strands during the process of protein folding. On the other hand, the information about inter-residue contacts is very useful to understand the transition state structures of proteins and the formation of α-helices, β-strands, hydrophobic core, etc. during protein folding. Hence, the combination of thermodynamic and kinetic experiments along with the knowledge of inter-residue contacts explores the mechanism of protein folding and to tackle the protein-folding problem.

The stability of proteins due to mutations mainly based on the nature of substitutions, which can be understood from the knowledge of inter-residue contacts at various environments, locations and secondary structures. Hence, the potentials developed from inter-residue interactions, taking into account of short-, medium- and long-range contacts, helix, strand, turn and coil segments, buried, partially buried and exposed regions, and the information about the preference of each amino acid residue surrounded by all the 20 amino acid residues due to medium- and long-range interactions will be helpful to understand the stability of proteins due to mutations. Further, these potentials can be used for predicting the stability of protein mutants, which could be used for protein engineering experiments for selecting the potential candidates.

Recent analysis shows the difference between thermophilic proteins and their mesophilic counterparts in terms of various interactions and residue contacts. The physical interactions, such as ion pairs, van der Waals, hydrophobic and hydrogen bonds reflect the influence of inter-residue contacts for the extreme stability of thermophilic proteins. Hence, the information gained from the analysis of inter-residue interactions may help to understand the mechanism for the stability of proteins at extreme environments.

It is envisaged that the insights obtained from the analysis of inter-residue interactions may be helpful in de novo protein design.

10. Conclusions

The environment around each residue in a globular protein as defined in a sphere of 8 Å radius has been effectively partitioned as comprising of residues that contribute to short-, medium- and long-range interactions. The medium-range interactions predominate in all-α class and the long-range interactions predominate in all-β class proteins. Accordingly, the performance of several structure prediction methods in different structural classes showed that all the methods predict the secondary structures of all-α proteins more accurately than other classes. Further, the preference of residue pairs influenced by medium- and long-range contacts in all structural classes globular and membrane proteins are delineated. This revealed the distinct features of interacting pattern in the formation of residue pairs in different environments. Hence, the information about inter-residue interactions will be very helpful for discriminating globular and membrane proteins and improving the secondary structure predictive accuracy of proteins.

Several parameters and potentials have been derived based on the information from short-, medium- and long-range interactions in proteins. The effective inter-residue contact potentials have been used for predicting the residue–residue contacts, domains, protein fold and sequence recognition, and three-dimensional structures of proteins. Further, the residue–residue contacts derived from distance criteria are very helpful to identify the key residues for protein folding.

The concept of contact order and long-range order, reflecting the importance of local and non-local interactions correlates well with folding rate of small proteins. Further, inter-residue interactions play an essential role for determining the transition state structures of two-state proteins.

The properties reflecting inter-residue interactions have good correlation with protein stability. The long-range interactions are very important to predict the stability of mutants in strand segments and other secondary structural regions are also influenced by such interactions. The effective role of inter-residue interactions is observed in predicting the stability of protein mutants in the core and at surface with significant accuracy. The torsion and distance potentials derived from the information about medium- and long-range interactions predicted the stability of protein mutants successfully.

In essence, the three-dimensional structures of proteins are dictated by short-, medium- and long-range interactions, and the information about inter-residue interactions in proteins will be very helpful to predict the secondary and tertiary structures of proteins and to understand their folding and stability.

Acknowledgements

We thank Professor Tom Blundell for providing an opportunity to contribute this review. MMG wishes to acknowledge Dr. Yutaka Akiyama and Dr. Makiko Suwa, and SS thanks Prof. M. Lakshmanan for their continuous encouragement.

References

Ahmad, S., Gromiha, M.M., 2002. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18, 819–824.

Ahmad, S., Gromiha, M.M., 2003. Design and training of a neural network for predicting the solvent accessibility of proteins. J. Comp. Chem. 24, 1313–1320.

Ahmad, S., Gromiha, M.M., Sarai, A., 2003a. Real value prediction of solvent accessibility from amino acid sequence. Proteins 50, 629–635.

Ahmad, S., Gromiha, M.M., Sarai, A., 2003b. RVP-Net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics 19, 1849–1851.

Andersson, H., Bakker, E., von Heijne, G., 1992. Different positively charged amino acids have similar effects on the topology of a polytopic transmembrane protein in *Escherichia coli*. J. Biol. Chem. 267, 1491–1495.

Bahar, I., Jernigan, R.L., 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J. Mol. Biol. 266, 195–214.

Bahar, I., Kaplan, M., Jernigan, R.L., 1997. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins 29, 292–308.

Baker, D., 2000. A surprising simplicity to protein folding. Nature 405, 39–42.

Barlow, D.J., Thornton, J.M., 1983. Ion-pairs in proteins. J. Mol. Biol. 168, 867–885.

Bowie, J.U., Luthy, R., Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.

Chandonia, J.-M., Karplus, M., 1996. The importance of larger data sets for protein secondary structure prediction with neural networks. Protein Sci. 5, 768–774.

Chirgadze, Y.N., Larionova, E.A., 1999. Spatial sign-alternating charge clusters in globular proteins. Protein Eng. 12, 101–105.

Chiti, F., Taddei, N., White, P.M., Bucciantini, M., Magherini, F., Stefani, M., Dobson, C.M., 1999. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. Nat. Struct. Biol. 6, 1005–1009.

Chiu, T.L., Goldstein, R.A., 1998. Optimizing energy potentials for success in protein tertiary structure prediction. Fold. Des. 3, 223–228.

Choe, S.E., Li, L., Matsudaira, P.T., Wagner, G., Shakhnovich, E.I., 2000. Differential stabilization of two hydrophobic cores in the transition state of the villin 14T folding reaction. J. Mol. Biol. 304, 99–115.

Chou, P.Y., Fasman, G.D., 1974. Prediction of protein conformation. Biochemistry 13, 222–245.

Cota, E., Steward, A., Fowler, S.B, Clarke, J., 2001. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. J. Mol. Biol. 305, 1185–1194.

Debe, D.A., Goddard, W.A., 1999. First principles prediction of protein folding rates. J. Mol. Biol. 294, 619–625.

Deber, C.M., Glibowicka, M., Woolley, G.A., 1990. Conformations of proline residues in membrane environments. Biopolymers 29, 149–157.

Dinner, A.R., Karplus, M., 2001. The roles of stability and contact order in determining protein folding rates. Nat. Struct. Biol. 8, 21–22.

Dosztanyi, Z., Fiser, A., Simon, I., 1997. Stabilization centers in proteins: identification, characterization and predictions. J. Mol. Biol. 272, 597–612.

Drablos, F., 1999. Clustering of non-polar contacts in proteins. Bioinformatics 15, 501–509.

Facchiano, A.M., Colonna, G., Ragone, R., 1998. Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. Protein Eng. 11, 753–760.

Fariselli, P., Olmea, O., Valencia, A., Casadio, R., 2001. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins S5, 157–162.

Fasman, G.D., 1989. The development of the prediction of protein structure. In: Fasman, G.D. (Ed.), Prediction of Protein Structure and Principles of Protein Conformation. Plenum Press, New York, pp. 193–316.

Fersht, A.R., 2000. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc. Natl. Acad. Sci. USA 97, 1525–1529.

Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., Sippl, M.J., 1995. Progress in fold recognition. Proteins 23, 376–386.

Fowler, S.B., Clarke, J., 2001. Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. Structure 9, 355–366.

Fulton, K.F., Main, E.R., Daggett, V., Jackson, S.E., 1999. Mapping the interactions present in the transition state for unfolding/folding of FKBP12. J. Mol. Biol. 291, 445–461.

Furuichi, E., Koehl, P., 1998. Influence of protein structure databases on the predictive power of statistical pair potentials. Proteins 31, 139–149.

Galzitskaya, O.V., Finkelstein, A.V., 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc. Natl. Acad. Sci. USA 96, 11299–11304.

Garnier, J., Osguthorpe, D.J., Robson, B., 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120, 97–120.

Gilis, D., Rooman, M., 1996. Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. J. Mol. Biol. 257, 1112–1126.

Gilis, D., Rooman, M., 1997. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J. Mol. Biol. 272, 276–290.

Gobel, U., Sander, C., Schneider, R., Valencia, A., 1994. Correlated mutations and residue contacts in proteins. Proteins 18, 309–317.

Grantcharova, V.P., Riddle, D.S., Santiago, J.V., Baker, D., 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nat. Struct. Biol. 5, 714–720.

Gromiha, M.M., 1999. A simple method for predicting transmembrane α helices with better accuracy. Protein Eng. 12, 557–561.

Gromiha, M.M., 2001. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. Biophys. Chem. 91, 71–77.

Gromiha, M.M., 2002. Influence of cation–pi interactions in mesophilic and thermophilic proteins. J. Liquid Chromat. Rel. Tech. 25, 3139–3147.

Gromiha, M.M., 2003a. Influence of cation–π interactions in different folding types of membrane proteins. Biophys. Chem. 103, 251–258.

Gromiha, M.M., 2003b. Inter-residue interactions in protein structures. J. Biophys. Soc. Jpn. 43, 87–92.

Gromiha, M.M., 2003c. Importance of native-state topology for determining the folding rate of two-state proteins. J. Chem. Inf. Comp. Sci. 43, 1481–1485.

Gromiha, M.M., 2003d. Factors influencing the thermal stability of buried protein mutants. Polymer 44, 4061–4066.

Gromiha, M.M., Ponnuswamy, P.K., 1993. Prediction of transmembrane beta-strands from hydrophobic characteristics of proteins. Int. J. Pept. Protein Res. 42, 420–431.

Gromiha, M.M., Ponnuswamy, P.K., 1995. Prediction of protein secondary structures from their hydrophobic characteristics. Int. J. Peptide Protein Res. 45, 225–240.

Gromiha, M.M., Ponnuswamy, P.K., 1996. Hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins. Int. J. Pept. Protein Res. 48, 452–460.

Gromiha, M.M., Selvaraj, S., 1997a. Influence of medium and long range interactions in different structural classes of globular proteins. J. Biol. Phys. 23, 151–162.

Gromiha, M.M., Selvaraj, S., 1997b. Influence of medium and long range interactions in $(\alpha/\beta)_8$ barrel proteins. J. Biol. Phys. 23, 209–217.

Gromiha, M.M., Selvaraj, S., 1998. Protein secondary structure prediction in different structural classes. Protein Eng. 11, 249–251.

Gromiha, M.M., Selvaraj, S., 1999a. Importance of long-range interactions in protein folding. Biophys. Chem. 77, 49–68.

Gromiha, M.M., Selvaraj, S., 1999b. Influence of medium and long range interactions in protein folding. Prep. Biochem. Biotech. 29, 339–351.

Gromiha, M.M., Selvaraj, S., 1999c. Amino acid clustering pattern and medium and long-range interactions in $(\alpha/\beta)_8$ barrel proteins. Period. Biolog. 101, 333–338.

Gromiha, M.M., Selvaraj, S., 2000. Inter-residue interactions in the structure, folding and stability of proteins. Recent Res. Dev. Biophys. Chem. 1, 1–14.

Gromiha, M.M., Selvaraj, S., 2001a. Comparison between long-range interactions and contact order in determining the folding rates of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. 310, 27–32.

Gromiha, M.M., Selvaraj, S., 2001b. Role of medium and long-range interactions in discriminating globular and membrane proteins. Int. J. Biol. Macromol. 29, 25–34.

Gromiha, M.M., Selvaraj, S. (Eds.), 2002a. Recent Research Developments in Protein Folding, Stability and Design. Research Signpost, Trivandrum, India.

Gromiha, M.M., Selvaraj, S., 2002b. Important amino acid properties for determining the transition state structures of two-state protein mutants. FEBS Lett. 526, 129–134.

Gromiha, M.M., Suwa, M., 2003. Variation of amino acid properties in all-β globular and outer membrane protein structures. Int. J. Biol. Macromol. 32, 93–98.

Gromiha, M.M., Thangakani, A.M., 2001. Role of medium- and long-range interactions to the stability of the mutants of T4 lysozyme. Prep. Biochem. Biotech. 31, 217–227.

Gromiha, M.M., Majumdar, R., Ponnuswamy, P.K., 1997. Identification of membrane spanning beta strands in bacterial porins. Protein Eng. 10, 497–500.

Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., Sarai, A., 1999a. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. Protein Eng. 12, 549–555.

Gromiha, M.M., Oobatake, M., Kono, H., Uedeira, H., Sarai, A., 1999b. Relationship between amino acid properties and protein stability: buried mutations. J. Protein Chem. 18, 565–578.

Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedeira, H., Sarai, A., 1999c. ProTherm: thermodynamic database for proteins and mutants. Nucl. Acids Res. 27, 286–288.

Gromiha, M.M., Oobatake, M., Sarai, A., 1999d. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem. 82, 51–867.

Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedeira, H., Prabakaran, P., Sarai, A., 2000a. ProTherm, version 2.0: thermodynamic database for proteins and mutants. Nucl. Acids Res. 28, 283–285.

Gromiha, M.M., Oobatake, M., Kono, H., Uedeira, H., Sarai, A., 2000b. Importance of surrounding residues for protein stability of partially buried mutations. J. Biomol. Struct. Dyn. 18, 281–295.

Gromiha, M.M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P., Sarai, A., 2002a. ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. Nucl. Acids Res. 30, 301–302.

Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., Sarai, A., 2002b. Importance of mutant position in Ramachandran plot for predicting protein stability of surface mutations. Biopolymers 64, 210–220.

Gromiha, M.M., Thomas, S., Santhosh, C., 2002c. Role of cation–pi interactions to the stability of thermophilic proteins. Prep. Biochem. Biotechnol. 32, 355–362.

Guerois, R., Serrano, L., 2000. The sh3-fold family: experimental evidence and prediction of variations in folding pathways. J. Mol. Biol. 304, 967–982.

Gugolya, Z., Dosztanyi, Z., Simon, I., 1997. Interresidue interactions in protein classes. Proteins 27, 360–366.

Hamill, S.J., Steward, A., Clarke, J., 2000. The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. J. Mol. Biol. 297, 165–178.

Heringa, J., Argos, P., 1991. Side-chain clusters in protein structures and their role in protein folding. J. Mol. Biol. 220, 151–171.

Islam, S.A., Luo, J., Sternberg, M.J., 1995. Identification and analysis of domains in proteins. Protein Eng. 8, 513–525.

Ito, M., Matsuo, Y., Nishikawa, K., 1997. Prediction of protein secondary structure using the 3D-1D compatibility algorithm. Comput. Appl. Biosci. 13, 415–423.

Itzhaki, L.S., Otzen, D.E., Fersht, A.R., 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation–condensation mechanism for protein folding. J. Mol. Biol. 254, 260–288.

Jackson, S.E., 1998. How do small single proteins fold? Fold. Des. 3, R81–R91.

Jaenicke, R., Bohm, G., 1998. The stability of proteins in extreme environments. Curr. Opin. Struct. Biol. 8 (6), 738–748.

Jäger, M., Nguyen, H., Crane, J.C., Kelley, J.W., Gruebele, M.W., 2001. The folding mechanism of a β-sheet: the WW domain. J. Mol. Biol. 311, 373–393.

Jewett, A.I., Pande, V.S., Plaxco, K.W., 2003. Cooperativity, smooth energy landscapes and the origins of topology-dependent protein folding rates. J. Mol. Biol. 326, 247–253.

Jiang, Z., Zhang, L., Chen, J., Xia, A., Zhao, D., 2002. Effect of amino acid on forming residue–residue contacts in proteins. Polymer 43, 6037–6047.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. A new approach to protein fold recognition. Nature 358, 86–89.

Kannan, N., Vishveshwara, S., 1999. Identification of side-chain clusters in protein structures by a graph spectral method. J. Mol. Biol. 292, 441–464.

Kannan, N., Selvaraj, S., Gromiha, M.M., Vishveshwara, S., 2001. Clusters in alpha/beta barrel proteins: implications for protein structure, function, and folding: a graph theoretical approach. Proteins 43, 103–112.

Karlin, S., Zhu, Z.Y., 1996. Characterizations of diverse residue clusters in protein three-dimensional structures. Proc. Natl. Acad. Sci. USA 93, 8344–8349.

Karlin, S., Zuker, M., Brocchieri, L., 1994. Measuring residue associations in protein structures. Possible implications for protein folding. J. Mol. Biol. 239, 227–248.

Karshikoff, A., Ladenstein, R., 1998. Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. Protein Eng. 11, 867–872.

Kneller, D.G., Cohen, F.E., Langridge, R., 1990. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214, 171–182.

Kocher, J.P., Rooman, M.J., Wodak, S.J., 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. Biol. 235, 1598–1613.

Kragelund, B.B., Osmark, P., Neergaard, T.B., Schiodt, J., Kristiansen, K., Kundsen, J., Poulsen, F.M., 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate-limiting in two-state protein folding of ACBP. Nature Struct. Biol. 6, 594–601.

Kumar, S., Nussinov, R., 2001. How do thermophilic proteins deal with heat? Cell Mol. Life Sci. 58, 1216–1233.

Kumar, S., Tsai, C.J., Nussinov, R., 2000. Factors enhancing protein thermostability. Protein Eng. 13, 179–191.

Kumarevel, T.S., Gromiha, M.M., Ponnuswamy, M.N., 1998. Analysis of hydrophobic and charged patches and influence of medium- and long-range interactions in molecular chaperones. Biophys. Chem. 75, 105–113.

Kumarevel, T.S., Gromiha, M.M., Ponnuswamy, M.N., 2000. Structural class prediction: an application of residue distribution along the sequence. Biophys. Chem. 88, 81–101.

Kumarevel, T.S., Gromiha, M.M., Selvaraj, S., Gayatri, K., Kumar, P.K., 2002. Influence of medium- and long-range interactions in different folding types of globular proteins. Biophys. Chem. 99, 189–198.

Ladenstein, R., Antranikian, G., 1998. Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. Adv. Biochem. Eng. Biotechnol. 61, 37–85.

Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.

Lopez-Hernandez, E., Serrano, L., 1996. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. Fold. Des. 1, 43–55.

Lorch, M., Mason, J.M., Clarke, A.R., Parker, M.J., 1999. Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state. Biochemistry 38, 1377–1385.

MacArthur, M.W., Thornton, J.M., 1991. Influence of proline residues on protein conformation. J. Mol. Biol. 218, 397–412.

Makarov, D.E., Plaxco, K.W., 2003. The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. Protein Sci. 12, 17–26.

Makarov, D.E., Keller, C.A., Plaxco, K.W., Metiu, H., 2002. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc. Natl. Acad. Sci. USA 99, 3535–3539.

Manavalan, P., Ponnuswamy, P.K., 1977. A study of the preferred environment of amino acid residues in globular proteins. Arch. Biochem. Biophys. 184, 476–487.

Manavalan, P., Ponnuswamy, P.K., 1978. Hydrophobic character of amino acid residues in globular proteins. Nature 275, 673–674.

Martinez, J.C., Pisabora, M.T., Serrano, L., 1998. Obligatory steps in protein folding and the conformational diversity of the transition state. Nature Struct. Biol. 5, 721–729.

Matouschek, A., Kellis Jr., J.T., Serrano, L., Fersht, A.R., 1989. Mapping the transition state and pathway of protein folding by protein engineering. Nature 340, 122–126.

Matthews, B.W., 1993. Structural and genetic analysis of protein stability. Ann. Rev. Biochem. 62, 139–160.

Miller, E.J., Fischer, K.F., Marqusee, S., 2002. Experimental evaluation of topological parameters determining protein-folding rates. Proc. Natl. Acad. Sci. USA 99, 10359–10363.

Mirny, L., Shakhnovich, E., 2001. Protein folding theory: from lattice to all-atom models. Annu. Rev. Biophys. Biomol. Struct. 30, 361–396.

Mirny, L.A., Shakhnovich, E.I., 1996. How to derive a protein folding potential? A new approach to an old problem. J. Mol. Biol. 264, 1164–1179.

Miyazawa, S., Jernigan, R.L., 1985. Estimation of interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18, 534–552.

Miyazawa, S., Jernigan, R.L., 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol. 256, 623–644.

Miyazawa, S., Jernigan, R.L., 1999a. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. Proteins 36, 347–356.

Miyazawa, S., Jernigan, R.L., 1999b. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins 36, 357–369.

Munoz, V., Eaton, W.A., 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc. Natl. Acad. Sci. USA 96, 11311–11316.

Nolting, B., 1998. Structural resolution of the folding pathway of a protein by correlation of phi-values with inter-residue contacts. J. Theor. Biol. 194, 419–428.

Nolting, B., 1999a. Protein Folding Kinetics. Springer, Heidelberg.

Nolting, B., 1999b. Analysis of the folding pathway of chymotrypsin inhibitor by correlation of phi-values with inter-residue contacts. J. Theor. Biol. 197, 113–121.

Northey, J.G., Di Nardo, A.A., Davidson, A.R., 2002. Hydrophobic core packing in the SH3 domain folding transition state. Nature Struct. Biol. 9, 126–130.

Oobatake, M., Crippen, G.M., 1981. Residue–residue potential function for conformational Analysis of proteins. J. Phys. Chem. 85, 1187–1197.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH—a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

Otzen, D.E., Itzhaki, L.S., ElMasry, N.F., Jackson, S.E., Fersht, A.R., 1994. Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. Proc. Natl. Acad. Sci. USA 91, 10422–10425.

Ouzounis, C., Sander, C., Scharf, M., Schneider, R., 1993. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. J. Mol. Biol. 232, 805–825.

Pace, C.N., Shirley, B.A., McNutt, M., Gajiwala, K., 1996. Forces contributing to the conformational stability of proteins. FASEB J. 10, 75–83.

Park, B.H., Huang, E.S., Levitt, M., 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J. Mol. Biol. 266, 831–846.

Plaxco, K.W., Simons, K.T., Baker, D., 1998. Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 277, 985–994.

Pollastri, G., Baldi, P., Fariselli, P., Casadio, R., 2002. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47, 142–153.

Ponnuswamy, P.K., 1993. Hydrophobic characteristics of folded proteins. Prog. Biophys. Mol. Biol. 59, 57–103.

Ponnuswamy, P.K., Gromiha, M.M., 1993. Prediction of transmembrane helices from hydrophobic characteristics of proteins. Int. J. Pept. Protein Res. 42, 326–341.

Ponnuswamy, P.K., Gromiha, M.M., 1994. On the conformational stability of folded proteins. J. Theor. Biol. 166, 63–74.

Ponnuswamy, P.K., Warme, P.K., Scheraga, H.A., 1973. Role of medium-range interactions in proteins. Proc. Natl. Acad. Sci. USA 70, 830–833.

Ponnuswamy, P.K., Prabakaran, M., Manavalan, P., 1980. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim. Biophys. Acta 623, 301–316.

Poupon, A., Mornon, J.P., 1999. Predicting the protein folding nucleus from sequences. FEBS Lett. 452, 283–289.

Qian, N., Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865–884.

Querol, E., Perez-Pons, J.A., Mozo-Villarias, A., 1996. Analysis of protein conformational characteristics related to thermostability. Protein Eng. 9, 265–271.

Reva, B., Finkelstein, A.V., Sanner, M., Olson, A.J., 1997. Residue–residue mean-force potentials for protein structure recognition. Protein Eng. 10, 865–876.

Rooman, M.J., Wodak, S.J., 1995. Are database-derived potentials valid for scoring both forward and inverted protein folding? Protein Eng. 8, 849–858.

Rooman, M.J., Kocher, J.-P.A., Wodak, S.J., 1992. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 31, 10226–10238.

Rost, B., Sandor, C., 1994a. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19, 55–72.

Rost, B., Sander, C., 1994b. Conservation and prediction of solvent accessibility in protein families. Proteins 20, 216–226.

Russ, W.P., Ranganathan, R., 2002. Knowledge-based potential functions in protein design. Curr. Opin. Struct. Biol. 12, 447–452.

Russell, R.B., Saqi, M.A., Bates, P.A., Sayle, R.A., Sternberg, M.J., 1998. Recognition of analogous and homologous protein folds—assessment of prediction success and associated alignment accuracy using empirical substitution matrices. Protein Eng. 11, 1–9.

Selbig, J., 1995. Contact pattern-induced pair potentials for protein fold recognition. Protein Eng. 8, 339–351.

Selbig, J., Argos, P., 1998. Relationships between protein sequence and structure patterns based on residue contacts. Proteins 31, 172–185.

Selvaraj, S., Gromiha, M.M., 1998a. An analysis of the amino acid clustering pattern in $(\alpha/\beta)_8$ barrel proteins. J. Protein Chem. 17, 407–415.

Selvaraj, S., Gromiha, M.M., 1998b. Importance of long-range interactions in $(\alpha/\beta)_8$ barrel fold. J. Protein Chem. 17, 691–697.

Selvaraj, S., Gromiha, M., 2000. Inter-residue interactions in protein structures. Curr. Sci. 78, 129–131.

Selvaraj, S., Gromiha, M.M., 2003. Role of hydrophobic clusters and long-range contact networks in the folding of $(\alpha/\beta)_8$ barrel proteins. Biophys. J. 84, 1919–1925.

Seno, F., Maritan, A., Banavar, J.R., 1998. Interaction potentials for protein folding. Proteins 30, 244–248.

Sippl, M.J., 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213, 859–883.

Sippl, M.J., 1995. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5, 229–235.

Skolnick, J., Kihara, D., 2001. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 42, 319–331.

Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P., Boniecki, M., 2001. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins S5, 149–156.

Szilagyi, A., Zavodszky, P., 2000. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct. Fold. Des. 8 (5), 493–504.

Tanaka, S., Scheraga, H.A., 1975. Model of protein folding: inclusion of short-, medium-, and long-range interactions. Proc. Natl. Acad. Sci. 72, 3802–3806.

Ternstrom, T., Mayor, U., Akke, M., Oliveberg, M., 1999. From snapshot to movie: phi analysis of protein folding transition states taken one step further. Proc. Natl. Acad. Sci. USA 96, 14854–14859.

Thornton, J.M., 1981. Disulphide bridges in globular proteins. J. Mol. Biol. 151, 261–287.

Tobi, D., Shafran, G., Linial, N., Elber, R., 2000. On the design and analysis of protein folding potentials. Proteins 40, 71–85.

Topham, C.M., Srinivasan, N., Blundell, T.L., 1997. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng. 10, 7–21.

Tudos, E., Fiser, A., Simon, I., 1994. Different sequence environments of amino acid residues involved and not involved in long-range interactions in proteins. Int. J. Pept. Protein Res. 43, 205–208.

Villegas, V., Martinez, J.C., Aviles, F.X., Serrano, L., 1998. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. J. Mol. Biol. 283, 1027–1036.

Vogt, G., Woell, S., Argos, P., 1997. Protein thermal stability, hydrogen bonds, and ion pairs. J. Mol. Biol. 269, 631–643.

Wilmanns, M., Eisenberg, D., 1993. Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. Proc. Natl. Acad. Sci. USA 90, 1379–1383.

Xiao, L., Honig, B., 1999. Electrostatic contributions to the stability of hyperthermophilic proteins. J. Mol. Biol. 289, 1435–1444.

Zehfus, M.H., 1995. Automatic recognition of hydrophobic clusters and their correlation with protein folding units. Protein Sci. 4, 1188–1202.

Zhang, C., Kim, S.-H., 2000. Environment-dependent residue contact energies for proteins. Proc. Natl. Acad. Sci. USA 97, 2550–2555.

Zhang, L., Skolnick, J., 1998. How do potentials derived from structural databases relate to "true" potentials? Protein Sci. 7, 112–122.

Zhang, L., Li, J., Jiang, Z., Xia, A., 2003. Folding rate prediction based on neural network model. Polymer 44, 1751–1756.

Zhou, H., Zhou, Y., 2002. Folding rate prediction using total contact distance. Biophys. J. 82, 458–463.