ELSEVIER

Biophysical Chemistry 77 (1999) 49–68

Biophysical Chemistry

# Importance of long-range interactions in protein folding†

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

[a] *The Institute of Physical and Chemical Research (RIKEN), Tsukuba Life Science Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan*

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

Received 12 October 1998; received in revised form 13 January 1999; accepted 13 January 1999

**Abstract**

Long-range interactions play an active role in the stability of protein molecules. In this work, we have analyzed the importance of long-range interactions in different structural classes of globular proteins in terms of residue distances. We found that 85% of residues are involved in long-range contacts. The residues occurring in the range of 4–10 residues apart contribute more towards long-range contacts in all-$\alpha$ proteins while the range is 11–20 in all-$\beta$ proteins. The hydrophobic residues Cys, Ile and Val prefer the 11–20 range and all other residues prefer the 4–10 range. The residues in all-$\beta$ proteins have an average of 3–8 long-range contacts whereas the residues in other classes have 1–4 long-range contacts. Furthermore, the preference of residue pairs to the folding and stability will be discussed. 

*Keywords:* Long-range contacts; Structural class; Tertiary structure; Protein folding

## 1. Introduction

The folding of a polypeptide chain into a compact, unique three-dimensional structure is directed and stabilized by intra molecular interactions between the constituent amino acid residues along the chain. Based on crystal structural data various investigations have been carried out to understand the role of different interactions in the folding and stability of globular proteins [1–8]. Furthermore, the importance of long-range interactions to the stability of proteins [9] and deriving potentials for fold recognition [10–12], the impact of local and non-local interactions on the folding of globular proteins [13], distance-dependent potentials [14] and relation between protein sequence and structure based on residue contacts [15] have been reported.

Recently, we have analyzed the influence of medium and long-range interactions in different structural classes of proteins and showed that while medium range interactions predominate in all-$\alpha$ class proteins, long-range interactions pre-

* Corresponding author. Tel.: +81 298 36 9082; fax: +81 298 36 9080; e-mail: gromiha@rtc.riken.go.jp

† This article is dedicated to our teacher, Professor P.K. Ponnuswamy on the occasion of his 60th birthday.

0301-4622/99/$ - see front matter 
PII: S0301-4622(99)00010-1

dominate the all-$\beta$ class [16]. Based on this fact, an evaluation of the performance of several secondary structure prediction methods revealed that all the methods predict the secondary structure of all-$\alpha$ proteins more accurately than other classes [17]. Furthermore, the influence of medium and long-range interactions and the importance of long-range interactions in terms of residue distances in $(\alpha/\beta)_8$ barrel proteins has been analyzed [18,19].

As residues which contribute towards long-range interactions need not be sequential neighbors and can exist far from the sequence, it is of interest to reveal the preference of residues towards these interactions in terms of intervening residue distances. Hence, in the present work, we have analyzed the role of long-range interactions in different structural classes of globular proteins in terms of residue distances.

## 2. Materials and methods

### 2.1. Database

The crystallographic data of 150 globular proteins taken from the Protein Data Bank (PDB) of Brookhaven National Laboratory [20,21] forms the source for our present study. The proteins selected were non-homologous and the structures were determined to a high resolution (resolution $< 2.5$ Å). The selected proteins are from four different structural classes, namely, all-$\alpha$, all-$\beta$, $\alpha+\beta$ and $\alpha/\beta$ with a set of 35, 38, 35 and 42 proteins, respectively. The PDB codes for all the proteins used in the present study along with their fold and length are given in Table 1. We obtained the information about the structural class and the fold of all proteins from SCOP [22] and CATH [23] databases, and used the DSSP algorithm [24] for the assignment of secondary structures.

### 2.2. Computation of surrounding residues and long-range contacts

The computation of surrounding residues in a protein molecule has been described in our earlier articles [16,25]. The residues in a protein molecule are represented by their $C_\alpha$ atoms. Using the $C_\alpha$ coordinates, a sphere of radius 8 Å is fixed around each residue and the composition of surrounding residues associated with all the residues is calculated. It has been shown that the influence of each residue over the surrounding medium extends effectively only up to 8 Å [26–28].

For a given residue, the composition of surrounding residues is analyzed in terms of the location at the sequence level and the contributions from $< \pm 3$ residues are treated as short range contacts, $\pm 3$ or $\pm 4$ residues as medium range contacts and $> \pm 4$ residues are treated as long-range contacts [16,18].

### 2.3. Long-range contacts in different ranges of residue distances

The long-range contacts ($> \pm 4$ residues) are further classified into several intervals with a step of 10 (4–10; 11–20; 21–30; 31–40; 41–50 and $> 50$). The number of long-range contacts in each interval for all the residues in 150 globular proteins belonging to four different structural classes were computed. Also, the percentage of long-range contacts for all the proteins in each interval were calculated. Furthermore, the contribution of all the 20 amino acid residues towards the long-range interactions in different intervals were estimated for the entire database. Moreover, the preference of the 20 amino acid residues in helical and strand segments to form long-range contacts were delineated.

### 2.4. Preference of surrounding residues influenced by long-range contacts

The residues coming within a sphere of 8 Å for each residue in all the four structural classes and for the complete set of proteins (comprising of 29 420 residues) were computed and the residues which contribute towards long-range contacts are selected as described above. For a given residue, the preference of all the 20 amino acid residues to form long-range contacts is computed and the total preference for all the 20 amino acid residues

is estimated. The average preference of surrounding residues is computed using the expression:

$$\langle N \rangle_{ij} = \frac{\Sigma N_{ij}}{\Sigma N_i + \Sigma N_j}$$

where $N_{ij}$ is the number of surrounding residues of type $j$ around residue $i$. $N_i$ and $N_j$ are, respectively the total number of residues of type $i$ and $j$. The top ten residue pairs were selected and used for further analysis. A similar analysis was carried out for residues influenced by medium range contacts.

## 3. Results and discussions

### *3.1. Occurrence of long-range contacts for different residue intervals in four structural classes*

The percentage of long-range contacts computed for different intervals in four structural classes are given in Table 1. Proteins belonging to the same fold have been placed together with increasing length. In Fig. 1, we show the average percentage in different intervals for each of the structural classes and the whole set of proteins.

A perusal of Fig. 1 (and Table 1) clearly reveals the opposite trends between the folding of all-$\alpha$ and all-$\beta$ proteins. The all-$\alpha$ class proteins have more long-range contacts in the 4–10 range and the all-$\beta$ class proteins have more long-range contacts in the 11–20 range. This may be due to the specific hydrogen bonding pattern of $\alpha$-helices and $\beta$-strands in these classes of proteins. The behavior of proteins in $\alpha+\beta$ and $\alpha/\beta$ classes are surprising. The range 4–10 is favored by $\alpha+\beta$ class of proteins while the $\alpha/\beta$ class of proteins prefer 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-$\alpha$ or all-$\beta$ type. In the present analysis we found that the features of $\alpha+\beta$ proteins are similar to that of all-$\alpha$ proteins. In the $\alpha/\beta$ class, the $\alpha$-helices and $\beta$-strands occur alternatively and some residue distances are necessary to form $\beta$-strand and barrel, which leads to having higher contacts in the 21–30 range. A similar trend was also observed in our previous study of $(\alpha/\beta)_8$ barrel proteins [19]. 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.

The overall analysis shows that 45% of proteins prefer the 4–10 range with a higher number of long-range contacts followed by 11–20 and 21–30 ranges, respectively, which were preferred by 27% and 19% of proteins (Table 1). Only an insignificant number of proteins are found to have long-range contacts in the 31–40 and 41–50 range.

Interestingly, we note from Fig. 1 that the limit of residual distances to form long-range contacts is 21–30. This is consistent with the recent analysis on the number of interactions per residue as a function of sequential distance between the interacting residues, which showed a significant margin after the 25th neighbor [29].

### *3.2. Effect of size at different residue intervals of long-range contacts*

We classified the proteins into three groups based on their size. Proteins with less than 100 residues were considered small, with residues between 100 and 200 were considered to be medium and with more than 200 residues were considered large. Our analysis shows that the larger proteins prefer the ranges 4–10 and 21–30; medium size proteins prefer the 4–10 range and small proteins prefer the three ranges 4–10, 11–20 and 21–30. This shows that the size of a protein may influence the long-range contact preferences to attain the stable tertiary structure.

### *3.3. Preference of amino acid residues in different intervals of long-range contacts*

The average long-range contacts in different intervals computed for all the 20 amino acid residues in a set of 150 globular proteins are given in Table 2. The sum of average contacts per residue for all the intervals are given in the last column which shows the total long-range contacts per residue.

From this table, we observe that the residues

Table 1
Occurrence of long range contacts for different residue intervals in four structural classes of proteins

| No. | PDB code | Fold | $N$ | $N_l$ | Percentage of long range contacts in different intervals | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 |
| All-$\alpha$ proteins | | | | | | | | | | |
| 1 | 1C5A | Anaphylotoxins | 65 | 148 | 17.57 | 9.46 | 37.84 | 17.57 | 10.81 | 6.76 |
| 2 | 1AVHA | Annexin | 318 | 786 | 26.46 | 11.45 | 1.27 | 31.04 | 9.16 | 20.61 |
| 3 | 2CTS | Citrate synthase | 437 | 1244 | 20.58 | 12.38 | 6.11 | 2.89 | 6.91 | 51.13 |
| 4 | 451C | Cytochrome C | 82 | 222 | 38.74 | 15.32 | 9.01 | 8.11 | 10.81 | 18.02 |
| 5 | 3CYT | Cytochrome C | 103 | 352 | 22.16 | 33.52 | 13.07 | 1.14 | 0.00 | 30.11 |
| 6 | 1BBL | Dihydrolipoamide transferase | 37 | 48 | 62.50 | 25.00 | 12.50 | 0.00 | 0.00 | 0.00 |
| 7 | 2PDE | Dihydrolipoamide transferase | 43 | 176 | 43.18 | 19.32 | 28.41 | 9.09 | 0.00 | 0.00 |
| 8 | 4CPV | EF-hand | 108 | 270 | 35.56 | 20.00 | 9.63 | 7.41 | 3.70 | 23.70 |
| 9 | 1FHA | Ferritin | 172 | 358 | 23.46 | 10.61 | 9.50 | 7.82 | 7.82 | 40.78 |
| 10 | 1FIAB | FIS protein | 78 | 54 | 81.48 | 11.11 | 3.70 | 3.70 | 0.00 | 0.00 |
| 11 | 1HIGA | Four-helical cytokines | 123 | 110 | 54.55 | 12.73 | 1.82 | 0.00 | 1.82 | 29.09 |
| 12 | 1IFA | Four-helical cytokines | 159 | 282 | 14.89 | 3.55 | 4.96 | 2.84 | 2.84 | 70.92 |
| 13 | 256B | Four-helical bundle | 100 | 204 | 19.61 | 17.65 | 21.57 | 16.67 | 10.78 | 13.73 |
| 14 | 2MHR | Four-helical bundle | 118 | 194 | 19.59 | 15.46 | 16.49 | 7.22 | 10.31 | 30.93 |
| 15 | 2CCYA | Four-helical bundle | 127 | 280 | 18.57 | 10.71 | 11.43 | 12.14 | 16.43 | 30.71 |
| 16 | 1LE4 | Four-helical bundle | 139 | 186 | 15.05 | 5.38 | 7.53 | 19.35 | 19.35 | 33.33 |
| 17 | 2LIG | Four-helical bundle | 157 | 286 | 16.78 | 11.19 | 11.19 | 14.69 | 15.38 | 30.77 |
| 18 | 1ECO | Globin | 136 | 218 | 22.94 | 2.75 | 4.59 | 11.93 | 6.42 | 51.38 |
| 19 | 2HCOA | Globin | 141 | 298 | 20.81 | 7.38 | 6.04 | 18.12 | 7.38 | 40.27 |
| 20 | 2HCOB | Globin | 146 | 312 | 19.87 | 5.77 | 7.69 | 16.67 | 12.82 | 37.18 |
| 21 | 1BABB | Globin | 146 | 326 | 20.25 | 6.75 | 7.98 | 15.34 | 11.66 | 38.04 |
| 22 | 1HBG | Globin | 147 | 370 | 16.76 | 7.03 | 7.57 | 11.89 | 6.49 | 50.27 |
| 23 | 2LHB | Globin | 149 | 310 | 22.58 | 7.74 | 6.45 | 7.10 | 17.42 | 38.71 |
| 24 | 4MBN | Globin | 153 | 270 | 27.41 | 5.19 | 4.44 | 10.37 | 8.89 | 43.70 |
| 25 | 1LH1 | Globin | 153 | 318 | 20.75 | 3.77 | 3.77 | 7.55 | 2.52 | 61.64 |
| 26 | 1MBS | Globin | 153 | 338 | 30.18 | 4.14 | 5.33 | 15.98 | 10.06 | 34.32 |
| 27 | 1FCS | Globin | 154 | 292 | 28.77 | 5.48 | 4.11 | 9.59 | 7.53 | 44.52 |
| 28 | 1CPCA | Globin | 162 | 344 | 29.65 | 6.98 | 6.40 | 6.98 | 8.72 | 41.28 |
| 29 | 1GCN | Glucagon | 29 | 2 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 30 | 2MLT | Mellitin | 26 | 2 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 31 | 1PPT | Peptide harmones | 36 | 48 | 33.33 | 29.17 | 37.50 | 0.00 | 0.00 | 0.00 |
| 32 | 1PP2 | Phospholipase A2 | 122 | 374 | 28.34 | 18.72 | 5.35 | 2.67 | 3.74 | 41.18 |
| 33 | 1RPRA | ROP protein | 63 | 110 | 16.36 | 16.36 | 18.18 | 16.36 | 9.09 | 23.64 |
| 34 | 1TROA | Trp repressor | 104 | 88 | 52.27 | 13.64 | 15.91 | 6.82 | 11.36 | 0.00 |
| 35 | 1UTG | Uteroglobin | 70 | 90 | 37.78 | 15.56 | 11.11 | 17.78 | 6.67 | 11.11 |
| Complete set | | | 4456 | 9310 | 24.88 | 11.19 | 8.38 | 10.93 | 8.21 | 36.41 |

Table 1 (*Continued*)

| No. | PDB code | Fold | $N$ | $N_l$ | Percentage of long range contacts in different intervals | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 |
| *All-β proteins* | | | | | | | | | | |
| 36 | 1HIVA | Acid proteases | 99 | 406 | 20.69 | 26.60 | 7.39 | 0.00 | 17.24 | 28.08 |
| 37 | 2APR | Acid proteases | 325 | 1760 | 17.61 | 19.20 | 6.82 | 5.45 | 2.05 | 48.86 |
| 38 | 1HOE | α-amylase inhibitor | 74 | 372 | 18.28 | 19.35 | 20.97 | 24.73 | 8.06 | 8.60 |
| 39 | 1TEN | β-sandwich | 89 | 428 | 14.02 | 26.17 | 14.02 | 15.89 | 15.89 | 14.02 |
| 40 | 1TLK | β-sandwich | 103 | 494 | 15.38 | 20.24 | 12.15 | 14.17 | 14.98 | 23.08 |
| 41 | 1ACX | β-sandwich | 107 | 524 | 17.56 | 22.90 | 17.18 | 11.83 | 8.02 | 22.52 |
| 42 | 1REI | β-sandwich | 107 | 566 | 15.90 | 26.86 | 4.59 | 3.89 | 10.60 | 38.16 |
| 43 | 1CD8 | β-sandwich | 114 | 516 | 17.05 | 28.29 | 6.98 | 0.00 | 2.33 | 45.35 |
| 44 | 2SOD | β-sandwich | 151 | 888 | 18.02 | 20.72 | 6.76 | 6.98 | 6.31 | 41.22 |
| 45 | 2GCR | β-sandwich | 174 | 974 | 15.20 | 17.25 | 27.72 | 18.07 | 6.16 | 15.61 |
| 46 | 1CID | β-sandwich | 177 | 938 | 19.40 | 30.49 | 11.94 | 9.81 | 2.77 | 25.59 |
| 47 | 3HHRC | β-sandwich | 194 | 934 | 18.42 | 21.41 | 14.56 | 13.06 | 11.78 | 20.77 |
| 48 | 1HILA | β-sandwich | 217 | 1050 | 17.33 | 24.76 | 8.00 | 6.29 | 12.19 | 31.43 |
| 49 | 1MAMH | β-sandwich | 217 | 1020 | 15.29 | 25.29 | 15.69 | 5.49 | 8.43 | 29.80 |
| 50 | 4FAB | β-sandwich | 219 | 1012 | 16.80 | 26.48 | 11.86 | 5.53 | 7.51 | 31.82 |
| 51 | 2ILA | β-trefoil | 145 | 768 | 25.26 | 25.26 | 4.95 | 9.64 | 12.50 | 22.40 |
| 52 | 8I1B | β-trefoil | 146 | 674 | 26.11 | 28.49 | 5.93 | 10.68 | 9.50 | 19.29 |
| 53 | 1TIE | β-trefoil | 166 | 844 | 21.56 | 29.15 | 8.29 | 9.72 | 7.35 | 23.93 |
| 54 | 2LALA | ConA-like lectins | 181 | 692 | 23.12 | 29.77 | 14.45 | 11.56 | 3.47 | 17.63 |
| 55 | 2AYH | ConA-like lectins | 214 | 1208 | 17.05 | 18.71 | 11.42 | 2.98 | 3.97 | 45.86 |
| 56 | 2PCY | Cupredoxins | 99 | 482 | 13.28 | 16.18 | 26.14 | 13.28 | 15.35 | 15.77 |
| 57 | 2AZA | Cupredoxins | 129 | 618 | 14.56 | 13.92 | 17.80 | 7.44 | 7.77 | 38.51 |
| 58 | 3CNA | ConA-like serine proteases | 237 | 1336 | 15.27 | 21.11 | 7.63 | 3.59 | 2.54 | 49.85 |
| 59 | 2PAB | Prealbumin | 114 | 524 | 18.70 | 22.52 | 18.70 | 13.74 | 11.83 | 14.50 |
| 60 | 1RDG | Rubredoxin-like | 52 | 182 | 45.05 | 19.78 | 7.69 | 7.69 | 19.78 | 0.00 |
| 61 | 1SHFA | SH3-like barrel | 59 | 258 | 25.58 | 40.31 | 12.40 | 6.20 | 6.98 | 8.53 |
| 62 | 1CDTA | Snake toxin-like | 60 | 290 | 25.52 | 24.14 | 24.83 | 15.86 | 0.69 | 8.97 |
| 63 | 3EBX | Snake toxin-like | 62 | 292 | 18.49 | 27.40 | 23.29 | 21.23 | 0.68 | 8.90 |
| 64 | 1CTX | Snake toxin-like | 71 | 324 | 17.28 | 22.84 | 20.7 | 22.22 | 6.17 | 11.11 |
| 65 | 2AVIA | Streptavidin-like | 121 | 580 | 21.03 | 41.38 | 24.48 | 1.38 | 0.00 | 11.72 |
| 66 | 1PPFE | Thrombin | 218 | 1166 | 15.44 | 19.38 | 11.15 | 7.72 | 10.81 | 35.51 |
| 67 | 1TPA | Trypsin-like | 58 | 254 | 14.17 | 25.98 | 38.58 | 11.81 | 0.79 | 8.66 |
| 68 | 1TGS | Trypsin-like serine proteases | 56 | 188 | 19.15 | 15.96 | 38.30 | 11.70 | 14.89 | 0.00 |
| 69 | 2SNV | Trypsin-like serine proteases | 151 | 740 | 27.84 | 23.78 | 15.68 | 11.62 | 8.65 | 12.43 |
| 70 | 2ALP | Trypsin-like serine proteases | 198 | 1244 | 17.68 | 20.58 | 12.86 | 11.41 | 7.88 | 29.58 |

Table 1 (*Continued*)

| No. | PDB code | Fold | $N$ | $N_l$ | Percentage of long range contacts in different intervals | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 |
| 71 | 4CHA | Trypsin-like serine proteases | 239 | 1336 | 12.43 | 20.51 | 11.83 | 7.34 | 8.53 | 39.37 |
| 72 | 3EST | Trypsin-like serine proteases | 240 | 1304 | 14.11 | 18.87 | 10.43 | 9.66 | 6.75 | 40.18 |
| 73 | 2BPA2 | Viral coat and capsid proteins | 175 | 912 | 5.26 | 7.46 | 20.61 | 4.17 | 8.77 | 53.73 |
| Complete set | | | 5558 | 28 098 | 17.59 | 22.59 | 13.23 | 8.77 | 7.56 | 30.27 |
| *$\alpha + \beta$ proteins* | | | | | | | | | | |
| 74 | 1LTSD | ADP ribosylation | 103 | 356 | 24.16 | 32.02 | 1.12 | 0.56 | 10.11 | 32.02 |
| 75 | 1PAX | ADP ribosylation | 350 | 1338 | 9.87 | 12.56 | 4.63 | 10.76 | 9.87 | 52.32 |
| 76 | 2PIA | $\beta$-grasp | 321 | 1394 | 15.64 | 14.78 | 21.23 | 7.60 | 5.88 | 34.86 |
| 77 | 4BLMA | $\beta$-lactamase | 256 | 1060 | 16.04 | 22.26 | 17.36 | 3.40 | 0.38 | 40.57 |
| 78 | 1EAF | CoA-dependent acetyl transferases | 243 | 866 | 17.09 | 17.09 | 13.63 | 8.78 | 3.00 | 40.42 |
| 79 | 1TFG | Cysteine knot cytokines | 112 | 506 | 17.79 | 24.51 | 19.37 | 13.44 | 0.00 | 24.90 |
| 80 | 2PAD | Cysteine proteinases | 108 | 270 | 35.56 | 20.00 | 9.63 | 7.41 | 3.70 | 23.70 |
| 81 | 1PPN | Cysteine proteinases | 212 | 1076 | 14.68 | 11.34 | 13.94 | 18.59 | 9.11 | 32.34 |
| 82 | 2ACT | Cysteine proteinases | 218 | 1102 | 15.79 | 12.89 | 12.52 | 11.07 | 14.34 | 33.39 |
| 83 | 2B5C | Cytochrome b5 | 85 | 234 | 24.79 | 7.69 | 17.95 | 13.68 | 15.38 | 20.51 |
| 84 | 4DNK | DNase-I type | 250 | 1134 | 17.11 | 13.58 | 17.28 | 23.10 | 11.82 | 17.11 |
| 85 | 4ENL | Enolase-like | 436 | 2046 | 18.96 | 11.53 | 15.35 | 6.16 | 7.23 | 40.76 |
| 86 | 1FDX | Ferredoxin-like | 54 | 188 | 25.53 | 13.83 | 19.15 | 15.96 | 19.15 | 6.38 |
| 87 | 1NRCA | Ferredoxin-like | 85 | 306 | 18.95 | 18.95 | 8.50 | 11.11 | 11.11 | 31.37 |
| 88 | 3RUBS | Ferredoxin-like | 123 | 360 | 21.11 | 18.33 | 15.00 | 6.67 | 6.67 | 32.22 |
| 89 | 8CAT | Heme-linked catalases | 498 | 1754 | 16.53 | 11.74 | 11.06 | 10.26 | 9.92 | 40.48 |
| 90 | 1HIP | HIPIP | 85 | 326 | 28.83 | 19.63 | 3.07 | 17.18 | 8.59 | 22.70 |
| 91 | 3IL8 | IL8-like | 68 | 192 | 29.17 | 30.21 | 15.62 | 13.54 | 11.46 | 0.00 |
| 92 | 3INS | Insulin-like | 102 | 388 | 17.01 | 28.35 | 24.23 | 0.00 | 18.56 | 11.86 |
| 93 | 3LYZ | Lysozyme-like | 129 | 472 | 31.36 | 26.69 | 7.63 | 9.32 | 3.39 | 21.61 |
| 94 | 2LZM | Lysozyme-like | 164 | 410 | 36.59 | 22.44 | 10.73 | 14.15 | 4.88 | 11.22 |
| 95 | 9RNT | Microboil ribonucleases | 104 | 392 | 28.57 | 31.12 | 9.69 | 10.20 | 4.59 | 15.82 |
| 96 | 1HSBA | MHC antigen recognition domain | 270 | 1076 | 24.54 | 25.65 | 12.64 | 5.95 | 8.55 | 22.68 |
| 97 | 2CDV | Multiheme cytochromes | 107 | 284 | 68.31 | 23.94 | 0.00 | 0.70 | 3.52 | 3.52 |
| 98 | 1OVB | Periplasmic binding protein | 159 | 740 | 24.05 | 18.65 | 8.38 | 0.81 | 10.00 | 38.11 |
| 99 | 2PRF | Profilin-like | 125 | 484 | 32.64 | 29.75 | 14.05 | 2.89 | 0.00 | 20.66 |
| 100 | 2RNS | Ribonuclease A-like | 124 | 510 | 21.18 | 21.18 | 9.41 | 24.31 | 4.31 | 19.61 |
| 101 | 1CTF | Ribosomal protein | 68 | 234 | 14.53 | 8.55 | 6.84 | 21.37 | 23.93 | 24.79 |
| 102 | 2ACHA | Serpins | 337 | 1380 | 17.39 | 13.19 | 13.77 | 5.65 | 6.23 | 43.77 |
| 103 | 1SHAA | SH2-like | 103 | 388 | 34.02 | 25.26 | 20.10 | 3.61 | 1.03 | 15.98 |

Table 1 (*Continued*)

| No. | PDB code | Fold | $N$ | $N_l$ | Percentage of long range contacts in different intervals | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 |
| 104 | 3SICI | Subtilisin inhibitor | 107 | 460 | 25.22 | 21.74 | 13.48 | 3.04 | 2.61 | 33.91 |
| 105 | 2SSI | Subtilisin inhibitor-like | 107 | 442 | 25.34 | 22.17 | 14.48 | 3.62 | 2.71 | 31.67 |
| 106 | 4TMS | Thymidylate synthase | 316 | 1082 | 17.93 | 17.01 | 6.47 | 14.79 | 4.62 | 39.19 |
| 107 | 2MS2A | Viral coat protein | 129 | 424 | 34.43 | 38.68 | 14.62 | 10.85 | 1.42 | 0.00 |
| 108 | 4TLN | Zincin-like | 316 | 1456 | 25.69 | 21.02 | 10.16 | 6.32 | 1.79 | 35.03 |
| Complete set | | | 6374 | 25 130 | 20.93 | 18.05 | 12.71 | 9.42 | 7.00 | 31.90 |
| *α / β proteins* | | | | | | | | | | |
| 109 | 2CAB | Carbonic anhydrases | 256 | 1364 | 15.69 | 14.08 | 12.76 | 8.21 | 1.47 | 47.80 |
| 110 | 1CRN | Crambin-like | 46 | 134 | 26.87 | 13.43 | 17.91 | 32.84 | 8.96 | 0.00 |
| 111 | 3DFR | Dihydrofolate reductase | 162 | 630 | 13.33 | 22.22 | 13.97 | 6.98 | 5.40 | 38.10 |
| 112 | 1FX1 | Flavodoxin | 147 | 630 | 14.29 | 5.71 | 13.33 | 39.68 | 14.60 | 12.38 |
| 113 | 1OFV | Flavodoxin-like | 169 | 722 | 14.40 | 8.31 | 15.79 | 33.24 | 13.57 | 14.68 |
| 114 | 1GPB | Glycogen phosphorylase | 823 | 3016 | 15.98 | 10.08 | 11.41 | 10.54 | 6.70 | 45.29 |
| 115 | 1Q21 | Isopropylmalate dehydrogenase | 171 | 666 | 17.42 | 12.61 | 14.71 | 17.72 | 11.71 | 25.83 |
| 116 | 1SBP | Periplasmic binding protein | 309 | 1212 | 20.63 | 8.75 | 2.64 | 8.58 | 7.43 | 51.98 |
| 117 | 1PFKA | Phosphofructokinase | 320 | 1374 | 13.25 | 11.64 | 11.64 | 10.19 | 6.40 | 46.87 |
| 118 | 1ULA | Phosphorylase-like | 289 | 1100 | 13.27 | 6.00 | 11.09 | 3.27 | 6.73 | 59.64 |
| 119 | 3CPA | Phosphorylase-like | 307 | 1386 | 16.45 | 10.53 | 7.65 | 5.92 | 11.11 | 48.34 |
| 120 | 1RHD | Rhodanase | 293 | 1170 | 16.75 | 7.69 | 20.34 | 5.64 | 1.71 | 47.86 |
| 121 | 1GLAG | Ribonuclease H-like motif | 489 | 2344 | 17.41 | 12.97 | 8.36 | 12.03 | 6.57 | 42.66 |
| 122 | 1CSEI | Rossmann fold | 63 | 216 | 16.67 | 48.15 | 4.63 | 0.93 | 14.81 | 14.81 |
| 123 | 3FXN | Rossmann fold | 138 | 508 | 14.17 | 5.51 | 24.80 | 23.62 | 11.81 | 20.08 |
| 124 | 1ETU | Rossmann fold | 177 | 680 | 18.24 | 12.06 | 14.71 | 17.06 | 13.24 | 24.71 |
| 125 | 3ADK | Rossmann fold | 194 | 540 | 21.48 | 4.07 | 1.11 | 3.33 | 1.11 | 68.89 |
| 126 | 1DHR | Rossmann fold | 236 | 988 | 15.38 | 19.84 | 12.75 | 3.04 | 29.15 | 19.84 |
| 127 | 2DRI | Rossmann fold | 271 | 1302 | 11.83 | 11.37 | 30.72 | 16.13 | 3.07 | 26.88 |
| 128 | 2SBT | Rossmann fold | 275 | 1568 | 13.78 | 11.73 | 18.24 | 10.84 | 6.38 | 39.03 |
| 129 | 5ABP | Rossmann fold | 306 | 1320 | 15.61 | 9.39 | 26.82 | 15.76 | 4.70 | 27.73 |
| 130 | 2HAD | Rossmann fold | 310 | 1220 | 17.05 | 6.72 | 26.07 | 5.74 | 3.77 | 40.66 |
| 131 | 3LDH | Rossmann fold | 329 | 1248 | 18.43 | 12.82 | 22.76 | 11.54 | 4.81 | 29.65 |
| 132 | 2GPD | Rossmann fold | 333 | 1514 | 15.85 | 15.98 | 22.32 | 3.04 | 4.49 | 38.31 |
| 133 | 2LIV | Rossmann fold | 344 | 1462 | 18.19 | 9.99 | 27.91 | 6.43 | 5.47 | 32.01 |
| 134 | 5ADH | Rossmann fold | 374 | 1910 | 14.03 | 13.61 | 26.91 | 3.98 | 3.98 | 37.49 |
| 135 | 4ICD | Rossmann fold | 414 | 1790 | 16.42 | 14.30 | 6.48 | 15.53 | 2.91 | 44.36 |
| 136 | 3PGK | Rossmann fold | 415 | 1816 | 15.53 | 8.15 | 20.81 | 14.10 | 8.04 | 33.37 |
| 137 | 2PGD | Rossmann fold | 473 | 1640 | 18.29 | 11.10 | 23.90 | 6.22 | 0.98 | 39.51 |
| 138 | 3COX | Rossmann fold | 500 | 2574 | 17.79 | 11.42 | 12.04 | 4.82 | 2.02 | 51.90 |
| 139 | 1ABA | Thioredoxin fold | 87 | 276 | 33.33 | 12.32 | 24.64 | 2.17 | 4.35 | 23.19 |
| 140 | 1SRX | Thioredoxin fold | 108 | 392 | 26.53 | 10.20 | 8.67 | 14.29 | 12.76 | 27.55 |

Table 1 (*Continued*)

| No. | PDB code | Fold | $N$ | $N_l$ | Percentage of long range contacts in different intervals | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 |
| 141 | 1WSYA | TIM barrel | 248 | 900 | 13.56 | 12.67 | 36.44 | 6.22 | 12.00 | 19.11 |
| 142 | 1TRE | TIM barrel | 255 | 1010 | 14.46 | 12.28 | 19.80 | 24.75 | 13.27 | 15.45 |
| 143 | 1GOX | TIM barrel | 350 | 1404 | 12.25 | 10.40 | 19.80 | 10.54 | 3.99 | 43.02 |
| 144 | 1MNS | TIM barrel | 357 | 1600 | 14.38 | 11.00 | 35.00 | 9.12 | 1.38 | 29.12 |
| 145 | 2AAA | TIM barrel | 476 | 2132 | 16.04 | 13.41 | 15.20 | 13.51 | 12.85 | 28.99 |
| 146 | 2TAA | TIM barrel | 478 | 2114 | 15.52 | 13.91 | 13.72 | 13.62 | 13.15 | 30.09 |
| 147 | 1TIM | TIM barrel | 494 | 2020 | 11.88 | 12.18 | 19.21 | 22.28 | 13.56 | 20.89 |
| 148 | 1FCB | TIM barrel | 494 | 1884 | 12.85 | 8.07 | 17.73 | 15.61 | 4.14 | 41.61 |
| 149 | 1CDG | TIM barrel | 686 | 3482 | 16.43 | 13.38 | 16.94 | 11.43 | 9.19 | 32.62 |
| 150 | 1CIS | Trypsin inhibitor | 66 | 268 | 14.18 | 34.33 | 20.90 | 2.24 | 11.94 | 16.42 |
| Complete set | | | 13 032 | 55 526 | 15.82 | 11.77 | 17.46 | 11.32 | 7.25 | 36.37 |

$N$ and $N_l$ are, respectively, the total number of residues and the total number of long-range contacts. The details about the fold are available in [22,23].

Cys, Ile and Val prefer the 11–20 range and all the other residues prefer the 4–10 range. Interestingly, Cys, Ile and Val are the three topmost hydrophobic residues [30]. These residues have the higher tendency of forming hydrophobic clusters and disulfide bridges due to long-range contacts and hence prefer the range of 11–20.

### *3.4. Preference of amino acid residues in different secondary structures to form long-range contacts at various intervals*

The number of long-range contacts in different intervals computed for the 20 amino acid residues in helical and strand segments of globular proteins are presented in Table 3. From this table, we observe that in all-$\alpha$ class, most of the residues in helical segments prefer the 4–10 range. Interestingly, the smallest residue, Gly in helical segments prefers the interval 31–40 with higher difference; the residues Cys and Arg prefer the ranges, 21–30 and 31–40, respectively.

The analysis on the residues in $\beta$-strands of all-$\beta$ proteins shows that all the residues except Lys and Asn prefer the 11–20 range; these two residues prefer the 4–10 range.

In the $\alpha+\beta$ class, all residues in helical segments prefer the 4–10 range; in strand segments, the residues Asp and Gln prefer the 4–10 range while all other residues prefer the 11–20 range. Interestingly, the aromatic residues Phe and Trp equally prefer these two ranges (4–10 and 11–20). Furthermore, comparison between helical and strand segments shows that in 4–10 range, the residues Ala and Met have higher long-range contacts in helical segments than those in strand segments; similar trend is observed for Glu and Ala in the intervals, 31–40 and 41–50, respectively. Also, we observed that some residues in specific ranges have a similar number of long-range contacts both in helical and strand segments; Lys and Leu in the 4–10 range, and Asn and Gln in the 41–50 range.

In $\alpha/\beta$ class, all the residues except Gly, His and Pro in helical segments prefer the 4–10 range; Gly and His prefer the 21–30 range and Pro prefers the 41–50 range. In strand segments, the residue Lys prefers the ranges 4–10 and 11–20; Arg, Ser, Thr and Trp prefer the 11–20 range and all other residues prefer the 21–30 range. In the 4–10 range, the hydrophobic residues, Cys, Ile and Val have higher long-range contacts in $\beta$-strand segments than those in $\alpha$-helical segments. Similar tendency is also observed for helix breaking residue, Pro; all other residues have higher long-range contacts in helical segments. In all

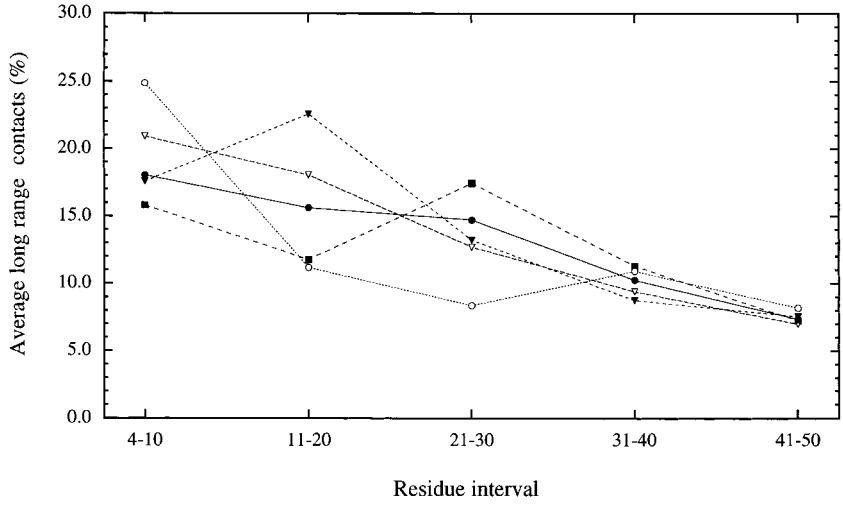


Fig. 1. Average percentage of long range contacts in different intervals for the four structural classes of globular proteins ○: all-$\alpha$; ▼: all-$\beta$; ▽: $\alpha+\beta$; ■: $\alpha/\beta$; ●: combined set.

Table 2
Average long-range contacts per residue for the 20 amino acid residues in different intervals

| Residue | Interval | | | | | | |
|---|---|---|---|---|---|---|---|
| | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 | Total |
| Ala | 0.73 | 0.56 | 0.62 | 0.39 | 0.31 | 1.44 | 4.05 |
| Asp | 0.52 | 0.38 | 0.44 | 0.29 | 0.23 | 0.94 | 2.80 |
| Cys | 1.03 | 1.16 | 0.99 | 0.78 | 0.51 | 1.58 | 6.05 |
| Glu | 0.52 | 0.40 | 0.37 | 0.29 | 0.22 | 0.83 | 2.63 |
| Phe | 0.89 | 0.74 | 0.63 | 0.45 | 0.30 | 1.54 | 4.54 |
| Gly | 0.64 | 0.59 | 0.61 | 0.47 | 0.34 | 1.51 | 4.17 |
| His | 0.63 | 0.56 | 0.52 | 0.37 | 0.32 | 1.11 | 3.52 |
| Ile | 0.81 | 0.92 | 0.79 | 0.55 | 0.41 | 1.78 | 5.26 |
| Lys | 0.64 | 0.44 | 0.38 | 0.28 | 0.19 | 0.84 | 2.77 |
| Leu | 0.86 | 0.71 | 0.61 | 0.43 | 0.31 | 1.59 | 4.52 |
| Met | 0.81 | 0.71 | 0.61 | 0.41 | 0.30 | 1.57 | 4.41 |
| Asn | 0.63 | 0.43 | 0.50 | 0.39 | 0.24 | 1.20 | 3.38 |
| Pro | 0.66 | 0.55 | 0.55 | 0.35 | 0.29 | 1.31 | 3.71 |
| Gln | 0.68 | 0.49 | 0.43 | 0.32 | 0.20 | 0.93 | 3.05 |
| Arg | 0.67 | 0.58 | 0.46 | 0.39 | 0.22 | 1.14 | 3.46 |
| Ser | 0.67 | 0.58 | 0.48 | 0.37 | 0.26 | 1.32 | 3.68 |
| Thr | 0.77 | 0.77 | 0.49 | 0.35 | 0.29 | 1.46 | 4.13 |
| Val | 0.91 | 0.92 | 0.82 | 0.51 | 0.44 | 1.84 | 5.45 |
| Trp | 1.01 | 0.89 | 0.44 | 0.44 | 0.31 | 1.51 | 4.60 |
| Tyr | 0.94 | 0.81 | 0.60 | 0.52 | 0.37 | 1.44 | 4.68 |

other ranges, most of the residues in $\beta$-strand segments have higher long-range contacts. We also observed that few charged residues have higher long-range contacts in helical segments for specific ranges; His in 31–40 and Glu in 41–50. The positively charged residues His, Lys and Arg have a similar tendency in helical and strand segments to form long-range contacts in the 41–50 range.

### *3.5. Influence of long-range contacts in different structural classes of globular proteins*

The number of long-range contacts vs. residue numbers for four typical proteins are displayed in Fig. 2. The proteins are selected in such a way that all are of the same size and from four different structural classes.

#### *3.5.1. (I) all-$\alpha$ class (4MBN)*

For the protein myoglobin (all-$\alpha$ class), we observed the highest number for the G65 residue with seven long-range contacts and six long-range contacts for G25 and Y146 (Fig. 2a). Both residues belong to helical segments. We also found that most of the residues have 0–4 long-range contacts.

#### *3.5.2. (II) all-$\beta$ class (2SOD)*

The maximum of 13 long-range contacts are observed for superoxide dismutase, an all-$\beta$ class of proteins (Fig. 2b). The higher long-range contacts are found for a cluster of residues near the highest peaks at G42, V116 and I147. They are present in the $\beta$-strand segments S4, S7 and S8 respectively. Furthermore, the analysis on the residues with more than 10 long-range contacts shows that Ile, Leu, Val and His are the members in this category. Surprisingly, we note the residue His has higher long-range contacts, as this is surrounded by hydrophobic residues.

#### *3.5.3. (III) $\alpha + \beta$ class (2LZM)*

We found two separate domains for $\alpha$-helices and $\beta$-strands in the $\alpha + \beta$ type of protein, lysozyme T4 (Fig. 2c). The N-terminal domain contains $\beta$-strands with highest number of long-range contacts. We observed a highest peak at Y25 with 12 long-range contacts, similar to all-$\beta$ class of proteins. The C-terminal domain contains $\alpha$-helices and the maximum number of contact is observed for T152, which has seven long-range contacts, similar to the all-$\alpha$ class of proteins.

#### *3.5.4. (IV) $\alpha / \beta$ class (3DFR)*

In Fig. 2d, we display the number of long-range contacts for the $\alpha/\beta$ class protein, dihydrofolate reductase. From this figure, we observed the maximum of 10 long-range contacts for the residues F3, I13, V41, L62, I96 and E156. Interestingly, all the residues are in $\beta$-strands and most of them are hydrophobic. The alternate position of $\alpha$-helical and $\beta$-strand segments are clearly seen in this figure with more and less numbers of long-range contacts. This pattern of high and low peaks differentiate the $\alpha/\beta$ type of proteins from the $\alpha + \beta$ class of proteins.

Table 3
Number of long-range contacts at different intervals for the 20 amino acid residues in helical and strand segments of globular proteins

| Residue | Number of long range contacts in helical segments | | | | | | |
|---|---|---|---|---|---|---|---|
| | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 | Total |
| (i) All-$\alpha$ proteins | | | | | | | |
| Ala | 164 | 76 | 78 | 125 | 96 | 406 | 945 |
| Asp | 42 | 24 | 23 | 17 | 18 | 72 | 196 |
| Cys | 28 | 18 | 33 | 15 | 9 | 38 | 141 |
| Glu | 63 | 46 | 24 | 60 | 25 | 107 | 325 |
| Phe | 98 | 45 | 26 | 30 | 13 | 111 | 323 |
| Gly | 41 | 21 | 33 | 116 | 73 | 82 | 366 |
| His | 46 | 13 | 14 | 22 | 19 | 27 | 141 |
| Ile | 58 | 48 | 38 | 44 | 36 | 185 | 409 |
| Lys | 108 | 58 | 42 | 66 | 30 | 123 | 427 |
| Leu | 185 | 127 | 99 | 119 | 114 | 361 | 1005 |
| Met | 75 | 30 | 34 | 13 | 16 | 68 | 236 |
| Asn | 37 | 26 | 24 | 11 | 11 | 53 | 162 |
| Pro | 24 | 12 | 8 | 17 | 6 | 42 | 109 |
| Gln | 51 | 28 | 23 | 40 | 9 | 69 | 220 |
| Arg | 44 | 11 | 24 | 53 | 20 | 77 | 229 |
| Ser | 37 | 34 | 34 | 13 | 24 | 109 | 251 |
| Thr | 53 | 28 | 23 | 23 | 20 | 100 | 246 |
| Val | 91 | 50 | 60 | 67 | 66 | 293 | 627 |
| Trp | 15 | 5 | 2 | 5 | 2 | 57 | 86 |
| Tyr | 53 | 21 | 23 | 27 | 22 | 115 | 261 |

Table 3 (*Continued*)

| Residue | Number of long range contacts in strand segments | | | | | | |
|---|---|---|---|---|---|---|---|
| | 4–10 | 11–20 | 21–30 | 31–40 | 41–50 | > 50 | Total |
| (ii) All-$\beta$ proteins | | | | | | | |
| Ala | 185 | 209 | 98 | 63 | 90 | 283 | 928 |
| Asp | 73 | 76 | 49 | 27 | 26 | 79 | 330 |
| Cys | 104 | 205 | 90 | 96 | 47 | 175 | 717 |
| Glu | 113 | 124 | 93 | 44 | 53 | 130 | 557 |
| Phe | 143 | 255 | 91 | 90 | 50 | 224 | 853 |
| Gly | 141 | 286 | 127 | 85 | 57 | 254 | 950 |
| His | 42 | 84 | 33 | 22 | 15 | 64 | 260 |
| Ile | 163 | 301 | 128 | 67 | 69 | 235 | 963 |
| Lys | 175 | 164 | 109 | 69 | 46 | 140 | 703 |
| Leu | 290 | 444 | 189 | 155 | 66 | 492 | 1636 |
| Met | 53 | 79 | 37 | 31 | 15 | 119 | 334 |
| Asn | 86 | 75 | 54 | 41 | 23 | 135 | 414 |
| Pro | 44 | 50 | 34 | 11 | 21 | 76 | 236 |
| Gln | 134 | 138 | 71 | 39 | 46 | 134 | 562 |
| Arg | 113 | 120 | 74 | 87 | 40 | 150 | 584 |
| Ser | 171 | 285 | 95 | 108 | 87 | 333 | 1079 |
| Thr | 216 | 374 | 98 | 87 | 74 | 322 | 1171 |
| Val | 304 | 525 | 250 | 166 | 164 | 525 | 1934 |
| Trp | 67 | 104 | 22 | 31 | 24 | 94 | 342 |
| Tyr | 191 | 242 | 110 | 115 | 58 | 143 | 859 |

Table 3 (*Continued*)

| Residue | Number of long range contacts in helical and strand segments | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4–10 | | 11–20 | | 21–30 | | 31–40 | | 41–50 | | > 50 | | Total | |
| | H | S | H | S | H | S | H | S | H | S | H | S | H | S |
| (iii) $\alpha + \beta$ proteins | | | | | | | | | | | | | | |
| Ala | 168 | 108 | 89 | 185 | 72 | 138 | 73 | 80 | 58 | 27 | 247 | 249 | 707 | 787 |
| Asp | 35 | 83 | 14 | 71 | 22 | 33 | 17 | 37 | 16 | 31 | 56 | 97 | 160 | 352 |
| Cys | 22 | 57 | 18 | 70 | 19 | 28 | 8 | 45 | 9 | 22 | 46 | 91 | 122 | 313 |
| Glu | 38 | 72 | 27 | 103 | 44 | 50 | 34 | 24 | 27 | 30 | 62 | 115 | 232 | 394 |
| Phe | 45 | 113 | 19 | 113 | 29 | 92 | 24 | 64 | 9 | 47 | 68 | 218 | 194 | 647 |
| Gly | 48 | 89 | 22 | 93 | 20 | 49 | 14 | 49 | 8 | 27 | 86 | 172 | 198 | 479 |
| His | 20 | 34 | 12 | 46 | 9 | 48 | 14 | 30 | 13 | 18 | 13 | 78 | 81 | 254 |
| Ile | 74 | 129 | 54 | 276 | 48 | 111 | 30 | 85 | 20 | 82 | 147 | 219 | 373 | 902 |
| Lys | 76 | 77 | 31 | 147 | 25 | 84 | 9 | 44 | 10 | 31 | 51 | 151 | 202 | 534 |
| Leu | 159 | 162 | 87 | 241 | 55 | 146 | 51 | 94 | 28 | 61 | 241 | 365 | 621 | 1069 |
| Met | 38 | 17 | 11 | 62 | 9 | 39 | 9 | 13 | 1 | 10 | 33 | 56 | 101 | 197 |
| Asn | 31 | 52 | 10 | 75 | 25 | 27 | 17 | 30 | 11 | 12 | 87 | 116 | 181 | 312 |
| Pro | 15 | 56 | 7 | 69 | 7 | 39 | 10 | 29 | 2 | 13 | 37 | 104 | 78 | 310 |
| Gln | 45 | 69 | 25 | 60 | 20 | 40 | 35 | 43 | 17 | 16 | 60 | 59 | 202 | 287 |
| Arg | 52 | 83 | 15 | 115 | 24 | 38 | 10 | 36 | 8 | 35 | 53 | 82 | 162 | 389 |
| Ser | 55 | 91 | 44 | 138 | 37 | 111 | 29 | 56 | 22 | 34 | 99 | 176 | 286 | 606 |
| Thr | 37 | 159 | 27 | 192 | 19 | 87 | 6 | 70 | 4 | 42 | 74 | 163 | 167 | 713 |
| Val | 92 | 243 | 60 | 369 | 43 | 172 | 33 | 130 | 41 | 91 | 138 | 517 | 407 | 522 |
| Trp | 25 | 49 | 9 | 48 | 3 | 24 | 12 | 26 | 13 | 13 | 32 | 27 | 94 | 187 |
| Tyr | 48 | 120 | 18 | 175 | 10 | 88 | 8 | 44 | 14 | 37 | 61 | 195 | 159 | 659 |

Table 3 (*Continued*)

| Residue | Number of long range contacts in helical and strand segments | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4–10 | | 11–20 | | 21–30 | | 31–40 | | 41–50 | | > 50 | | Total | |
| | H | S | H | S | H | S | H | S | H | S | H | S | H | S |
| (iv) $\alpha/\beta$ proteins | | | | | | | | | | | | | | |
| Ala | 334 | 118 | 207 | 170 | 196 | 283 | 143 | 120 | 104 | 122 | 645 | 451 | 1629 | 1264 |
| Asp | 68 | 47 | 42 | 84 | 61 | 86 | 35 | 53 | 38 | 41 | 153 | 182 | 397 | 493 |
| Cys | 24 | 32 | 19 | 35 | 14 | 58 | 14 | 63 | 11 | 23 | 84 | 65 | 166 | 276 |
| Glu | 124 | 80 | 44 | 76 | 49 | 94 | 45 | 69 | 35 | 31 | 187 | 110 | 484 | 460 |
| Phe | 142 | 71 | 65 | 84 | 63 | 180 | 41 | 69 | 27 | 41 | 186 | 335 | 524 | 780 |
| Gly | 129 | 83 | 70 | 155 | 135 | 221 | 116 | 157 | 53 | 98 | 363 | 317 | 866 | 1031 |
| His | 24 | 49 | 8 | 43 | 35 | 53 | 31 | 14 | 15 | 16 | 61 | 86 | 174 | 261 |
| Ile | 137 | 161 | 106 | 234 | 101 | 431 | 103 | 238 | 60 | 169 | 379 | 727 | 886 | 1960 |
| Lys | 147 | 88 | 58 | 88 | 51 | 78 | 51 | 82 | 35 | 36 | 159 | 189 | 501 | 561 |
| Leu | 275 | 146 | 109 | 233 | 129 | 374 | 112 | 204 | 84 | 173 | 437 | 651 | 1146 | 1781 |
| Met | 73 | 38 | 53 | 74 | 42 | 74 | 31 | 54 | 29 | 25 | 154 | 148 | 382 | 413 |
| Asn | 63 | 52 | 33 | 51 | 44 | 87 | 26 | 54 | 30 | 30 | 150 | 139 | 346 | 413 |
| Pro | 16 | 59 | 23 | 61 | 25 | 79 | 18 | 65 | 35 | 32 | 97 | 119 | 214 | 415 |
| Gln | 87 | 50 | 29 | 67 | 51 | 71 | 41 | 48 | 22 | 19 | 143 | 92 | 373 | 347 |
| Arg | 108 | 49 | 33 | 94 | 52 | 91 | 71 | 82 | 26 | 27 | 161 | 208 | 451 | 551 |
| Ser | 99 | 72 | 35 | 115 | 64 | 106 | 66 | 78 | 31 | 47 | 249 | 215 | 544 | 633 |
| Thr | 127 | 118 | 81 | 166 | 49 | 140 | 77 | 86 | 49 | 59 | 301 | 334 | 684 | 903 |
| Val | 182 | 335 | 96 | 368 | 115 | 560 | 104 | 244 | 80 | 169 | 381 | 890 | 958 | 2566 |
| Trp | 53 | 41 | 19 | 49 | 16 | 23 | 19 | 22 | 9 | 21 | 98 | 96 | 214 | 252 |
| Tyr | 84 | 74 | 47 | 98 | 49 | 111 | 47 | 97 | 19 | 78 | 194 | 206 | 440 | 664 |

H and S are, respectively, helical and strand segments.

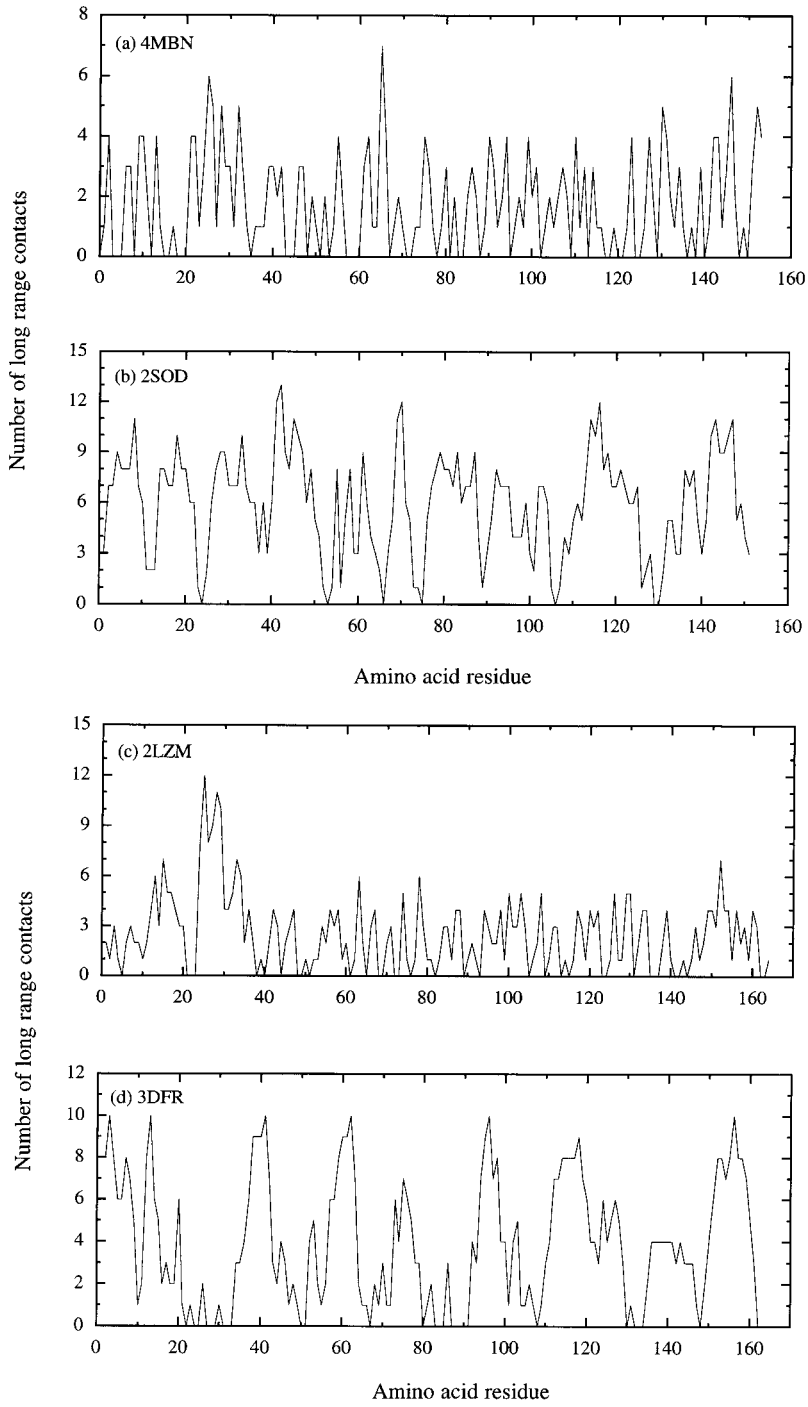


Fig. 2. Number of long-range contacts profile for four typical globular proteins in different structural classes (a) all-$\alpha$ (4MBN, Myoglobin) (b) all-$\beta$ (2SOD, Superoxide dismutase) (c) $\alpha+\beta$ (2LZM, Lysozyme T4) (d) $\alpha/\beta$ (3DFR, Dihydrofolate reductase).

### 3.6. Preference of amino acid residues to form long-range contacts

We analyzed the features of all the amino acid residues to have long-range contacts. Surprisingly, 84.8% of residues have at least one long-range contact. Of the considered 29 420 residues, 24 955 residues form long-range contacts. We note a variation of this tendency in different structural classes. The percentage of residues having long-range contacts in different structural classes, all-$\alpha$, all-$\beta$, $\alpha+\beta$ and $\alpha/\beta$ are, respectively 71.5, 90.2, 84.6 and 87.1%. These results suggest that while long-range contacts are crucial in the folding of proteins belonging to all the four structural classes, their effect is more pronounced in the all-$\beta$ class of proteins followed by $\alpha/\beta$ class of proteins.

The preference of the 20 amino acid residues to have at least one long-range contact is given in Table 4. From this table, we observe that more than 90% of hydrophobic and aromatic residues have at least one long-range contact in all structural classes. In the all-$\alpha$ class, the residues Ile and Cys are mostly influenced by long-range contacts. Interestingly, most of the positively charged residues, His, Lys and Arg in all-$\beta$ class proteins have at least one long-range contact.

The analysis on residues having more than 10 long-range contacts showed that the residues Cys, Gly, Ile, Ser and Val have higher preferences. Interestingly, the residues Val, Ser and Ile are the most preferred residues in the $\beta$-strand segment [31] and the residues Cys, Ile and Val fall under the category beta former and strong beta former [32].

### 3.7. Number of long-range contacts vs. number of residues

We computed the percentage of residues for different numbers of long-range contacts in all the structural classes and combined set of proteins and the results are presented in Table 5.

From this table, we note that the residues in all-$\beta$ class of proteins will have an average of 3–8 long-range contacts. All other classes will have an average of 1–4 long-range contacts and more than 50% of residues have this level of contacts.

Table 4
Preference of residues with atleast one long-range contact in four structural classes

| Residue | All-$\alpha$ | All-$\beta$ | $\alpha+\beta$ | $\alpha/\beta$ | Combined set |
|---|---|---|---|---|---|
| Ala | 69.39 | 88.01 | 83.33 | 86.78 | 82.73 |
| Asp | 57.30 | 86.51 | 70.05 | 76.30 | 73.53 |
| Cys | 86.54 | 98.74 | 98.60 | 97.26 | 97.00 |
| Glu | 57.01 | 86.00 | 72.03 | 74.63 | 72.30 |
| Phe | 80.65 | 94.71 | 96.80 | 93.88 | 92.48 |
| Gly | 68.30 | 82.44 | 80.55 | 86.13 | 81.77 |
| His | 68.87 | 94.51 | 85.99 | 84.55 | 82.59 |
| Ile | 90.58 | 95.42 | 94.60 | 95.39 | 94.60 |
| Lys | 69.23 | 89.00 | 75.85 | 75.62 | 76.44 |
| Leu | 84.63 | 96.46 | 94.41 | 93.94 | 92.51 |
| Met | 84.82 | 97.47 | 88.68 | 92.71 | 90.99 |
| Asn | 64.67 | 84.48 | 80.26 | 83.79 | 80.46 |
| Pro | 64.00 | 89.93 | 84.42 | 86.65 | 83.96 |
| Gln | 61.14 | 86.84 | 80.17 | 81.77 | 78.77 |
| Arg | 70.81 | 88.29 | 83.22 | 84.33 | 82.55 |
| Ser | 60.64 | 85.60 | 82.18 | 84.75 | 81.20 |
| Thr | 65.64 | 92.42 | 83.76 | 88.84 | 85.60 |
| Val | 85.05 | 96.38 | 93.68 | 95.17 | 93.73 |
| Trp | 79.66 | 91.92 | 94.51 | 95.98 | 92.43 |
| Tyr | 85.42 | 97.20 | 93.45 | 93.90 | 93.29 |

A plot connecting number of long-range contacts and percentage of residues for four structural classes and combined set of proteins is displayed in Fig. 3. From this figure we observed high peaks at three and four long-range contacts for all the structural classes. It is evident that more than 22% of residues have three or four long-range contacts irrespective of structural classes. This shows the importance of long-range interactions to the folding and stability of all classes of globular proteins.

### 3.8. Residue pairs influenced by long- and medium-range contacts

The preference of each amino acid residue to be surrounded by all the 20 amino acid residues due to long-range contacts are computed and the topmost 10 residue pairs are given in Table 6a. From this table, approximately 50% of residue pairs are observed with the same residue (C-C;

Table 5
Percentage of residues with different long-range contacts in four structural classes

| $N_{long}$ | Percentage of residues | | | | |
|---|---|---|---|---|---|
| | All-$\alpha$ | All-$\beta$ | $\alpha+\beta$ | $\alpha/\beta$ | Combined set |
| 0 | 28.47 | 9.75 | 15.36 | 12.88 | 15.49 |
| 1 | 19.03 | 6.81 | 10.88 | 11.18 | 11.64 |
| 2 | 14.99 | 6.53 | 10.97 | 10.34 | 10.54 |
| 3 | 14.56 | 10.23 | 11.08 | 11.20 | 11.54 |
| 4 | 10.85 | 12.14 | 13.19 | 11.20 | 11.74 |
| 5 | 5.84 | 10.07 | 9.34 | 9.82 | 9.01 |
| 6 | 3.34 | 9.16 | 7.48 | 8.10 | 7.30 |
| 7 | 1.52 | 10.60 | 6.76 | 6.95 | 6.63 |
| 8 | 0.82 | 10.23 | 5.91 | 6.74 | 6.26 |
| 9 | 0.35 | 6.99 | 4.25 | 5.22 | 4.55 |
| 10 | 0.08 | 3.83 | 2.42 | 3.31 | 2.71 |
| 11 | 0.11 | 1.80 | 1.35 | 1.81 | 1.44 |
| 12 | 0.00 | 0.94 | 0.74 | 0.90 | 0.07 |

$N_{long}$, number of long range contacts.

V-V; G-G; L-L, A-A and I-I). The highest preference is observed for C-C, which may be due to the formation of disulfide bridges. The hydrophobic residues Ala, Val and Leu contribute more for long-range contacts, which may be due to the formation of hydrophobic clusters. Note that the contribution of Ile is less compared to other hydrophobic residues. In all-$\alpha$ proteins, the charged residues Glu and Lys have higher influence to form long-range contacts.

In medium-range contacts (Table 6b), we found that the effect of C-C is much less; hydrophobic and polar residues have an equal role in forming medium-range contacts. The polar residue, Ser and the charged residues, Asp, Glu and Lys have more medium-range contacts with other residues. Interestingly, D-K and E-K are one of the top-most three preferred residue pairs for the classes, $\alpha+\beta$ and $\alpha/\beta$, respectively. This may be due the formation of ion-pairs [33]. The residue pair S-S is the second most preferred one in the all-$\beta$ class of proteins.

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 contribution. In long-range contacts, most of the contribution is influenced by hydrophobic residues and the role of polar residues are minimal.

The information about the preference of each amino acid residue surrounded by all the 20 amino acid residues due to medium and long-range contacts may be helpful to understand the stability of proteins due to mutations.

### *3.9. Comparison between short- and long-range interactions*

The importance of long-range interactions has been determined by comparing it with short-range

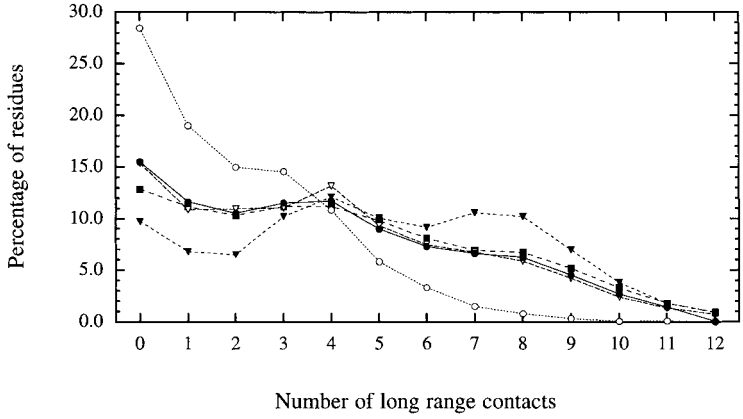


Fig. 3. Number of long range contacts vs. percentage of residues in four structural classes of proteins ○: all-$\alpha$; ▼: all-$\beta$; ▽: $\alpha+\beta$; ■: $\alpha/\beta$; ●: combined set.

Table 6
Topmost 10 residue pairs influenced by medium and long-range contacts in four structure classes

| No. | All-$\alpha$ | All-$\beta$ | $\alpha+\beta$ | $\alpha/\beta$ | Combined set |
|---|---|---|---|---|---|
| *A. Long range contacts* | | | | | |
| 1 | A-L | C-C | C-C | V-V | C-C |
| 2 | L-L | V-V | V-V | G-G | V-V |
| 3 | C-C | L-V | L-L | A-V | G-G |
| 4 | A-A | G-G | A-V | I-V | L-V |
| 5 | L-V | L-L | A-L | L-V | A-V |
| 6 | A-V | S-S | L-V | I-L | L-L |
| 7 | V-V | A-V | I-L | A-L | A-L |
| 8 | E-K | T-V | I-I | A-A | A-A |
| 9 | I-L | G-T | I-V | I-I | I-V |
| 10 | A-I | A-A | A-I | L-L | I-L |
| *B. Medium range contacts* | | | | | |
| 1 | A-A | A-A | L-V | A-A | A-A |
| 2 | L-L | S-S | A-A | E-K | L-L |
| 3 | A-K | C-C | D-K | L-L | E-K |
| 4 | E-K | D-G | G-G | A-G | A-K |
| 5 | D-K | G-G | A-L | G-G | D-K |
| 6 | L-V | I-V | E-K | A-L | A-L |
| 7 | F-F | L-V | C-C | A-D | L-V |
| 8 | I-L | D-R | A-D | I-L | G-G |
| 9 | A-E | G-S | A-S | A-K | A-G |
| 10 | G-L | G-V | W-W | L-V | I-L |

Table 7
Number of short and long range interactions in different structural classes and in different sizes of globular proteins

| Structural class/ size | $N_s$ | $N_l$ | $N_l/N_s$ |
|---|---|---|---|
| Structural class | | | |
| All-$\alpha$ proteins | 17562 | 9310 | 0.530 |
| All-$\beta$ proteins | 21974 | 28098 | 1.279 |
| $\alpha+\beta$ proteins | 25256 | 25130 | 0.995 |
| $\alpha/\beta$ proteins | 51834 | 55526 | 1.071 |
| Size | | | |
| Short | 8216 | 6954 | 0.846 |
| Medium | 35154 | 31260 | 0.889 |
| Large | 73256 | 79850 | 1.090 |

$N_s$ and $N_l$ are, respectively, number of short and long range contacts.

interactions in different aspects, namely (1) in four different structural classes and; (2) in different size of proteins. The results are shown in Table 7. The ratio between the number of interactions in long- and short-range is a good measure to estimate the importance of these interactions. We found that the long-range interactions are more important than short-range in all-$\beta$ class of proteins and they are equally important in $\alpha+\beta$ and $\alpha/\beta$ classes. The short-range interactions play a dominant role in all-$\alpha$ proteins. Among different size of proteins, we found that the influence of long-range interactions are higher than that of short range in proteins of large size.

These results suggest that the long-range interactions play an important role in the folding and stability of protein molecules.

## 4. Conclusions

Protein structures are stabilized by both local (short, medium) and long-range interactions. The analysis on different structural classes of proteins

shows that the residues in all-$\beta$ class of proteins have more long-range contacts than that in all-$\alpha$ proteins. Most of the long-range contacts are found to be in the distance of 4–10 residue far apart from the central residue as well as 11–20 and 21–30 ranges. The hydrophobic and aromatic residues have at least one long-range contact in all structural classes and the positively charged residues His, Lys and Arg have significant long-range contacts in all-$\beta$ class of proteins. The C-C residue pair and hydrophobic pairs are dominated in long-range contacts and the charged residue pairs D-K and E-K are, respectively, one of the most influenced pairs for medium range contacts in $\alpha+\beta$ and $\alpha/\beta$ class of proteins. The knowledge about favored residue distances and preferred residue pairs for long- and medium-range interactions in different structural classes obtained in the present study may help to improve the secondary/tertiary structure predictions and in the de novo design of proteins.

## References


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

[2] M.Ya. Karpeisky, V.A. Ilyin, Analysis of non-polar regions in proteins. J. Mol. Biol. 224 (1992) 629–638.

[3] N. Vtyurin, The role of local tight packing of hydrophobic groups in $\beta$-structure. Proteins 15 (1993) 62–70.

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

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

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

[7] C. Zhang, J.L. Cornette, C. Delisi, Consistency in structural energetics of protein folding and peptide recognition. Protein Sci. 6 (1997) 1057–1064.

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

[9] D.S. Gottfried, E. Haas, Nonlocal interactions stabilize compact folding intermediates in reduced unfolded bovine pancreatic trypsin inhibitor. Biochemistry 31 (1992) 12353–12362.

[10] A. Nayeem, H.A. Scheraga, A statistical analysis of side-chain conformations in proteins: comparison with ECEPP predictions. J. Protein Chem. 13 (1994) 283–296.

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

[12] J. Skolnick, A. Kolinski, A.R. Ortiz, MONSSTER: a method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265 (1997) 217–241.

[13] V.I. Abkevich, A.M. Gutin, E.I. Shakhnovich, Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. J. Mol. Biol. 252 (1995) 460–471.

[14] M.G. Reese, O. Lund, J. Bohr, H. Bohr, J.E. Hansen, S. Brunak, Distance distributions in proteins: a six-parameter representation. Protein Eng. 9 (1996) 733–740.

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

[16] M.M. Gromiha, S. Selvaraj, Influence of medium and long-range interactions in different structural classes of globular proteins. J. Biol. Phys. 23 (1997) 151–162.

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

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

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

[20] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, et al., The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112 (1977) 535–542.

[21] E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle, J. Weng, Protein Data Bank, in: F.H. Allen, G. Bergerhoft, R. Sievers editors. Crystallographic Databases — Information Content, Software Systems, Scientific Applications, 1987, pp. 107–132.

[22] T.J.P. Hubbard, B. Ailey, S.E. Brenner, A.G. Murzin, C. Chothia, SCOP: a structural classification of proteins database. Nucl. Acids Res. 27 (1999) 254–256.

[23] C.A. Orengo, F.M.G. Pearl, J.E. Bray, et al., The CATH database provides insights into protein structure/function relationships. Nucl. Acids Res. 27 (1999) 275–279.

[24] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (1983) 2577–2637.

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

[26] P. Manavalan, P.K. Ponnuswamy, A study of the pre-

ferred environment of amino acid residues in globular proteins. Arch. Biochem. Biophys. 184 (1977) 476–487.
[27] P. Manavalan, P.K. Ponnuswamy, Hydrophobic character of amino acid residues in globular proteins. Nature 275 (1978) 673–674.
[28] P.K. Ponnuswamy, Hydrophobic characteristics of folded proteins. Prog. Biophys. Mol. Biol. 59 (1993) 57–103.
[29] Z. Gugolya, Z. Dosztanyi, I. Simon, Interresidue interactions in protein classes. Proteins 27 (1997) 360–366.
[30] P.K. Ponnuswamy, M.M. Gromiha, Prediction of transmembrane helices from hydrophobic characteristics of proteins. Int. J. Pept. Protein Res. 42 (1993) 326–341.
[31] M.M. Gromiha, P.K. Ponnuswamy, Prediction of protein secondary structures from their hydrophobic characteristics. Int. J. Pept. Protein Res. 45 (1995) 225–240.
[32] G.D. Fasman, The development of the prediction of protein structure, in G.D. Fasman editor. Prediction of Protein Structure and Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 193–316.
[33] D.J. Barlow, J.M. Thornton, Ion-pairs in proteins. J. Mol. Biol. 168 (1983) 867–885.