

STATISTICAL PROPERTIES OF LONG-RANGE CONTACTS IN GLOBULAR PROTEINS*

Xiang-hong Wang**, Jian-hong Ke and Mang-xiao Hu

Department of Physics, Wenzhou Normal College, Wenzhou 325000, China

Abstract The analysis of residue-residue contacts in protein structures can shed some light on our understanding of the folding and stability of proteins. In this paper, we study the statistical properties of long-range and short-range residue-residue contacts of 91 globular proteins using CSU software and analyze the importance of long-range contacts in globular protein structure. There are many short-range and long-range contacts in globular proteins, and it is found that the average number of long-range contacts per residue is 5.63 and the percentage of residue-residue contacts which are involved in long-range ones is 59.4%. In more detail, the distribution of long-range contacts in different residue intervals is investigated and it is found that the residues occurring in the interval range of 4–10 residues apart in the sequence contribute more long-range contacts to the stability of globular protein. The number of long-range contacts per residue, which is a measure of ability to form residue-residue contacts, is also calculated for 20 different amino acid residues. It is shown that hydrophobic residues (including Leu, Val, Ile, Met, Phe, Tyr, Cys and Trp) having a large number of long-range contacts easily form long-range contacts, while the hydrophilic amino acids (including Ala, Gly, Thr, His, Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro) form long-range contacts with more difficulty. The relationship between the Fauchere-Pliska hydrophobicity scale (FPH) and the number of short-range and long-range contacts per residue for 20 amino acid residues is also studied. An approximately linear relationship between the Fauchere-Pliska hydrophobicity scale (FPH) and the number of long-range contacts per residue C_L is found and can be expressed as

$$C_L = a + b \times \text{FPH}$$

where $a = 5.04$ and $b = 1.23$. These results can help us to understand the role of residue-residue contacts in globular protein structure.

Keywords Long-range and short-range contact, Structural class, Globular protein, CSU software

INTRODUCTION

In computational biophysical chemistry, based on the crystal structures and residue-contacts available in the Protein Data Bank (PDB), a number of investigations have been carried out to understand the folding and stability of globular proteins^[1–7]. In protein molecules, short-range and long-range contacts are very important because there are effective attractions among the contacts. Tanaka *et al.* first categorized the inter-residue interactions into short, medium and long-range ones and proposed a hypothesis for protein folding by a three-step mechanism based on these interactions^[8]. The importance of long-range interactions has been stressed on various aspects such as the folding and stability of proteins^[9]. There are many publications about the relationship between protein sequence and structure based on residue-residue contacts. Gromiha and Selvaraj have investigated the

* This work was supported by the National Natural Science Foundation of China (Nos. 29874012, 20174036, and 20274040), and the Natural Science Foundation of Zhejiang Province (10102) and Science Technology Development Plan of Wenzhou City (S2002A014).

** Corresponding author: Xiang-hong Wang (王向红), E-mail: Wangxh@wznc.zj.cn, wzwdl@263.net

Received September 18, 2003; Revised November 19, 2003; Accepted November 27, 2003

preference of residues towards long-range interaction in different structural classes of globular proteins in terms of residue distance, and investigated the importance of long-range interaction in protein folding^[10].

In previous works, every residue is represented by a single point at the center of its side-chain atom position; and the position of the C^α atom is used for the glycine residue. Residues whose centers are closer than R_C are defined to be in contact^[2, 11–12]. In fact, each amino acid residue has a different structure. If the distance between two C^α atoms of residues is the same, the interactions between the two residues may be different for different residues. There are many types of contacts in proteins, such as hydrogen bonds, hydrophobic-hydrophobic, aromatic-aromatic, aromatic–polar, *etc.* In order to determine which type of contact contributes most to stabilizing the structure, a “contacts of structural units” (CSU) software package is adopted for calculation. There are many short-range and long-range contacts in globular proteins, and there are different influences on the stability of proteins for short-range and long-range contacts of amino acid residues. As amino acid residues which contribute towards long-range contacts need not be sequential neighbors and can exist far away from the chain sequence, it is of interest to determine the average probability of long-range contact for 20 amino acid residues and to establish the relationship between the ability to form short-range and long-range contacts and the Fauchere-Pliska hydrophobicity scale (FPH)^[13].

MATERIALS AND METHODS

Database

The crystallographic data of 91 globular proteins taken from the Protein Data Bank (PDB) of Brookhaven National Bank^[14, 15] forms the source for our present study. The PDB codes of the protein samples used in the present study are given in Table 1. We obtain the information about structural class and contacts of all proteins from SCOP^[16] and CSU^[17] databases.

Table 1. The PDB code of protein samples used in this paper

α class proteins								
1: 1CCR	2: 2MHR	3: 2END	4: 2HCO-A	5: 1BAB-A	6: 1HBG	7: 2FAM	8: 2PGD	
9: 2LHB	10: 2CMM	11: 2MYA	12: 4MBN	13: 1ABS	14: 1FCS	15: 1IFA	16: 2LIG-A	
17: 2LIG-B	18: 1FHA	19: 2ABK	20: 1PPN	21: 2CTS				
α/β class proteins								
22: 1AKT	23: 2ATC-B	24: 8ATC-B	25: 3DFR	26: 1OFV	27: 1Q21	28: 1GKY	29: 1AJE	
30: 3ADK	31: 1CLA	32: 3PGM	33: 1DHR	34: 2DRI	35: 2SBT	36: 1ULA	37: 1RHD	
38: 1ABF	39: 5ABP	40: 3CPA	41: 1SBP	42: 2HAD	43: 1AAX	44: 2LIV	45: 1MNS	
46: 1GOX	47: 3PGK	48: 4ICD	49: 4ENL	50: 3COX				
$\alpha + \beta$ class proteins								
51: 1FXI-A	52: 3SSI	53: 2RNS	54: 4TLN	55: 3LYZ	56: 1LHI	57: 9PAP	58: 1AIM	
59: 2ACT	60: 1EML	61: 2EMD	62: 2EMN	63: 2BAA	64: 1CDG			
β class proteins								
65: 1AKP	66: 2BFV-H	67: 2BFH	68: 1AHK	69: 2SNV	70: 1COB-B	71: 8I1B	72: 2SOD-O	
73: 2ALP	74: 2AYH	75: 2HFT	76: 5PTP	77: 2ENR	78: 3CNA	79: 2CNA	80: 3EST	
81: 1EST	82: 2GCH	83: 2BRD	84: 2CAB	85: 2PIA	86: 2-APR	87: 2AAA	88: 5ADH	
89: 3RN3	90: 1TFG	91: 3LDH						

Computation of Short-range and Long-Range Contacts

The computation of surrounding residues in a protein molecule has been described in references^[18–20]. If two amino acid residues are postulated to form a contact, intra- and inter-molecular interactions such as (1) the closest distance between their atoms, (2) the solvent- accessible surface of every atom; (3) the interaction of hydrophobic-hydrophobic, aromatic-aromatic, aromatic-polar, *etc.* and (4) hydrogen bonds^[21–30] should be considered. We calculate the residue-residue contacts using “contacts of structural units” (CSU) software because CSU software can provide more insights into the contacts^[17]. In CSU software, the PDB formatted file is again used as input. As output, CSU lists all chains, helices, sheets and strands indicted by a PDB file. The program can output all residue-residue contacts formed by any helix or strand therein, pointing out their specific contacts.

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, $\geq \pm 4$ residues as long-range contacts^[10, 28].

In order to know the statistical properties of long-range contacts in globular proteins, we investigate the distribution of long-range contacts in a large scale of residue length. The long-range contacts ($\geq \pm 4$ residues) are classified into several intervals with a step of 10 in the range of 4–100, and with a step of 20 (or 50) in the range of 100–500 (201–250; 251–300;; 451–500; > 500). In previous studies^[10, 28] the residue intervals are only classified into 6 intervals (4–10; 11–20; 21–30; 31–40; 41–50 and > 50). If the number of residues of proteins is large, the percentage (or the number) of long-range contacts in the interval of > 50 may be large, and the percentage (or the number) of long-range contacts in the other intervals (< 50) becomes small. In the same structural class of proteins, the number of residues for different proteins may be different, and this means that the average percentage of long-range in the interval may be incorrect.

The Average Probability of Contacts in Each Interval

In order to know the distribution of contacts in the different residue intervals, we define the probability of residue-residue contacts of the k -th interval in all the contacts, P_k as

$$P_k = \frac{N_k}{N_C} \quad (1)$$

Here N_k and N_C are the number of residue-residue contacts in the k -th interval and the total number of residue-residue contacts, respectively. The index k represents the residue interval, namely 1–3 (short-range contacts); 4–10, 11–20, 21–30, 31–40, ..., 91–100; 101–120, 121–140, ..., 181–200; 201–250, 251–300;; 451–500; > 500 (long-range contacts).

We also define the average probability of residue-residue contacts in the k -th interval $\bar{P}_k$ as

$$\bar{P}_k = \frac{\sum_{v=1}^M P_{k,v}}{M} \quad (2)$$

Here M is the total number of proteins, and $M = 91$ in this paper. The index v represents a protein molecule in all sample proteins. $P_{k,v}$ is the probability of residue-residue contacts of the k -th interval P_k for certain protein molecule. We calculate the average probability of residue-residue contacts $\bar{P}_k$ in the interval of 10 residues for different residue intervals.

We investigate the average probability of residue-residue contacts for four structural classes of globular proteins. The selected proteins are from four different structural classes of α , β , $\alpha + \beta$ and α / β proteins.

The Average Number of Contacts Per Residue for 20 Amino Acid Residues

The average number of residue-residue contacts per residue is different for the 20 amino acid residues. The average number of long-range contacts per residue C_L is calculated according to

$$C_{L,i} = \frac{\sum_j N_{ij}}{N_i} \quad (i = \text{Ala, Asp, } \dots, \text{Tyr}) \quad (3)$$

where $N_{i,j}$ is the number of long-range contacts between residue i and residue j . $\sum_j N_{ij}$ represents the total number of long-range contacts between residue i and all the 20 amino acid residues. N_i is the total number of residues. C_L describes the tendency to form long-range contacts. If C_L has a large value, this means that this residue is easily forms long-range contacts.

RESULTS AND DISCUSSIONS

Occurrence of Long- and Short-range Contacts for Different Residue Intervals

Distributions of probability of long-range contacts in different residue intervals for 91 proteins are presented in Fig. 1. The average probabilities of long-range contacts observed in different intervals are shown in Fig. 2. Here the average probability of long-range contacts is calculated in the interval of 10 residues. For example, in the range of 201–250 the total probability of long-range contacts is 1.55%, therefore, the average probability of long-range contacts in the interval of 10 residues is 0.31% because there are 5 intervals (5×10) in the range of 201–250. In Fig. 2, about 15.48% of long-range contacts is observed in the interval of 4–10, and about 2.95% of the contacts is in the range of > 200 . It is obvious that the average probability of long-range contacts decreases with increasing residue distance.

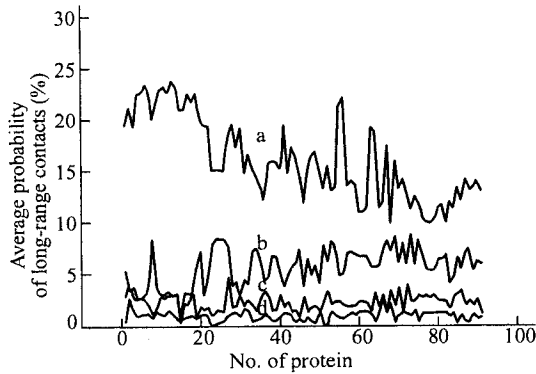


Fig. 1 Distribution of average probability of long-range contacts in different intervals
a) 4–10, b) 11–50, c) 51–100, and d) 101–200 for the 91 proteins given in Table 1
The average probability of long-range contacts is calculated in intervals of 10 residues.

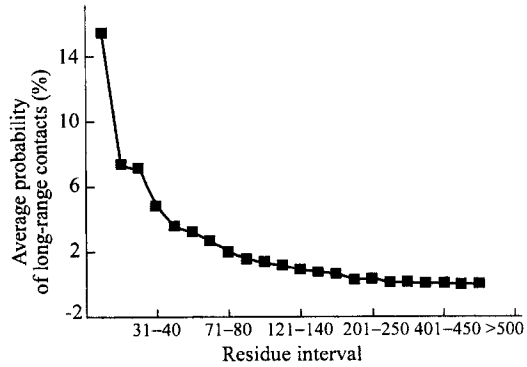


Fig. 2 Distribution of average probability of long-range contacts in different intervals
The probability is calculated in intervals of 10 residues.

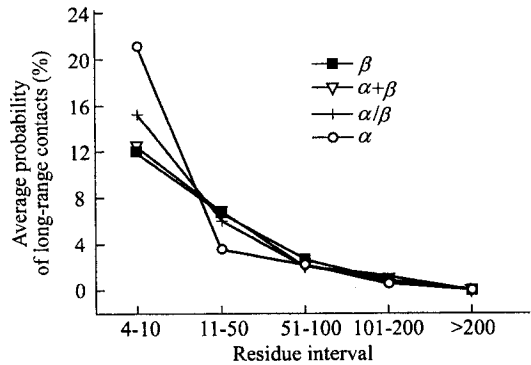


Fig. 3 Average probability of long-range contacts in different intervals for the four structure classes of proteins
The probability is calculated in the interval of 10 residues.

The Average Probability of Long-range Contacts for Different Structure Classes

Analysis of the average probability of long-range contacts for different structural classes of globular proteins reveals the importance of several physical interactions for their folding and stability. In Fig. 3, the average probability of long-range contacts in the different intervals for four structure classes is shown; here it is calculated in the interval of 10 residues. In the range of 4–10, the average probability of long-range contacts has the highest values for all the four structural classes. In the range of 4–10, the average probability of long-range

contacts in all α class proteins has the highest value 21.1% (in Fig. 3). However, it has the lowest value 3.62% in the range of 11–50. The average probability of long-range contact is 2.64% in the range of 51–100 for β class proteins, and proteins in α , $\alpha + \beta$, α/β classes have few contacts in the range of 51–100. This implies that in the range of 4–10 there are more residues to form long-range contacts for the four classes. The β class proteins have more long-range contacts in the range of 11–50, and this may be due to the specific hydrogen bonding pattern of α -helices and β -strands in these classes. It is interesting to note that the average probability of long-range contact varies with the residue interval.

In Fig. 4, we give the average probabilities of long-range and short-range contacts for the four structure classes of proteins. We find that the average probability of long-range contacts over all the β class proteins is 64.0%, and this is the highest value. However, the average probability is only 51.8% for the α class proteins. In Fig. 4(b) the average probabilities of short-range contacts for the four structure classes of proteins are also shown. Contrary to the average probability of long-range contacts, the average probability of short-range contacts is largest for the α class proteins.

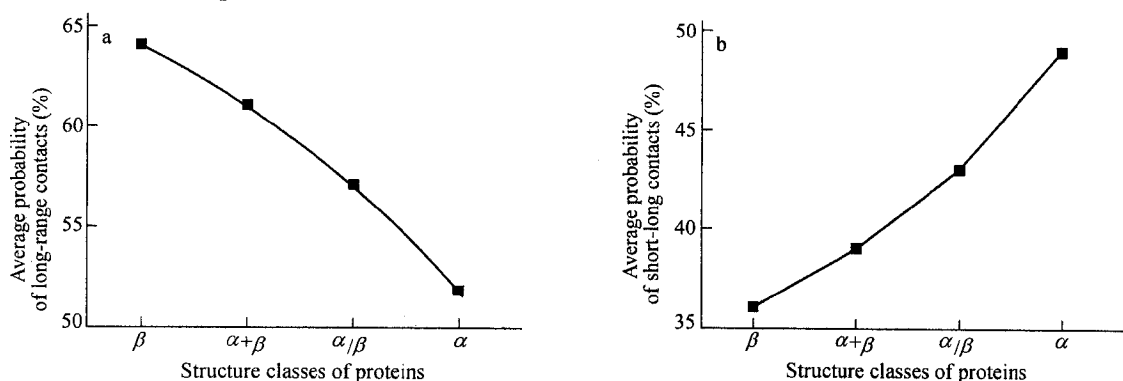


Fig. 4 Average probability of all long-range and short-range contacts for the four structure classes of proteins
a) Long-range contacts, b) Short-range contacts

The Ability to Form Long-range Contacts for Different Amino Acid Residues

For different amino acid residues, the ability to form residue-residue contacts is different. We first calculate the average number of long-range contacts per residue in different intervals, and the results are given in Fig. 5. In Fig. 5, we find that Leu, Val, Ile, Met, Phe, Tyr, Cys, and Trp have large average numbers of long-range contacts in different intervals, and Ala, Gly, Thr, His, Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro have small average numbers of long-range contacts in different intervals. Experiments and theoretical studies have shown that the amino acids are subdivided into two kinds of residues: hydrophobic (H) and hydrophilic (P)^[31]. In general, hydrophobic residues include Leu, Val, Ile, Met, Phe, Tyr, Cys, and Trp, and hydrophilic residues include Ala, Gly, Thr, His, Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro. In Fig. 5, we find that hydrophobic residues have large average numbers of long-range contacts and hydrophilic residues have small average numbers of long-range contacts in the different intervals.

In Fig. 6, we plot the average number of long-range contacts in the different residue intervals for hydrophobic and hydrophilic residues. Figure 6(a) shows the case of hydrophobic residues, and Fig. 6(b) the case of hydrophilic residues. We find that the average number of long-range contacts per residue $\bar{C}_L$ is largest in the range of 4–10, and it decreases with increasing residue interval. However, the values of hydrophobic residues are greater than those of hydrophilic residues in the same residue interval. For example, $\bar{C}_L = 2.90$ for hydrophobic residues and $\bar{C}_L = 1.67$ for hydrophilic residues in the interval range of 4–10. These values could help us establish clearly that the long-range contacts play an important role in protein folding.

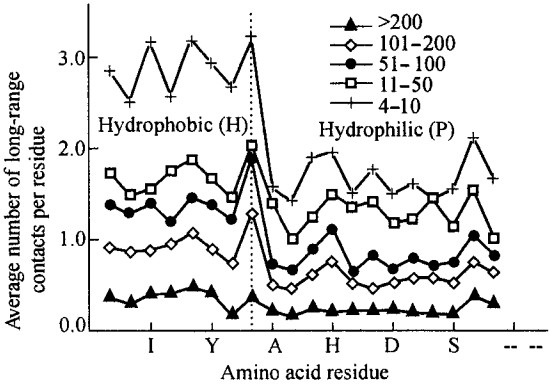


Fig. 5 Average number of long-range contacts per residue for the 20 amino acid residues

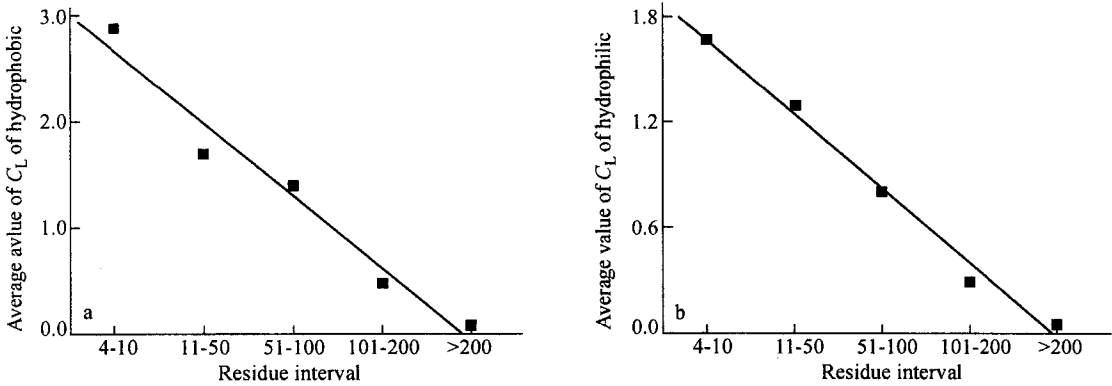


Fig. 6 Average value of $\overline{C}_L$ in different intervals for hydrophobic (a) and hydrophilic (b) residues

Table 2. The average number of long-range contacts per residue in different intervals*

20 amino acids	C_L	Case A		Case B		Case C		Case D		Case E	
		C_L	$\overline{C}_L$	C_L	$\overline{C}_L$	C_L	$\overline{C}_L$	C_L	$\overline{C}_L$	C_L	$\overline{C}_L$
LEU	7.213	0.366		0.907		1.380		1.725		2.835	
VAL	6.439	0.293		0.865		1.280		1.487		2.514	
ILE	7.384	0.396		0.879		1.399		1.544		3.166	
MET	6.862	0.402	0.358	0.952	0.945	1.185	1.394	1.742	1.690	2.581	2.889
PHE	8.080	0.474		1.064		1.457		1.882		3.203	
TYR	7.275	0.417		0.891		1.375		1.655		2.937	
CYS	6.209	0.160		0.717		1.202		1.459		2.671	
TRP	8.743	0.355		1.285		1.873		2.027		3.203	
ALA	4.380	0.208		0.496		0.721		1.386		1.570	
GLY	3.666	0.157		0.446		0.652		0.995		1.416	
THR	4.886	0.239		0.617		0.888		1.246		1.896	
HIS	5.537	0.203		0.769		1.110		1.496		1.959	
GLU	4.218	0.216		0.520		0.632		1.345		1.505	
GLN	4.685	0.217	0.222	0.461	0.574	0.826	0.802	1.416	1.283	1.765	1.666
ASP	4.099	0.226		0.530		0.676		1.171		1.496	
ASN	4.386	0.196		0.573		0.794		1.214		1.609	
LYS	4.367	0.181		0.583		0.710		1.447		1.446	
SER	4.104	0.169		0.518		0.744		1.128		1.545	
ARG	5.828	0.374		0.745		1.047		1.542		2.120	
PRO	4.408	0.281		0.629		0.825		1.010		1.663	

*The cases A, B, C, D, and E are in the intervals of > 200, 101–200, 51–100, 11–50, and 4–10, respectively. C_L represents the total average number of long-range contacts per residue for different amino acid residues

As the number of short-range and long-range contacts per amino acid residues is a measure of capability of forming short-range and long-range contacts, we calculate the average number of short-range and long-range contacts per residue C_S and C_L for 20 amino acid residues. The results are given in Table 2. We could find that the hydrophobic residues such as Leu, Val, Ile, Met, Phe, Tyr, Cys, and Trp have larger values of C_L . We also find that the hydrophilic residues such as Ala, Gly, Thr, His, Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro have smaller values of C_L . This means that hydrophilic residues are more efficient than hydrophilic residues in forming long-range contacts.

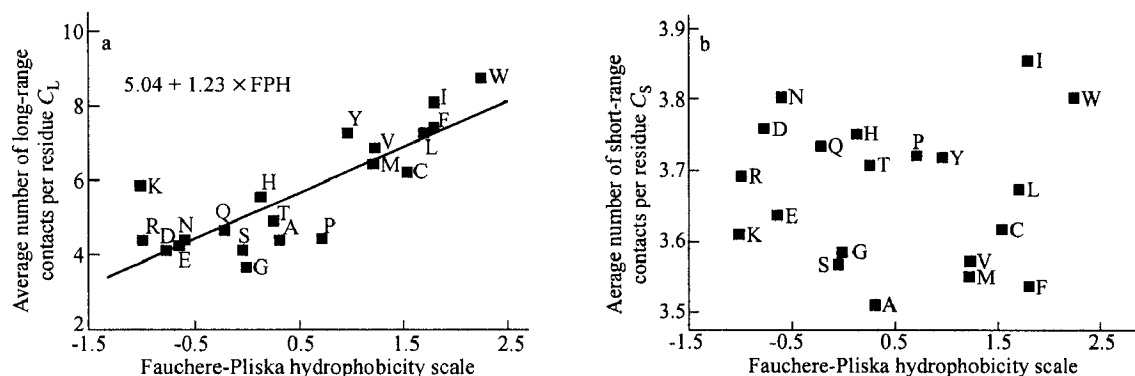


Fig. 7 Average number of long-range and short-range contacts per residue versus Fauchere-Pliska hydrophobicity scale of amino acid residues
a) Long-range contact, b) Short-range contact

We also find that the average number of long-range contacts per residue is in accord with the free energies of transfer of the amino acid residues from aqueous to nonpolar environments by Faucheres and Plisska^[13]. In Fig. 7(a) we plot the average number of long-range contacts per residue C_L versus the Fachere-Pliska hydrophobicity scale (FPH), and find that the value of C_L increases with increasing FPH values. An approximately linear relationship between C_L and the Fachere-Pliska hydrophobicity scale (FPH) was obtained, which can be expressed as

$$C_L = a + b \times \text{FPH} \quad (4)$$

where $a = 5.04$ and $b = 1.23$. A plot of the average number of short-range contact per residue C_S versus the Fachere-Pliska hydrophobicity scale (FPH) is given in Fig. 7(b), and no clear regular relationship between them can be found. Therefore, it is supposed that the free energy of transfer of the amino acid residues from aqueous to nonpolar environments depends mainly on the number of long-range contacts per residue, and is independent of short-range contacts. These results may help us to understand the importance of long-range contacts in protein stability.

REFERENCES

- 1 Ponnuswamy, P.K. and Gromiha, M.M., *J. Theory Biol.*, 1994, 166: 63
- 2 Bahar, M. and Kaplan, R.L., *Proteins*, 1997, 29: 309
- 3 Zhang, C., Cornette J.L. and Delisi C., *Protein Sci.*, 1997, 6: 1057
- 4 Kumareve, T.S. and Ponnuswamy, M.N., *Biophys. Chem.*, 1998, 75: 105
- 5 Russell, R.B., Saqi, M.A., Sayle, P.A., Sternberg, M.J., *Protein Eng.*, 1998, 11: 1
- 6 Gromiha, M.M., Oobatatke, M., Kono, H., Uedaira, H. and Sarai, A., *Protein Eng.*, 1999, 12: 557
- 7 Gromiha, M.M. and Selvaraj, S., *Recent Res. Devel. Biophys. Chem.*, 2000, 1: 1
- 8 Tanaka, S. and Scheraga, H.A., *Proc. Natl. Acad. Sci.*, 1975, 72: 3802

- 9 Gottfried, D.S. and Hass, E., *Biochemistry*, 1992, 31: 12353
- 10 Gromiha, M.M. and Selvaraj, S., *Biophys. Chem.*, 1999, 77: 49
- 11 Miyazawa, S. and Jernigan, R.L., *J. Mol. Biol.*, 1996, 256: 623
- 12 Miyazawa, S. and Jernigan, R.L., *Macromolecules*, 1985, 18: 534
- 13 Fauchere, J.L. and Pliska, V., *Eur. J. Medicinal. Chem.*, 1983, 18: 369
- 14 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., *J. Mol. Biol.*, 1977, 112: 535
- 15 Berman, H.M., Westbrook, J. and Feng, Z., *Nucleic Acids Res.*, 2000, 28: 235
- 16 Hubbard, T.J.P., Ailey, B., Brenner, S.E., Murzin, A.G. and Chothia, C., *Nucleic Acids Res.*, 1999, 27: 254
- 17 Sobolev, V. and Sorokine, A., *Prilusky, J., Bioinformatics*, 1999, 1: 5327
- 18 Selvaraj, S. and Gromiha, M.M., *J. Protein Chem.*, 1998, 17: 407
- 19 Jiang, Z.T., Zhang, L.X. and Xia, A.G., *Polymer*, 2002, 43: 6037
- 20 Chen, J., Zhang, L.X. and Li, J., *Biophys. Chem.*, 2003, 105: 11
- 21 Miyazawa, S. and Jernigan, R.L., *Protein Eng.*, 1993, 6: 267
- 22 Russell, R.B. and Barton, G.J., *J. Mol. Biol.*, 1994, 244: 332
- 23 Wallace, A.C., Laskowski, R.A. and Thornton, J.M., *Protein Eng.*, 1995, 8: 127
- 24 Swindells, M.B., *Protein Sci.*, 1995, 4: 93
- 25 Fiser, A., Dosztányi, Zs. and Simon, I., *Comp. App. Biosci.*, 1997, 13: 297
- 26 Selbig, J. and Argos, P., *Proteins*, 1998, 31: 172
- 27 McDonald, I. K. and Thornton, J.M., *J. Mol. Biol.*, 1994, 238: 77
- 28 Gromiha, M.M. and Selvaraj, S., *J. Biol. Phys.*, 1997, 23: 151
- 29 Chen, J., Zhang, L.X., Jiang, Z.T., Xia, A.G., Ye, G.X., and Zhao, D.L., *Acta Polym. Sin.*, 2003, (2): 171
- 30 Jiang, Z.T., Zhang, L.X., Chen, J., Xia, A.G., Ye, G.X. and Zhao, D.L., *Acta Polym. Sin.*, 2003, (2): 207
- 31 Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. and Mecht, M.H., *Science*, 1993, 262: 1680