RESEARCH PAPERS

Acta Cryst. (1997). D53, 627–637

Determinants of Backbone Packing in Globular Proteins: an Analysis of Spatial Neighbours

SANTOSH K. PANJIKAR,[a]† MARGARET BISWAS[b] AND SARASWATHI VISHVESHWARA[c]*

[a]*The School of Biotechnology, Devi Ahilya University, Indore, India,* [b]*Bioinformatics Centre, Indian Institute of Science, Bangalore 560012, India, and* [c]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India. E-mail: sv@mbu.iisc.ernet.in*

(*Received 30 September 1996; accepted 2 April 1997*)

Abstract

This study attempts to examine the pattern and variability of backbone packing density in protein structures. A carefully selected non-redundant data set of known protein structures is analyzed in terms of amino-acid composition and the preference of individual amino acids to fall into regions of low, medium or high density depending on the number of observed non-sequence spatial neighbours. The relationship of the backbone packing density to a number of properties such as the hydrophobicity, non-bonded energies and secondary structural features of the amino acids are examined. The correlation between the average percentage composition and the percentage composition in regions corresponding to different levels of packing density of the proteins is evaluated. These studies are extended to the family of globins whose amino-acid sequences have diverged retaining the same three-dimensional structure during evolution. The significance of high-backbone-density regions in this family has become apparent as due to helix/helix packing. Further, the variation in the amino-acid composition in different contact regions of globin proteins follows the same pattern found for the general data set.

1. Introduction

The amino-acid sequence of a protein is known to contain the information required for it to fold into its native three-dimensional structure. A large number (>4000) of protein crystal structures have been solved at atomic resolution and the coordinates of the atoms are stored in the Brookhaven Protein Data Bank (PDB). A number of workers have used these coordinates in analyses which aim to decode the message for protein folding. These studies have resulted in some degree of success in predicting secondary structures such as helix, β-sheet and turn from sequence information. While accurate secondary structures can be predicted for some sequences (Rooman, Kocher & Wodak, 1992; Barton, 1995) there are also several examples of the same stretch of amino-acid sequences adopting different conformations in different proteins (Cohen, Presnell & Cohen, 1993). This reflects the conformational flexibility of small peptide fragments and suggests that in many cases the conformations they adopt are dictated by the protein environment.

In protein folding it is clear, from many observations, that non-local interactions play a major role in determining the three-dimensional structure of the protein (Dill *et al.*, 1995). For example, there are instances of proteins with no sequence similarity taking up similar three-dimensional structures (Flores, Orengo, Moss & Thornton, 1993). Theoretically, it has been possible to design folding sequences of heteropolymers by controlling the non-local interactions (Go & Abe, 1981; Shrivastava, Vishveshwara, Cieplak, Maritan & Banavar, 1995). Protein data analyses from this point of view have been carried out in the context of the development of potential functions (Tanaka & Scheraga, 1976; Warme & Morgan, 1978; Miyazawa & Jernigan, 1985; Crippen & Vishwanadhan, 1985; Maiorov & Crippen, 1992; Johnson, Overington & Blundell, 1993; Bryant & Lawrence, 1993; Johnson, Srinivasan, Sowdhamini & Blundell, 1994; Sippl, 1990; Kocher, Rooman & Wodak, 1994), the analysis of side-chain packing (Heringa & Argos, 1991) and in redefining hydrophobicity (surrounding hydrophobicity) (Ponnuswamy, 1993). However, a reliable algorithm for obtaining the tertiary structure from sequence information alone is not yet in sight (Eisenhaber, Persson & Argos, 1995). An unambiguous distinction between the native conformation and the incorrect alternatives cannot, as yet, be satisfactorily made. A recent evaluation of current techniques for *ab inito* protein structure prediction (Defay & Cohen, 1995) makes it clear that accurate tertiary structure prediction is not yet possible.

In the present study we have carried out a systematic analysis of non-local interactions from the point of view

† Part of this work was carried out at the Bioinformatics Centre, Indian Institute of Science, Bangalore, India, towards partial fulfillment of an MSc thesis. Present address: Institute for Medical Physics and Biophysics, University of Münster, Robert Koach-Strass 31, D-48149 Münster, Germany.

of the packing of the protein backbone around the Cα atoms. Although a large number of sophisticated analyses, mentioned above, are available, the present work gives new insight into the problem of protein folding and its relation to the composition and distribution of amino acids in different regions of proteins. Specific amino-acid residues are classified as low-, medium- and high-contact residues depending on the number of non-sequential Cα atoms which fall within a sphere of radius 6.5 Å centred on each of the Cα atoms. This is similar to the concept of association indices developed by Karlin, Zuker & Brocchieri (1994). Our results are analyzed in terms of the preference for each of the amino-acid residues, when taken as the central residue, to be in different contact regions. The relation of this preference to a number of amino-acid properties, like hydrophobicity and long-range interaction energies, is discussed. Using this approach, we have obtained a novel result which relates the percentage composition of amino acids in proteins to different levels of backbone density (compositional equilibrium/non-equilibrium). This has been shown to hold in globins, a family of related proteins. Previous studies have shown (Nakashima, Nashikawa & Ooi, 1986; Chou, 1989, 1995; Dubchak, Holbrook & Kim, 1993) a correlation between amino-acid composition and the protein class. However, for the first time, the present study demonstrates a correlation between amino-acid composition and backbone density in certain parts of a given protein. The implications of these results are discussed in the context of protein folding which is generally considered to take place in two stages: (1) the hydrophobic collapse (a sequence-independent step) and (2) final attainment of unique structure (a sequence-dependent phenomenon).

2. Methods

2.1. *Data set*

The protein structures used in the analyses were taken from the Brookhaven Protein Data Bank using the PDB SELECT subdatabase (August 1994) (Sander & Schneider, 1991). Proteins with less than 25% sequence homology and with resolutions of 3.0 Å or better were selected for use in the present study. The 187 proteins which form this data set are listed in Table 1(*a*) along with the number of amino-acid residues which they contain. The globin structures were selected based on samples from different evolutionary species (Lesk & Chothia, 1980) and are presented in Table 1(*b*).

2.2. *Identification of spatial neighbours*

The number of Cα atoms falling within a sphere of chosen radius around each Cα (C_i) along the chain (excluding the two sequence neighbours, $C_i - 2$, $C_i - 1$, $C_i + 1$ and $C_i + 2$) were identified as spatial neighbours (referred to as contacts) for each residue of the protein. For proteins with more than one subunit, only one of the subunits was considered. However, all subunits of globins are independently considered since they have distinctly different sequences. A sphere of radius 6.5 Å was used in the present investigation, which was found to be the distance corresponding to the first peak in the radial distribution of residues in the interior of proteins in similar analysis (Miyazawa & Jernigan, 1985, 1996).

2.3. *Classification of the central residue (C_i) on the basis of the number of neighbours (contacts)*

The identification of spatial neighbours around C_i, as described above, was carried out on all the residues of the proteins in the data set (Tables 1*a* and 1*b*). The number of spatial neighbours (contacts) found for the central residues (C_i) ranged from 0 to 10. The total number of each amino acid in a given contact was obtained by summing up the respective values for that amino acid in all the proteins. The results of this classification on the general data set (Table 1*a*) are presented in Table 2.

2.4. *Data analysis*

2.4.1. *Percentages.* (i) The total number of each of the 20 different amino acids in contacts 0–8 are shown in Table 2. Each number a_{ij}, corresponds to the number of amino acids of type i in jth contact, where i varies from 1 to 20 and j from 0 to 8.

$$A_{ij} = (a_{ij})(100)/\sum_{p=0}^{8\ \text{contacts}} a_{ip} \qquad (1)$$

is the percentage of the ith amino acid in the jth contact with respect to the sum of all contacts (see Fig. 1).

(ii) The parameter B_{ij} defined as

$$B_{ij} = (a_{ij})(100)/\sum_{q=1}^{20\ \text{amino acids}} a_{qj} \qquad (2)$$

is the percentage of the ith acid in the jth contact, with respect to the sum over all the amino acids.

(iii) The total number of a given amino acid i, summed over all the contacts is given by

$$A_i = \sum_{p=0}^{8\ \text{contacts}} a_{ip}. \qquad (3)$$

The average percentage composition of each amino acid i in all contacts is given by

$$\bar{A}_i = (A_i)(100)/\sum_{k=1}^{20\ \text{amino acids}} A_k. \qquad (4)$$

(iv) The total number of amino acids in contact j is given by

$$B_j = \sum_{q=1}^{20\ \text{amino acids}} A_{qj} \qquad (5)$$

Table 1. *Tertiary structure data set and globin structures*

(*a*) Tertiary structure data set from PDB SELECT database, August 1994 (Sander & Schneider, 1991). Homology <25%, number of residues in the chain (*N*res) > 100 and resolution of the crystal structure <3.0 Å.

PDB code	*N*res	Resolution (Å)	PDB code	*N*res	Resolution (Å)	PDB code	*N*res	Resolution (Å)	PDB code	*N*res	Resolution (Å)
102L	163	1.74	1GOX	350	2.00	1RHD	293	2.50	2STV	184	2.50
1AAJ	105	1.80	1GP1A	185	2.00	1RND	124	1.50	2TBVA	287	2.90
1AAK	150	2.40	1GPB	823	1.90	1RVEA	244	2.50	2TMVP	154	2.90
1ABH	321	1.70	1GPR	158	1.90	1S01	275	1.70	3ADK	194	2.10
1ABK	211	2.00	1GRCA	193	3.00	1SGT	223	1.70	3APP	323	1.80
1ABMA	198	2.20	1GSTA	217	2.20	1SHAA	103	1.50	3CBH	365	2.00
1ABMB	198	2.20	1HGEB	175	2.60	2SIC	275	1.80	3CHY	128	1.66
1ACE	526	2.80	1HILA	217	2.00	1SNC	135	1.65	3CLA	213	1.75
1ADA	349	2.40	1HSDA	255	2.60	1SPA	396	2.00	3DFR	163	1.70
1ADS	315	1.60	1IFA	158	2.60	1TFG	112	1.95	3GAPA	208	2.50
1ARB	263	1.20	1LAP	481	2.70	1THO	109	2.30	3GBP	305	2.40
1ASOA	552	2.20	1LPE	144	2.25	1TIE	166	2.50	3GRS	461	1.54
1ATNA	372	2.80	1LTSA	185	1.95	1TLK	103	2.80	3INKC	122	2.50
1AVHA	318	2.30	1LTSD	103	1.95	1TNFA	152	2.60	3PGK	415	2.50
1AYH	214	2.00	1LZ3	129	1.50	1TPT	440	2.80	3RUBS	123	2.00
1BAA	243	2.80	1MAMH	217	2.50	1TRB	316	2.00	3SC2A	254	2.20
1BBHA	131	1.80	1MBD	153	1.40	1TROA	104	1.90	3SODO	151	2.10
1BBPA	173	2.00	1MDC	132	1.80	1ULA	289	2.75	3TGL	265	1.90
1BBT1	186	2.60	1MINA	468	2.20	1VSGA	362	2.90	4BLMA	256	2.00
1BBT2	210	2.60	1MINB	522	2.20	1WSYA	248	2.50	4BP2	117	1.60
1BTC	491	2.00	1MUP	157	2.40	1WSYB	385	2.50	4ENL	436	1.90
1CAJ	258	1.90	1NIPB	287	2.90	256BA	106	1.40	4FGF	124	1.60
1CBX	307	2.00	1NRD	333	2.30	2AAA	476	2.10	4FXN	138	1.80
1CCR	111	1.50	1OFV	169	1.70	2AZAA	129	1.80	4GCR	185	1.50
1CD8	114	2.60	1OMF	340	2.40	2CDV	107	1.80	4GPD1	333	2.80
1CID	177	2.80	1OMP	370	1.80	2CMD	312	1.87	4ICD	414	2.50
1CLM	144	1.80	1OVB	159	2.30	2CTS	437	2.00	4RCRH	237	2.80
1CMBA	104	1.80	1PAFA	262	2.50	2CYP	293	1.70	4SBVA	199	2.80
1CPCA	162	1.66	1PBXA	142	2.50	2DNJA	253	2.00	4TMS	316	2.35
1CPL	165	2.50	1PCDA	201	2.80	2HAD	310	1.90	4TS1A	317	2.50
1DHR	236	2.30	1PDA	296	1.80	2ILA	145	2.30	5FBPA	314	2.10
1DPI	546	2.80	1PFKA	320	2.40	2LBP	346	2.40	5NN9	388	2.30
1DRI	271	1.70	1PGD	469	2.50	2MADL	124	2.25	5P21	166	1.35
1EAF	243	2.30	1PHG	405	1.60	2MEV1	268	3.00	7TIMA	247	1.90
1ECO	136	1.40	1PHH	394	2.30	2MHR	118	1.70	7XIA	387	1.90
1END	137	1.60	1PHS	364	3.00	2PF2	145	2.20	8ABP	305	1.49
1ETU	177	2.90	1PHY	126	2.40	2PIA	321	2.00	8ACN	753	2.00
1EZM	298	1.50	1PPFE	218	1.80	2PLV1	288	2.88	8ADH	374	2.40
1FBAA	360	1.90	1PPL	323	1.70	2PLV3	235	2.88	8ATCA	310	2.50
1FC1A	207	2.90	1PPN	212	1.60	2PMGA	561	2.70	8ATCB	146	2.50
1FDD	106	1.90	1PTE	348	2.80	2POR	301	1.80	8CATA	498	2.50
1FHA	172	2.40	1PYAB	228	2.50	2REN	320	2.50	8I1B	146	2.40
1FNR	296	2.20	1PYP	281	3.00	2RN2	155	1.48	9LDTA	331	2.00
1GKY	186	2.00	1RBP	175	2.00	SCPA	174	2.00	9RNT	104	1.50
1GLAG	489	2.60	1RCB	129	2.20	2SGA	181	1.50	9RUBB	458	2.60
1GLY	470	2.20	1REA	304	2.70	2SICI	107	1.80	9WGAA	171	1.80
1GMFA	119	2.40	1RHB	110	1.90	2SNV	151	2.80			

(*b*) The globin structures selected for analysis

PDB code	Protein	Resolution (Å)
1ECO	Larval insect erythrocrurin	1.4
1MBA	Sea hare myoglobin	1.6
1MBN	Sperm whale myoglobin	2.0
1MBS	Seal myoglobin	2.5
2DHB	Horse hemoglobin	2.8
1FDH	Human fetus hemoglobin	2.5
1HDS	Deer sickle cell hemoglobin	1.98
1NIH	Human carbonmonoxy hemoglobin	2.6
1PBX	Antartic fish hemoglobin	2.5
2LHB	Sea lamprey hemoglobin	2.0

Table 2. *Frequency of occurrence of different amino acids in the various contacts 0–8 in the data set given in Table 1*

$\bar{A}_i$ = average percentage composition of amino acid. M_i = the mean contact of the amino acid of the type (i). B_j = percentage of total amino acids in the contact j.

Contact / Amino acid	0	1	2	3	4	5	6	7	8	$\bar{A}_i$	M_i
C	15	41	66	106	151	124	98	80	30	1.43	4.43
V	132	174	379	626	718	628	529	217	53	6.95	4.05
I	76	156	325	486	611	518	324	151	21	5.37	3.93
A	229	280	603	680	826	689	486	365	115	8.59	3.92
Y	76	125	258	321	406	319	210	96	24	3.69	3.81
L	131	248	558	785	1018	775	449	170	27	8.37	3.80
M	46	66	144	177	258	188	124	47	13	2.14	3.79
W	29	47	94	138	158	122	83	36	7	1.44	3.78
F	98	110	295	388	435	334	266	90	11	4.08	3.75
G	433	454	767	617	544	529	403	261	130	8.32	3.41
T	232	286	517	539	532	395	289	113	26	5.89	3.34
H	77	100	204	216	215	152	87	34	12	2.21	3.33
R	138	199	385	449	493	277	156	58	10	4.36	3.29
S	356	345	580	514	464	356	291	171	58	6.31	3.25
Q	151	174	347	370	371	229	101	34	8	3.59	3.09
N	264	275	446	432	383	261	176	68	15	4.67	3.00
K	295	379	605	570	626	280	156	43	8	5.96	2.87
E	301	376	693	613	604	240	133	48	14	6.08	2.81
D	390	393	643	549	515	280	155	63	17	6.04	2.76
P	342	410	454	430	294	184	95	38	5	4.53	2.48
B_j	7.7	9.3	16.8	18.1	19.4	13.8	9.3	4.4	1.2		

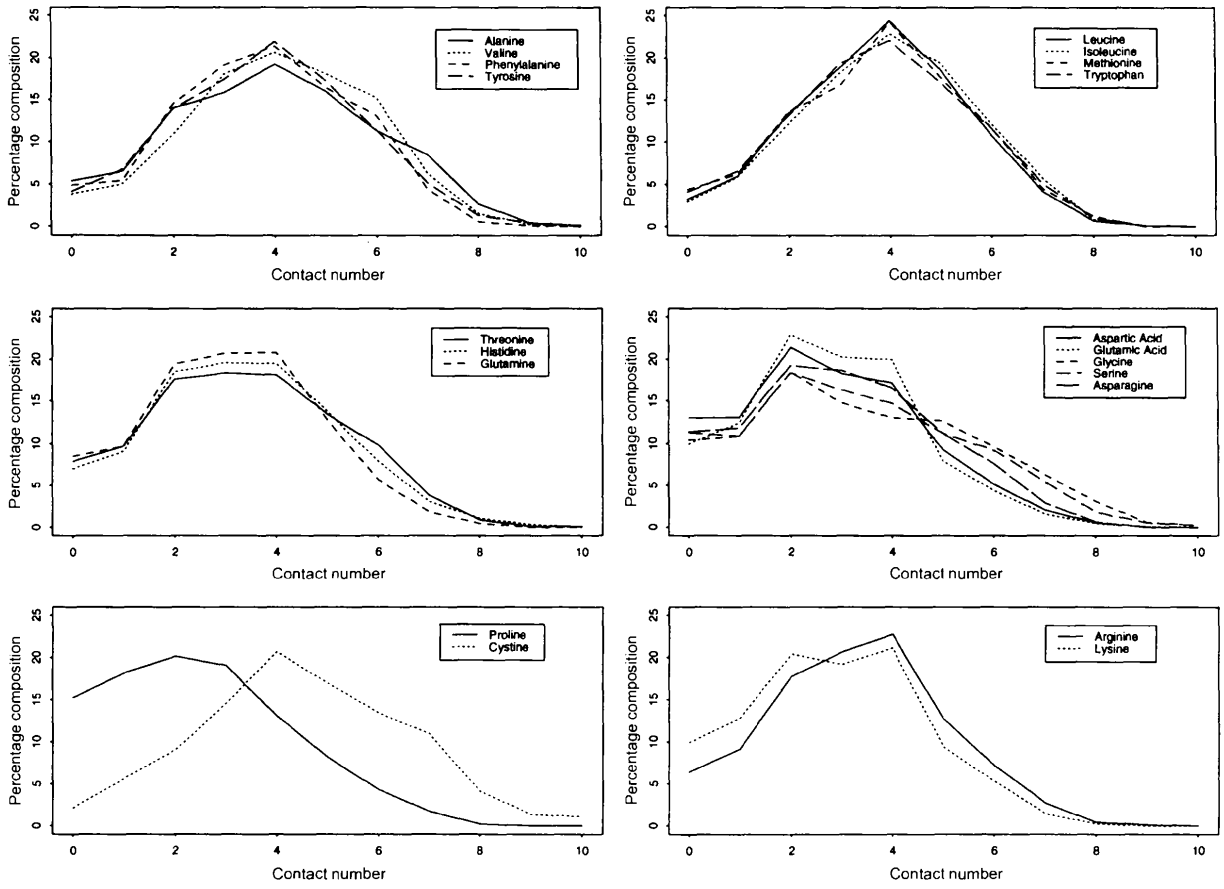

Fig. 1. The percentage composition of amino acids in different contacts as calculated using equation (1).

and the average percentage of amino acids in each of the j contacts is given by

$$\bar{B}_j = (B_j)(100)/\sum_{l=0}^{8\ \text{contacts}} B_l. \quad (6)$$

It may be noted that the denominators in (4) and (6) are identically equal to the total number of all amino acids in all contacts, *i.e.*, $\sum_{i=1}^{20}\sum_{j=0}^{8} a_{ij}$.

The parameter $(B_{ij} - A_i)$ represents the deviation of the percentage composition of the ith amino acid in the jth contact from the average composition and this quantity in contacts 0–8 for the 20 amino acids is given as bar diagrams in Fig. 2.

2.4.2. *Mean contact.* The mean contact for a given Cα, C_i, is evaluated as

$$M_i = \sum_{j=0}^{8\ \text{contacts}} j a_{ij}/A_i. \quad (7)$$

2.4.3. *Distance in compositional space.* The distance in compositional space between the pth and qth contact is given by

$$d_{pq} = [\sum_{i=1}^{20\ \text{amino acids}} (B_{ip} - B_{iq})^2]^{1/2}. \quad (8)$$

where B_{ip} and B_{iq} are as described in (2).

2.4.4. *Correlation coefficient.* The Pearson correlation coefficient is evaluated between various parameters. These are: (i) the correlation between average amino-acid composition of the data set and the composition in a given contact; (ii) the correlation between the mean contact (M_i) and some of the amino-acid properties; and (iii) the correlation between the percentage composition of amino acids in a given protein with the percentages in a given contact of that protein. The correlation coefficients obtained in different contact regions for the general data set are presented as bar diagrams in Fig. 3,

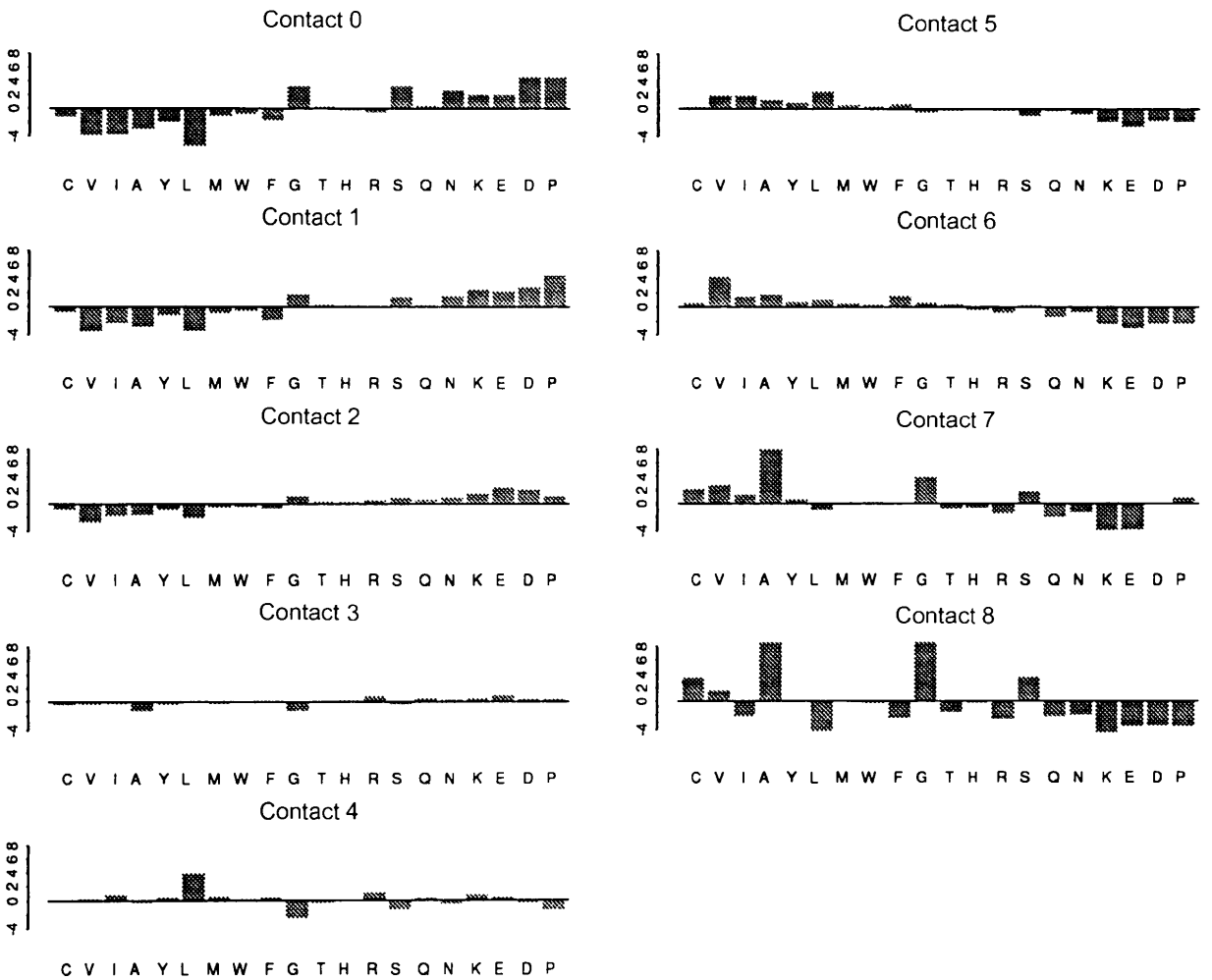

Fig. 2. Bar diagrams of the difference between the composition of C_i in contacts 0–8 [using equation (2)] and the average composition [using equation (4)].

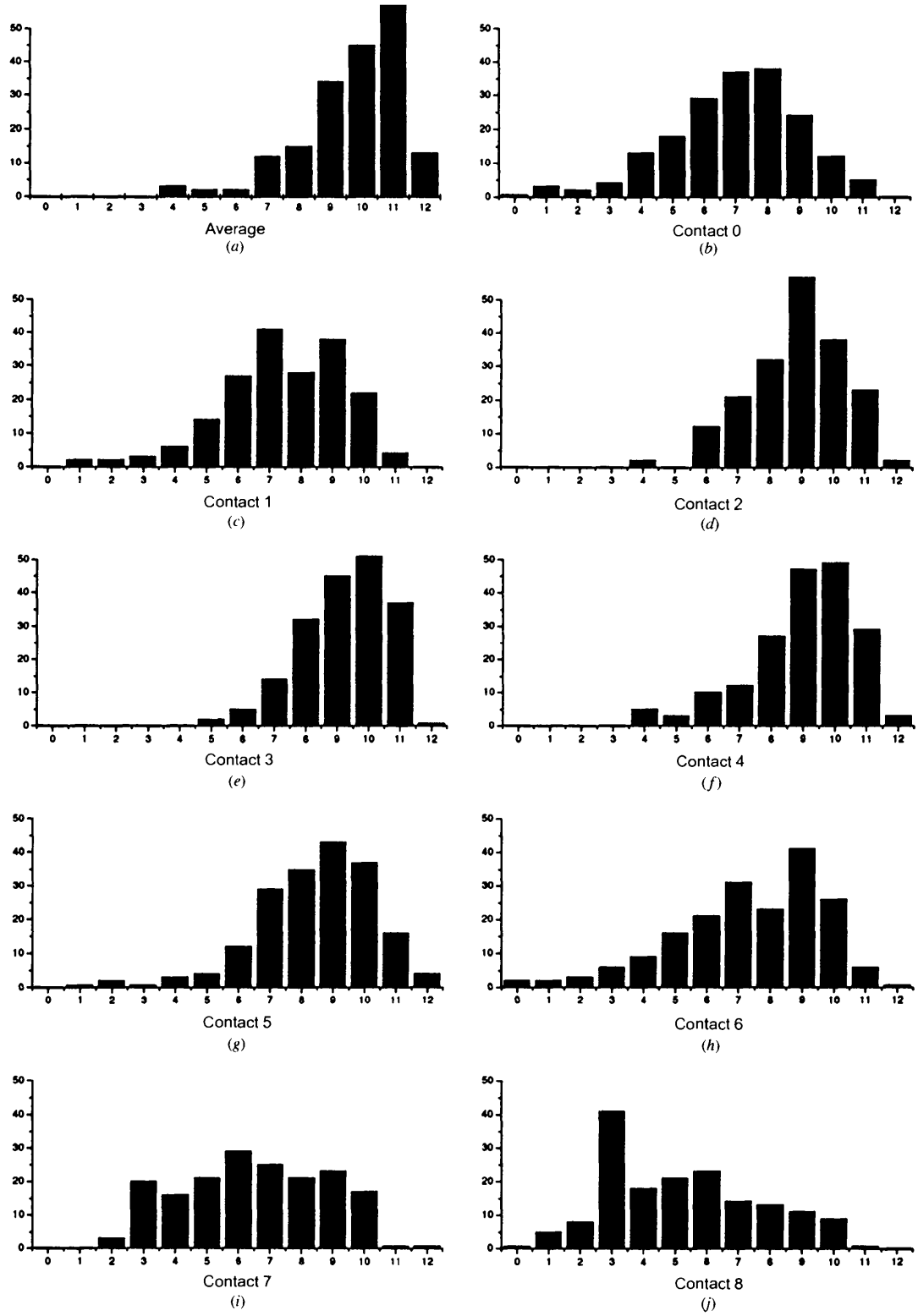

Fig. 3. Bar diagram of the number of proteins in different ranges of correlation coefficients. (*a*) The correlation between the percentage composition of the amino acids of individual proteins with the average in the data set. (*b*)-(*j*) The correlations between the amino-acid composition of a given protein in contact regions 0–8. Numbers 0–12 on the *x* axis refer to a range for the correlation coefficient from −0.3 to 1.0 in increments of 0.1. The *y* axis gives the number of proteins in each range.

and for the globins are presented in Fig. 4 along with standard deviation.

3. Results and discussion

3.1. *Contact preferences of amino acids*

The number of amino-acid residues in different contacts are given in Table 2. The number of contacts lies between 2 and 5 for most of the amino-acid residues constituting 68% of all residues (B_j in Table 2). It is interesting to note that the small residues, A, G, S and C, can have up to nine or ten neighbours. This, however, forms a small fraction (<0.4%) of the total number of amino-acid residues and is represented by only a few of the proteins in the data set, and therefore has been omitted from further analysis. About 17% of the amino acids are found in the low-density region (0–1 contact) and 15% in the high-density (6–8) region. As can be seen from Table 2, the mean contact [M_i in equation (7)] for the 20 amino acids varies from 2.5 (Pro) to 4.4 (Cys).

The data in Table 2 were converted to the percentage composition of each residue in different contacts using (1) and the results are plotted in Fig. 1. For each of the amino-acid residues the maximum occurrence is in the medium contact region. The total percentage varies from 17 to 19% in contacts 2–4. However, distinct features are exhibited by residues with different properties. The hydrophobic amino acids (Figs. 1*a*, 1*b*) and Cys (Fig. 1*e*) have maximum occurrence in the 4 contact region. The acidic residues D, E, the polar residues N, S, the conformationally rigid P and flexible G have maximum occurrence in the 2 contact region (Fig. 1*d*, 1*e*) but with high representation also in the 3 and 4 contact regions.

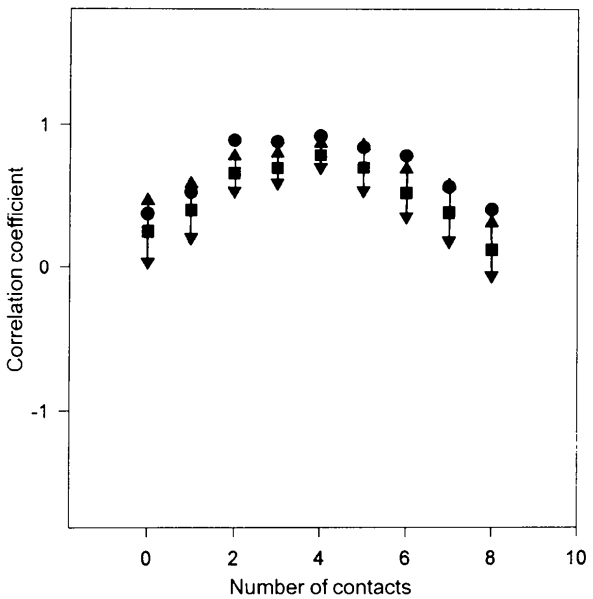

Fig. 4. Correlation coefficients between the percentage composition of amino acids in globins (Table 1*b*) and the percentages in different contact regions of globins. The closed squares represent the average correlations and the arrows represent the spread of correlation coefficients as evaluated by standard deviation. The closed circles are the correlation coefficients between the amino-acid percentage composition and the percentage composition in different contacts obtained collectively for the set of globins in Table 1(*b*).

The basic residues K, H and the polar residues Q, T (Figs. 1*c*, 1*f*) have a high frequency of occurrence in all three contact regions (2–4). Arg resembles the hydrophobic residues in having a maximum contact of 4 and resembles the charged and polar residues in having high occurrence also in contacts 2 and 3.

The percentage variations are more striking in the low-contact (0–1) regions. Those residues with maximum percentage in contact 2 also occur more frequently in zero contact with P (15.2%) being highest, followed by D, N, S, G and E (10–13%). These results are consistent with what one would expect for residues at the surface of proteins. Most of the residues at the surface will occur in short loops. The turns which these loops take depend primarily on the positions of certain residues in the loop. G, N, P, D and S are the most common amino acids found in turns allowing the chain to take up the unusual conformations required to reverse the direction of the polypeptide chain at the surface of proteins (Creighton, 1993).

In the high-contact region, the hydrophobic residues and the amino acids with small volume occur with high frequency (9–15%) in contact 6, although the presence of the bulky hydrophobic residues drastically reduces in contacts 7 and 8. The cysteine residue which appears with high frequency in the 4 contact region, has a low occurrence with zero contact and high occurrence in regions with greater than six contacts. This may be due to the fact that sequentially separated cysteine residues are often covalently linked by disulfide bonds.

Fig. 1 confirms the well established fact that the contact preference of the amino acids is related to their hydrophobicity with hydrophilic residues preferentially having low contacts. However, it also shows that it is not a simple linear relationship. Further P prefers a lower number of contacts than the other amino-acid residues. Although G and S have a preferential maximum of two contacts, they are comfortably accommodated in all contact regions of the proteins (Fig. 1*d*). Qualitatively similar results are also obtained by Miyazawa & Jernigan (1996) where a slightly different criterion is used to define the number of contacts.

3.2. *Correlation of the mean contact with some of the properties of the amino acids*

To quantify the relationship between the mean contact (M_i in Table 2) and some of the properties of the amino acids the correlation coefficient was evaluated (see §2).

3.2.1. *Hydrophobicity.* A number of hydrophobicity scales for amino acids are available (Cornette *et al.*, 1987). The concept of hydrophobicity from transfer of

free energy originated from Tanford (1962) and modified values based on this concept form the NTJ scale (Jones, 1975). The protein environment is incorporated to a great extent in the 'surrounding hydrophobicity' scale of Ponnuswamy (1993). Others take into account protein environment to varying extents. The mean contact (M_i) is best correlated with the scale derived from surrounding hydrophobicity [correlation coefficient (c.c.) = 0.91] while the lowest correlation (c.c. = 0.37) is seen with the scale derived from transfer free energy. The good correlation with the surrounding hydrophobicity scale may be due to a similarity in the method of analysis in Ponnuswamy's studies and in the present work. However, in the present study, the hydrophobicity is evaluated *ab initio* without using an already known hydrophobicity value as was the case in the earlier work (Ponnuswamy, 1993). Properties related to hydrophobicity were clustered by Kidera, Konishi, Oka, Ooi & Scheraga (1985) which is also moderately correlated (c.c. = 0.64) with the mean contact (M_i).

3.2.2. *Other properties*. The long-range non-bonded energy (Ponnuswamy, 1979) correlates well with the mean contact (c.c. = 0.81). It is interesting to note that this property, along with surrounding hydrophobicity, was shown by Joshi, Korde & Sitaramam (1993) to be conserved in the genetic code. Preferences for secondary structural features such as β-sheet (c.c. = 0.69) and α-helix (c.c. = 0.53) are moderately correlated while the short-range non-bonded interactions (c.c. = 0.3) are weakly correlated with M_i. Other properties such as refractive index, pH, pK exhibit only moderate correlation. Interestingly, the mean contact does not correlate with bulkiness and the average percentage composition of amino-acid residues in the data set.

3.3. *Percentage amino-acid composition in various contacts*

The percentage amino-acid composition of a protein plays a role in deciding its gross three-dimensional structure. This has been shown in the correlation between the composition and the class of protein (Nakashima *et al.*, 1986; Chou, 1989, 1995; Dubchack *et al.*, 1993). A given percentage composition of amino acids in a protein may reflect uniform distribution throughout the protein or uneven distribution in different parts of the protein. The relationship between the spatial distribution of composition and average composition is explored below using three different approaches.

3.3.1. *Deviation from average composition of all the proteins in the data set in a given contact region*. The difference between the average percentage composition ($\bar{A}_i$) and the percentage composition in various contacts (Fig. 2) shows that while the distribution of amino acids in contacts 3 and 4 is very close to the average percentage composition there is a drift away from the average percentage composition in the high- and low-contact regions. Towards the zero contact region, the number of residues with high mean contact (L, A, I, V, Y, C) drops and those with low mean contact (P, D, E, K, N, S, Q) increases. For residues with medium mean contact (T, H, R, Q), however, the values remain close to the average. The highest increase and decrease in zero and one contacts are seen in the percentages of P and L, respectively. Towards higher contact regions, an opposite effect is seen. There is a decrease in the polar residues (P, K, D, E) and an increase in the hydrophobic residues (A, L, I, V, F) when compared with the average composition. Further, the amino acids with small volume (A, C, G, S) and the smaller of the hydrophobic residues (V) occur more frequently in seven and eight contact regions of the protein.

3.3.2. *Distance in compositional space*. The similarity of any two protein molecules can be deduced from their distance in composition space (Nakashima *et al.*, 1986; Chou, 1989, 1995). The similarity of any two regions of a protein can also be deduced through this parameter. As expected the evaluated distances [equation (8)] in low and high contacts are long. The 3 contact of the ith residue has the lowest distance (0.9 distance units) from the average composition, exhibiting near compositional equilibrium and the longest distance (7.6 distance units) is exhibited in the 8 contact.

3.3.3. *Correlation between the average amino-acid composition of a given protein and the composition in a given contact for that protein*. The deviation from the average composition in a given contact for the total data set was presented in Fig. 2. The extent to which this trend is exhibited by the individual proteins separately was evaluated by computing the correlation coefficient between the percentage composition of residues in different contact regions for each protein and the percentage amino-acid composition of that individual protein. The correlation coefficients thus obtained were plotted as histograms and are displayed in Fig. 3.

To begin with, the percentage compositions of most of the proteins have a correlation coefficient greater than 0.6 with the average percentage composition (Fig. 3*a*), indicating that the amino-acid composition of individual proteins is close to the average composition for a large number of the present data set. Similarly, a large number of proteins show good correlation with the percentage composition of the given protein in the medium-contact (2–4) region (Figs. 3*d*, 3*e* and 3*f*). On the other hand, in the low-contact region (Figs. 3*b* and 3*c*) a significant number of proteins show poor correlation with amino-acid composition. This is also true for high-contact regions (Figs. 3*g*–3*j*). (For seven and particularly eight contacts, the large peak with correlation coefficients close to zero is due to the fact that there are many cases where there are no examples of neighbours greater than six.) The poor correlation with the percentage composition in high- and low-contact regions observed in a number of proteins is manifested as the deviation of

compositional equilibrium from the overall average, presented in Fig. 2. Thus, backbone packing in the medium-density region (contacts 2, 3 and 4) is in compositional equilibrium implying that the region is not highly sequence specific whereas those in low (0 and 1 contact) and high (>5 contacts) density regions are not, implying sequence specificity of peptides in these regions.

3.3.4. *Analysis of globin structures*. Globins are one of the best examples of proteins which have evolved into divergent amino-acid sequences while retaining the same three-dimensional structure (Ptitsyn, 1974; Lesk & Chothia, 1980; Perutz, 1992), with only a small percentage of amino acids being invariant in different species. The concepts presented in this paper were used to examine this family of structures. A list of globins selected for analysis is presented in Table 1(*b*). The results, presented in Fig. 4, clearly indicate that the medium-contact region (2–5 contacts) is in compositional equilibrium and the low (0–1 contacts) and high-contact (6–8 contacts) regions exhibit poor correlation with the percentage composition of amino acids in the globins. Thus, the features which we observed in the general data set are seen in the family of evolutionarily related globins where the three-dimensional structures are maintained in spite of divergence in amino-acid sequences.

It is interesting to note that in the aligned globin sequences (Fig. 5) the small residues which are found in the high-contact regions are size-conserved. The significance of these high-contact regions are apparent when the helix/helix interactions are examined. The highlighted residues (G, A, S, C) in Fig. 5 belong to the high-contact regions. The high contacts of G34 of helix *B* and G74 of helix *F* are due to interhelix crossing. Similarly the high contacts of S101 and C/A105 of helix *F* are due to the crossing of helices *F* and *H*, which appropriately positions the H104 of helix *F* to ligand with haem. Further, many other interhelical packings identified by Lesk & Chothia (1980) involve residues of small volume which appear in high-contact regions from our analysis. An examination of the significance of such high-contact regions in packing of sequentially distant residues in other protein families or types is in progress.

3.3.5. *Implications to protein folding*. Our observation of compositional equilibrium in the medium-contact region both in the general data set as well as in the evolutionarily diverged globin sequences may have implications to the mechanism of protein folding. It can be postulated that the hydrophobic collapse which has been described as a sequence-independent property (Bryngelson, Onuchic, Socci & Wolynes, 1995; Dill, Alonso & Hutchinson 1989; Socci & Onuchic, 1994; Kuwajima, 1992; Baldwin, 1993) perhaps leads to a compact object of medium density and could correspond to our observed medium-density regions. The collapsed protein may then reorganize itself to its native conformation through a sequence-dependent step. It is at this stage that properties of the amino acids such as hydrophobicity/polarity, geometric flexibility/rigidity and the amino-acid size play an important role in redistributing the residues onto the surface (low density) and into the dense core (high contact) of the protein (such as helix/helix packing regions of globins) leading to its unique native structure. It is suggested that the departure from the compositional equilibrium in the high- and low-contact regions, found in the present study could be a manifestation of the sequence-dependent reorganization step after hydrophobic collapse. Further, the observation that certain regions of proteins depart from compositional equilibrium may have implications in refinement of potential functions used for structure predictions.

4. Summary and conclusions

By examining the spatial neighbour patterns in a set of carefully selected proteins, the present study attempts to evaluate the role of non-local interactions in protein folding. The protein molecules are partitioned into low-, medium- and high-backbone-density regions depending

```
Helices--<----- B ----->  <--- C & D --->  <------- E ------->  <------ F ------>

1ECO   -----GDPVGILYAVFKADPSIMAKFTQFAGKDLES-IKGTAPFETHANRIVGFFSKIIGELPN---IEADVNTFVASHKP-RGVTH-
1MBA   -ANKNANGLDFLVALFEKFPDSANFFADFKGKSVAD-IKASPKLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVG-FGVGS-
1MBN   -ADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH---HEAELKPLAQSHATKHKIPI-
1MBS   -TDLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGH---HEAELKPLAQSHATKHKIPI-
2DHBA  -GHAGEYGAEALERMFLGFPTTKTYFPHF-DLSH----G-SAQVKAHGKKVADGLTLAVGHLDD---LPGALSDLSNLHAHKLRVDP-
1FDHA  -AHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH----G-SAQVKGHGKKVADALTNAVAHVDD---LPNALSALSDLHAHKLRVDP-
1HDSA  -GNAPAYGAQALQRMFLSFPTTKTYFPHF-DLSH----G-SAQQKAHGQKVANALTKAQGHLND---LPGTLSNLSNLHAHKLRVNP-
1NIHA  -AHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH----G-SAQVKGHGKKVADALTNAVAHVDD---MPNALSALSDLHAHKLRVDP-
1PBXA  -KSADAIGNDALSRMIVVYPQTKTYFSHWPDVTP----G-SPHIKAHGKKVMGGIALAVSKIDD---LKTGLMELSEQHAYKLRVDP-
2DHBB  ---EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN----G-SPHIKAHGKKVMGGIALAVSKIDD---LKTGLMELSEQHAYKLRVDP-
1HDSB  ---VDVVGAQALGRLLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFTQGLKHLDD---LKGAFAQLSGLHCNKLHVNP-
1NIHB  ---VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN---LKGTFATLSELHCDKLHVDP-
1PBXB  ---YDDIGPKALSRCLIVYPWTQRHFSGFGNLYNAEAIIGNANVAAHGIKVLHGLDRGVKNMDN---IAATYADLSTLHSEKLHVDP-
2LHB   -STYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADELKKSADVRWHAERIINAVDDAVASMDDTEKMSMKLRNLSGKHAKSFQVDP-
1FDHG  ---VEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKHLDD---LKGTFAQLSELHCDKLHVDP-
1HDSC  -GNAPAYGAQALQRMFLSFPTTKTYFPHF-DLSH----G-SAQQKAHGQKVANALTKAQGHLND---LPGTLSNLSNLHAHKLRVNP-
1NIHC  -AHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH----G-SAQVKGHGKKVADALTNAVAHVDD---MPNALSALSDLHAHKLRVDP-
1HDSD  ---VDVVGAQALGRLLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFTQGLKHLDD---LKGAFAQLSGLHCNKLHVNP-
1NIHD  ---VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN---LKGTFATLSELHCDKLHVDP-
              .   *       *     *  .                 ..  .                       ..  *     .
              ↑   ↑             ↑                     ↑                              ↑
              34  38            52                    74                             104
```

Fig. 5. Part of the aligned sequences [using the program *Multalin* (Carpet, 1989)] listed in Table 1(*b*). In the five-letter entries, the last letter corresponds to the type of chain. The top line indicates the regions of helices *B* to *F*. The (*) and (.) below the aligned sequences indicate the completely and partially conserved residues, respectively. The numbers in the last line correspond to the numbering for globin sequences (Dayhoff, 1972). The highlighted residues are those which fall in the high-contact region.

on the number of non-sequence neighbours in a given volume when a limiting distance of 6.5 Å around the Cα residue is considered. The low-, medium- and high-density regions are denoted to have spatial neighbours (contacts) (0–1), (2–4) and (5–8), respectively. It is noted that most of the amino acids (~70%) are in medium backbone density regions while the remaining ones are present in the low- and high-density regions. Hydrophilic residues show a maximum preference for having two or three neighbours and occur more frequently with zero neighbours, whereas the hydrophobic residues exhibit a preference maximum of four neighbours. The evaluated mean contact compares well with widely used hydrophobicity scales particularly with surrounding hydrophobicity scale. The mean contact is also well correlated with long-range non-bonded interaction energies and moderately correlated with secondary structural features. Further, the results indicate that the high-contact regions are dominated by G and S, and other amino acids of small volume.

The present data are also analyzed in terms of the percentage composition of amino acids in various contacts and compared with the average percentage amino-acid composition for the proteins in the data set. It reveals that the amino acids which fall in the medium-contact region (with 2–4 neighbours) are in compositional equilibrium. This is not true for the amino acids which fall in regions of low and high contact indicating that factors other than amino-acid composition of the protein play a role both in the dense core (such as the helix/helix packing regions of globins) and on the periphery of the protein. The low-contact regions described in this paper can be associated with the surface region of the protein, which has been well characterized from various points of view such as hydrophobicity, loop regions, solvent-accessible area and so on. The high-contact region on the other hand is a complex manifestation of various properties of the amino acids. From the present study it appears that the properties such as hydrophobicity, backbone flexibility/rigidity and molecular volume play an important role in shaping this region of the proteins. Analysis of the globin family of structures yields similar results. In addition, the significance of small residues in high-contact region has become apparent as due to their involvement in helix/helix packing. Based on the present analysis, one could postulate that the sequence-independent step of hydrophobic collapse in protein folding is guided by the percentage composition of amino acids in the proteins and in the second stage, highly selected stretches play a role in imparting uniqueness to the tertiary structure of the proteins. The results also have implications in improving the methods of protein structure prediction.

We thank Dr N. V. Joshi for useful discussions on statistical methods and for valuable comments on the manuscript, Dr J. R. Banavar and the Protein Folding Group at the Molecular Biophysics Unit, Indian Institute of Science have provided hours of lively discussion on protein folding. Programming and manuscript preparation help from D. Satyan, Y. N. Shamala, Indira Shrivastava and S. Vinayasree are acknowledged. Most of the computations were carried out on the computing facilities of the Bioinformatics Centre and the Interactive Graphics Facility, Indian Institute of Science, Bangalore. Some of the figures were made when one of us (MB) was a visitor to the laboratory of Dr G. J. Barton in Oxford. Financial support to SV from DST Scheme No. SP/SO/D44/93 is acknowledged.

References

Baldwin, R. L. (1993). *Curr. Opin. Biotechnol.* **3**, 84–91.
Barton, G. J. (1995). *Curr. Opin. Struct. Biol.* **5**, 372–376.
Bryant, S. H. & Lawrence, C. E. (1993). *Proteins Struct. Funct. Genet.* **16**, 92–112.
Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). *Proteins Struct. Funct. Genet.* **21**, 167–195.
Carpet, F. (1989). *Nucleic Acid Res.* **16**, 10881–10890.
Chou, K.-C. (1995). *Proteins Struct. Funct. Genet.* **21**, 319–344.
Chou, P. Y. (1989). *Prediction of Protein Structure*, edited by G. D. Fasman, pp. 549–586. New York: Plenum.
Cohen, B. I., Presnell, S. R. & Cohen, F. E. (1993). *Protein Sci.* **2**, 2134–2145.
Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzopsky, J. A. & DeLisi, C. (1987). *J. Mol. Biol.* **195**, 659–685.
Creighton, T. E. (1993). *Proteins, Structures and Molecular Properties*, 2nd ed., p. 225. New York: W. H. Freeman.
Crippen, G. M. & Vishwanadhan, V. N. (1985). *Int. J. Peptide Protein Res.* **25**, 487–509.
Dayhoff, M. O. (1972). Editor. *Atlas of Protein Sequence and Structure*, Vol. 5. Maryland, USA: National Biomedical Research Foundation.
Defay, T. & Cohen, F. E. (1995). *Proteins Struct. Funct. Genet.* **23**, 431–445.
Dill, K. A., Alonso, D. O. V. & Hutchinson, K. (1989). *Biochemistry*, **28**, 5439–5449.
Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). *Protein Sci.* **4**, 561–602.
Dubchack, I., Holbrook, S. R. & Kim, S.-H. (1993). *Proteins Struct. Funct. Genet.* **16**, 79–91.
Eisenhaber, F., Persson, B. & Argos, P. (1995). *Curr. Rev. Biochem. Mol. Biol.* **30**, 1–94.
Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). *Protein Sci.* **2**, 1181–1826.
Go, N. & Abe, H. (1981). *Biopolymers*, **20**, 991–1011.
Heringa, J. & Argos, P. (1991). *J. Mol. Biol.* **220**, 151–171.
Johnson, M. S., Overington, J. P. & Blundell, T. L. (1993). *J. Mol. Biol.* **231**, 735–752.
Johnson, M. S., Srinivasan, N., Sowdhamini, R. & Blundell, T. L. (1994). *Crit. Rev. Biochem. Mol. Biol.* **29**, 1–68.
Jones, D. D. (1975). *J. Theor. Biol.* **50**, 167–183.
Joshi, N. V., Korde, V. V. & Sitaramam, V. (1993). *J. Genet.* **72**, 47–58.

Karlin, S., Zuker, M. & Brocchieri, L. (1994). *J. Mol. Biol.* **239**, 227–248.
Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H. A. (1985). *J. Protein Chem.* **4**, 23–55.
Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994). *J. Mol. Biol.* **235**, 1598–1713.
Kuwajima, K. (1992). *Curr. Opin. Biotechnol.* **3**, 462–467.
Lesk, A. M. & Chothia, C. (1980). *J. Mol. Biol.* **136**, 225–270.
Maiorov, V. N. & Crippen, G. M. (1992). *J. Mol. Biol.* **227**, 876–888.
Miyazawa, S. & Jernigan, R. L. (1985). *Macromolecules*, **18**, 534–552.
Miyazawa, S. & Jernigan, R. L. (1996). *J. Mol. Biol.* **256**, 623–644.
Nakashima, H., Nashikawa, K. & Ooi, T. (1986). *J. Biochem.* **99**, 153–162.
Perutz, M. F. (1992). *Faraday Discuss.* **93**, 1–11.
Ptitsyn, O. B. (1974). *J. Mol. Biol.* **88**, 287–300.
Ponnuswamy, P. K. (1993). *Prog. Biophys. Mol. Biol.* **59**, 57–103.
Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1992). *Biochemistry*, **31**, 10226–10238.
Sander, C. & Schneider, R. (1991). *Proteins*, **9**, 56–68.
Shrivastava, I., Vishveshwara, S., Cieplak, M., Maritan, A. & Banavar, J. R. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 9206–9209.
Sippl, M. J. (1990). *J. Mol. Biol.* **213**, 859–883.
Socci, N. D. & Onuchic, J. N. (1994). *J. Chem. Phys.* **101**, 1519–1528.
Tanaka, S. & Scheraga, H. A. (1976). *Macromolecules*, **9**, 945–950.
Tanford, C. (1962). *J. Am. Chem. Soc.* **84**, 240–4247.
Warme, C. & Morgan, J. (1978). *J. Mol. Biol.* **118**, 270–287.