Acta Crystallographica Section D
Biological
Crystallography
ISSN 0907-4449

Logos for amino-acid preferences in different backbone packing density regions of protein structural classes

N. Kannan, Thomas D. Schneider and S. Vishveshwara

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Logos for amino-acid preferences in different backbone packing density regions of protein structural classes

N. Kannan,[a] Thomas D. Schneider[b] and S. Vishveshwara[a]*

[a]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India, and [b]National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Experimental and Computational Biology, Frederick, MD 21702, USA

Correspondence e-mail: sv@mbu.iisc.ernet.in

Received 28 February 2000
Accepted 22 June 2000

A protein sequence can be classified into one of four structural classes, namely α, β, $\alpha + \beta$ and α/β, based on its amino-acid composition. The present study aims at understanding why a particular sequence with a given amino-acid composition should fold into a specific structural class. In order to answer this question, each amino acid in the protein sequence was classified to a particular neighbor density based on the number of spatial residues surrounding it within a distance of 6.5 Å. Each of the four structural classes showed a unique preference of amino acids in each of the neighbor densities. Residues which show a high compositional bias in a structural class are also found to occur in high neighbor densities. This high compositional bias towards specific residues in the four different structural classes of proteins appears to be caused by structural and functional requirements. The distribution of amino acids in different neighbor densities is graphically presented in a novel logo form which incorporates several features such as composition, the frequency of occurrence and color code for amino acids. The spatial neighbors of the residues in different neighbor densities and their secondary structural location are also represented in the form of logos. This representation helped in the identification of specific details of the whole data which may otherwise have gone unnoticed. It is suggested that the data presented in this study may be useful in knowledge-based structure modelling and *de novo* protein design.

1. Introduction

A knowledge of the amino-acid composition of a protein sequence is believed to be sufficient to predict its structural class (Chou, 1989). Previously reported methods predict the structural class of the sequence with an accuracy ranging from 80 to near 100% (Metfessel *et al.*, 1991; Chou, 1989, 1995). Though the prediction accuracy of the structural class from sequence information is very high, the reason as to why a sequence with a given composition should fold into a specific structural class is still largely unexplored. To our knowledge, the first attempt at addressing this question was made by Bahar *et al.* (1997), who proved by lattice-model studies that the four structural classes differed in the distribution of neighbors of the residues. The amino-acid residues in protein structures are classified into different density regions based on the number of neighbors of the residue (Panjikar *et al.*, 1997). This analysis, on a set of proteins belonging to the globin family, showed that each of the neighbor densities has a specific preference for certain amino acids which are important for the structure and function of the proteins belonging to the globin fold.

In the present study, we have adopted the same methodology for classifying the residues into different neighbor densities ranging from zero to eight based on the number of spatial neighbors. A non-homologous data set of 120 proteins, in which each structural class contained 30 proteins, was used in the present study. We find a distinct preference of amino acids in different neighbor densities for the four structural classes. The results are displayed in a unique graphical form (logos). This method of presenting the data helps one to capture the similarities and differences in the distribution of amino acids for the four structural classes. The non-covalently bonded residues surrounding a given residue, called the spatial neighbors, are also shown in the form of logos for the proteins belonging to different structural classes. The logos show a distinct preference in the distribution of neighbors for proteins belonging to different structural classes. The preference of residues to occur in different secondary structural locations are also shown in the form of logos. This representation led us to focus on turns/loops in high neighbor densities of β and α/β proteins. A detailed analysis of the structures showed that the turns/loops were close to the active site of the protein and are located in the interface regions of two domains. A specific set of amino acids are preferred in these turns/loops occurring in the high neighbor densities of β and α/β proteins. This amino-acid preference information is used along with sequences to predict turns/loops which occur in high backbone packing density regions. The nature of spatial neighbors and their secondary structural location is analysed for the turns/loops occurring in eight-density neighborhoods, which also provides new lessons for knowledge-based protein structure modelling.

The present study shows that the bias for specific residues in the four structural classes could be a consequence of structural and functional requirements. Furthermore, the data presented in this paper may be useful for protein modelling and *de novo* protein design.

2. Methods

2.1. Protein data analysis

2.1.1. Data set. A non-homologous data set of 120 proteins were selected for the present analysis. Each structural class had a representative set of 30 proteins. This data set was used previously by Chou (1995). A separate testing set of 15 proteins in each of the four structural classes was also used. The proteins were classified into different structural classes based on the percentage of secondary structural elements (Chou, 1995). The α proteins had more than 40% of the residues occurring in helices and less than 5% in sheets. β proteins had greater than 40% of the residues occuring in sheets and less than 5% in helices. $\alpha + \beta$ proteins constituted more than 15% of the residues occurring in α-helices and more than 15% in β-sheets, of which more than 60% are in antiparallel sheets. α/β proteins had greater than 15% of the residues in α-helices and greater than 15% in sheets, of which more than 60% of the residues are in parallel β-sheets. The secondary structural assignment of the residues in the protein was made using the program *DSSP* (Kabsch & Sander, 1983).

2.1.2. Identification of spatial neighbors. The procedure adopted for the identification of spatial neighbors is the same as that used in the previous study by Panjikar *et al.* (1997). The number of C_α atoms falling within a sphere of radius 6.5 Å around each C_α (*i.e.* a C_i atom along the chain, excluding the four sequence neighbors C_{i-2}, C_{i-1}, C_{i+1} and C_{i+2}) were identified as spatial neighbors and the four amino-acid residues C_{i-2}, C_{i-1}, C_{i+1} and C_{i+2} as sequence neighbors for each residue of the protein. This is shown schematically in Fig. 1. A radius of 6.5 Å was used in the present investigation, as this value was found to be the distance corresponding to the first peak in the radial distribution of residues in the interior of proteins (Miyazawa & Jernigan, 1985) and a recent study on residue packing showed that the neighboring residues in a protein structure conform almost perfectly to a sphere radius of 6.5 Å (Raghunathan & Jernigan, 1997).

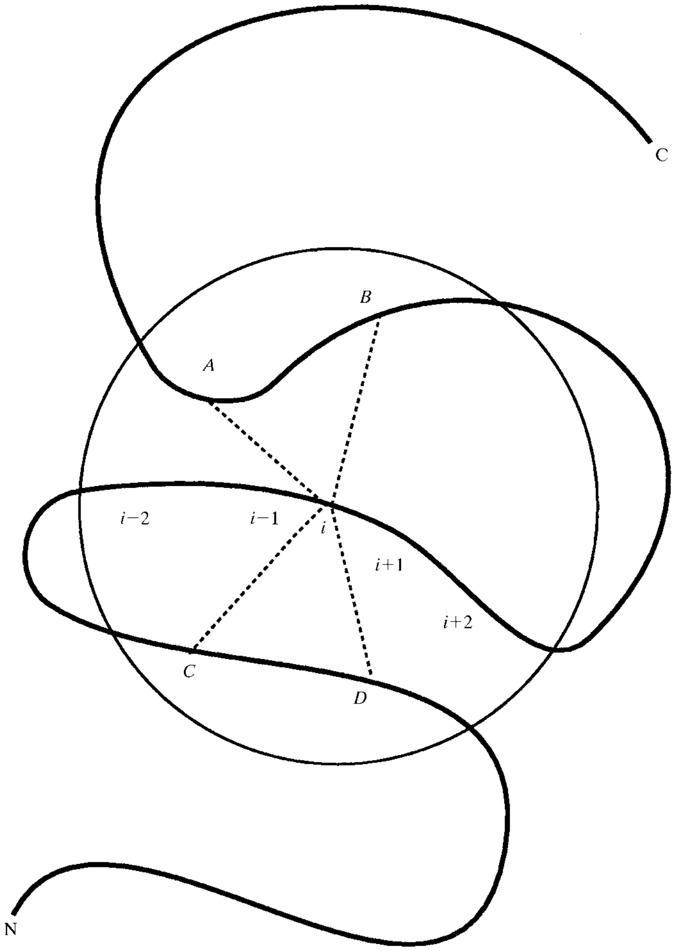

Figure 1
A schematic representation of spatial and sequence neighbors of a residue '*i*' in a polypeptide chain. *A*, *B*, *C* and *D* are the spatial neighbors of residue '*i*' as they fall within a sphere of 6.5 Å and are shown by dotted lines. $i - 1$, $i - 2$, $i + 1$ and $i + 2$ are the sequence neighbors of residue *i*.

Table 1
Percentage amino-acid composition in the data set for the four structural classes of proteins.

Residue type	α	β	α/β	$\alpha+\beta$
Ala	10.979	6.933	9.355	8.263
Arg	4.798	3.326	4.443	4.793
Asn	3.263	5.602	4.134	4.544
Asp	5.835	5.287	6.454	6.135
Cys	0.883	1.996	1.074	2.128
Gln	4.146	3.992	3.974	4.045
Glu	7.217	4.377	6.158	5.445
Gly	6.641	9.489	8.429	6.883
His	2.917	1.085	2.407	2.588
Ile	4.607	4.272	6.134	4.563
Leu	11.708	6.828	8.133	7.841
Lys	7.754	5.392	6.195	6.077
Met	2.572	1.541	2.308	1.879
Phe	3.762	4.797	3.764	3.796
Pro	2.572	4.272	4.097	5.119
Ser	5.374	8.648	5.442	6.691
Thr	4.875	8.999	5.245	6.231
Trp	1.075	1.541	1.407	1.515
Tyr	3.263	4.412	3.678	4.755
Val	5.758	7.213	7.170	6.710

2.1.3. Classification of the central residue (C_i) to different neighbor densities. The identification of spatial neighbors around C_i, as described above, was carried out for all the residues of the proteins in the data set. The number of spatial neighbors found for the central residue (C_i) ranges from zero to ten and therefore, based on the number of spatial neighbors, every residue in a protein was classified to a specific spatial neighbor density region or 'neighbor density'. As the number of residues which occurred in the nine- and ten-neighbor densities was very small, they were grouped into the eight-neighbor density. The total number of occurrences of each amino acid in a given spatial neighbor-density region was obtained by summing the respective values for that amino acid in all the proteins. Furthermore, the residues which fall into the zero-density neighborhood, one-density neighborhood and two-density neighborhood regions were grouped as 'low-density neighborhood', the residues which fall within three-, four- and five-density neighborhood regions were grouped as 'medium-density neighborhood' and the residues with greater than five spatial neighbors were classified as 'high-density neighborhood' regions.

2.2. Identification of residues close to the active/binding site

A residue in a protein molecule was considered to be close to the active/binding site of the protein if any of the residue atoms was within a distance of 6.8 Å from any of the ligand atoms.

2.3. Classification of residues to different regions of secondary structure

The residues in a protein structure were assigned to a particular secondary structure using the *DSSP* program (Kabsch & Sander, 1983). The length of the secondary structure (helix or strand) was evaluated based on the number of residues constituting the secondary structure. If the length of the helix or strand was greater than seven residues then the first three residues of a helix (towards the N-terminus) were assigned the term 'N' and those of a strand were designated as 'S'. The three residues towards the C-terminal region of the secondary structure were assigned 'C' and 'E' for helix and strand, respectively. The remaining residues in the middle were assigned 'M' for helix and 'B' for strand. If the length of the secondary structure was less than or equal to seven residues then only one residue in the N- and C-terminal regions was assigned to 'N/S' or 'C/E' based on the secondary structure in which the residue occurs; the remaining residues were classified as 'M' or 'B' for helix and strand, respectively. The loops and turns were denoted by the symbol 'L'.

2.4. Logo representation

Logo representation has been used previously for showing sequence-alignment data where, in a sequence logo for proteins, stacks of letters represent the amino-acid composition at different positions. The height of each letter is proportional to its frequency in the aligned set of proteins, while the height of the entire stack of letters is the information content, a good measure of the sequence conservation. The letters are sorted to place the most frequently occurring ones at the top (Schneider & Stephens, 1990). However, it should be noted that although the letters which occur with nearly the same frequency will appear to have the same size, their ordering within a stack is sorted by frequency.

In the present study, logos have been used to represent the occurrence of each amino acid in different neighbor-density regions of the proteins. The *y* axis represents the uncertainty index *H* in bits given by Shannon's formula (Shannon, 1948*a*,*b*; Shannon & Weaver, 1949; Pierce, 1980; Abramson, 1963; Singh, 1966; Gatlin, 1972; Schneider *et al.*, 1986),

$$H = -\sum_{i=1}^{20} P_i \log_2 P_i, \tag{1}$$

where P_i is the frequency (an estimate of the probability) of occurrence of an amino acid of the *i*th type in a given spatial neighbor-density region. The neighbor-density regions from zero to eight are plotted on the *x* axis.

An uncertainty index of $\log_2 20 = 4.32$ bits on the *y* axis would mean that the composition of amino acids in the particular neighbor density is close to uniform, *i.e.* each of the 20 amino acids is found to occur 5% of the time.

The different color codes used for the amino acids represent the nature of the amino-acid residues. All the hydrophobic residues are represented in black and the polar residues with a hydroxyl group (serine, threonine and tyrosine) in green. The basic residues lysine, arginine and histidine are shown in blue and the acidic residues aspartic acid and glutamic acid in red. The neutral residues asparagine and glutamine are shown in purple. The special residues glycine and proline are shown in orange and cyan, respectively.

The data for the logos can be obtained from the websites http://www.lecb.ncifcrf.gov/~toms/vishveshwara.html and http://www.mbu.iisc.ernet.in/~sv/logo.html.

3. Results and discussion

3.1. Amino-acid preferences in different neighbor densities

3.1.1. *α* and *β* class. Table 1 provides the percentage composition of the 20 amino acids evaluated for the data set of proteins considered in each structural class. A high composition of hydrophobic residues A and L and charged residues E and K is observed for the *α* class. The location of amino acids in different neighbor densities for the *α* structural class is shown in Fig. 2. Alanine, which has a high composition, is frequently present in all neighbor densities. On the other hand, leucine is preferred only in neighbor densities three to eight and is rarely seen in neighbor densities zero to two. Since zero- to two-density neighborhoods would correspond to locations on the protein surface, hydrophobic moieties such as leucine with a large hydrophobic side chain are not preferred in these regions; instead, a high preference for G (high occurrence in loops) and the charged residue D is observed. Glycine, which shows a moderate composition in the *α* class (Table 1), is also prominent in the eight-neighbor densities. Our previous study on the globin family (Panjikar *et al.*, 1997) and the present study show that glycine residues having eight-density neighborhoods are located in helix-crossing regions. Alanine is highly preferred in the seven-density neighborhood region as is shown by the size of the letter 'A'. Alanine residues in seven-density neighborhood regions occur in the middle of the helix in 95% of cases and were also found to be located, like glycine, in helix-crossing regions (Fig. 3). Charged residues E and K which show a high composition in the *α* class are found to be preferred in the three to five neighbor densities apart from the low neighbor density. Lysine, however, also occurs in the eight-neighbor density. Most of these features were also observed in the separate data set (test set) of proteins belonging to the *α* class.

In the case of proteins belonging to the *β* structural class, a high composition of glycine, the polar residues serine and threonine and the hydrophobic residues valine, alanine and leucine is observed (Table 1). An examination of the logo for

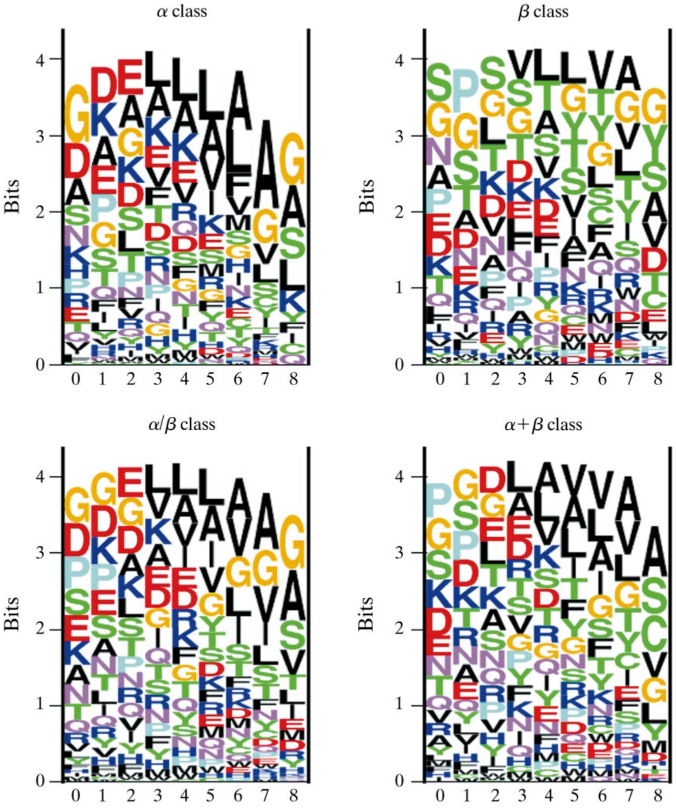

Figure 2
Logos of the amino-acid distribution in the different neighbor densities (0–8) for the four structural classes of proteins. The uncertainty index is plotted as bits on the *y* axis. The vertical lines of the *y* axis are 4.3 bits high.

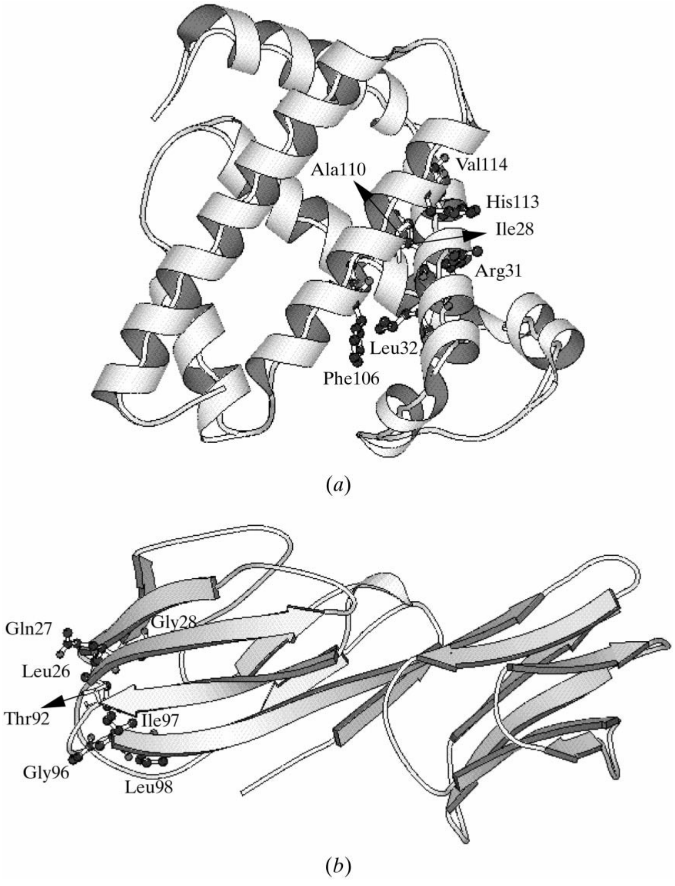

Figure 3
A *MOLSCRIPT* diagram (Kraulis, 1991) showing the structural location of residues in high-density neighborhoods and their spatial neighbors. (*a*) Residue Ala110 shown in BONDS representation and located in a helix-crossing region in protein myoglobin (PDB code 4mbn; Takano, 1977) of the *α* class. The spatial neighbors are shown in CPK representation. (*b*) Residue Thr93 shown in BONDS representation and located in the C-terminal region of a sheet in T-cell surface glycoprotein (PDB code 1cid; Brady *et al.*, 1993) of the *β* class. The spatial neighbors are shown in CPK representation.

the β class (Fig. 2) shows that the occurrence of the polar residues serine and threonine are not restricted only to the surface regions; these residues, along with the residue glycine, are preferred in almost all neighbor densities. A distinction occurs, however, in the preference of hydrophobic residues in different neighbor densities. It is seen from the logo that valine is highly preferred in either the three-neighbor density or the six- to seven-neighbor density, whereas leucine is preferred in the four- to five-neighbor density and alanine in the seven- to eight-neighbor density. This specific preference of hydrophobic residues for specific neighbor densities is in striking contrast to that observed for the α class.

About 60% of threonine and 40% of serine residues in high neighbor densities occur in the terminal region of sheets, which are known to be highly packed (Beardsley & Kauzmann, 1996). The high occurrence of serine and tyrosine in eight-neighbor densities of the β class and threonine and tyrosine in the six- and seven- neighbor densities of the β class as seen from the logo (Fig. 2) indicates that these residues are involved in high backbone packing regions, which are likely to be essential for structural stability. For example, the occurrence of Thr92 in a six-density neighborhood is seen in the T-cell surface glycoprotein (1cid) shown in Fig. 3. The Thr92 residue occurs in the C-terminus of the sheet and is surrounded by spatial residues which are mostly hydrophobic in nature. A preference for glycine as spatial neighbor is also seen in this example. Glycine helps in the high backbone packing because of its smaller size. All the spatial neighbors are also found to emanate from the terminal regions of different sheets.

Apart from the difference in the preference of residues with high composition in different neighbor densities, another prominent feature observed from the logos (Fig. 2) is the high occurrence of tyrosine (Y) in the high neighbor density of β proteins. This feature also holds in the test set, where a high preference of tyrosine for the high neighbor density of the β class was consistently observed.

3.1.2. α/β and α + β class. The trend in the preference of amino acids to occur in different neighbor densities of the α/β class is in general similar to that in the α class. Glycine shows a preference for the high and low neighbor densities in the α/β class which is similar to that exhibited in α proteins. Among the hydrophobic residues alanine, leucine and valine which show a high composition, alanine and leucine show a strong preference for occurring in medium or high neighbor density, as is the case in the α class. On the other hand, valine shows a specific preference for either three- to four-density neighborhoods or six- to seven-density neighborhoods, which is similar to that observed in the β class. The charged D, E and K residues which are compositionally high show a preference similar to that in the α class.

The α + β class shows a high composition of hydrophobic residues A and L (Table 1), as in the α class, and polar residues S and T, as in the β class. However, the distribution of these residues in different neighbor densities of the α + β class is different from that exhibited in either the α or β class (Fig. 2). Alanine and leucine occur more frequently in the medium and high neighbor densities than they do in the α class, where alanine is preferred even in low neighbor densities. The polar residue serine is preferred in the low neighbor density in both classes. In both α + β and α/β classes, a high preference for leucine and valine is seen in medium and high neighbor densities (Fig. 2), consistent with the fact that these residues were shown to be important in helix-sheet packing (Janin & Chothia, 1980).

In general, the preferences of amino acids to occur in different neighbor densities are different for the four structural classes. The medium neighbor density (three to five) shows a distribution close to the equiprobable distribution in all four classes of proteins; the uncertainty index plotted on the y axis is close to 4.32 (Fig. 2). The low-density neighborhood (zero to two) and the high-density neighborhood (six to eight) regions show more deviation from the equiprobable composition (less obvious in the β class). The residues which show high composition in each structural class are also found to occur prominently in the high neighbor density. Since the high neighbor densities correspond to highly packed main-chain regions, the preference of specific amino acids in these

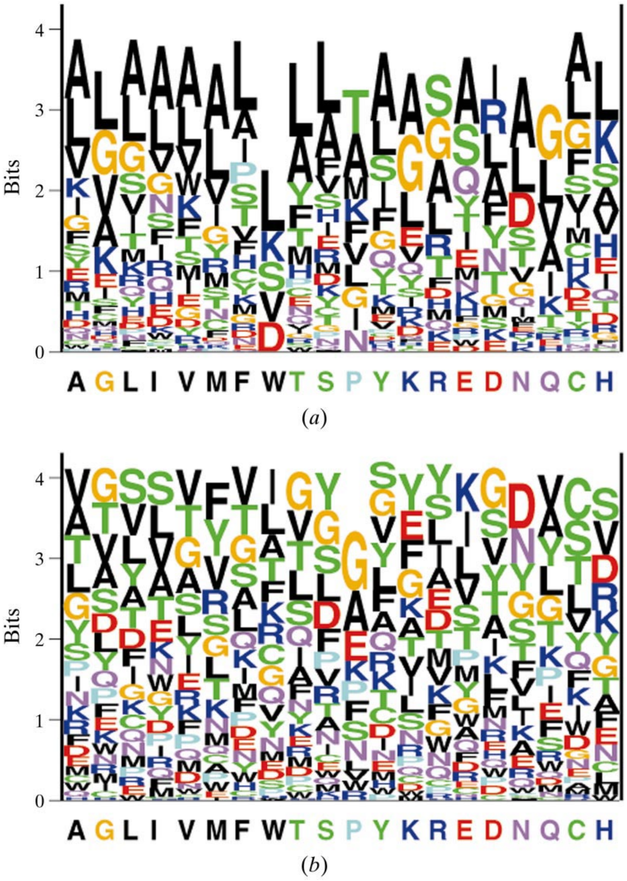

Figure 4
Distribution of spatial neighbors around each amino acid in the high neighbor density (six to eight) of proteins belonging to (a) the α class and (b) the β class. N-terminal residues are green: N, α-structure; S, β-structure. Middle residues are black: M, α-structure; B, β-structure. C-terminal residues are red: C, α-structure; E, β-structure. L (red) represents loops.

regions should be important for optimal packing of the protein and for its structural stability. The specific differences in the distribution of amino acids in different neighbor densities seem to be important in maintaining a particular fold.

3.2. Spatial neighbors

Non-local interactions are known to play a major role in determining the three-dimensional structure of proteins (Dill *et al.*, 1995) and hence the preference of an amino acid to occur in a specific neighbor density should be influenced to an extent by the nature of spatial neighbors surrounding it. Since high neighbor densities (six to eight) would correspond to highly packed backbone regions in the protein, we have analysed the nature of spatial neighbors surrounding the residues classified to a high neighbor density. The spatial neighbor preference for the α and β classes for all 20 amino acids occurring in high neighbor densities is shown in Fig. 4. In the α class, we can generally observe a preference for residues A and L (which are compositionally high) as spatial neighbors by most of the residues. Although this is expected, the preferential pattern of the spatial neighbors just below these residues in the logo (Fig. 4) is different for the 20 amino acids in the high neighbor density. Further, T and S are preferred as neighbors of P and R, respectively. The uncertainty index of the neighbors of W is rather low and only five types of residues appear as its neighbors. However, this may not be significant because of the small sample size ($n_\beta = 14$ and $n_\alpha = 28$).

The logo for spatial neighbor preference of high-density neighborhood residues in the β class is quite different from that of the α class. The sizes of the letter for many residues in a vertical stack in the logo are the same, indicating a lack of preference, with a few exceptions such as the preference for G as a spatial neighbor of P, which also has a low uncertainty index. The other important features observed in the logo are a high preference for Y as a spatial neighbor of residues K, R and S. Charge-complementarity preference appears to lead to the occurrence of K as a neighbor of E.

In order to test the sensitivity of the spatial neighbors logo to the distance criteria used and the inclusion of residue '$i+2$' or '$i-2$' as a spatial neighbors of residue 'i', logos were produced for both structural classes by increasing the cutoff distance criteria to 7.5 Å and also including the '$i+2$' and '$i-2$' residues as spatial neighbors. Qualitatively, the same features were observed (data not shown). However, a noticeable feature was the appearance of the residues with a high amino-acid composition as spatial neighbors of most of the residues. This feature was particularly seen in the β class, where the residue valine, having the highest composition, also appeared as the spatial neighbors of most of the residues.

3.3. Secondary structural preferences

The residues in different neighbor-density regions were assigned to different secondary structural locations (as mentioned in §2.3). With these seven symbols, the secondary structural logos were created for all four classes of proteins and are shown in Fig. 5. Since the number of symbols is seven, the maximum uncertainty is $\log_2 7 = 2.81$ bits. On visualizing the secondary structure logos, it is evident that the four structural classes show different preferences for secondary structural location in different neighbor densities. A very prominent feature observed is in the high neighbor density of the α class, where most of the residues occurring in high-density neighborhood occur in the middle of the helix. This was also observed in our previous analysis on the globin family, in which the high neighbor densities correspond to helix-crossing regions (Panjikar *et al.*, 1997). All these features were retained when tested on a separate data set (test set), emphasizing the fact that the observed features are not an artifact of the data set used.

A different preference for secondary structural location is found for the residues in high-density neighborhood regions of the β class. Although the middle of strands (B) dominated the high neighbor density, a significant number of residues are also found in the N-terminus (S) and the C-terminus (E) of the sheet. This observation is in agreement with the results reported previously (Beardsley & Kauzmann, 1996), where an investigation of the efficiency of packing near β-pleated sheets showed high packing density around the end regions of the β-sheet where regular backbone–backbone hydrogen bonding was interrupted.

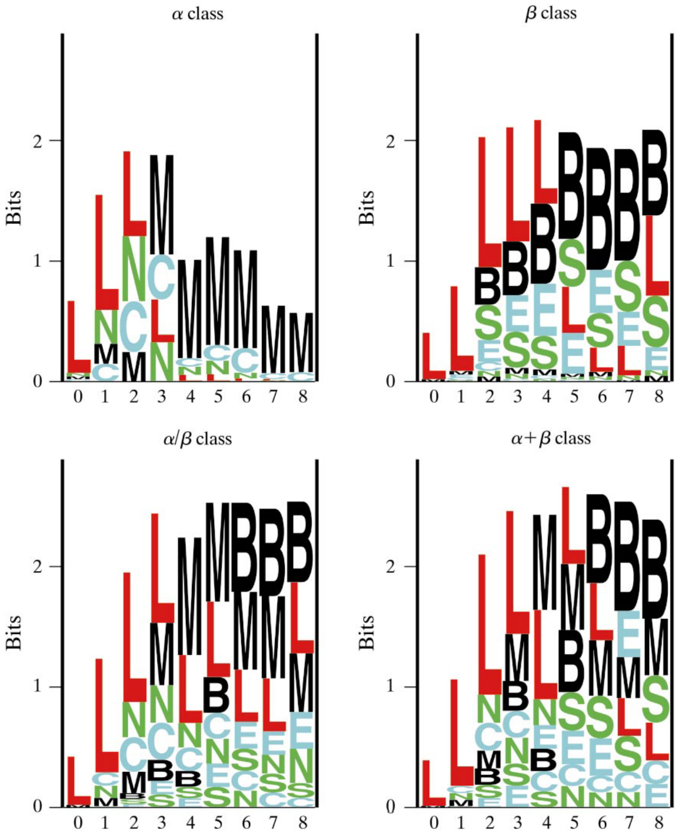

Figure 5
Logos for secondary structural preference in different neighbor densities (zero to eight) for the four structural classes of proteins. The vertical lines of the *y* axis are 2.8 bits high.

Table 2
Structural location of turns/loops in eight neighbor density regions in the data set of proteins studied.

Structural class	β	α/β
Number of turns/loops in eight neighbor density regions	12	28
Number of cases near active/binding site	6	13
Number of cases not near active/binding site	5	8
Number of cases where the active-site information is not known	1	7

The other prominent feature in the secondary structure logo is the frequent occurrence of turns/loops in the eight-density neighborhood region of the β class and the α/β class. We examined the location of the these turns/loops in each structure and found that more than 50% of the turns in eight-density neighborhoods occurred close to the active/binding site of the protein. The details are given in Table 2.

3.3.1. Loops in high neighbor densities of the β and α/β classes. The residues occurring in the turn/loop regions in eight-density neighborhoods and their sequence neighbors were analysed in the β and α/β classes. In both classes, the residues which were preferred in the ith position were residues G, A, S and D. An analysis of the sequence neighbors of the ith residue ($i-2$, $i-1$, $i+1$, $i+2$) in the β class showed a preference for D and R as sequence neighbors in the $i-2$ position, a preference for polar residues S or T in the $i-1$ position, R or S in the $i+1$ position and D, R, S and N in the $i+2$ position (Fig. 6). Most of these pentapeptides were found in the protease family and were located close to the catalytic site of the protein. However, this trend in amino-acid preferences could also be the result of a small data set (Table 2). Further, the turn/loop regions in the eight neighbor density of the β class were specifically examined. It was found that in six out of ten cases residues occurring in an eight-density neighborhood were close to the active/binding site of the protein or in the domain interfaces. Fig. 7(*a*) shows the case of Sindbis virus capsid protein (PDB code 2snv), a serine protease in which Ser215 and Gly216 occur in a loop in the high neighbor density. Ser215 is also part of the catalytic triad of the serine protease (Fig. 7*a*).

A different trend is seen in the α/β class (Fig. 6). Hydrophobic residues such as alanine, valine and isoleucine are preferred in the $i-2$, $i-1$ and $i+1$ positions. A preference for glycine in the four positions except the $i-1$ position and a strong preference for serine or aspartic acid in the $i-1$ position is found in the turn regions occurring in eight neighbor densities of α/β proteins. A detailed analysis of all the loops in eight-density neighborhoods of the α/β class showed that most of them occurred in the interface regions of the domains. Furthermore, 13 of the 28 total cases (Table 2) were found close to the active/binding site. Fig. 7(*b*) shows a case of a sulfate-binding protein (PDB code 1sbp) in which the pentapeptide sequence is located in the interface region of the domains and occurs close to the active site.

Since the turns/loops in a protein structure are always expected at the protein exterior (Rose, 1978; Levitt & Chothia, 1976), a rare occurrence of a turn/loop in a highly packed region should be of some functional significance (Rose *et al.*, 1983). Information derived from protein structure and sequence has proved to be very useful in predicting the functional sites of a protein (Blom *et al.*, 1999). In the present study, a characteristic preference of residues in the five positions is seen for the residues occurring in the loops of the eight neighbor density in the α and α/β classes. This is in contrast to the loops in the low neighbor densities, where the preferential pattern of sequence neighbors is less specific in general and the residues proline and glycine show a high preference in the low-density neighborhood loops (Fig. 6).

Identifying loop regions of high-density neighborhoods from sequence information can be indirectly useful in locating the functional regions of the proteins. In order to predict pentapeptide sequences from the amino-acid sequence information in which the middle residue (zeroth position) of the pentapeptide occurred in a loop or a turn as well as in the eight-neighbor density, the frequently occurring residues in the five positions ($i-2$ to $i+2$) were extracted from a sequence logo of pentapeptides belonging to the high neighbor densities of the β and α/β classes. Pentapeptide patterns were generated from these residues and a pattern-matching algorithm was developed. The algorithm was used to

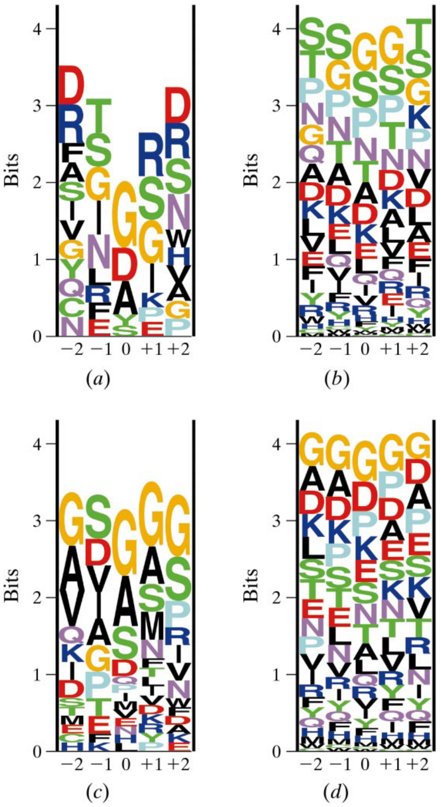

Figure 6
Sequence neighbor preferences of amino acids in positions −2 ($i-2$), −1 ($i-1$), 0 (i), +1 ($i+1$) and +2 ($i+2$) of residue 'i' that are in eight neighbor density regions and occurring in loop regions of (*a*) the β class and (*c*) the α/β classes. Sequence neighbor preferences in the four positions for residue 'i' in low neighbor density regions (zero to two) and occurring in loop regions of (*b*) the β class and (*d*) the α/β classes.

match pentapeptide stretches within the protein sequences which form the non-homologous data set of the August 1999 PDB release. The matched pentapeptide sequences were analysed for their neighbor density and structural location. The structures which were included in obtaining the logos were excluded for the pentapeptide sequence pattern-matching procedure. A pentapeptide sequence was generated using the frequently occurring residues in the five positions. Based on this, D, R, A, F or S in the $i - 2$ position, T, S, G, I or N in the $i - 1$ position, G, D, A or Y in the ith position, R, S, G, I or K in the $i + 1$ position and D, R, S, N or W in the $i + 2$ position with the residue in the ith position in the eight-density neighborhood region were searched for in all the non-homologous structures. Four such cases were found and these are listed in Table 3.

Similarly, for the α/β class a pentapeptide sequence pattern search was made with either residues G, A, V, Q or K in the $i - 2$ position, S, D, V, I, A or G in the $i - 1$ position, G, A, S or D in the ith position, G, A, S, M or N in the $i + 1$ position and G, S, P, R or I in the $i + 2$ position. 11 cases were found in which the middle residue of the pentapeptide sequence occurred in a turn/loop in an eight-neighbor density. Seven of the 11 turns/loops were close to the substrate-binding site and are shown along with the protein name in Table 3.

Further, the spatial neighbors of the residues in loops in eight-density neighborhoods and the secondary structural location of all the spatial neighbors were analysed for the β and α/β classes of proteins. Since the turns/loops in both classes occur mostly in the domain-interface regions, a knowledge of the spatial neighbors would be useful in protein structure modelling. The secondary structural location of the spatial neighbors show that most of the spatial neighbors occur in loops connecting regular secondary structures, particularly in the α/β proteins. These loops were classified into four types, L_{ss}, L_{hh}, L_{sh} and L_{hs}, based on the two

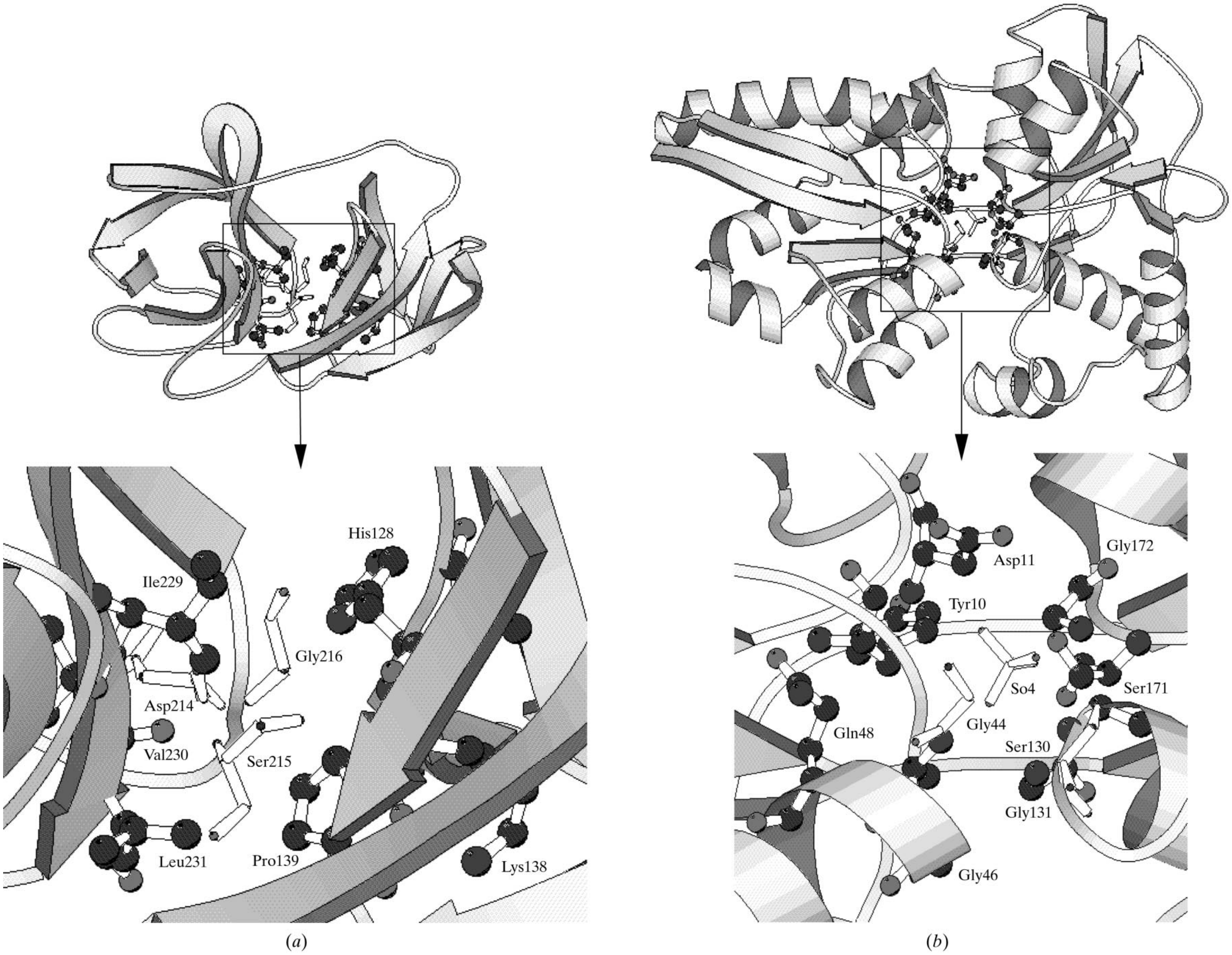

Figure 7
A *MOLSCRIPT* diagram (Kraulis, 1991) showing residues in an eight-density neighborhood and occurring in a loop in (*a*) Sindbis virus capsid protein, belonging to the β class, (*b*) sulfate-binding protein, belonging to the α/β class. Each loop forms part of the active site and occurs in domain-interface regions. This region is shown enlarged below with the residue in eight-density neighborhood of the loop regions shown in BONDS representation and the spatial neighbors in CPK representation.

Table 3
Predicted pentapeptides in eight neighbor densities of other proteins not included in the data set.

* denotes that the pentapeptide in the protein is close to the active/binding site. Residue numbers within parentheses are in eight-density neighborhoods.

PDB code	Protein name	Pentapeptide	Class
1abr	Abrin-A	RI**G**GR	β
1got*	Gt-α/gu-α chimera gt-β gt-γ	AT**G**SD	β
1p04	α-Lytic protease	DS**G**GS	β
3sil	Sialidase	RS**D**IS	β
1b2n*	Methanol dehydrogenase	GS**G**NP	α/β
1agn*	Human σ alcohol dehydrogenase	AG**A**SR	α/β
1ysc	Serine carboxypeptidase	AG**D**KD	α/β
2ctc*(156)	Carboxypeptidase	AG**A**SS	α/β
2ctc*(252)	Carboxypeptidase	AS**G**GS	α/β
2tys*	Tryptophan synthase	AG**A**IS	α/β
1fdi	Formate dehydrogenase	KS**A**AI	α/β
1gso	Glyanamide ribonucleotide synthease	VI**G**NG	α/β
1ixh*	Phosphate-binding protein	GI**G**SS	α/β
1rus	Rubisco	GA**S**GI	α/β
2pia*	Pthalate deoxygenase reductase	VA**G**GI	α/β

secondary structures [helix (h) or sheet (s)] they were connecting. In the α/β proteins, the residues in loops which occur as spatial neighbors are found to be mostly of the type L_{hh} (connecting two helices, $n = 21$) or L_{sh} (connecting an N-terminal sheet to a C-terminal helix, $n = 30$). Loops of type L_{hs} (connecting an N-terminal helix to a C-terminal sheet, $n = 6$) are very uncommon. No specific residue pattern is seen in these spatial neighbors. However, A and G appear more often, which helps in high backbone packing. It would be interesting to see if these observed features detected in a small data set emerge as a regular pattern in a large data set of high-density neighborhood loop regions when it becomes available.

4. Conclusions

Residues are classified into different neighbor densities based on the number of spatial neighbors. Specific preferences of amino acids are seen for different neighbor densities characteristic of each structural class. In all four structural classes the medium neighbor density shows an amino-acid composition close to the average composition, while the low and high neighbor densities exhibit deviations from the average composition. The results are provided in a novel graphical logo form. Further, each type of amino acid in a given density neighborhood region has a unique distribution of spatial neighbors for different structural classes. The differences observed in amino-acid preferences and spatial neighbor preferences help in understanding why sequences with a given composition fold into a specific structural class.

The logo representation provides an excellent means of representing the vast amount of data in an elegant and easily understandable manner. Using these representations, the reader can directly extract information by visualization and can identify both gross features and specific aspects which otherwise may go unnoticed. The use of logos in showing such specific aspects is illustrated in the logos created for secondary structural preference, which show a significantly high prominence of turns/loops in the eight-neighbor densities of the β class. Further, investigation of these loops revealed their importance in the function and structure of the proteins and also helped in identifying such features in a data set which were not used in the logo generation.

The secondary structural logo also provides some new structural insights which can be useful in comparative protein modelling and *de novo* protein design. Since the spatial neighbor preferences are different for the 20 amino acids in different structural classes, statistical potentials for protein folding can be improved if potentials are derived separately for the four structural classes. Our present study shows that the bias towards specific amino acids in the four structural classes of proteins exists because the preferred amino acids are important for the structure and function of the protein.

The authors are thankful to Dr Margaret Biswas and Denise Rubens for useful comments on the manuscript and S. Vinaysree for her help in installing the program on a Sun workstation. The computational facilities of the Supercomputer Education Research Centre, the Interactive Graphics Facility of the Indian Institute of Science, Bangalore and the NCI-FBSC at Frederick, Maryland are acknowledged. SV thanks NCI for a visiting position during the tenure of which the logo-production work was started. Partial support to SV from the Department of Science and Technology is acknowledged.

References

Abramson, N. (1963). *Information Theory and Coding.* New York: McGraw–Hill.

Bahar, I., Atilgan, A. R., Jernigan, R. L. & Erman, B. (1997). *Proteins*, **29**, 172–185.

Beardsley, D. S. & Kauzmann, W. J. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 4448–4453.

Blom, N., Gammeltoft, S. & Brunak, S. (1999). *J. Mol. Biol.* **294**, 1351–1362.

Brady, R. L., Dodson, E. J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). *Science*, **260**, 979–983.

Chou, K. C. (1995). *Proteins*, **21**, 319–344.

Chou, P. Y. (1989). *Prediction of Protein Structures and Principles of Protein Conformation.* New York: Plenum Press.

Dill, K. A., Bronberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). *Protein Sci.* **4**, 561–602.

Gatlin, L. L. (1972). *Information Theory and the Living System.* New York: Columbia University Press.

Janin, J. & Chothia, C. (1980). *J. Mol. Biol.* **143**, 95–128.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.

Levitt, M. & Chothia, C. (1976). *Nature (London)*, **261**, 552–558.

Metfessel, B. A., Saurugger, P. N., Connelly, D. P. & Rich, S. S. (1991). *Protein Sci.* **2**, 1171–1182.

Miyazawa, S. & Jernigan, R. L. (1985). *Macromolecules*, **18**, 534–552.

Panjikar, S. K., Biswas, M. & Vishveshwara, S. (1997). *Acta Cryst.* D**53**, 627–637.

Pierce, J. R. (1980). *An Introduction to Information Theory: Symbols, Signs and Noise*, 2nd ed. New York: Dover.

Raghunathan, G. & Jernigan, R. L. (1997). *Protein Sci.* **6**, 2072–2083.

Rose, G. D. (1978). *Nature (London)*, **272**, 586–590.

Rose, G. D., Young, W. B. & Gierasch, L. M. (1983). *Nature (London)*, **304**, 654–657.

Schneider, T. D. & Stephens, R. M. (1990). *Nucleic Acids Res.* **18**, 6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). *J. Mol. Biol.* **188**, 415–431.

Shannon, C. E. (1948*a*). *Bell Syst. Tech. J.* **27**, 379–423.

Shannon, C. E. (1948*b*). *Bell Syst. Tech. J.* **27**, 623–656.

Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

Singh, J. (1966). *Great Ideas in Information Theory, Language and Cybernetics.* New York: Dover.

Takano, T. (1977). *J. Mol. Biol.* **110**(3), 569–584.