ELSEVIER

Biophysical Chemistry 55 (1995) 167–184

Biophysical Chemistry

The α-helix as seen from the protein tertiary structure: a 3-D structural classification

Tom L. Blundell *, Zhan-Yang Zhu [1]

ICRF Unit of Structural Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Abstract

Helices are selected from globular protein structures defined at high resolution by X-ray analysis. We cluster α-helices in two ways: according to their position in the tertiary structure by considering patterns of solvent inaccessible residues and according to the arc of the solvent inaccessible face. For each class of helices we have defined propensities for amino acids at each position; these can be used to calculate templates for recognition of a member of that class. The analysis provides a basis for the prediction of α-helices and estimation of their approximate position in a protein tertiary structure. It also provides an approach to estimating the probability of finding amino acid sequences as helices in solution and in a folded protein, thus indicating those helices that might be involved in nucleation of protein folding.

Keywords: Protein classification; Amphipathic α-helix; Amino acid pattern; Secondary structure

1. Introduction

In the classical work of Pauling et al. [1] the α-helix was defined as a regular, right-handed helix with 3.6 residues per turn. In general this description is satisfactory, but in detail helices have proved to be less regular. For example, amphipathic helices are generally curved with the centre of curvature towards the hydrophobic core [2]. This curvature can be associated with hydrogen binding water molecules or amino acid sidechains to the carbonyl groups on the solvent accessible side of the helix. As a consequence the peptide planes are tilted so that the carbonyls are inclined away from the helix axis and towards the hydrogen bonded water molecules. Thus the C_i–O_i...N_{i+4} angle is 148.5° for the accessible surface compared to 159.0° on the solvent inaccessible side. This leads to ψ_i and ϕ_{i+1} of $-41°$ and $-66°$ for the solvent accessible side and $-44°$ and $-59°$ for the solvent inaccessible side. The periodic alternation of these angles along the helix is consistent with the observed curvature of helices in globular proteins and the supercoiling in helix duplexes and higher order coiled-coils.

In the first analysis of α-helical segments of myoglobin and haemoglobin, Perutz et al. [3] observed that non-polar residues repeat on the average every 3.6 residues, making the interior face of the helix non-polar. This feature was well described by

* Corresponding author.

[1] Current address: Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA.

SSDI 0301-4622(94)00149-9

Schiffer and Edmundson [4] using the "helical wheel" and Ptitsyn [5] noted that leucine tends to occupy internal turns of α-helical regions. These analyses were extended by Palau and Puigdoménech [6] and Lim [7] who found that hydrophobic triplets 1–2–5 and 1–4–5 are required for stabilising helices. More recently King and Sternberg [8] have used a machine learning approach to derive α-helix forming rules, and they confirm that hydrophobic residues occur at positions n, $n+3$ and $n+4$ or at positions $n-4$, $n-3$ and n or at n, $n+3$. Torgerson et al. [9] defined an infinite template in which hydrophobic residues occur at positions n, $n+4$, $n+7$, $n+11$, $n+14$, etc. and they fitted 247 α-helices to the template to maximise the strip-of-helix hydrophobicity index. They described significant quadrant distributions of amino acids with respect to N- and C-termini, interiors and entire helices. Lüthy et al. [10] obtained the probabilities of amino acids occurring in accessible and inaccessible positions in α-helices.

Amino acid propensities also differ at the N- and C-termini of the helices. Schellman [11] was the first to observe that glycine often adopts a positive ϕ conformation at the C-cap position. Argos and Palau [12] studied the compositional distribution of amino acids for particular positions within and surrounding elements of secondary structure and indicated that Ser and Thr frequently occur at N-terminal helical positions and confirmed that glycine is often adjacent to the C-terminus. Richardson and Richardson [13] found a 3.5:1 preference for Asn at the N-cap position, 2.6:1 for Pro at N-cap + 1, a strong preference (34%) for Gly at the C-cap and hydrophobic residues with a high probability at N-cap + 4 and C-cap − 4. Preißner and Bork [14] found that 30% of C-caps contain a Gly with the positive ϕ conformation and hydrophobic residues especially Leu and Ala occur very frequently at a position four residues before the Gly at the terminus. In an attempt to predict the position of the terminus in this type of helix, they derived six consensus sequence patterns for 13 positions around the C-capping residue Gly, three residues after and nine residues before it [15]. They were able to use the six patterns to identify 501 out of 575 helix ends in the Brookhaven Protein Data Bank.

From the alignments of homologous protein structures, Overington et al. [16,17] and Bowie et al. [18] calculated environment-specific amino acid substitution tables which give the probabilities of a residue in a specific structural environment (e.g. secondary structure, accessibility, side-chain hydrogen bonds etc.) to be replaced by all possible amino acid residues. Donnelly et al. [19] and Wako and Blundell [20] have extended the use of such amino acid propensities and substitution tables to provide a systematic prediction of helices. Their approach involves (1) prediction of the solvent inaccessible residues (2) use of Fourier transform approaches to predict the solvent inaccessible face and (3) N-cap and C-cap patterns to identify the helix termini.

Such statistical analyses, often carried out with the purpose of improving predictions, have been complemented by experimental studies. For example, the helix-forming tendencies of amino acids were studied by the host–guest method [21,22] and by making mutations in monomeric helix-forming peptides [23] or in dimeric helix-forming peptides [24]. Kemp et al. [25] obtained the propagation parameter of alanine in water from template-nucleated helices. Baldwin and co-workers found that straight non-polar amino acids are good helix-formers in water [26] and that the helical propensities of amino acids depend on their positions with respect to the termini [27,28]. Fersht and co-workers [29,30] have studied helices in a globular protein by site-directed mutagenesis. They concluded that it is not valid to assign to each amino acid a unique helix-forming propensity that is generally applicable to all positions in all helices. In order to simplify the protein folding problem and to investigate the role of alanine as a helix-stabilising residue, Matthews and co-workers [31,32] substituted residues within the helix 126–134 of T4 lysozyme by alanine, either singly or in selected combinations. They found that in five cases, the substitution of a buried leucine with alanine decreased the protein stability; the other four which are substitutions of Asp, Glu, Val and Asn with alanines increased the stability. Matthews and co-workers [33] have also demonstrated that insertions in helices often do not lead to extensions but rather the helices terminate with a similar length to the wild type, presumably influenced by packing of the core.

The results of the statistical analysis of amino acid propensities and the experimental work have

been used as a basis for speculation on the mechanism of protein folding. Chou and Fasman [34] suggested that helices were mediated by nucleation at the inner region by strong helix formers, whereas Argos and Palau [12] maintained that nucleation occurs at the termini. Other workers have assumed that in protein folding most helices are stabilized only when there is a loose association as in a molten globule [35]. Although it had been assumed that most sequences are unstable in aqueous solution, Baldwin and co-workers [26] have shown that helices can be stabilised in short peptides without association, usually by the presence of alanines and ion pairs between adjacent charged residues on the helix surface. Such helices may be importent nuclei in protein folding. Helices can be stabilised by non-aqueous solvents such as chloroethanol indicating that the propensity for helix formation is much often higher in a non-polar environment, an observation that is consistent with the stabilisation of helical regions in a molten globule.

In conclusion, helices have different conformations and amino acid propensities according to the position of the helix in the tertiary structure. Furthermore amino acid propensities differ in the capping and central regions. In order to study this further we have constructed a database of helices selected from globular protein structures defined at high resolution by X-ray structure analysis. We have clustered helices of the same length on the basis of patterns of solvent inaccessible residues. Noting that the major factor in such clustering is the size of the arc of the solvent inaccessible face, we have further clustered helices according to the number of residues in the face. By comparing helices in the same class we have been able to define propensities for amino acids at each position. These templates can be used to recognise a member of such a helical class.

2. Methods

Protein structures were obtained from the Brookhaven Databank [36]. Homologous structures in the database were aligned using COMPARER [37,38]. α-Helices and 3_{10}-helices were defined by DSSP [39]. The minimum lengths of an α-helix and a 3_{10}-helix were defined as 4 and 3, respectively. The accessibilities of residues were calculated [40]. If the side chain accessibility of a residue is less than 7%, the residue is defined as inaccessible [41].

Fig. 1 defines helix extensions and capping regions. The four residues pre N- and post C-terminus of an α-helix are the *N-extension* and *C-extension*, respectively. The *N-capping regions* of helices are defined as their N-extensions and their first four residues and *C-capping regions* are their C-extensions and their last four residues. The first and last residues in an α-helix are N and C respectively. A position which is n residues from the first residue of an α-helix in the N-capping region is $N + n$ if it is inside the helix; it is $N - n$ if it is in the N-extension. A position which is n residues from the last residue of the helix in the C-capping region is $C + n$ if it is inside the helix and it is $C - n$ if it is in C-extension. n is 1, 2, 3, or 4.

α-Helices with their extensions were extracted from the protein structures. For each element of secondary structure, the amino acid sequence, the length, the average side chain accessibility, side chain hydrogen bond to main chain or side chain etc. were also calculated and included in a database.

A pair of helices in two homologous proteins is equivalent if over 70% of the residues in the two segments are equivalent as defined using COMPARER [37,38]. If, in a group of three homologous proteins, the segment A is equivalent to both segment B and segment C, then we define B and C

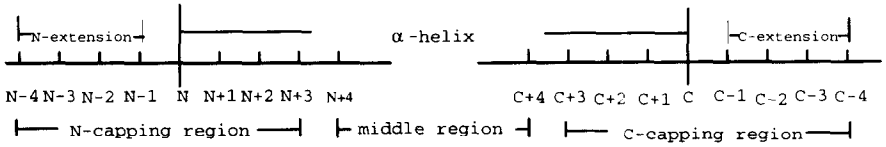

Fig. 1. The definition of sequence positions in an α-helix. An α-helix is shown as a line between N and C. Its N-extension is from N − 4 to N − 1. Its N-capping region is from N − 4 to N + 3. The C-extension and C-capping region are also defined similarly. The middle region is between N + 4 and C + 4 inclusive.

equivalent. Thus, groups of equivalent segments are defined and labelled in the database.

In the calculation of frequencies of amino acids, all segments in the databases are used but proper weights are set for equivalent segments, instead of including one representative. Each equivalent group is given weight 1. If there are n segments in an equivalent group, then each segment has weight $1/n$. The frequency of each amino acid in each segment is $1/n$. Hence the frequencies are not always integer.

We first classify α-helices according to their lengths. Suppose the side chain accessibilities of residues in the elements A and B are $(a_1, a_2,..., a_n)$ and $(b_1, b_2,..., b_n)$ where n is the length. Then the difference of side-chain accessibilities, or accessibility distance between A and B is

$$D_{a,b} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Each group of helices is used to calculate an accessibility distance matrix from which dendrograms can be generated using KITSCH [42]. Subgroups are defined by examining the dendrograms. We also examine the similarities of subgroups across the groups with different lengths.

This clustering of α-helices indicated that the size of the inaccessible face is the most important feature. This is defined as the arc of its helical wheel made by inaccessible residues (Fig. 2) and is measured by the number of residues in the arc. For an α-helix with less than 18 residues, we extend the helix in order to determine its inaccessible face. At each of the 18 positions in a helical wheel, if there is a residue, we use the side chain accessibility of the residue; if there is no residue, we calculate it using linear interpolation from neighbouring residues in the wheel. For an α-helix with more than 18 residues, where there are more than one residue at a position in the wheel, we use the average of their side-chain accessibilities. The largest inaccessible face was calculated for each α-helix in our database. Helices are classified together as type-n helices if they have n residues in their inaccessible face. Helices in the same type are then aligned by matching residues in the inaccessible faces. No gap is allowed in the alignments.

For any type of α-helix, we define its pattern by considering eight positions in its N-capping region: N − 4, N − 3, N − 2, N − 1, N, N + 1, N + 2 and N + 3; eight positions in its C-capping region: C + 3, C + 2, C + 1, C, C − 1, C − 2, C − 3 and C − 4; and its middle region between and including N + 4 and C + 4. Each position can be matched with each of the 18 helical wheel positions. Hence, for any helix class, a 20 × 18 table defines probabilities for each of the 20 amino acids at each of the 18 positions of the helical wheel, omitting the first four and last four residues of a helix. Eight such tables, one

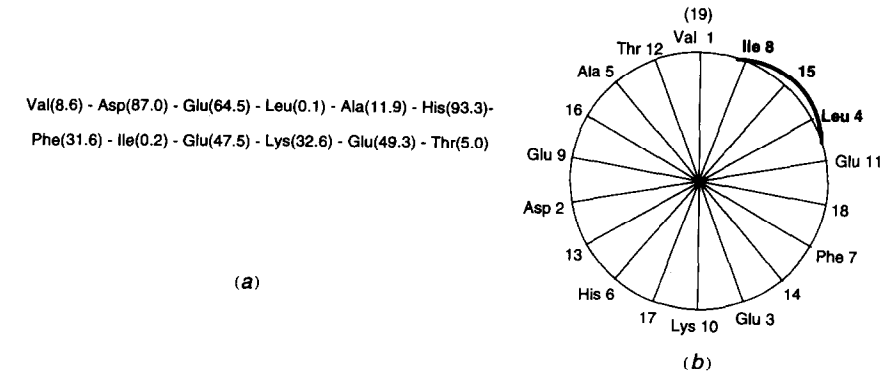

Fig. 2. The definition of the inaccessible face of α-helix. An inaccessible face of an α-helix is defined as the arc of its helical wheel [4] made by inaccessible residues in the helix. It is measured by the number of residues in the arc. (a) shows the amino acid sequence of the α-helix at 227–238 in protein 1pfk (PDB code). Side chain accessibilities are given in brackets. In (b), the α-helix is projected into a helical wheel. Side chain accessibilities for added dummy residues at positions such as 13, 14 etc. are calculated using linear interpolation from neighbouring residues in the wheel. For example, the accessibilities for dummy residues at 13 and 15 are 90.1% and 0.15%, respectively. Thus, the inaccessible face for the helix is at positions 4–15–8 as shown in bold line. Its size is 3 residues.

for each of the eight N-capping residues occurring at any position of the helical wheel, are obtained. There is a similar set of eight tables for the C-capping residues. Thus, there are 17 tables of dimensions 18×20 for each class of helix. These conditional probabilities are calculated from frequencies of occurrences of amino acids in corresponding conditions in the sequence alignment of each helix type.

To obtain propensities, probabilities of amino acids in various conditions were normalised by the probabilities of amino acids occurring in our protein structure database irrespective of any conditions.

In general, the probability P of amino acid X_i occurring in condition f is calculated as

$$P(X_i|f) = F(X_i|f) / \sum_{k=1}^{20} F(X_k|f)$$

where $F(X_i|f)$ is the frequency of occurrence of the amino acid X_i in the condition f. The condition variable f for an α-helix can be sequence position, helical wheel position or their proper combination. We also obtained the probability of the amino acid X_i occurring anywhere in a protein.

When the probability of an event is approximated by the observed frequency of the event, the more data we have the better the approximation and vice versa. Because the frequencies of occurrence of amino acids are counted in relatively strict conditions, some of the frequency data are very sparse, and the probabilities calculated will not be reliable. The smoothing procedure [43–45] used to overcome this problem will be described elsewhere (Zhu and Blundell, in preparation).

3. Results and discussion

3.1. General results of structural database analysis

Our analysis defined α-helices with 4 or more residues. Although α-helices with 4 residues occur very often in protein structures, it is shown that

Table 1
Distribution of amino acids in helices

	α-Helix				3_{10}-Helix			Protein	
AA [a]	N/O	P/C	PP	ΔPP	N/O	P/C	PP	N/O	P/C
I	428	5.6	1.14	+0.06	32	2.7	0.54	1561	4.9
F	340	4.5	1.22	+0.09	46	3.8	1.03	1166	3.7
V	506	6.7	0.97	−0.09	47	3.9	0.56	2171	6.9
L	836	11.0	1.40	+0.19	88	7.2	0.92	2480	7.8
W	133	1.8	1.26	+0.18	24	2.0	1.45	440	1.4
M	210	2.8	1.53	+0.08	14	1.2	0.64	574	1.8
A	937	12.3	1.47	+0.05	123	10.1	1.20	2659	8.4
G	293	3.9	0.44	−0.13	76	6.3	0.72	2759	8.7
C	108	1.4	0.82	+0.12	22	1.8	1.06	545	1.7
Y	253	3.3	0.93	+0.24	51	4.2	1.17	1130	3.6
P	167	2.2	0.42	−0.15	79	6.5	1.26	1644	5.2
T	359	4.7	0.73	−0.10	52	4.3	0.66	2057	6.5
S	353	4.6	0.62	−0.15	111	9.1	1.22	2381	7.5
H	157	2.1	0.95	−0.05	29	2.4	1.09	688	2.2
E	626	8.2	1.52	+0.01	107	8.9	1.64	1713	5.4
N	245	3.2	0.68	+0.01	42	3.5	0.73	1504	4.7
Q	353	4.6	1.29	+0.18	41	3.4	0.93	1142	3.6
D	395	5.2	0.85	−0.16	113	9.3	1.52	1936	6.1
K	525	6.9	1.15	−0.01	70	5.7	0.96	1898	6.0
R	380	5.0	1.29	+0.31	46	3.8	0.99	1227	3.9
Total	7603	–	–	–	1212	–	–	31676	–

[a] *AA*, amino acid; *N/O*, the number of occurrences in the secondary structure of this amino acid; *P/C*, the percentage composition of the amino acid in the secondary structure; *PP*, calculated propensity of the amino acid to form this secondary structure; *ΔPP*, the difference of the amino acid propensities between the values listed here and those published by Chou and Fasman (1978).

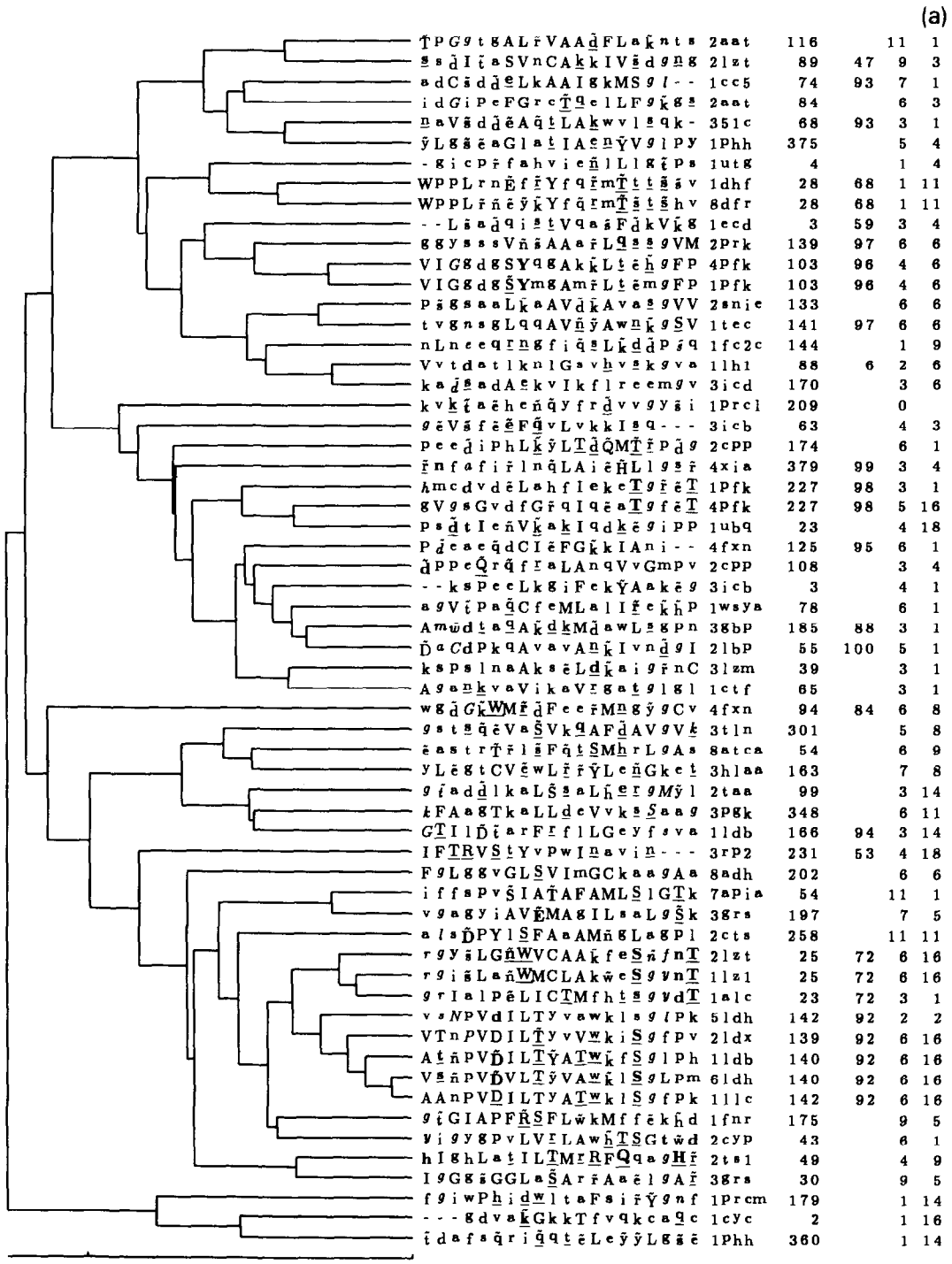

Fig. 3. The cluster of the α-helices of length (a) 12 residues and (b) 13 residues. Single letter codes of amino acids are presented according to their structural features [16]: *italic*, for positive ϕ; UPPER CASE, for solvent inaccessible residues (less 7%); lower case, for solvent accessible residues; **bold**, for hydrogen bonds to main chain amide nitrogen; underline, for hydrogen bonds to main chain carbonyl oxygen; ~, for side chain and side chain hydrogen bonds. Four residues before N-terminus and after C-terminus, respectively, are shown with the amino acid sequence of each α-helix. The PDB code, index of the first residue and the family number are given after amino acid sequence. The number of residues between the first residue in an α-helix to its inaccessible face and the size of the inaccessible face are also shown.

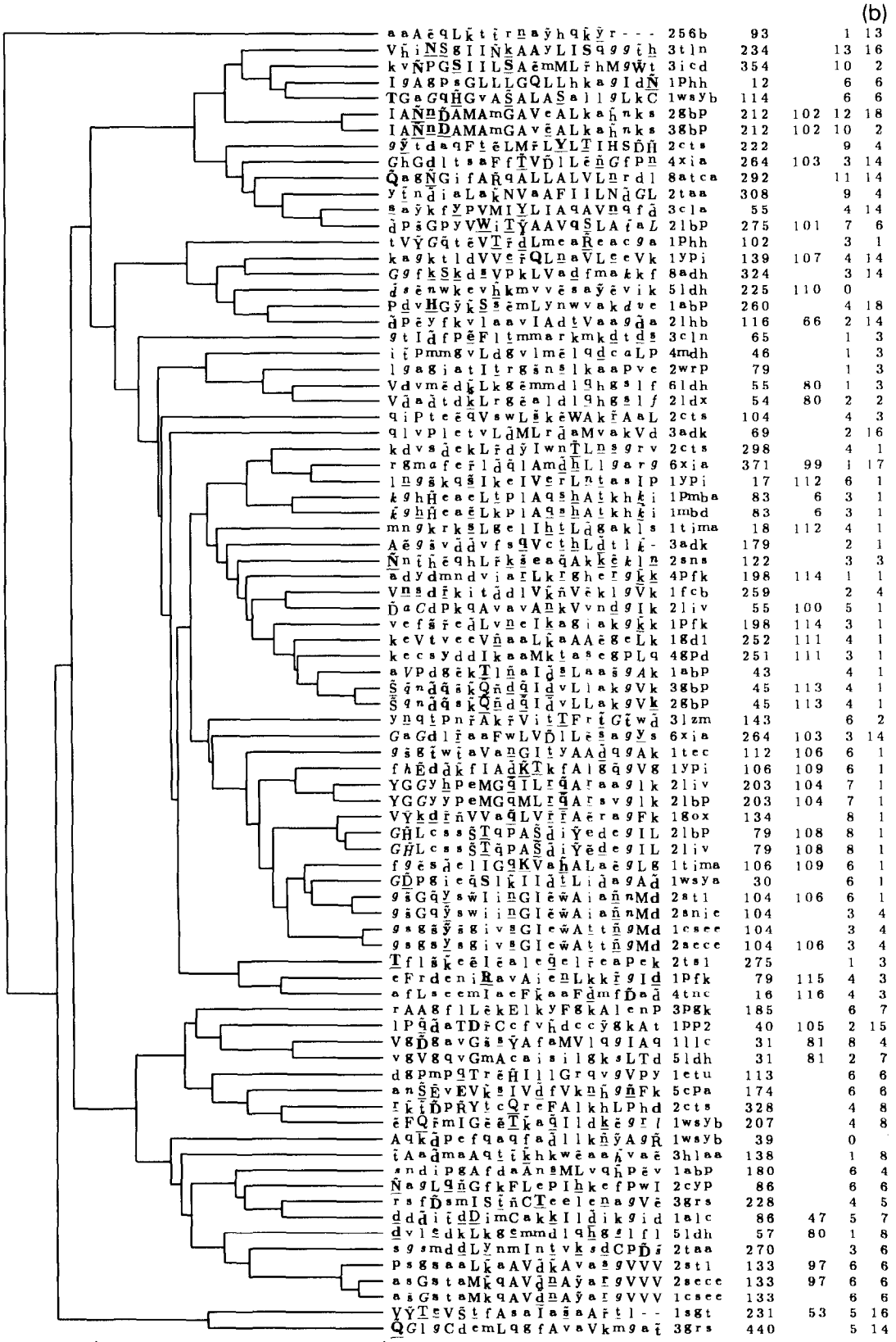

Fig. 3 (continued).

α-helices with 5 to 13 residues occur with almost equal frequency (data not shown). However, helices longer than 13 residues occur much less often. A similar trend can also be seen from the distribution of α-helices given by Barlow and Thornton [46]. In contrast to the distribution shown by Srinivasan [47], there are no relative maxima at lengths of 7, 11, 15 residues which are corresponding to 2, 3 and 4 full turns of helix respectively. We found 370 3_{10}-helices, of which 302 are three residues long and 46 are four residues long. None was greater in length than nine residues. The distributions of average side-chain accessibilities for α-helices and 3_{10}-helices are roughly normal with means at about 35% for α-helices and 50% for 3_{10}-helices, showing that 3_{10}-helices – like loops – are usually exposed to solvent.

The distributions of amino acids in α-helices and 3_{10}-helices in proteins are shown in Table 1. The correlation coefficients between these amino acid propensities and those published by Chou and Fasman [48] are 0.90 for α-helices. Some significant differences for some amino acids are observed. The two largest changes for α-helix are Arg ($\Delta PP = +031$: $0.98 \rightarrow 1.29$) and Tyr ($\Delta PP = 0.24$: $0.69 \rightarrow 0.93$). The new values show that Arg is not indifferent to α-helix formation but rather a strong α-helix former. Tyr is not an α-helix breaker but is indifferent to α-helix formation. The other significant changes in α-helix are Leu ($\Delta PP = +0.19$: $1.21 \rightarrow 1.40$), Gln ($\Delta PP = +0.18$: $1.11 \rightarrow 1.29$), and Asp ($\Delta PP = -0.16$: $1.01 \rightarrow 0.85$). Thus the strong α-helix formers are Met > Glu > Ala > Leu > Gln = Arg and those disfavouring α-helices are Pro > Gly > Ser > Asn.

3.2. Classification using accessibility distance

For each length we have clustered helices on the basis of the differences between patterns of solvent accessibility. Fig. 3 shows cladograms which cluster the helices of length 12 and 13 residues. Although the clustering is objective, it poses many subjective problems for classification of helices. If fairly broad sub-clusters are selected as a class, there are many outliers that appear to be structurally distinct. Examination of cladograms for helices of different length shows that there are strong similarities between helices with equivalent positions in the tertiary structure and there are also apparent ambiguities in the lengths arising from poor definitions of N- and C-termini. In general the dominant features within clusters seem to arise from two sources: first the size and position of the inaccessible face and second the position of the N- and C-capping regions. This indicated that each of these features might be analysed independently and then brought together combinatorially.

1. $(\square\ 17X)_n$
2. $(\square\ 10X\ \square\ 6X)_n$
3. $(\square\ 3X\ \square\ 6X\ \square\ 6X)_n$
4. $(\square\ 3X\ \square\ 6X\ \square\ 3X\ \square\ 2X)_n$
5. $(\square\ 3X\ \square\ 2X\ \square\ 3X\ \square\ 2X\ \square\ 3X)_n$
6. $(\square\square\ 2X\ \square\ 3X\ \square\ 2X\ \square\ 3X\ \square\ 2X)_n$
7. $(\square\square\ 2X\ \square\ 3X\ \square\ 2X\ \square\square\ 2X\ \square\ 2X)_n$
8. $(\square\square\ 2X\ \square\square\ 2X\ \square\ 2X\ \square\square\ 2X\ \square\ 2X)_n$
9. $(\square\square\ 2X\ \square\square\ 2X\ \square\ 2X\ \square\square\ 2X\ \square\square\ 1X)_n$
10. $(\square\square\ 2X\ \square\square\ 2X\ \square\square\ 1X\ \square\square\ 2X\ \square\square\ 1X)_n$
11. $(\square\square\square\ 1X\ \square\square\ 2X\ \square\square\ 1X\ \square\square\ 2X\ \square\square\ 1X)_n$
12. $(\square\square\square\ 1X\ \square\square\ 2X\ \square\square\ 1X\ \square\square\square\ 1X\ \square\square\ 1X)_n$
13. $(\square\square\square\ 1X\ \square\square\square\ 1X\ \square\square\ 1X\ \square\square\square\ 1X\ \square\square\ 1X)_n$
14. $(\square\square\square\ 1X\ \square\square\square\ 1X\ \square\square\ 1X\ \square\square\square\ 1X\ \square\square\square)_n$
15. $(\square\square\square\ 1X\ \square\square\square\ 1X\ \square\square\square\square\square\ 1X\ \square\square\square)_n$
16. $(\square\square\square\square\square\square\square\ 1X\ \square\square\square\square\square\ 1X\ \square\square\square)_n$
17. $(\square\square\square\square\square\square\square\ 1X\ \square\square\square\square\square\square\square\square\square)_n$
18. $(\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square)_n$ — buried helix
19. $(18X)_n$ — exposed helix

Fig. 4. The patterns of α-helices with various sizes of inaccessible face. $\square$ represents an inaccessible residue; kX represents k exposed residues ($k = 1,2,...,18$); n is a repeating number.

3.3. Classification using the size of the inaccessible face

We derived all possible sequence patterns of infinite length for the regular α-helices classified according to their inaccessible faces as shown in Fig. 4. The inaccessible residues in each sequence pattern constitute an inaccessible face with various sizes. Hydrophobic pairs (1–5) and triplets (1–2–5; 1–4–5) defined earlier by Lim [49] and others are sub-patterns. For example, 1–5 is equivalent to the sub-pattern $\square\ 3X\ \square$; 1–2–5 and 1–4–5 are equivalent to sub-patterns $\square\ \square\ 2X\ \square$ and $\square\ 2X\ \square\ \square$, respectively. The most populated class is the type-4 α-helix.

Totally exposed helices are usually short (4 or 5 residues). We observed some totally buried helices for example helix 64–74 in protein subtilisin (2st1).

These helices are usually in the centre of the proteins, surrounded by all other secondary structure elements.

The size of the inaccessible face gives a quantitative measure of how tightly a helix packs with other parts of a protein. Fig. 5 shows the superposition of the C-terminal domains of 1pfk and 4pfk two homologous structures. The two structures are very similar but there is a large shift of two equivalent helices at 227–238 in 1pfk and 4pfk. This gives a different packing "strength" of the helices with the core of protein. This difference in "strength" can be described by the size of the inaccessible face: the closely packed helix in 4pfk has 5 residues in its inaccessible face while its equivalent helix in 1pfk has only 3 residues. Fig. 6 gives examples of helices with different inaccessible faces viewed along the helix axis.

The probabilities of twenty amino acids in various conditions were calculated. For individual classes, the conditions are (1) 17 sequence positions (eight at each of the termini and one central) and (2) 18 positions in the helical wheel.

There are 7603 amino acids in the 769 independent helices. Smoothing has been used to overcome the sparsity problem of the data as described [50]. Tables 2 and 3 show the propensities of amino acids in the 18 different helical wheel positions of 3-type helices in which positions 4, 11 and 15 constitute the inaccessible face. In Table 2, only middle residues were included, the first four and last four residues in a helix were not considered. In Table 3, only N residues were used. There are 17 such tables for each class of helix.

As shown in Tables 2 and 3, the propensity of an amino acid varies with the helical wheel position, as each wheel position has different average accessibility. This is shown in Fig. 7 which is the plot of values in Table 2. The propensities of Leu at 18 helical wheel positions in the middle regions of 3-type helices were extracted from Table 2 and are highlighted in Fig. 8. The Figure shows clearly that propensities of Leu vary for different helical wheel positions. Where the value is greater than 1, Leu is a helical former and where value is less than 1, Leu is a helical breaker. The average value of 1.40 for the

Table 2
Distribution of amino acids in the middle region of type-3 α-helix

	1	2	3	↓ 4	5	6	7	8	9	10	↓ 11	12	13	14	↓ 15	16	17	18
I	1.39	0.49	1.71	2.53	1.71	0.47	1.41	1.71	0.51	0.78	2.41	1.84	0.49	1.67	2.39	0.78	0.49	1.41
F	1.49	0.35	1.08	1.97	1.08	0.35	1.51	1.08	0.38	0.70	1.97	1.19	0.43	1.11	1.92	0.70	0.38	1.51
V	1.09	0.49	0.88	1.30	0.88	0.48	1.09	0.88	0.51	0.65	1.26	0.91	0.45	0.90	1.25	0.67	0.48	1.07
L	1.46	0.58	1.40	2.01	1.38	0.58	1.50	1.41	0.58	0.81	1.90	1.24	0.62	1.38	1.94	0.79	0.59	1.47
W	1.86	0.50	1.64	1.79	1.64	0.43	1.86	1.64	0.57	1.64	2.00	1.93	0.79	1.86	1.93	1.64	0.57	1.86
M	1.89	1.00	1.44	3.06	1.50	1.00	1.89	1.44	1.06	2.22	3.00	1.39	1.11	1.72	3.17	2.28	1.06	1.89
A	1.57	1.86	1.33	1.50	1.32	1.87	1.58	1.33	1.82	1.38	1.43	1.25	1.68	1.24	1.42	1.38	1.82	1.56
G	0.52	0.29	0.31	0.75	0.30	0.28	0.52	0.30	0.29	0.36	0.78	0.32	0.31	0.31	0.72	0.37	0.29	0.53
C	1.82	0.65	1.18	1.47	1.24	0.59	1.82	1.18	0.65	1.18	1.53	1.47	0.88	1.29	1.59	1.18	0.65	1.82
Y	1.08	0.50	1.50	0.64	1.50	0.47	1.06	1.50	0.53	0.81	0.67	1.50	0.58	1.47	0.67	0.78	0.53	1.08
P	0.23	0.23	0.17	0.23	0.17	0.21	0.23	0.17	0.23	0.33	0.27	0.27	0.29	0.23	0.29	0.35	0.23	0.25
T	0.51	1.02	0.60	0.69	0.58	1.02	0.51	0.58	1.02	0.92	0.68	0.57	1.02	0.60	0.69	0.92	1.03	0.52
S	0.59	0.83	0.40	0.47	0.40	0.83	0.60	0.40	0.83	0.77	0.48	0.41	0.83	0.40	0.47	0.76	0.84	0.60
H	0.91	0.68	0.45	1.05	0.45	0.68	0.86	0.45	0.73	1.18	1.14	0.59	0.91	0.59	1.14	1.18	0.73	0.91
E	0.89	1.72	1.31	0.48	1.33	1.74	0.87	1.31	1.69	1.37	0.50	1.17	1.76	1.33	0.50	1.37	1.70	0.87
N	0.55	1.26	0.49	0.36	0.49	1.28	0.55	0.49	1.28	0.96	0.43	0.55	1.19	0.51	0.43	0.96	1.26	0.55
Q	1.33	1.83	1.50	0.44	1.53	1.81	1.33	1.50	1.81	1.61	0.53	1.42	1.81	1.36	0.53	1.58	1.78	1.33
D	0.39	1.56	0.70	0.46	0.70	1.57	0.38	0.70	1.54	1.15	0.49	0.72	1.49	0.79	0.49	1.15	1.54	0.39
K	1.05	1.98	1.37	0.30	1.37	2.02	1.05	1.38	1.95	1.50	0.33	1.33	1.87	1.33	0.35	1.50	1.97	1.05
R	1.62	1.64	2.10	0.44	2.10	1.67	1.59	2.13	1.64	1.85	0.51	2.18	1.62	1.97	0.51	1.87	1.64	1.59
→	19.1	64.2	37.0	1.5	37.6	63.4	21.7	27.5	67.0	49.1	1.3	27.0	63.6	27.9	1.5	42.8	63.4	24.8

→ Average side chain accessibility of the residues observed at each position of the helical wheel. ↓ Shows the positions which comprise the inaccessible face.

propensity of Leu in α-helices (Table 1) indicates that Leu is a helical former. This depends on the structural environment of the amino acid (helical wheel position).

The average side chain accessibility at each position of the helical wheel increases gradually on each side of the inaccessible face and usually reaches its highest value opposite the inaccessible face. We divided the average side chain accessibility into five different ranges 0–7%, 7–25%, 25–40%, 40–55% and over 55% and used these to obtain the probabilities of amino acids for different sequence positions of an α-helix; these were used in smoothing procedures [50]. The smoothed probabilities were converted into propensities and are shown in Fig. 9.

The propensities of amino acids at the N − 4, N − 3 and N − 2 positions are close to 1 and do not vary very much, indicating that these position do not contain significant information for helix formation. The apparent preference of Met for certain positions in this region of a helix is probably a consequence of the very small sample. However, the very significant

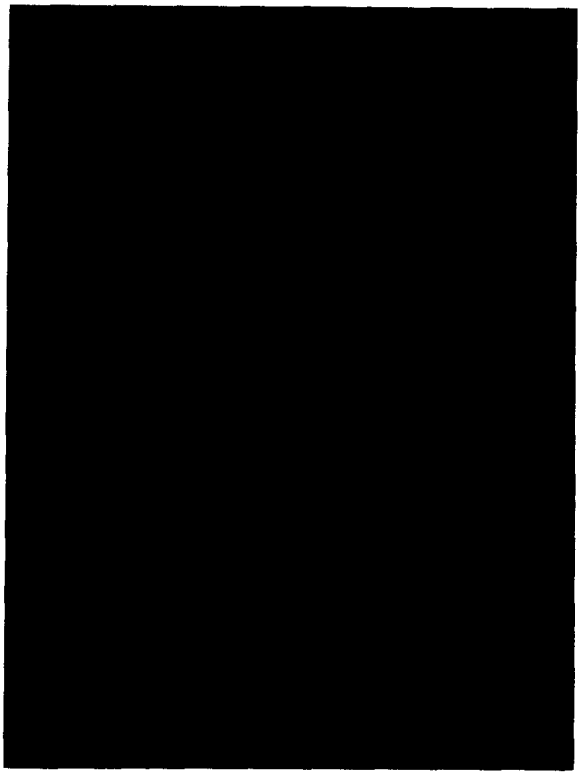

Fig. 6. Helices with different sizes of inaccessible face. The six helices are (1) 6×ia-64, (2) 256b-56, (3) 1pfk-227, (4) 6×ia-108, (5) 2mhr-19, (6) 1mba-126 (here the numbers in parentheses give the size of the inaccessible face; PDB code and the index of the first residue are used). The red dot surface gives their inaccessible faces. The figure was obtained using program SETOR [53].

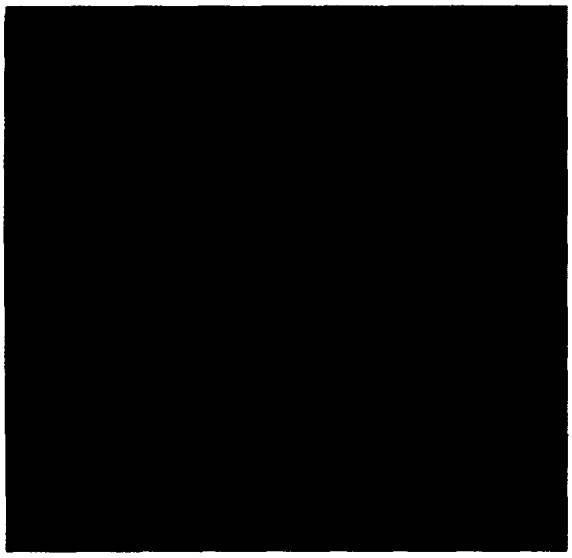

Fig. 5. Two types of α-helix in homologous protein structures. The structures are C-terminal domains of 1pfk in white line and 4pfk in green line. The two structures are superposed. There is a large shift of two equivalent helices at 227–238 in the two structures. The size of the inaccessible face of the helix is three residues in 1pfk and five residues in 4pfk. Hence, the size of the inaccessible face of α-helix gives a quantitative measure of the strength of packing.

patterns shown at N − 1 confirm that Ser, Asp, Asn, Gly and Thr are capping residues irrespective of their position relative to the inaccessible face; there is a tendency for Thr and Ser to be more strongly capping on the inaccessible side and Asp on the accessible side. The propensity of hydrophobic residues such as Ile, Phe, Val, Leu at N − 1 is very low. Pro can often be found at the first position (N) of an α-helix, especially in an exposed environment; but it is rare at other positions of an α-helix. Glu occurs often at N and N + 1 and most often at N + 2. Hydrophobic residues tend to occur at N and more often at the N + 3 position; this confirms previous results [13]. The C-terminus of an α-helix does not have such distinct patterns as the N-terminus. Hydrophobic residues are quite common at all posi-

tions, although their propensities, especially for Leu, are greater at C + 1 and C + 2. Tyr has some preference for the C position. Positively charged Lys and Arg do not show such a strong preference for the C-terminus of a helix as negatively charged Glu and Asp do at the N-terminus; in fact Glu is quite common at the C-terminus. Gly is frequently found at the C − 1 position where it acts as a C-capping residue. His is also found often at the C − 1 position but this might be a result of its function, such as binding haem iron, rather than a structural feature. Pro is very rare at C − 1, but with Gly is often found at C − 2. C − 3 and C − 4 positions contain little significant information for α-helix formation.

In the central region hydrophobic residues such as Phe, Leu, Val, Ile, Met are strongly preferred in the inaccessible regions, and Glu and Lys in the accessible regions. Gly has a preference for the inaccessible side in the central region of helices. Trp, Tyr, Arg and Gln show slight preference for the regions of intermediate accessibility reflecting the hydrophobicity of the region of the side chain close to the main chain.

Ala is the only amino acid which has a propensity greater than 1 in an α-helix irrespective of its solvent accessibility and sequence position. Ala can be in the middle of an α-helix or at the C-terminus in a buried or exposed environment. No other amino acid is a general helix-former. The propensities of other amino acids depend mainly on their positions relative to the inaccessible face and on the sequential position in the helix. The characterisation of the propensity with respect to the position of the helix in the tertiary structure is the major difference between this study and previous studies that have sought to define the amino acid distributions [12,13,34,51]. In general this analysis supports the conclusion of Serrano et al. [30] that single values of propensities are not generally applicable to all positions in the helix

3.4. Amino acid propensities and protein folding

During the process of protein folding the local environment of the amino acids in the polypeptide chain will vary. At times during the folding process amino acids will be completely surrounded by the

Table 3
Distribution of amino acids at the N position in type-3 α-helices

	1	2	3	↓ 4	5	6	7	8	9	10	↓ 11	12	13	14	↓ 15	16	17	18
I	0.78	0.76	0.67	1.63	0.69	0.76	0.78	0.69	0.78	0.67	1.63	0.69	0.76	0.67	1.63	0.67	0.76	0.78
F	1.08	1.19	1.46	1.49	1.46	1.19	1.08	1.38	1.19	0.78	1.49	1.46	1.19	1.46	1.49	0.78	1.19	1.08
V	1.13	0.72	1.03	1.38	1.01	0.72	1.13	1.04	0.72	1.01	1.36	1.01	0.72	1.03	1.39	1.01	0.72	1.14
L	1.13	0.74	0.74	1.10	0.73	0.74	1.13	0.78	0.73	0.91	1.09	0.73	0.74	0.74	1.10	0.91	0.74	1.15
W	1.50	1.57	1.57	2.21	1.64	1.57	1.57	1.86	1.64	1.29	2.29	1.71	1.57	1.57	2.21	1.29	1.57	1.43
M	1.72	1.28	0.94	0.94	1.06	1.28	1.72	1.11	1.33	1.67	1.00	1.06	1.28	0.94	0.94	1.67	1.28	1.67
A	0.89	1.02	1.12	1.43	1.08	1.01	0.89	1.06	1.00	1.20	1.40	1.10	1.02	1.12	1.44	1.19	1.02	0.92
G	0.38	0.37	0.43	0.77	0.44	0.37	0.38	0.44	0.38	0.40	0.76	0.44	0.37	0.43	0.77	0.40	0.37	0.37
C	1.24	1.65	0.88	1.59	1.00	1.65	1.29	1.12	1.71	1.41	1.65	1.00	1.65	0.88	1.59	1.41	1.65	1.18
Y	1.42	0.89	1.28	0.92	1.31	0.89	1.42	1.25	0.92	1.19	0.94	1.28	0.89	1.28	0.92	1.19	0.89	1.42
P	1.94	2.25	1.90	1.50	1.85	2.23	1.94	1.79	2.17	2.35	1.48	1.85	2.25	1.90	1.50	2.35	2.25	2.00
T	0.68	0.86	0.63	0.85	0.65	0.86	0.69	0.63	0.86	0.62	0.85	0.65	0.86	0.63	0.85	0.62	0.86	0.68
S	0.92	0.59	0.53	0.52	0.53	0.59	0.92	0.51	0.60	0.87	0.52	0.55	0.59	0.53	0.52	0.87	0.59	0.93
H	1.09	0.95	0.64	0.73	0.73	0.95	1.09	0.82	1.05	0.77	0.77	0.73	0.95	0.64	0.73	0.77	0.95	1.05
E	1.33	1.69	1.54	0.98	1.50	1.67	1.31	1.50	1.65	1.72	0.98	1.50	1.69	1.54	0.98	1.70	1.69	1.35
N	0.55	0.57	0.85	0.36	0.87	0.57	0.57	0.89	0.60	0.49	0.38	0.85	0.57	0.85	0.36	0.49	0.57	0.53
Q	0.94	1.61	1.47	0.72	1.47	1.61	0.94	1.47	1.58	1.08	0.75	1.47	1.61	1.47	0.72	1.08	1.61	0.92
D	0.97	1.30	0.85	0.57	0.85	1.30	0.97	0.84	1.28	0.84	0.59	0.87	1.30	0.85	0.57	0.84	1.30	0.97
K	0.80	0.90	0.98	0.60	0.97	0.90	0.80	0.95	0.90	0.82	0.60	0.98	0.90	0.98	0.60	0.82	0.90	0.80
R	1.08	1.08	1.92	0.87	1.87	1.08	1.08	1.90	1.08	1.13	0.90	1.90	1.08	1.92	0.87	1.13	1.08	1.08
→	19.1	64.2	39.0	1.5	37.6	63.4	21.7	27.5	67.0	49.1	1.3	27.0	63.6	27.9	1.5	42.8	63.4	24.8

↓ Average side chain accessibility of the residues observed at each position of helical wheel. ↓ Shows the positions which comprise the inaccessible face.

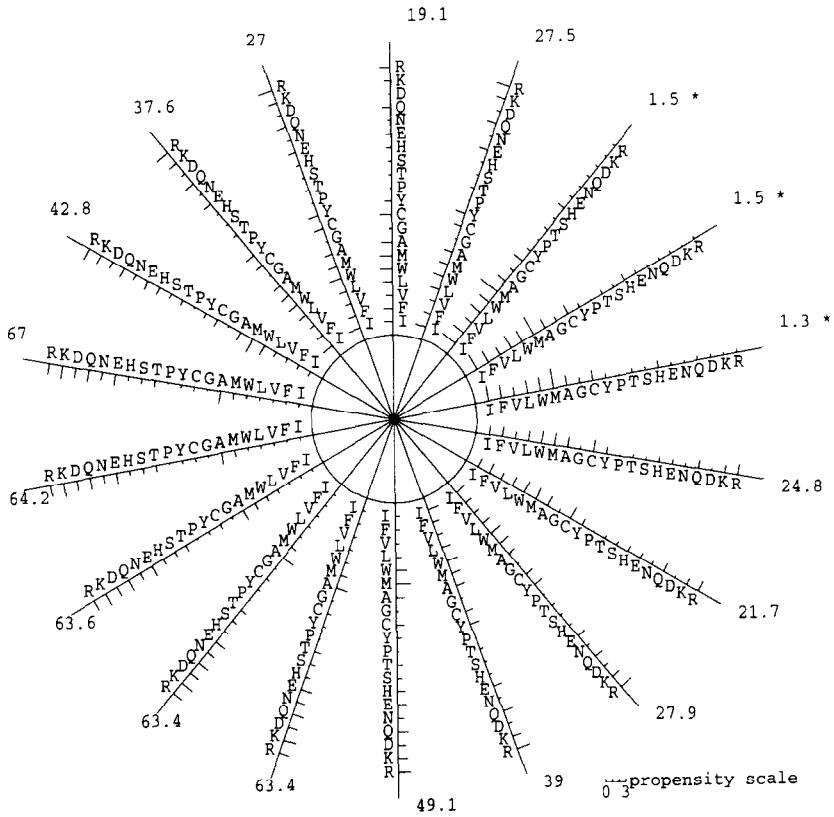

Fig. 7. Propensity plot of amino acids occurring in the middle region of a type-3 helix. The helical wheel positions marked by (*) indicate the inaccessible face. The propensities of the twenty amino acids occurring at each helical position are shown by small bars on the lines pointing out of the helical wheel. The propensity scale is plotted. The average side chain accessibilities of residues observed at each position are given.

aqueous solvent. At other times, for example, in a "molten globule" the environment may be partially hydrophobic. Because the propensities of amino acids for helices vary with the local environment, we might expect that nucleation of the secondary structure will depend on the degree of folding and be very

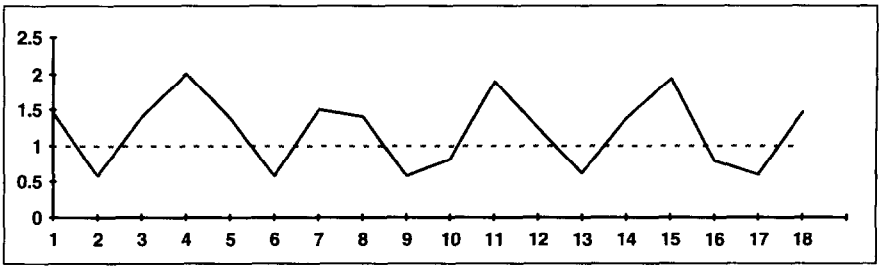

Fig. 8. The variation of the propensity of the amino acid Leu at helical wheel positions in the middle region of the type 3 helix. The data are taken from Table 4. The positions 4, 11, 15 are in an inaccessible face.

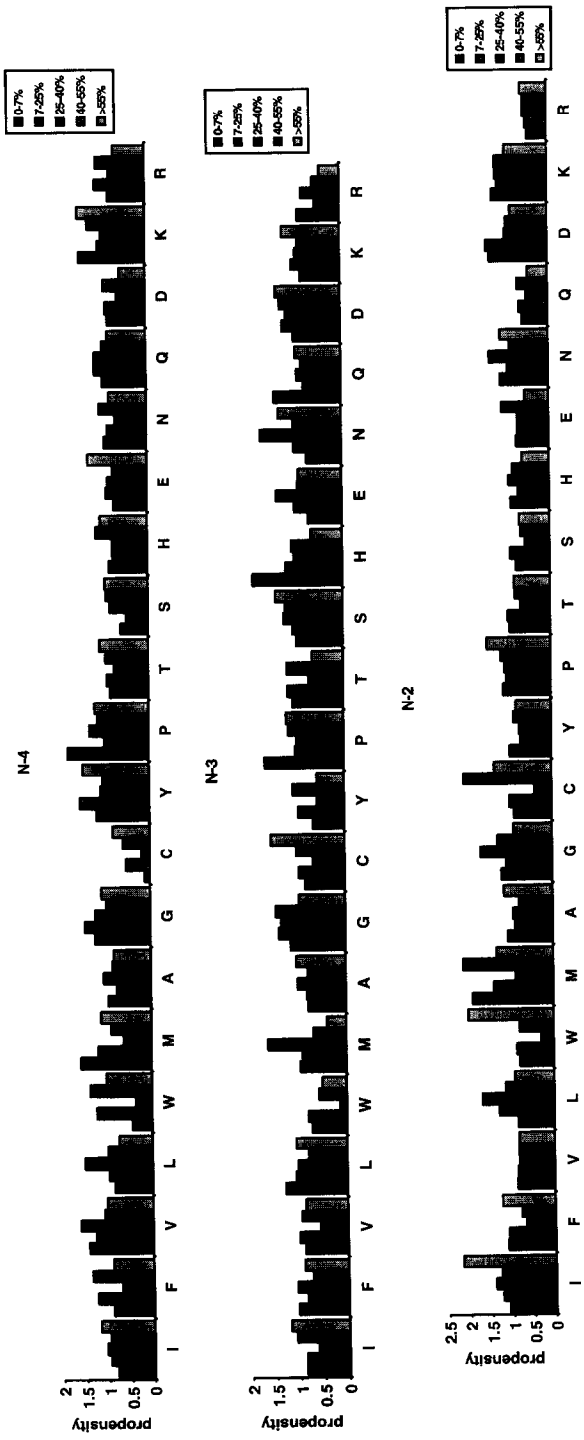

Fig. 9. The propensities of amino acids in five different accessibility environments at different positions in or near an α-helix irrespective of type of helix.

Fig. 9 (continued).

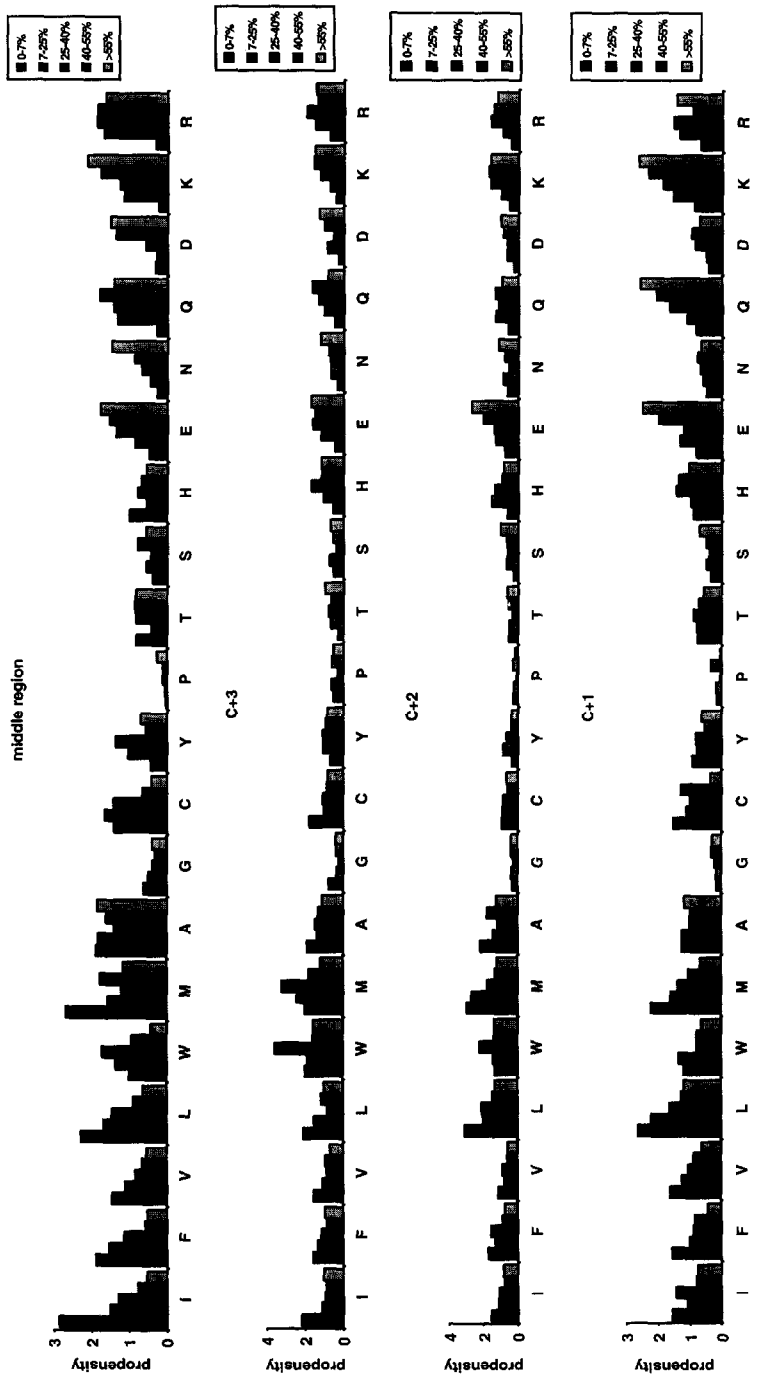

Fig. 9 (continued).

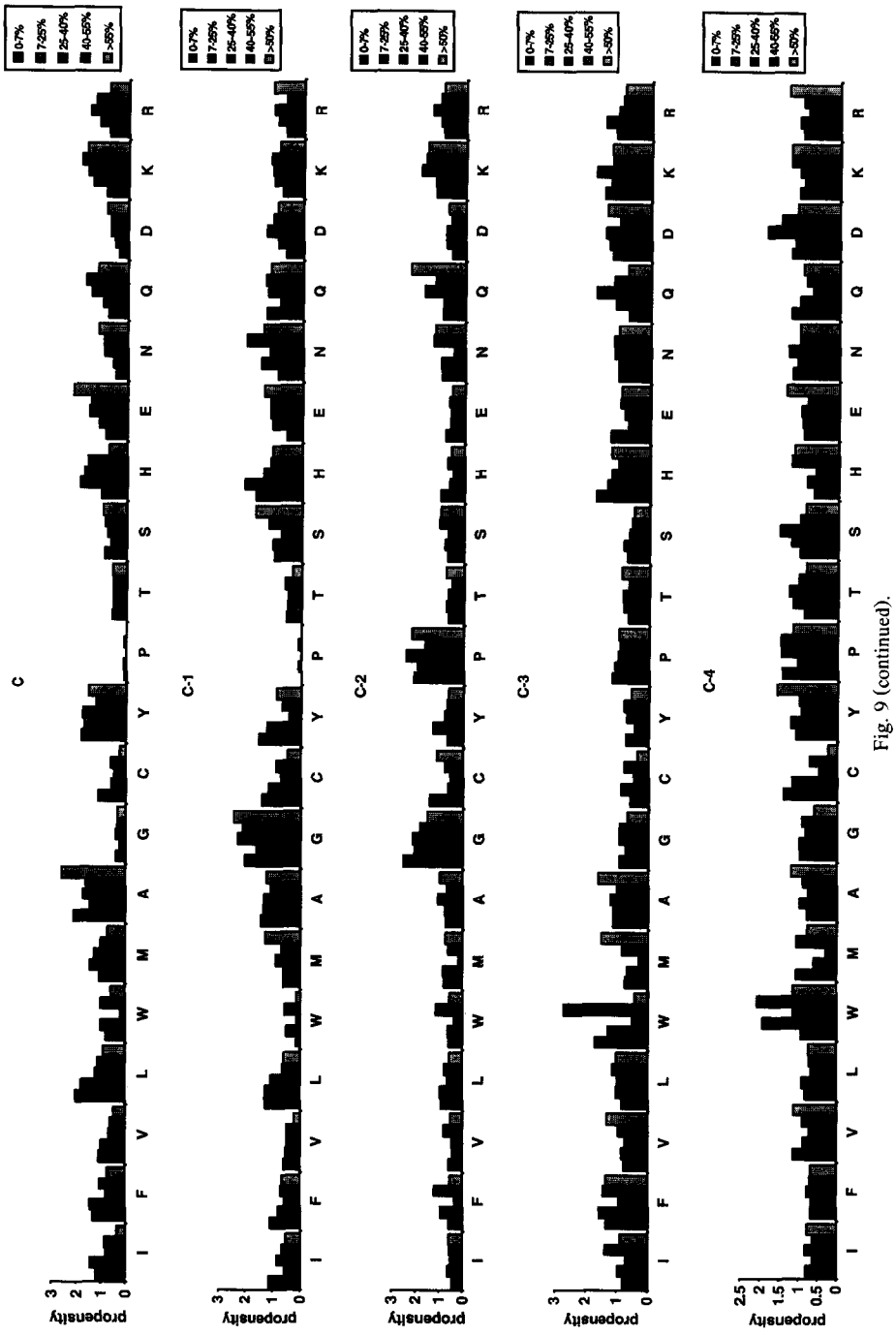

Fig. 9 (continued).

different for a random coil exposed to solvent when compared to the polypeptide chain in a molten globule or optimally folded native state.

The availability of propensities for different solvent-exposed states allows estimates of relative stabilities in these different environments. This may give clues to those regions that will nucleate helices in an aqueous environment. For example, a helix which is predicted to be stable in a fully exposed position to solvent might be expected to adopt a helical conformation at an early stage in the folding and, therefore, might contribute to nucleation of the folding of the protein. On the other hand, those helices that will require a hydrophobic core for stabilisation, i.e. those that are either amphipathic or completely hydrophobic, are unlikely to exist until at least a molten globule is formed.

We are now carrying out analyses of such effects. These hopefully can be correlated with NMR studies on isolated peptides and proteins in aqueous environments and in non-aqueous solvents. Such measurements should give clues about the stability of particular secondary structures during the folding process of the protein intermediate.

3.5. Application on secondary structure predictions

The propensities of amino acids at specific positions in different classes of helices allow us to develop new methods of secondary structure prediction (Zhu and Blundell, 1994 in preparation). For each class of helix a pattern of propensities for each position in the helix can be calculated. This enables us to develop a database of profiles or templates for each class of helix defined by its length and accessibility. Each of these profiles or templates can be compared with an amino acid sequence and the probability of secondary structure evaluated by the products of propensities of the residues at each position. This allows us to predict which secondary structures are likely to be assumed by each segment of the sequences. Where more than one secondary structure element is predicted we take that with the highest probability. The approach can be used to predict not only the position of a secondary structure element in a protein sequence, but also the orientation and accessibility with respect to the protein core.

3.6. Application on solving inverse protein folding problems

Johnson and co-workers [16,52], Eisenberg and co-workers [10] have developed methods to identify protein sequences that adopt a known protein 3-D fold. These methods depend on the calculation of amino acid propensities and/or amino acid substitution probabilities for each position in the sequence for proteins of known 3-D structure. These propensities/substitution patterns are then compared with the sequences of unknown structure. Such propensities that characterise specific secondary structural elements with respect to the tertiary structure provide a more accurate and characteristic profile. The inclusion of this information should improve profile and template search procedures.

Acknowledgements

We are grateful to Dr. David Moss and our colleague in the modelling group at Birkbeck for many stimulating discussions, especially Dr. Chris Topham for his help on smoothing procedures.

References

[1] L. Pauling, R.B. Corey and H.R. Branson, Proc. Natl. Acad. Sci. USA, 37 (1951) 205.
[2] T.L. Blundell, D. Barlow, N. Borkakoti and J. Thornton, Nature, 306 (1983) 281.
[3] M.F. Perutz, J.C. Kendrew and H.C. Watson, J. Mol. Biol. 13 (1965) 669.
[4] M. Schiffer, and A.B. Edmundson, Biophys. J., 7 (1967) 121.
[5] O.B. Ptitsyn, J. Mol. Biol., 42 (1969) 501.
[6] J. Paulau and P. Puigdoménech, J. Mol. Biol., 88 (1974) 457.
[7] V.I. Lim, J. Mol. Biol., 88 (1974) 857.
[8] R.D. King and M.J. Sternberg, J. Mol. Biol., 216 (1990) 441.
[9] R.R. Torgerson, R.A. Lew, V.E. Reyes, L. Hardy and R.E. Humphreys, J. Biol. Chem., 266 (1991) 5521.
[10] R. Lüthy, A.D. McLachlan and D. Eisenberg, Proteins, 10 (1991) 229.
[11] C. Schellman, in R. Jaenicke (Editor), Protein Folding, Elsevier, Amsterdam, 1980, p. 53.
[12] P. Argos and J. Palau, Int. J. Pep. Protein Res., 19 (1982) 380.
[13] J.S. Richardson and D.C. Richardson, Science, 240 (1988) 1648.

[14] R. Preißner and P. Bork, Biochem. Biophys. Res. Commun. 180 (1991) 660.
[15] P. Bork and R. Preißner, Biochem. Biophys. Res. Commun. 180 (1991) 666.
[16] J.P. Overington, M.S. Johnson, A. Šali and T.L. Blundell, Proc. R. Soc. London, Ser. B, 241 (1990) 132.
[17] J.P. Overington, D. Donnelly, M.S. Johnson, A. Šali and T.L. Blundell, Protein Sci., 1 (1992) 216.
[18] J.U. Bowie, R. Lüthy and D. Eisenberg, Science, 253 (1991) 164.
[19] D. Donnelly, J.P. Overington and T.L. Blundell, Protein Eng., 7 (1994) 645.
[20] H. Wako and T.L. Blundell, J. Mol. Biol., 238 (1994) 693.
[21] M. Sueki, S. Lee, S.P. Powers, J.B. Denton, Y. Konishi and H.A. Scheraga, Macromolecules, 17 (1984) 148.
[22] H.A. Scheraga, Proc. Natl. Acad. Sci. USA, 82 (1985) 5585.
[23] P.C. Lyu, M.I. Liff, L.A. Marky and N.R. Kallenbach, Science, 250 (1990) 669.
[24] K.T. O'Neil and W.F. Degrado, Science, 250 (1990) 646.
[25] D.S. Kemp, J.G. Boyd and C.C. Muendel, Nature, 352 (1985) 451.
[26] S. Padmanabhan and R.L. Baldwin, J. Mol. Biol., 219 (1991) 135.
[27] R. Fairman, K.M. Armstrong, K.R. Shoemaker, E.J. York, J.M. Stewart and R.L. Baldwin, J. Mol. Biol., 221 (1991) 1395.
[28] A.J. Doig, A. Chakrabartty, T.M. Klingler and R.L. Baldwin, Biochemistry, 33 (1994) 3396.
[29] L. Serrano and A.R. Fersht, Nature, 342 (1989) 296.
[30] L. Serrano, J.-L. Neira, J. Sancho and A.R. Fersht, Nature, 356 (1992) 453.
[31] X.-J. Zhang, W.A. Baase and B.W. Matthews, Biochemistry, 30 (1991) 2012.
[32] X.-J. Zhang, W.A. Baase and B.W. Matthews, Protein Sci., 1 (1993) 761.
[33] D. Heinz, W.A. Baase, F.W. Dahlquist and B.W. Matthews, Nature, 361 (1993) 561.
[34] P.Y. Chou and G.D. Fasman, Biochemistry, 13 (1974) 211.
[35] D. Barrick and R.L. Baldwin, Protein Sci., 2 (1993) 869.
[36] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanovichi and M. Tasumi, J. Mol. Biol., 112 (1977) 535.
[37] A. Šali and T.L. Blundell, J. Mol. Biol., 212 (1990) 403.
[38] Z.-Y. Zhu, A. Šali and T.L. Blundell, Protein Eng., 5 (1992) 43.
[39] W. Kabsch and C. Sander, Biopolymers, 22 (1983) 2577.
[40] T.J. Richmond and F.M. Richards, J. Mol. Biol., 119 (1978) 537.
[41] T.J.P. Hubbard and T.L. Blundell, Protein Eng., 1 (1987) 159.
[42] J. Felsenstein, Evolution, 39 (1985) 783.
[43] M.J. Sippl, J. Mol. Biol., 213 (1990) 859.
[44] A. Šali, PhD thesis, University of London, 1991.
[45] C.M. Topham, A. McLeod, F. Eisenmenger, J.P. Overington, M.S. Johnson and T.L. Blundell, J. Mol. Biol., 229 (1993) 194.
[46] D.J. Barlow and J.M. Thornton, J. Mol. Biol., 201 (1988) 601.
[47] R. Srinivasan, Ind. J. Biochem. Biophys., 13 (1976) 192.
[48] P.Y. Chou and G.D. Fasman, Adv. Enzymol., 47 (1978) 45.
[49] V.I. Lim, J. Mol. Biol., 88 (1974) 873.
[50] Z.-Y. Zhu, PhD Thesis, University of London, 1994.
[51] P.Y. Chou and G.D. Fasman, Biochemistry, 13 (1974) 222.
[52] M.S. Johnson, J.P. Overington and T.L. Blundell, J. Mol. Biol., 231 (1993) 735.
[53] S.V. Evans, J. Mol. Graphics, 11 (1993) 134.