J. Mol. Biol. (1992) **223**, 1053–1061

Sequence Alignment Approach to Pick Up Conformationally Similar Protein Fragments

A. S. Kolaskar and Urmila Kulkarni-Kale

Biotechnology Training Programme
Department of Zoology, University of Poona
PUNE-411 007, India

(Received 3 July 1991; accepted 8 November 1991)

Crystal structure data of globular proteins were used to prepare (ϕ, ψ) probability maps of 20 proteinous amino acids. These maps were compared grid-wise with each other and a conformational similarity index was calculated for each pair of amino acids. A weight matrix, called Conformational Similarity Weight (CSW) matrix, was prepared using the conformational similarity index. This weight matrix was used to align sequences of 21 pairs of proteins whose crystal structures are known. The aligned regions with more than seven contiguous amino acids were further analysed by plotting average weight ($\bar{W}$) values of overlapping heptapeptides in these regions and carrying out curve fitting by Fourier series having TEN harmonics. The protein fragments corresponding to the half-linewidth of peaks were predicted as fragments having similar conformation in the protein pair under consideration. Such an approach allows us to pick up conformationally similar protein fragments with more than 67% accuracy.

Keywords: conformationally similar protein fragments; alignment method with conformational similarity weight matrix; alignment; conformational similarity

1. Introduction

In the last few years, due to the concerted efforts of a few protein crystallographers, the size of the protein crystal structure data bank (PDB†) has increased considerably. This data bank can be used to derive properties of proteinous amino acids at single residue level, which was not possible earlier. Such information can be integrated into algorithms for predicting the three-dimensional structure of proteins. Several successful attempts have been made to derive useful information from the analysis of protein crystal structure data (Blundell *et al.*, 1990; Jones & Thirup, 1986; Kolaskar & Ramabrahmam, 1981; Ramakrishnan & Srinivasan, 1990; Pastore & Lesk, 1990). Crystal structure data have also been used to derive properties such as accessibility, hydrophobicity, flexibility and the volume occupied by the side-chain inside the globule of the protein as well as on the surface of the protein (Parker *et al.*, 1986). In fact, Froemmel has created a data bank of properties of amino acids, several of which are derived using the PDB data bank (personal communication). Protein crystal structure data have also been used to develop algorithms for predicting secondary structures of proteins, though with limited success (Chou & Fasman, 1978; Gibrat *et al.*, 1987; Lim, 1974; Schultz *et al.*, 1974).

On the other hand, primary structure data are used to find out sequence homology by alignment studies. Whenever sequence similarity is high and extended in length, it can be correlated easily with structural similarity, but the assessment of structural significance based on sequence similarity becomes difficult if sequence similarity is weak or is restricted to short regions. Therefore protein sequence data have been used to find out patterns of oligopeptides that are signatures to particular classes of proteins. The existence of such patterns has been used to postulate the functions of proteins (Brown *et al.*, 1982; Dixon *et al.*, 1986). Such an approach has limitations mainly because the existence of a particular oligopeptide may not necessarily be a good indicator of the function of the protein. It has been established that oligopeptides having the same sequence can exist in different conformations and can be part of different secondary structures (Kabsch & Sander, 1984; Kolaskar & Ramabrahmam, 1984; Argos, 1987).

† Abbreviation used: PDB, protein crystal structure data bank.

0022-2836/92/041053-09 $03.00/0

Further, protein fragments having very different sequences were found to have similar local conformations (unpublished results). Therefore, attempts were made to combine the protein structure data bank and the sequence data bank to extract useful information. Sander & Schneider (1991) have used the secondary structure information and sequence alignment approach to prepare a database of Homology-Derived Protein Structures. Risler *et al.* (1988) have used homologous proteins and the information of their three-dimensional structures to derive structurally similar patterns. Similarly, Bairoch (1990) has derived a data bank of structural motifs called PROSITE. This suggests that the PDB data bank can be used as a powerful research tool to gain insight into the problems related to protein structure and protein engineering.

In this work, we describe a method developed to pick up conformationally similar regions, which uses the weight matrix derived from protein crystal structure data, an approach of sequence alignment and analysis of the aligned regions. The accuracy of this method was studied by its application to proteins whose three-dimensional structures are known. From these studies, it is shown that the method can be used to pick up, in the test protein whose three-dimensional structure is not known, regions conformationally similar to those proteins whose three-dimensional structures are known.

2. Method

The method developed can be divided into 3 parts:

(1) preparation of the Ramachandran probability maps of amino acid residues and their comparisons;

(2) preparation of the weight matrix used in the sequence alignment studies;

(3) picking up conformationally similar regions.

(a) *Preparation of Ramachandran probability maps and their comparison*

The protein data bank (PDB) has more than 535 entries (release dated 20th April, 1990). 102 proteins were selected from this data bank. These proteins have different 3-dimensional structures and different sequences. Thus, each protein entry that was chosen was the one having the best-resolved structure among other similar proteins. In the case of multi-chain proteins only 1 chain was used in the present study. The co-ordinates of the main-chain atoms were used to calculate dihedral angles (ϕ, ψ). Taking into consideration the R-factor of the proteins, the standard deviation for the (ϕ, ψ) was found to be $\geqslant \pm 10°$; therefore a grid interval of 20° was chosen to prepare (ϕ, ψ)-Ramachandran probability maps. A simple computer program was written to sort out each type of amino acid residue along with its (ϕ, ψ) values. The occurrence of each type of amino acid in each grid $20° \times 20°$ was counted by the program. These occurrence values were normalized to obtain $P_{(i,j)}$, the probability values (see Fig. 1). Each map was compared grid-wise with all other maps.

ΔP_{AB} values were calculated using the following formula:

$$\Delta P_{AB} = \sum_{i=1}^{18} \sum_{j=1}^{18} |P_{A(i,j)} - P_{B(i,j)}|, \qquad (1)$$

where $P_{A(i,j)}$ and $P_{B(i,j)}$, respectively, are the percent frequencies of amino acid residues A and B in the grid (i, j) in the (ϕ, ψ)-plane. The summation was carried over the entire (ϕ, ψ)-plane. ΔP_{AB} values were used as an index of conformational similarity between residues A and B. Such ΔP_{AB} values were calculated for every pair of amino acids. These values are given in the lower triangle of Table 1. Standard deviation (σ) associated with each ΔP_{AB} was calculated; these values are given in the upper triangle of Table 1.

(b) *Preparation of the weight matrix*

ΔP_{AB} and σ_{AB} values given in Table 1 were used to create the weight matrix. It should be mentioned here that the reference (ϕ, ψ)-probability map, with which the remaining 19 amino acid residue maps were compared, changes with the reference amino acid. Due to this $\Delta P_{AB_{min}}$ and the corresponding σ_{AB} values are different in the case of each of the reference amino acids (see Table 1).

In other words, $(\Delta P_{AB_{min}} + n\sigma_{AB}) \neq (\Delta P_{BC_{min}} + n\sigma_{BC})$, necessarily. This creates a problem during weight assignment though ΔP_{AB} and ΔP_{BA} are equal. They fall in different class intervals and thus are given different weight (W) values. The problem was overcome by assigning the minimum possible weight value to the amino acid pair AB and BA. The criteria used to assign the weight (W) are mentioned below:

(1) An amino acid replaced by itself

$$(\Delta P_{AA}) = 0 \qquad W = 1{\cdot}0;$$

or

(2) an amino acid replaced by any of those amino acids having:

(a) $(\Delta P_{AB}) \leqslant (\Delta P_{AB_{min}} + \sigma_{AB})$ $\quad W = 0{\cdot}9;$

or

(b) $(\Delta P_{AB_{min}} + \sigma_{AB}) \leqslant (\Delta P_{AB}) \leqslant (\Delta P_{AB_{min}} + 2\sigma_{AB})$ $\quad W = 0{\cdot}66;$

or

(c) $(\Delta P_{AB_{min}} + 2\sigma_{AB}) \leqslant (\Delta P_{AB}) \leqslant (\Delta P_{AB_{min}} + 3\sigma_{AB})$ $\quad W = 0{\cdot}5;$

or

(d) $(\Delta P_{AB}) \geqslant (\Delta P_{AB_{min}} + 3\sigma_{AB})$ $\quad W = 0{\cdot}0;$

The weight matrix thus formed is termed as the Conformational Similarity Weight (CSW) Matrix and is given in Table 2. The above-mentioned method of assigning weight values to various amino acids was supported by the data on single linkage cluster analysis of ΔP_{AB} values. The results of the cluster analysis are shown in Fig. 2.

The CSW matrix was then used in the ALIGN program, which is based on the Needleman & Wunsch (1970) algorithm. The optimal bias (B) and penalty (P) values for this matrix were determined using the procedure discussed by Schwartz & Dayhoff (1979) and George *et al.* (1990). The optimal bias and penalty values calculated for CSW matrix were found to be 6,6 respectively. Similarly, in the FASTP program of Wilbur & Lipman (1983), the cut off value was found to be 33 for this matrix. This cut off value was calculated using the Pearson (1990) procedure.

(c) *Picking up conformationally similar regions*

Sequences of the human α-chain of haemoglobin (HAHU), Skipjack tuna cytochrome *c* (CCBN), hen egg

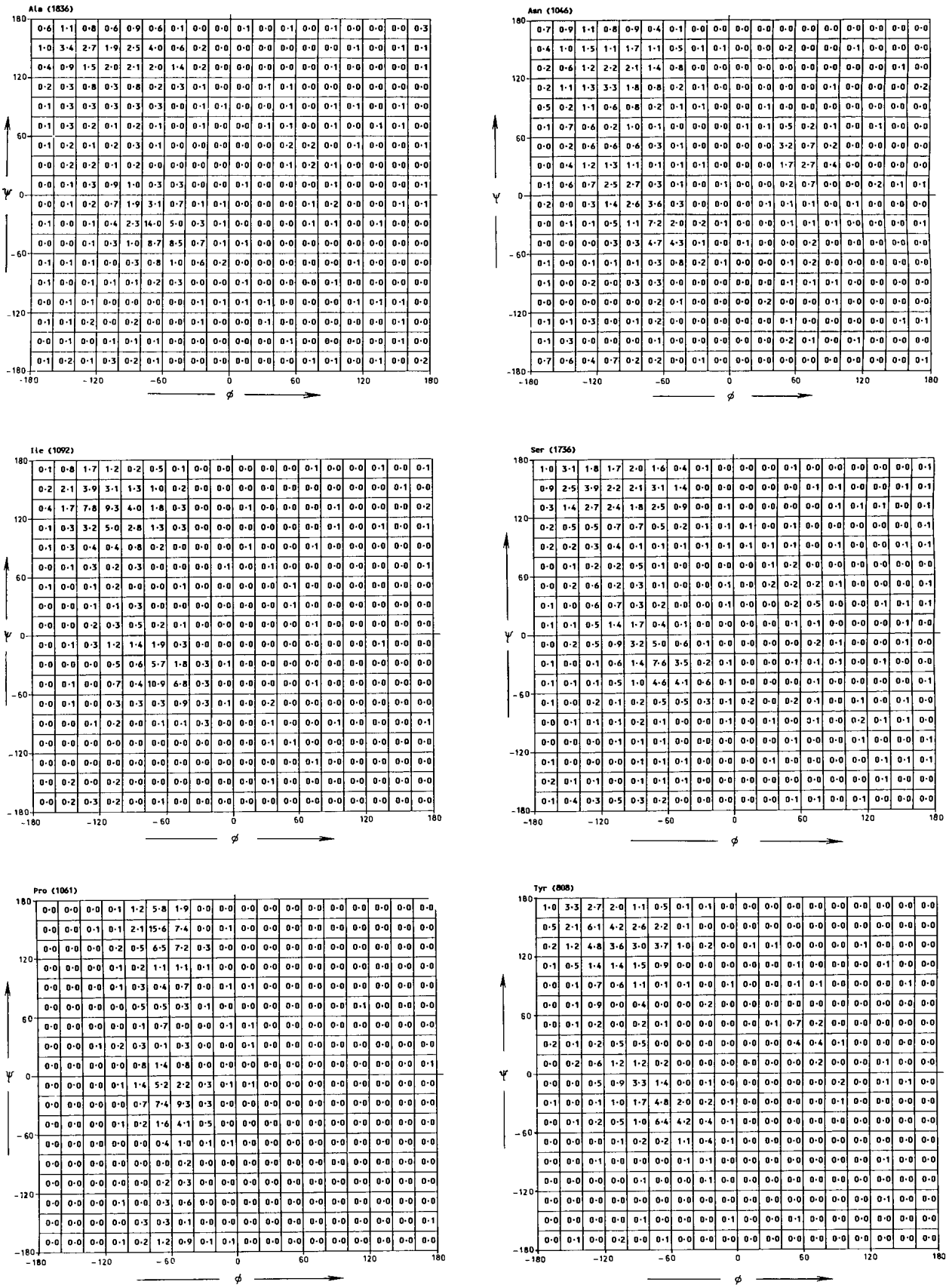

Figure 1. (ϕ, ψ)-probability maps for a representative selection of amino acid residues obtained from crystal structure data of 102 different globular proteins. Number of each type of residue used to draw these maps is given in parentheses, along with the name of the residue. Note very different probability distribution for the Pro residue.

Table 1

ΔP_{AB} and σ_{AB} values calculated from comparison of (ϕ, ψ)-probability maps of amino acid residues

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0·0	6·9	12·5	9·6	11·8	6·5	**4·8**	18·4	10·4	15·6	7·1	6·9	6·2	10·1	21·3	10·3	12·3	9·5	13·5	13·7
Arg	47·0	0·0	9·2	7·3	8·6	4·8	6·4	16·0	7·2	11·8	6·1	**4·9**	**5·8**	**5·7**	22·8	7·7	8·0	6·1	9·1	9·8
Asn	78·9	68·7	0·0	**6·4**	9·0	9·4	11·7	**12·6**	8·0	14·6	10·9	8·6	11·4	9·7	22·6	8·3	9·3	10·0	10·5	13·1
Asp	63·9	61·1	**55·5**	0·0	8·9	7·2	8·7	13·9	7·0	13·7	7·8	**6·1**	9·4	8·8	21·2	7·8	8·9	8·9	10·2	12·4
Cys	74·0	56·6	71·9	72·5	0·0	8·6	11·8	14·2	7·3	12·8	9·8	8·9	9·6	7·5	22·1	7·1	**6·1**	8·1	7·1	10·1
Gln	46·1	39·3	72·4	88·8	62·7	0·0	5·9	16·5	7·3	12·0	**5·0**	4·7	6·3	7·2	22·5	7·9	8·0	7·1	9·3	10·0
Glu	**37·4**	44·8	73·7	60·7	74·2	45·7	0·0	18·3	9·6	14·3	6·1	5·7	6·9	10·3	23·0	11·0	11·8	9·3	13·3	12·8
Gly	124·6	129·4	**112·3**	115·7	122·2	125·9	127·2	0·0	14·2	20·1	17·9	15·5	17·8	15·5	24·1	13·9	15·1	16·4	15·4	18·8
His	66·6	52·6	66·9	58·3	56·8	53·5	62·8	122·8	0·0	13·0	8·7	6·9	9·0	**7·0**	23·0	7·2	7·7	7·7	7·4	11·0
Ile	78·9	64·5	85·8	81·2	71·1	64·6	70·9	137·1	73·9	0·0	10·3	11·9	12·2	10·9	27·9	14·3	10·0	10·9	11·0	**5·4**
Leu	47·9	47·3	74·8	56·6	66·3	**39·0**	47·3	132·3	64·1	55·4	0·0	5·5	6·2	8·1	22·7	10·0	9·1	7·5	10·4	9·0
Lys	44·7	**40·7**	64·8	**54·4**	69·3	41·0	40·7	123·1	56·5	67·2	43·8	0·0	6·7	7·6	22·1	7·8	8·0	6·9	9·6	10·5
Met	50·9	**43·9**	80·6	68·2	68·1	50·8	54·7	135·1	63·1	65·6	48·6	55·0	0·0	8·0	22·6	9·3	9·4	7·9	10·5	9·5
Phe	56·0	**39·2**	69·3	62·5	53·2	48·1	58·0	128·8	**51·0**	62·5	50·1	52·3	50·1	0·0	23·1	**7·6**	6·8	**6·1**	**6·2**	8·8
Pro	113·9	121·3	127·2	116·6	122·7	120·2	118·3	151·7	124·7	140·2	119·8	118·2	123·8	127·1	0·0	**19·9**	23·5	22·6	23·7	26·3
Ser	57·7	53·5	65·4	59·6	57·9	54·6	67·1	115·7	56·7	81·4	63·2	52·4	66·8	**51·7**	**110·2**	0·0	6·9	7·3	7·6	11·9
Thr	68·9	54·7	70·2	63·8	**52·9**	51·6	69·3	121·1	58·1	89·2	56·2	54·0	64·4	50·9	128·0	51·8	0·0	·7·8	**5·3**	7·5
Trp	61·0	45·2	74·1	68·7	59·4	51·9	61·4	136·2	59·0	61·9	53·5	56·5	53·5	**44·0**	116·3	58·7	58·3	0·0	8·0	9·2
Tyr	71·6	53·2	74·1	65·5	56·6	52·5	71·2	126·8	51·8	62·5	57·3	58·7	60·3	**41·0**	128·3	51·9	**42·7**	54·0	0·0	8·6
Val	73·6	58·0	83·1	73·7	63·8	57·5	69·8	135·0	65·2	**33·9**	52·7	61·1	58·9	53·4	133·9	70·9	51·9	55·8	51·4	0·0

ΔP_{AB} values, in percentage, are given in the lower triangle (below the 0·0 to 0·0 diagonal) and corresponding σ_{AB} values are given in upper triangle (above the 0·0 to 0·0 diagonal). $\Delta P_{AB_{min}}$ and corresponding σ_{AB} values are shown in bold type.

white lysozyme (EC 3.2.1.17) (LZCH), pig pancreatic tissue kallikrein (EC 3.4.21.35) (KQPG) and human calmodulin (MCHU) were chosen from the NBRF PIR data bank since the crystal structures of these proteins are known and they have different 3-dimensional structures. Further, these proteins belong to different families having very little sequence similarity.

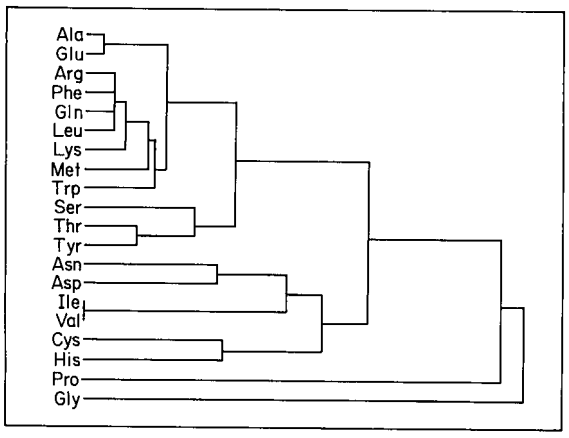

Figure 2. Dendrogram obtained for ΔP_{AB} values using Single Linkage Cluster Analysis.

FASTP runs using CSW matrix on each of these 5 sequences were carried out on the complete data bank of PIR. Those proteins, which have an int score > cutoff and opt score > int score were chosen (for explanation of int and opt scores, see Pearson, 1990). Those proteins for which the crystal structure data are available were selected from this set to carry out pairwise alignment.

From the output of the ALIGN program (obtained by using CSW matrix) aligned regions having more than 7 continguous amino acid residues (without gaps) were picked up. For these regions the average weight ($\bar{W}$) was calculated using the following procedure:

Each region was broken into overlapping heptapeptides such as $(i, i+6)$, $(i+1, i+7)$, $(i+2, i+8)$, etc. The weight values corresponding to each amino acid pair given in the CSW matrix (Table 2) were used to calculate the average weight ($\bar{W}$) of each heptapeptide. These $\bar{W}$ values were assigned to N-terminal residues of the corresponding heptapeptide, for example, ith of $(i, i+6)$ heptapeptide, to the $(i+1)$th residue in case of $(i+1, i+7)$ etc. These $\bar{W}$ values were plotted against overlapping heptapeptides.

Table 2
Conformational Similarity Weight matrix of proteins

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	10	6·6	0	0	0	6·6	9	0	0	0	5	6·6	5	0	0	0	0	0	0	0
Arg	6·6	10	0	0	0	9	6·6	0	5	0	6·6	9	9	9	0	5	5	6·6	5	0
Asn	0	0	10	9	0	0	0	0	5	0	0	0	0	0	0	6·6	0	0	0	0
Asp	0	0	9	10	0	0	0	0	6·6	0	0	5	0	0	0	6·6	0	0	0	0
Cys	0	0	0	0	10	0	0	0	9	0	0	0	0	5	0	9	6·6	5	5	0
Gln	6·6	9	0	0	0	10	6·6	0	5	0	9	9	5	6·6	0	0	5	5	5	0
Glu	9	6·6	0	0	0	6·6	10	0	0	0	5	9	0	0	0	0	0	0	0	0
Gly	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
His	0	5	5	6·6	9	5	0	0	10	0	0	0	0	5	0	9	5	5	6·6	0
Ile	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	9
Leu	5	6·6	0	0	0	9	5	0	0	0	10	9	6·6	5	0	0	0	5	0	0
Lys	6·6	9	0	5	0	9	9	0	0	0	9	10	5	5	0	5	5	0	0	0
Met	5	9	0	0	0	5	0	0	0	0	6·6	5	10	6·6	0	0	0	6·6	0	0
Phe	0	9	0	0	5	6·6	0	0	5	0	5	5	6·6	10	0	5	5	9	9	0
Pro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
Ser	0	5	6·6	6·6	9	0	0	0	9	0	0	5	0	5	0	10	6·6	5	6·6	0
Thr	0	5	0	0	6·6	5	0	0	5	0	0	5	0	5	0	6·6	10	5	9	0
Trp	0	6·6	0	0	5	5	0	0	5	0	5	0	6·6	9	0	5	5	10	5	0
Tyr	0	5	0	0	5	5	0	0	6·6	0	0	0	0	9	0	6·6	9	5	10	0
Val	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	10

$(W)\max = 10$ and $(W)\min = 0$

Since this plot gave several peaks and troughs, it was smoothed using Fourier series having the following equation:

$$f(x_r) = \tfrac{1}{2}a_0 + \sum_{n=1}^{10} (a_n \cos nx_r + b_n \sin nx_r), \qquad (2)$$

where

$$a_0 = 2 \sum_{r=1}^{m} y_r,$$

$$a_n = 2 \sum_{r=1}^{m} y_r \cos (nx_r),$$

$$b_n = 2 \sum_{r=1}^{m} y_r \sin (nx_r),$$

n is number of harmonics,
x_r is starting position of rth heptapeptide,
y_r is average weight ($\bar{W}$) of rth heptapeptide,
m is total number of data points.
$f(x_r)$ is weight value calculated using Fourier series having n harmonics.

From trial-and-error studies, the cut off at harmonic 10 was found to be a good approximation. This plot for (MCHU and TPCHCS) is given in Fig. 3. The half line-width at each peak (at 66% height of peak), assuming that the peak corresponds to normal distribution, was calculated; this region was assigned as conformationally similar region to the corresponding region in the reference protein. Using this approach, conformationally similar regions in 21 pairs of proteins were picked up.

3. Results and Discussion

It can be seen from the probability distribution Ramachandran plots (Fig. 1) that these maps are unique for Gly and Pro and $\Delta P_{AB_{min}}$ values are large for these maps, as expected. Conformational similarity index, measured in terms of the ΔP_{AB} values, is given in Table 1. As mentioned in Method, during comparison of amino acid residue (ϕ, ψ)-maps, the reference amino acid changes. Because of this, amino acid pairs AB and BA can fall into different class intervals formed using $\Delta P_{AB_{min}}$ and corresponding σ_{AB} values. For example, $\Delta P_{Ala\text{-}Glu} = 37{\cdot}4$ is a minimum value among ΔP_{AB} values with Ala as reference amino acid; the corresponding σ is 4·8. Therefore the boundaries of three class intervals

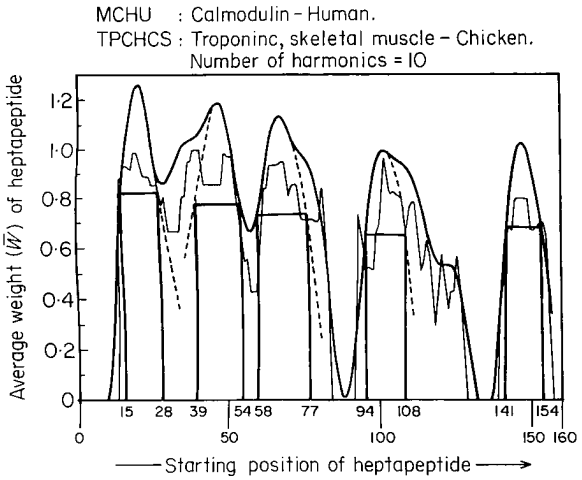

Figure 3. Graph of overlapping heptapeptides in aligned regions in calmodulin (MCHU) and troponin *c* (TPCHCS) against average weight ($\bar{W}$) values (thin line). There being large numbers of peaks and troughs it is smoothed using Fourier Series having 10 harmonics (thick line). Assuming that the peaks have normal distribution half linewidth is determined. The broken line shows normal distribution around each peak. Half linewidth, which gives the length of conformationally similar protein fragments, are also shown.

will be 42·2, 47·0 and 51·8, respectively. It can be seen from Table 1 that $\Delta P_{\text{Ala-Ser}}(=57{\cdot}7) > (\Delta P_{\text{Ala-Glu}}+3\sigma)$. On the other hand, when Ser is taken as the reference then $\Delta P_{\text{Ser-Phe}}(=51{\cdot}7)$ is the minimum having $\sigma = 7{\cdot}6$ and thus, $\Delta P_{\text{Ser-Ala}} (=57{\cdot}7) \leqslant (\Delta P_{\text{Ser-Phe}}+\sigma)$. This suggests that (ϕ, ψ)-probability distribution of Ser, when compared with Ala, map is similar but Ala has (ϕ, ψ) distribution similar to Glu. This is due to the fact that two-dimensional distribution is compared by a single index, which depends on the reference map. This aspect of closeness among (ϕ, ψ) distribution was also studied by carrying out Single Linkage Cluster analysis. The results are given in Figure 2. It can be seen from Figure 2 that (Ala, Glu), (Arg, Phe, Gln, Leu), (Thr, Tyr), (Asn, Asp), (Ile, Val), (Cys, His) have most similar conformational distribution in the (ϕ, ψ)-plane. Figure 2 also indicates that if Ala is replaced by Glu, or *vice versa*, in the protein fragments, conformational change will be minimal. Similar conclusions can be drawn for other pairs of amino acids. Thus, in short, the results of cluster analysis or $(\Delta P_{\text{AB}}, \sigma_{\text{AB}})$ values can be used to generate a weight matrix that can be used in sequence alignment studies that will give, during alignment, maximum importance to the property of conformational similarity. In the present study $(\Delta P_{\text{AB}}, \sigma_{\text{AB}})$ values are used to create a weight matrix. In order to make the weight matrix symmetrical, lower weight values assigned to pair AB and BA gave results that were in accordance with the dendrogram obtained from Cluster analysis in most cases. Therefore, as can be seen from Table 2, $W = 0$ was assigned to pairs Ala-Ser and Ser-Ala even though Ser-Ala can be assigned $W = 0{\cdot}9$, assuming Ser as the reference.

It can also be seen from Table 1 that the pairs of amino acids that have low conformational similarity index (ΔP_{AB}) are not necessarily related by any one of the physicochemical properties such as bulkiness of side-chain, hydrophobicity, aliphatic or aromatic nature, secondary structure forming capacity, etc. Thus, the conformational similarity of the main chain seems to be a result of several physicochemical properties of amino acids that have a different net effect on the conformational similarity index. This property, therefore, can be considered as a new property. Therefore, the CSW matrix (Table 2), formed using conformational similarity index is different as compared to the property matrix (AAAM) of McLachlan (1971). The CSW matrix is also quite different from the structural similarity matrix of Risler *et al.* (1988), obtained using a set of homologous proteins. Further, the CSW matrix remained essentially unchanged when created using data from 90 proteins (PDB Release Oct, 1988) and 102 proteins (PDB Release Nov, 1990). This indicated that the CSW matrix is robust and thus the weight values derived can be used with reasonable confidence.

From the FASTP output of five proteins (haemoglobin α, cytochrome *c*, lysozyme, calmodulin and kallikrein) 21 pairs of proteins whose crystal structures are known were selected (Table 3). The pairwise alignment using CSW matrix was carried out for each of the 21 pairs. As an example, the results of MCHU and TPCHCS alignment are discussed below. From the output of the ALIGN program of these proteins, the average weight ($\bar{W}$) values were calculated and plotted for overlapping continguous heptapeptides (Fig. 3). It can be seen from Figure 3 that the number of peaks (thin line curve) are too many and sharp. Therefore smoothing is essential. Initial attempts to fit the polynomial equation were not successful. Therefore, the Fourier series equation was chosen and various harmonics were tried. For 10 harmonics, as can be seen in Figure 3, the curve gave a reasonable number of peaks with a minimum number of shoulders and broadness. The half linewidth of each peak was determined using standard Gaussian distribution. Gaussian curves that fit various peaks are also shown by a dotted line in Figure 3. From this analysis the half linewidths, which are suggested as the lengths of the regions to be conformationally similar in test and reference proteins, were found to be 15–28, 39–54, 58–77, 94–108, 141–154. These numbers correspond to the aligned output of MCHU and TPCHCS. Similar studies were carried out for the remaining 20 pairs of proteins mentioned in Table 3. From these studies 135 pairs of protein fragments having similar conformation were picked up.

The following approach was used to check the correctness of our prediction. (ϕ, ψ)-values were calculated for each of the predicted pair of conformationally similar oligopeptides. Those $(\phi_{\text{ref}}, \psi_{\text{ref}})$ of

Table 3

List of the proteins, obtained in the FASTP output using CSW matrix, for which crystal structure data is available

No.	Reference protein		Test proteins		% Identity
1	HAHU	Haemoglobin α-chain—Human	HAHO	—Haemoglobin α-chain—Horse	87·9
			GGLMS	—Globin v—Sea lamprey	35·8
			JGECA	—L-Arabinose binding protein—*E. coli*	27·6
			PVCAB	—Parvalbumin—Carp	25·5
			TRBOTR	—Trypsinogen—Bovine	17·8
			KQRTTN	—Tonin—Rat	19·2
			HYBST	—Neutral proteinase—*Bacillus* species	18·4
2	KQPG	—Tissue kallikrein—Pig	TRBOTR	—Trypsinogen–Bovine	40·5
			TRSMG	—Trypsin—*Streptomyces*	29·2
			KYBOA	—Chymotrypsinogen—Bovine	36·4
			TRPGTR	—Trypsin—Pig	41·7
			ELPG	—Elastase—Pig	35·5
			PRRTG	—Mast cell proteinase—Rat	31·1
3	CCBN	—Cytochrome *c*—Tuna	CCPC5O	—Cytochrome *c*550—*Paracoccus*	42·4
			CCQF2R	—Cytochrome *c*2—*Rhodospirillum*	21·0
			CCRZ	—Cytochrome—Rice	33·0
4	LZCH	Lysozyme—Chicken	LZHU	—Lysozyme—Human	59·6
			LZTK	—Lysozyme—Turkey	83·6
			CPBOA	—Carboxypeptidase A—Bovine	21·2
5	MCHU	Calmodulin—Human	TPCHCS	—Troponin *c*—Chicken	52·7
			PVCAB	—Parvalbumin–Carp	23·8
			KLBOI	—Calcium binding protein—Bovine	15·5

% Identities obtained using mutation data matrix while aligning the reference proteins with test proteins are also given. Please note that for most pairs % identity is not >40%.

the reference protein were compared with (ϕ_{test}, ψ_{test}) of the test protein.

If, $|\phi_{ref}-\phi_{test}|$ and $|\psi_{ref}-\psi_{test}| \leqslant 30°$ then the conformation of the amino acid pair under consideration was said to be similar. The regions under consideration were said to be conformationally similar to each other when more than 60% amino acid pairs were conformationally similar by the above criterion. We are aware that there are other criteria to check the conformational similarity of protein fragments. Kabsch (1978) has used root-mean-square differences of equivalent C^{α} positions in the three-dimensional space after optimal superimposition to define conformational similarity. However, the above mentioned criterion was chosen to avoid subjectivity, which might arise due to optimal superimposition. In general, this criterion agrees well with other criteria used to study structural similarity.

By our alignment studies, the protein fragment 16–28 of TPCHCS (chicken troponin *c*) was predicted to be conformationally similar to the protein fragment 6–18 of MCHU (human calmodulin). The correctness of this prediction was checked by the application of the above mentioned (ϕ, ψ) criterion. It can be seen from Figure 4(a), where main-chain atoms are used to show the conformation, that the fragment 16–28 of TPCHCS and fragment 6–18 of MCHU have very similar conformations. Only, the conformation around residue C^{α}_{26} of TPCHCS is slightly different as compared to that around C^{α}_{16} of MCHU. In this particular case most of the amino acids are identical (9 out of 13). Additional examples of correctly predicted conformationally similar regions are shown in Figure 4(a). The reference protein chosen is KQPG (pig tissue kallikrein) and the test proteins are TRPGTR (pig trypsinogen), TRSMG (*Streptomyces griseus* trypsin) and TRBOTR (bovine trypsinogen). These examples indicate that not only the criterion chosen to check the correctness of the predicted conformation is good but also point out that the method discussed above has the ability to pick up conformationally similar regions. This fact can be further supported from the results shown in Figure 4(b). For example, JGECA (*Escherichia coli* L-arabinose binding protein) has, in all, ten helices and two β-sheets (Gilliland & Quiocho, 1981). The total number of residues in this protein, in helix and sheet are 128 (41·8%) and 69 (22·5%), respectively. On the other hand, HAHU (human α-haemoglobin) is known to be an α-helical protein with most of the residues in the helical conformation (Fermi *et al.*, 1984). Figure 4(b) shows the conformation of the protein fragment 73–79 of JGECA (*E. coli* L-arabinose binding protein) and the conformation of the corresponding protein fragment 27–33 of HAHU (human α-haemoglobin). It can be seen that both these fragments, which were picked up by our method, have α-helical conformation. This is particularly interesting because JGECA and HAHU have 306 and 141 amino acid residues, respectively. Alignment of such sequences is a difficult proposition using any weight matrix. In fact, one has to use high penalty (P) values, which mainly allows sliding operation. Even under such drastic conditions it

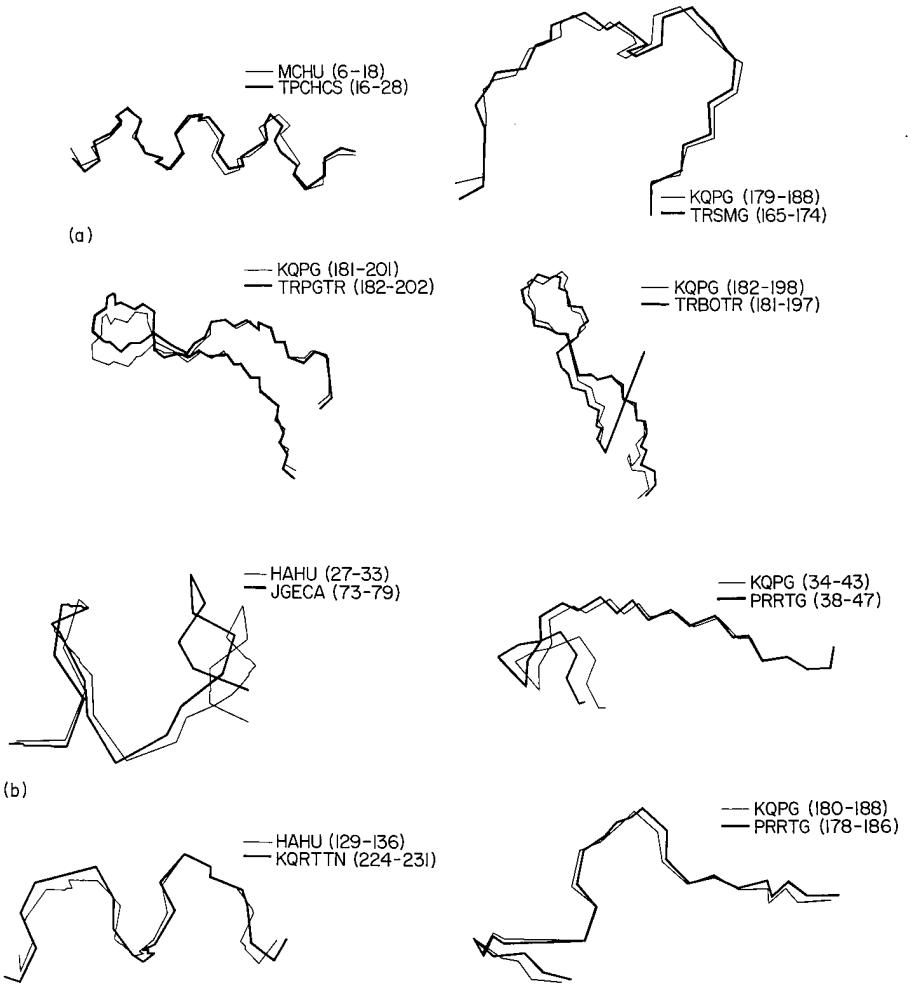

Figure 4. Plot of only main-chain atoms in conformationally similar protein fragments, which are predicted by our method. (a) Conformationally similar protein fragments from the pairs of structurally similar proteins. (b) Conformationally similar fragments from pairs of structurally dissimilar proteins.

was possible to pick up three conformationally similar regions. Further the α-helix and β-sheet content in JGECA is low while HAHU has only α-helix as secondary structure. Other helical regions in JGECA are not getting aligned with helical regions in HAHU. Another example taken from Figure 4(b) is KQRTTN (rat tonin). This protein belongs to β-protein class (Fuginaga & James, 1987) and has a very different length as compared to HAHU. As can be seen, the conformation of the fragment 224–231 in KQRTTN is similar to the fragment 129–136 of HAHU. Similar examples from proteins having different three-dimensional structure and little sequence similarity are given. These fragments are 34–43 and 180–188 of KQPG (pig pancreatic tissue kallikrein) and regions 38–47 and 178–186 of PRRTG (rat mast cell proteinase II), respectively. These fragments take coil conformation and are still picked up, showing that our method does not necessarily pick up regions having any particular secondary structure. In fact, most correctly predicted regions have conformation in the coil state.

Out of 135 predicted regions, 91 regions were correctly predicted giving an overall accuracy of around 67%. In the overpredicted 44 regions, almost every region has at least seven or eight residues that are conformationally similar but are spread over the whole region and, therefore, do not belong to the class of correctly predicted conformationally similar fragments. Change in the cut off value of average weight ($\bar{W}$) was tried in order to improve prediction accuracy. The noise level was reduced by about 25% when the condition ($\bar{W}$) $\geqslant 0{\cdot}5$ was introduced in addition to the present criterion. But then about 25% conformationally similar fragments were missed among the aligned regions. On the other hand, by the use of the present criterion not a single conformationally similar fragment, longer than seven amino acid residues, was missed from aligned segments. Therefore the present criterion is suggested for use in such studies.

To validate the results further, the alignment studies were carried out on the same set of 21 pairs of proteins using Mutation Data (MD) matrix (Dayhoff *et al.*, 1979). The aligned regions were

analysed to find out whether MD matrix can give conformationally similar regions. We found that only about 37% of the aligned regions by MD matrix were conformationally similar. Thus it is clear that the weight matrix that we have developed has an inherent property that allows us to pick up conformationally similar regions in proteins.

Thus, in short, a method to pick up conformationally similar protein fragments was developed and its accuracy was checked. It has been shown that this method can be applied to pick up conformationally similar regions through sequence comparison.

We acknowledge financial support from Department of Biotechnology, Govt. of India.

References

Argos, P. A. (1987). A sensitive procedure to compare amino acid sequences. *J. Mol. Biol.* **193**, 385–396.

Bairoch, A. (1990). *Prosite: A Dictionary of Protein Sites and Patterns*, 5th edit. Département de Biochimie Médicale, Université de Genève, Geneva.

Blundell, T. M., Lapatto, R., Wilderspin, A. F., Hemmings, A. M., Hobert, P. M., Danley, D. E. & Whittle, P. J. (1990). The 3-D structure of HIV-1 proteinase and design of antiviral agents for the treatment of AIDS. *Trends Biochem. Sci.* **15**, 425–430.

Brown, J. P., Hewick, R. M., Hellstrom, I., Hellstrom, K. E., Doolittle, R. F. & Dreyer, W. J. (1982). Human melanoma associated antigen p-97 is structurally and functionally related to transferrin. *Nature (London)*, **296**, 171–173.

Chou, P. Y. & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251–276.

Dixon, R. A. F., Kobilka, B. K., Strader, D. J., Benovic, J. L., Dohlman, H. J. & Frielle, T. (1986). Cloning, sequencing and expression of complementary DNA encoding the muscarinic acetylecholine receptor. *Nature (London)*, **323**, 411–416.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1979). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), **5** (supp. 3), 345–353.

Fermi, G., Perutz, M. F. & Shaanan, B. (1984). The crystal structure of human deoxyhaemoglobin at 1·74 Å resolution. *J. Mol. Biol.* **175**, 159–174.

Fuginaga, M. & James, M. N. G. (1987). Rat submaxillary gland serine protease, Tonin. Structure solution and refinement at 1·8 Å resolution. *J. Mol. Biol.* **195**, 373–396.

George, D. G., Barker, W. C. & Hunt, L. T. (1990). Mutation Data Matrix and its uses. *Methods Enzymol.* **183**, 333–351.

Gibrat, Garnier, J. & Robson, B. (1987). Future developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **198**, 425–443.

Gilliland, G. L. & Quiocho, F. A. (1981). Structure of the L-Arabinose binding protein from *Escherichia coli* at 2·4 Å resolution. *J. Mol. Biol.* **146**, 341–362.

Jones, T. A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.

Kabsch, W. (1978). Discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. sect. A*, **34**, 827–828.

Kabsch, W. & Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Nat. Acad. Sci., U.S.A.* **81**, 1075–1078.

Kolaskar A. S. & Ramabrahmam, V. (1981). Conformational similarity among amino acid residues: I. Analysis of crystal structure data. *Int. J. Biol. Macromol.* **3**, 171–178.

Kolaskar, A. S. & Ramabrahmam, V. (1984). Are secondary structures secondary? *Int. J. Pept. Protein Res.* **24**, 392–401.

Lim, V. I. (1974). Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* **88**, 873–894.

McLachlan, A. D. (1971). Tests for comparing related amino acid sequences. *J. Mol. Biol.* **61**, 409–424.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Parker, J. M. R., Guo, D. & Hodges, R. S. (1986). New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, **25**, 5425–5432.

Pastore, A. & Lesk, A. M. (1990). Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. *Proteins: Struct. Funct. Genet.* **8**, 133–155.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.

Ramakrishnan, C. & Srinivasan, N. (1990). Glycyl residues in proteins and peptides: an analysis. *Curr. Sci.* **59**, 851–861.

Risler, J. L., Delorme, M. O., Delacroix, H. & Henaat, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol.* **204**, 1019–1029.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B. & Nagano, K. (1974). Protein structure prediction. *Nature (London)*, **250**, 140–142.

Schwartz, R. M. & Dayhoff, M. O. (1979). Matrices for detecting distant relationships. *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), **5** (supp. 3), 353–358.

Wilbur, W. J. & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Nat. Acad. Sci., U.S.A.* **80**, 726–730.

Edited by W. A. Hendrickson