ELSEVIER

Biophysical Chemistry 103 (2003) 35–49

Biophysical Chemistry

www.elsevier.com/locate/bpc

Statistical analysis of pair-wise compatibility of spatially nearest neighbor and adjacent residues in α-helix and β-strands: application to a minimal model for secondary structure prediction

Srikanta Sen*

Human Genetics and Gemonics Division, Indian Institute of Chemical Biology, 4, Raja S. C. Mullick Road, Jadavpur, Calcutta 700032, India

Received 8 May 2002; received in revised form 26 June 2002; accepted 1 July 2002

Abstract

Secondary structural elements like α-helix and β-strands posses distinctly different structural features and thus the relative positioning of the nearest neighbor residues, and also the sequence-wise adjacent residues is important in determining the structural preference. In the present work we have statistically examined the pair-wise compatibility pattern of physically nearest neighbors and separately the adjacent residue pairs along the sequence in between the nearest neighbor partners in α-helices and β-strands. It has been demonstrated that the patterns and hence, the physical basis of the compatibility of adjacent residue pairs and the spatially nearest neighbors are significantly different in most cases. The influence of tertiary contacts on the pair-wise compatibility is shown to be significant for β-strands while it is small for α-helices. Based on the compatibility of physically nearest neighbors and the sequence-wise adjacent residue pairs, a minimal model has been constructed to predict the α-helices, β-strands and coils of a protein from its sequence. Application of this method to 100 sequences shows that it has a predictive capability comparable to that of other more sophisticated statistical methods.

Keywords: Characteristic structural features; Compatibility patterns; Local compatibility index; Effect of tertiary contacts

1. Introduction

Secondary structural elements like a α-helix and β-strand are quite common in proteins. Understanding the physical basis of forming secondary and tertiary structures is a major issue of modern molecular biology and biophysics. Over these years significant insights on the physical basis of structure determination have been achieved [1–6]. A number of novel methods for predicting the structure from a given sequence have also been developed. These include secondary structure predictions using statistical methods, fold recognition, sequence homology analysis, etc., and are successful to varying degrees [7–20].

In the present work we have analyzed the pair-wise compatibility based on the structural characteristics of α-helix and β-strand. Compatibility of

*Corresponding author. Chembiotek Research International, Block-BN, Plot-7, Sector-V, Salt Lake City, Calcutta 700091, India. Tel.: +91-33-367-3150/51/52; fax: +91-33-367-3058.
E-mail address: srikanta@chembiotek.com (S. Sen).

0301-4622/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0301-4622(02)00230-2

residue pairs for a given structure may be due to steric interaction, mutual H-bonding capability, hydrophobicity, etc. Such factors can be important only for residues physically close to each other. Again, different secondary structures have quite different nearest neighbor patterns with diverse relative positioning and orientations and different local backbone conformations. Thus, for a given sequence the steric characteristics and H-bonding capabilities differ considerably under different secondary structures and should influence the structural preference of the local sequence. In the present work we have performed a detailed statistical analysis on the pair-wise compatibility of physically nearest neighbors (in space) and adjacent residues (along the sequence) and have applied this knowledge in constructing a minimal model for predicting secondary structure from the sequence. Thus, our work consists of the following three major aspects that have not been adequately addressed before in connection to the specific secondary structural preferences of given sequences:

i. We have considered the pair-wise compatibility of a residue with only its spatially nearest neighbor residues characteristic of the specific secondary structure considered (see Section 2). Standard window methods consider the compatibility of the central residue with respect to each of the individual residues within the window.
ii. We have introduced the concept that the adjacent residue pairs along the sequence in between the central residue and its nearest neighbor residues should be adequately compatible to a specific secondary structure such that the central residue and the nearest neighbor can be properly placed to adopt the specific secondary structure (see Section 2). The statistical compatibility data on all possible kinds of sequence-wise adjacent residue pairs and spatially nearest neighbor residue pairs have been presented.
iii. We have used a minimal sequence window that is justified on the basis of the unique structural characteristics of the α-helices and β-strands (see method). The minimal window sizes are different for α-helices and β-strands.

In the context of secondary structure prediction it may be mentioned that over the past years a number of very useful approaches based on similar kind of nearest neighbor residue pairs, have been presented [21–28]. The conventional statistical methods of predicting the secondary structural status of the *I*th residue use statistical propensity data and all individual residues or residue pairs with respect to the central one in a sequence window $I-N$ to $I+N$, where N varies from 5 to 10 [21–28]. The same window size is used for all kinds of secondary structures. However, the characteristic structural properties of the secondary structures are not considered in these methods. In contrast to those works, here, we have considered explicitly the structural features that define a minimal window size for the respective secondary structure. In this work we have included the compatibility of the spatially nearest neighbor residue pairs only and the influences of the other residues within the minimal window have also been taken care of separately through the compatibility of the sequence-wise adjacent residue pairs within the window as pointed out above [(i)–(iii)]. Thus, our approach is significantly different from the previously presented nearest neighbor models.

In the present work we have reported new results, for example, it has been demonstrated here that the patterns and the physical basis of the pair-wise compatibility of physically nearest neighbors and the adjacent residue pairs in-between the nearest neighbor partners in α-helices and β-strands are significantly different from each other. The influence of tertiary contacts on the pair-wise compatibility is shown to be significant for β-strands while it is small for α-helices. A new and simple scheme has also been constructed to use these compatibility indices in predicting the α-helices and β-strands of any given sequence. Results indicate that this minimal model provides a prediction level close to that of more sophisticated schemes using statistical and neural network.

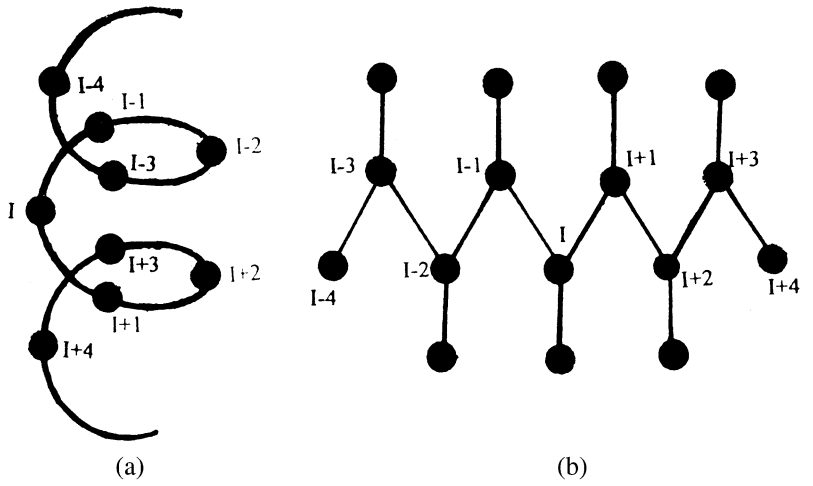

Fig. 1. Schematic diagram of the nearest-neighbor patterns for (a) α-helix and (b) β-strand structures in proteins. In the case of α-helix (a), the residues at positions $I-4$, $I-3$, $I+3$, $I+4$ are spatially nearest neighbor to the residue at position I while the residue pairs ($I-4$,$I-3$), ($I-3$,$I-2$), ($I-2$,$I-1$) and ($I-1$,I) are the adjacent residue pairs in between the nearest neighbor residue pairs at I and $I-4$. In the case of β-strand structures (b) the residues at $I-2$ and $I+2$ are spatially nearest to the residue at I, while the residue pairs ($I-2$,$I-1$) and ($I-1$,I) are the sequence-wise adjacent residues for the nearest neighbor pair (I–$I-2$).

2. Methods

2.1. Formulation of the basic scheme

In an α-helix the residues at position pairs I and J, where $J=I\pm3$ or $I\pm4$, are spatially nearest to each other, while in a β-strand, such nearest neighbors have $J=I\pm2$ (Fig. 1) [29]. Thus, for a given sequence, the nearest-neighbor patterns in terms of the relative positions of the residues on the primary sequence are quite different for different secondary structures. It is then most logical to expect that a sequence will preferably adopt a specific secondary structure that corresponds to the most compatible nearest-neighbor pair pattern compared to the other structures. Again, the sequence-wise adjacent residue pairs should allow the local backbone to adopt the correct torsional angles specific for the preferred structure. Based on this concept, the potential for α-helix of the residue at position I depends on two factors, (i) the physically compatibility of its spatially nearest neighbors to the residue at I for α-helix conformation, and (ii) the physically compatibility of the sequence-wise adjacent residue pairs (now on will be referred as adjacent residue pair only) in the position range ($I-4$ to $I+4$) for α-helix. Thus, only the residues within the range ($I-4$ to $I+4$) are important in directly determining the secondary structural preference of the residue at position I. Consideration of larger sequence window overestimates the effect of sequence context and a smaller one underestimates it. In this sense, the present method is 'minimal'. For a β-strand a minimal sequence window in the range ($I-2$ to $I+2$) is required (Fig. 1) [29].

The compatibility of the partners in a residue pair may require complementarity in different physical properties such as shape, size, interactions with the partner, etc., and hence, it only represents a measure of the overall complementarity of a pair of residues. The compatibility index $CI^{x}_{p,q}(I,J)$ of a residue 'p' at position I to be involved in a secondary structure X having a nearest-neighbor partner 'q' also in state X at position J is given by

$$CI^{x}_{p,q}(I,J)=n^{x}_{p,q}/n_{p,q} \quad (1)$$

$n_{p,q}$ is the number of cases where the residue 'p' (at position I) has a nearest-neighbor partner 'q' (at position J, characteristic of the specific second-

ary structure X) and $n^x_{p,q}$ represents the number of such p–q pairs where, both 'p' (at position I) and 'q' (at position J) are involved in the secondary structure X. Using a 'training' pool of 293 randomly selected proteins of known structure (a list is given in the Appendix A), we have computed the compatibility indices for each of the 400 ($=20\times20$) possible amino acid pairs for α-helix and β-strand separately. For each type of pair we have considered two different situations: (i) 'p' at position I and its partner 'q' at position $J1$ (upstream along the primary sequence); and (ii) in the other case, 'p' at I is followed by 'q' at $J2$ (downstream). The compatibility factors are represented by $CI^x_{p,q}(I,J1)$ and $CI^{xb}_{p,q}$ $(I,J2)$, respectively. In the case of a α-helix, $J1=I+3$ or $I+4$ and $J2=I-3$ or $I-4$ while for a β-strand, $J1=I+2$ and $J2=I-2$, respectively. Similarly, the compatibility indices for the adjacent residue pairs to be involved in a secondary structure X is given by

$$CIA^x_{p,q}(I,J)=m^x_{p,q}/m_{p,q} \tag{2}$$

$m_{p,q}$ is the total number of adjacent p–q pairs and $m^x_{p,q}$ is the number of p–q adjacent pairs where both 'p' and 'q' are involved in the secondary structure X. Additional four (two for α-helix and two for β-strand) have been prepared to take into account the pair-wise compatibility information of the adjacent residue pairs.

According to our scheme, the total local compatibility index $LCI^\alpha(I)$ that a particular residue at position I will be in a α-helix can be represented as,

$$\begin{aligned}LCI^\alpha(I)=&\,w^\alpha_{1b}CI^{\alpha b}_{p,q}(I,I-4)+w^\alpha_{2b}CI^{\alpha b}_{p,q}(I,I-3)\\&+w^\alpha_{1f}CI^{\alpha f}_{p,q}(I,I+3)+w^\alpha_{2f}CI^{\alpha f}_{p,q}(I,I+4)\end{aligned} \tag{3}$$

where each w^α_{xy} represents a weight factor that depends on the compatibility scores of the adjacent residues in the sequence in between the partners involved in a nearest neighbor pair. The weight factors for α-helix are given by

$$w^\alpha_{1b}=\sum\gamma^\alpha(i,j)/4,\quad j=i-1;\quad i\rightarrow\{I,I-3\}, \tag{4}$$

$$w^\alpha_{2b}=\sum\gamma^\alpha(i,j)/3,\quad j=i-1;\quad i\rightarrow\{I,I-2\}, \tag{5}$$

$$w^\alpha_{1f}=\sum\gamma^\alpha(i,j)/3,\quad j=i+1;\quad i\rightarrow\{I,I+2\}, \tag{6}$$

$$w^\alpha_{2f}=\sum\gamma^\alpha(i,j)/4,\quad j=i+1;\quad i\rightarrow\{I,I+3\}, \tag{7}$$

where $\gamma^\alpha(i,j)=1$ if the compatibility factor $CIA^\alpha(i,j)>CIA^\alpha_r$, and equals 0 otherwise.

In the above, $i\rightarrow\{I,I-3\}$ represents i that takes all the values in the range I to $(I-3)$, etc., and Σ implies summation over all values of i. CIA^α_r is the critical value of the $CIA^\alpha(i,j)$ score below which $\gamma^\alpha(i,j)=0$. Each of the weight factors can vary in the range {0,1} modulating the contribution of the associated nearest-neighbor pair accordingly. Thus, according to this scheme, in order to assume a helix state for the residue at I, not only must the nearest-neighbor partners must be significantly compatible for a helix, but also most of the intervening adjacent residue pairs must be reasonably compatible to a helix structure. Thus, the effects of the nearest-neighbors as well as the in between adjacent residue pairs are accounted for in our scheme. In the case of a β-strand we have,

$$\begin{aligned}LCI^\beta(I)=&\,w^\beta_{1b}CI^{\beta b}_{p,q}(I,I-2)\\&+w^\beta_{1f}CI^{\beta f}_{p,q}(I,I+2)\end{aligned} \tag{8}$$

The weight factors are given by

$$w^\beta_{1b}=\sum\gamma^\beta(i,j)/2,\quad j=i-1;\quad i\rightarrow\{I,I-1\}, \tag{9}$$

$$w^\beta_{1f}=\sum\gamma^\beta(i,j)/2,\quad j=i+1;\quad i\rightarrow\{I,I+1\}, \tag{10}$$

where $\gamma^\beta(i,j)=1$ if the adjacent compatibility factor $CIA^\beta(i,j)>CIA^\beta_r$,and equals 0 otherwise.

$CIA^\beta(i,j)$ and CIA^β_r are terms similar to $CIA^\alpha(i,j)$ and CIA^α_r for a helix, respectively. The preference to an α-helix or β-strand structures was made following the criteria in Section 2.2.

2.2. Choice of criteria

In an α-helix, each residue at the position I has four nearest-neighbors (at positions $I-4$, $I-3$, $I+3$ and $I+4$). So the total average local compatibility score $LCI^\alpha(I)$ for the residue at I is $4\times CI^\alpha_{\rm ave}$ where $CI^\alpha_{\rm ave}$ is the average of all 400 CI^α values. Each of these four pairwise compatibility indices is further modulated by the weight factors related to the compatibility of the adjacent residue pairs. So, as a working criterion, we have chosen 50% of this value to be the reference value LCI^α_r for

helix assignment to a residue, i.e. $LCI_r^{\alpha} = 2 \times CI_{ave}^{\alpha}$. Thus, if a residue has a $LCI^{\alpha}(I)$ score greater than LCI_r^{α}, the residue will be primarily assigned to a helix state. Similarly, the average compatibility score for the adjacent residues (CIA_{ave}^{α}) is chosen as the critical compatibility value CIA_r^{α} (because it occurs only once in the calculation) for the adjacent residue pairs. In the case of β-strand analysis, the average values are CI_{ave}^{β} and CIA_{ave}^{β} and following the same arguments, the critical values are chosen as $LCI_r^{\beta} = CI_{ave}^{\beta}$ and $CIA_r^{\beta} = CIA_{ave}^{\beta}$.

For residues that are predicted to have considerable compatibility for both the α-helix and β-strand [i.e. $LCI^{\alpha}(I) \geq LCI_r^{\alpha}$ and $LCI^{\beta}(I) \geq LCI^{\beta}$), we have used the following criterion to decide which one will be favored over the other to take into account the difference in the number of nearest neighbors pairs in these two cases. By definitions, if a residue has $LCI^{\alpha}(I) = LCI_r^{\alpha}$ and $LCI^{\beta}(I) = LCI_r^{\beta}$, then it physically means that the residue has an equal possibility for adopting an α-helix or a β-strand status. Hence, these two scores are equivalent. Thus for direct comparison, we can calibrate the calculated $LCI^{\alpha}(I)$ and $LCI^{\beta}(I)$ scores for the same residue by the relation

$$LCI_e^{\beta}(I) = LCI^{\alpha}(I) \times (LCI_r^{\beta}/LCI_r^{\alpha}) \quad (11)$$

where $LCI_e^{\beta}(I)$ is the equivalent $LCI^{\beta}(I)$ score corresponding to the $LCI^{\alpha}(I)$ score of a residue. So, if a residue has a $LCI_e^{\beta}(I) < LCI^{\beta}(I)$, the predicted compatibility of the residue for a β-strand, then the residue will prefer an α-helix state and vice versa.

2.3. Secondary structure prediction procedure

For each residue in a given sequence, we first calculate the LCI^{α} value using Eqs. (3)–(7), and then α-helix status is assigned to the residue following our criterion. Then, we calculate the LCI^{β} values using Eqs. (3)–(7), and separately assign the β-strand status to each residue satisfying our criterion for β-strand status. We then apply the equivalence criterion (Eq. (11)) to select the most favored statuses of the residues that are predicted (in separate α and β assignment steps) to have significant potential to adopt both α-helix and β-strand status. In this way we get a grand assigned sequence. Finally, the fractions of correctly predicted three types of structures α-helix, β-strand and non-α-non-β structures in the given sequences is calculated according to the definition $f_g = (n_{\alpha}^c + n_{\beta}^c + n_{n\alpha\beta}^c)/N$, where, n_x^c is the number of residues correctly assigned to status X and N is the total number of residues in the sequence. The parameter X can be α, β and nαβ (non-α, non-β) [8]. The predictions have been tested on 100 different 'test' sequences (not in the 'training' set) to assess the generality of the procedure.

Here, it may be pointed out that the secondary structural feature assigned to the individual residues in the crystal structure (sequence details menu for the structure at http://www.rcsb.org.pdb/) was considered as the experimentally obtained correct secondary structure assignments. In the present work, we have considered only three state descriptions, i.e. α-helices, β-strands and coils (non-α-helices non-β-strands) and we have not considered the helix ends separately.

3. Results

3.1. Analysis of the compatibility properties of spatially nearest neighbor pairs and adjacent residue pairs

The compatibility values are given in Tables 1–4. The average CI values over all the possible 400 nearest neighbor pairs for α-helix was obtained as $CI_{ave}^{\alpha} = 0.20$ and for the adjacent pairs it is $CIA_{ave}^{\alpha} = 0.26$, while for β-strand structures these are, respectively, $CI_{ave}^{\beta} = 0.13$ and $CIA_{ave}^{\beta} = 0.18$. So, according to our choice of criteria, the different reference values are considered as $LCI_r^{\alpha} = 0.40$, $LCI_r^{\beta} = 0.13$, $CIA_r^{\alpha} = 0.26$ and $CIA_r^{\beta} = 0.18$.

The linear correlation coefficient between two similar data series can be calculated by the relation

$$C_{xy} = \langle (x^i - x_{ave}^i)(y^i - y_{ave}^i) \rangle / \{ \langle (x^i - x_{ave}^i)^2 \rangle \langle (y^i - y_{ave}^i)^2 \rangle \}^{1/2} \quad (12)$$

where x^i and y^i are the ith elements in the two data series, respectively [30]. In the present work, the correlation coefficient between CI^{α} (backward) and CI^{β} (backward) is obtained as 0.26, and between CIA^{α} (backward) and CIA^{β} (backward)

Table 1
The pair-wise compatibility indices (*CI* values) of all the nearest neighbor pairs in α-helical structures (downstream) computed from a large ensemble of protein sequence and structure

	G	A	V	L	I	P	F	Y	W	C	M	S	T	K	R	H	D	E	N	Q
G	0.05	0.12	0.06	0.07	0.06	0.04	0.07	0.06	0.06	0.05	0.09	0.05	0.08	0.09	0.09	0.07	0.10	0.08	0.04	0.08
A	0.19	0.44	0.31	0.39	0.35	0.22	0.34	0.26	0.34	0.34	0.40	0.24	0.21	0.32	0.37	0.28	0.31	0.45	0.22	0.41
V	0.10	0.31	0.21	0.28	0.24	0.09	0.24	0.22	0.30	0.21	0.21	0.10	0.15	0.17	0.16	0.17	0.16	0.23	0.10	0.20
L	0.16	0.42	0.35	0.45	0.42	0.12	0.39	0.32	0.36	0.23	0.43	0.21	0.20	0.27	0.33	0.21	0.21	0.34	0.19	0.35
I	0.12	0.36	0.28	0.37	0.38	0.12	0.31	0.25	0.38	0.24	0.35	0.18	0.17	0.16	0.27	0.22	0.14	0.27	0.17	0.23
P	0.01	0.07	0.02	0.05	0.03	0.02	0.09	0.07	0.03	0.01	0.05	0.03	0.03	0.09	0.05	0.04	0.04	0.09	0.04	0.04
F	0.11	0.30	0.20	0.33	0.36	0.07	0.18	0.22	0.25	0.25	0.25	0.13	0.11	0.19	0.19	0.14	0.13	0.24	0.14	0.23
Y	0.08	0.27	0.17	0.25	0.31	0.13	0.25	0.17	0.18	0.17	0.23	0.11	0.11	0.14	0.15	0.12	0.12	0.19	0.10	0.30
W	0.09	0.38	0.28	0.34	0.20	0.17	0.24	0.20	0.18	0.19	0.16	0.08	0.16	0.17	0.35	0.18	0.17	0.31	0.07	0.23
C	0.07	0.26	0.20	0.23	0.16	0.09	0.22	0.15	0.08	0.12	0.17	0.15	0.09	0.15	0.20	0.17	0.18	0.26	0.11	0.15
M	0.12	0.48	0.33	0.43	0.35	0.10	0.44	0.39	0.25	0.27	0.41	0.16	0.21	0.25	0.24	0.28	0.20	0.36	0.22	0.52
S	0.07	0.19	0.14	0.18	0.11	0.13	0.14	0.08	0.20	0.06	0.22	0.10	0.08	0.17	0.17	0.15	0.13	0.25	0.13	0.19
T	0.04	0.23	0.14	0.17	0.17	0.08	0.13	0.11	0.16	0.14	0.21	0.12	0.12	0.13	0.20	v15	0.12	0.18	0.08	0.27
K	0.17	0.40	0.20	0.24	0.23	0.13	0.21	0.19	0.22	0.20	0.21	0.21	0.20	0.20	0.33	0.17	0.29	0.46	0.23	0.36
R	0.14	0.41	0.24	0.32	0.27	0.17	0.23	0.24	0.23	0.24	0.26	0.25	0.21	0.28	0.30	0.20	0.28	0.46	0.19	0.29
H	0.05	0.35	0.22	0.26	0.17	0.11	0.16	0.14	0.15	0.16	0.26	0.12	0.21	0.11	0.22	0.20	0.20	0.28	0.06	0.19
D	0.08	0.21	0.07	0.12	0.10	0.09	0.12	0.07	0.13	0.14	0.17	0.10	0.08	0.26	0.15	0.14	0.13	0.19	0.10	0.30
E	0.10	0.39	0.21	0.26	0.23	0.12	0.16	0.17	0.21	0.23	0.24	0.20	0.19	0.36	0.35	0.20	0.27	0.30	0.20	0.35
N	0.07	0.21	0.13	0.12	0.15	0.06	0.12	0.11	0.19	0.12	0.09	0.12	0.16	0.15	0.18	0.11	0.15	0.26	0.11	0.30
Q	0.17	0.35	0.23	0.30	0.28	0.22	0.23	0.33	0.27	0.14	0.33	0.19	0.21	0.35	0.33	0.18	0.23	0.37	0.19	0.43

The element at the *I*th row and *J*th column represents the compatibility of the residue at the *I*th row in the first column and the residue at the *J*th column as its nearest neighbor. The same pattern is followed for each of the tables presented in this paper.

data sets it is 0.06. Thus it is indicated that the relative patterns are quite different for the α-helices and β-strands and the differences are more pronounced in the case of adjacent residue pairs. However, the C_{xy} value between the backward and forward data set for an α-helix is approximately 0.81 while it is 0.79 for β-strand.

Analysis of the Tables 1–4 provides new data on the compatibility of spatially nearest neighbor pairs and adjacent residue pairs, and reproduces

Table 2
The pair-wise compatibility indices (*CI* values) of all the nearest neighbor pairs in β-strand structures (downstream) computed from a large ensemble of protein sequence and structure

	G	A	V	L	I	P	F	Y	W	C	M	S	T	K	R	H	D	E	N	Q
G	0.06	0.08	0.13	0.16	0.18	0.02	0.12	0.10	0.21	0.15	0.13	0.04	0.09	0.04	0.06	0.11	0.02	0.03	0.01	0.05
A	0.10	0.06	0.19	0.12	0.18	0.03	0.22	0.13	0.16	0.23	0.03	0.07	0.09	0.07	0.06	0.10	0.02	0.04	0.05	0.08
V	0.17	0.17	0.52	0.34	0.43	0.09	0.45	0.34	0.29	0.36	0.28	0.16	0.29	0.14	0.13	0.21	0.05	0.12	0.09	0.12
L	0.10	0.13	0.34	0.28	0.34	0.07	0.26	0.30	0.29	0.14	0.22	0.09	0.14	0.07	0.08	0.10	0.04	0.04	0.06	0.09
I	0.13	0.20	0.46	0.39	0.47	0.08	0.41	0.39	0.34	0.33	0.30	0.13	0.24	0.16	0.14	0.14	0.04	0.07	0.14	0.08
P	0.01	0.02	0.11	0.03	0.13	0.02	0.10	0.03	0.14	0.03	0.07	0.03	0.08	0.03	0.04	0.03	0.01	0.05	0.01	0.05
F	0.10	0.18	0.50	0.23	0.36	0.07	0.42	0.29	0.33	0.25	0.32	0.14	0.21	0.13	0.17	0.17	0.06	0.06	0.09	0.17
Y	0.18	0.15	0.24	0.21	0.25	0.06	0.32	0.31	0.34	0.20	0.21	0.17	0.30	0.16	0.20	0.27	0.08	0.06	0.09	0.19
W	0.10	0.18	0.32	0.35	0.42	0.09	0.53	0.24	0.00	0.13	0.26	0.06	0.11	0.08	0.20	0.05	0.03	0.09	0.05	0.23
C	0.10	0.14	0.24	0.37	0.41	0.02	0.43	0.49	0.00	0.06	0.56	0.11	0.06	0.08	0.05	0.09	0.00	0.05	0.00	0.04
M	0.08	0.15	0.26	0.20	0.35	0.04	0.26	0.15	0.22	0.14	0.25	0.11	0.10	0.09	0.08	0.31	0.00	0.08	0.04	0.15
S	0.08	0.08	0.19	0.13	0.16	0.01	0.11	0.11	0.10	0.14	0.14	0.21	0.20	0.10	0.10	0.08	0.04	0.08	0.07	0.13
T	0.06	0.10	0.27	0.13	0.23	0.04	0.24	0.23	0.15	0.24	0.14	0.17	0.23	0.13	0.08	0.16	0.03	0.16	0.07	0.11
K	0.05	0.05	0.17	0.10	0.07	0.01	0.13	0.14	0.06	0.15	0.06	0.09	0.18	0.05	0.10	0.06	0.05	0.09	0.05	0.02
R	0.09	0.04	0.16	0.10	0.10	0.05	0.17	0.14	0.14	0.17	0.06	0.08	0.21	0.08	0.15	0.08	0.03	0.09	0.03	0.15
H	0.05	0.06	0.27	0.12	0.12	0.04	0.15	0.14	0.24	0.17	0.03	0.05	0.20	0.05	0.09	0.29	0.06	0.17	0.03	0.26
D	0.07	0.05	0.11	0.06	0.13	0.01	0.06	0.08	0.07	0.02	0.07	0.11	0.10	0.06	0.14	0.11	0.02	0.06	0.01	0.02
E	0.05	0.05	0.17	0.05	0.09	0.02	0.10	0.10	0.11	0.15	0.11	0.09	0.15	0.11	0.10	0.05	0.04	0.08	0.11	0.08
N	0.04	0.06	0.13	0.11	0.09	0.00	0.10	0.08	0.12	0.05	0.05	0.09	0.09	0.05	0.13	0.16	0.02	0.06	0.09	0.07
Q	0.06	0.04	0.16	0.12	0.10	0.00	0.11	0.19	0.17	0.19	0.06	0.16	0.13	0.08	0.01	0.02	0.07	0.07	0.04	0.07

Table 3
The pair-wise compatibility indices for the adjacent residue pairs (downstream) for α-helix structure

	G	A	V	L	I	P	F	Y	W	C	M	S	T	K	R	H	D	E	N	Q
G	0.06	0.17	0.11	0.20	0.11	0.06	0.15	0.14	0.15	0.06	0.10	0.07	0.08	0.08	0.11	0.21	0.09	0.14	0.08	0.15
A	0.17	0.55	0.38	0.54	0.49	0.21	0.35	0.45	0.47	0.27	0.50	0.25	0.28	0.44	0.50	0.34	0.34	0.54	0.26	0.53
V	0.14	0.38	0.19	0.28	0.22	0.12	0.24	0.18	0.16	0.27	0.35	0.18	0.15	0.28	0.24	0.30	0.25	0.28	0.19	0.30
L	0.16	0.54	0.36	0.43	0.40	0.19	0.48	0.44	0.40	0.28	0.49	0.22	0.25	0.33	0.37	0.40	0.30	0.47	0.20	0.50
I	0.16	0.51	0.20	0.34	0.39	0.14	0.28	0.25	0.33	0.30	0.38	0.24	0.21	0.26	0.30	0.15	0.31	0.36	0.18	0.43
P	0.08	0.08	0.06	0.07	0.06	0.02	0.05	0.09	0.05	0.05	0.17	0.02	0.06	0.08	0.06	0.05	0.04	0.07	0.02	0.07
F	0.12	0.46	0.19	0.33	0.20	0.12	0.30	0.24	0.46	0.31	0.33	0.18	0.17	0.36	0.25	0.17	0.22	0.40	0.20	0.22
Y	0.07	0.36	0.20	0.23	0.26	0.16	0.28	0.27	0.24	0.21	0.26	0.19	0.16	0.31	0.32	0.14	0.21	0.34	0.12	0.32
W	0.03	0.35	0.35	0.43	0.25	0.15	0.38	0.13	0.43	0.22	0.33	0.28	0.14	0.33	0.37	0.14	0.20	0.47	0.22	0.45
C	0.16	0.39	0.19	0.28	0.35	0.07	0.22	0.13	0.05	0.44	0.42	0.16	0.09	0.22	0.32	0.16	0.19	0.26	0.29	0.28
M	0.15	0.58	0.33	0.50	0.34	0.11	0.27	0.28	0.47	0.23	0.31	0.37	0.23	0.42	0.47	0.34	0.26	0.50	0.31	0.39
S	0.06	0.27	0.20	0.24	0.25	0.07	0.20	0.14	0.24	0.09	0.18	0.10	0.15	0.26	0.26	0.26	0.19	0.38	0.14	0.26
T	0.10	0.26	0.15	0.25	0.19	0.06	0.21	0.09	0.18	0.16	0.34	0.12	0.14	0.23	0.26	0.16	0.14	0.32	0.14	0.34
K	0.11	0.53	0.29	0.38	0.34	0.08	0.31	0.25	0.37	0.33	0.39	0.21	0.18	0.42	0.37	0.23	0.23	0.38	0.23	0.41
R	0.10	0.44	0.36	0.48	0.39	0.11	0.36	0.28	0.40	0.20	0.35	0.21	0.25	0.35	0.38	0.25	0.28	0.44	0.28	0.43
H	0.18	0.27	0.16	0.41	0.35	0.14	0.09	0.15	0.48	0.14	0.32	0.23	0.20	0.20	0.25	0.19	0.21	0.30	0.15	0.35
D	0.08	0.26	0.22	0.31	0.23	0.12	0.18	0.18	0.22	0.22	0.30	0.20	0.16	0.26	0.34	0.17	0.20	0.26	0.19	0.37
E	0.16	0.53	0.33	0.44	0.44	0.28	0.30	0.26	0.39	0.38	0.43	0.29	0.26	0.43	0.46	0.36	0.37	0.42	0.23	0.42
N	0.09	0.29	0.21	0.27	0.25	0.08	0.24	0.15	0.05	0.18	0.24	0.08	0.11	0.21	0.33	0.23	0.13	0.26	0.10	0.15
Q	0.17	0.49	0.33	0.46	0.44	0.14	0.43	0.25	0.51	0.16	0.38	0.24	0.30	0.41	0.35	0.26	0.27	0.56	0.28	0.49

most of the known data validating our minimal model. Fig. 2a represents the average CI^{α} value ($CI^{\alpha}_{\text{mean}}$) of each residue type when its nearest neighbor partner is varied over all possible ways. It is found that each of the residues A, L, I, M, Q, R, K and E has a $CI^{\alpha}_{\text{mean}}$ value above the CI^{α}_{ave} value and thus has potential to form an α-helix irrespective of its nearest neighbor partner. Similarly, the residues V, I, F, Y, L, M and T are potentially compatible for adopting a β-strand structure (Fig. 2b). These average trends of preference are in general agreement with other studies [31–34]. It may also be noticed that the CI^{α} values for the residues like A and L, are always large

Table 4
The pair-wise compatibility indices for the adjacent residue pairs (down stream) for β-strand structure

	G	A	V	L	I	P	F	Y	W	C	M	S	T	K	R	H	D	E	N	Q
G	0.07	0.07	0.23	0.09	0.26	0.02	0.18	0.12	0.09	0.17	0.12	0.10	0.09	0.08	0.03	0.06	0.03	0.03	0.02	0.07
A	0.10	0.07	0.24	0.12	0.21	0.01	0.27	0.17	0.14	0.25	0.12	0.15	0.18	0.13	0.13	0.09	0.05	0.06	0.06	0.07
V	0.16	0.29	0.48	0.40	0.46	0.15	0.43	0.40	0.43	0.29	0.37	0.32	0.42	0.31	0.36	0.30	0.14	0.28	0.22	0.37
L	0.13	0.14	0.36	0.20	0.35	0.09	0.27	0.28	0.19	0.39	0.17	0.20	0.29	0.24	0.25	0.24	0.07	0.15	0.10	0.16
I	0.20	0.22	0.52	0.36	0.38	0.10	0.49	0.40	0.37	0.24	0.31	0.27	0.47	0.26	0.37	0.43	0.14	0.32	0.16	0.12
P	0.05	0.11	0.15	0.09	0.09	0.01	0.04	0.12	0.13	0.02	0.12	0.03	0.02	0.06	0.08	0.02	0.02	0.05	0.04	0.08
F	0.18	0.15	0.41	0.29	0.43	0.06	0.29	0.37	0.17	0.47	0.35	0.32	0.44	0.20	0.22	0.37	0.09	0.23	0.15	0.46
Y	0.17	0.17	0.41	0.40	0.44	0.04	0.38	0.37	0.45	0.21	0.29	0.17	0.39	0.10	0.23	0.21	0.08	0.18	0.20	0.19
W	0.30	0.14	0.39	0.22	0.25	0.03	0.27	0.39	0.10	0.67	0.20	0.23	0.34	0.19	0.18	0.43	0.04	0.20	0.25	0.14
C	0.15	0.04	0.38	0.28	0.29	0.05	0.46	0.35	0.37	0.22	0.33	0.32	0.38	0.20	0.12	0.14	0.13	0.11	0.10	0.28
M	0.08	0.15	0.41	0.20	0.31	0.11	0.29	0.16	0.16	0.23	0.31	0.19	0.25	0.18	0.18	0.07	0.04	0.14	0.10	0.26
S	0.11	0.12	0.30	0.20	0.31	0.05	0.28	0.19	0.22	0.28	0.15	0.11	0.15	0.13	0.14	0.11	0.03	0.07	0.05	0.09
T	0.11	0.17	0.44	0.24	0.30	0.01	0.30	0.39	0.42	0.22	0.23	0.19	0.14	0.12	0.09	0.18	0.08	0.11	0.09	0.11
K	0.05	0.05	0.24	0.15	0.22	0.07	0.18	0.13	0.16	0.11	0.10	0.06	0.16	0.05	0.09	0.07	0.03	0.06	0.03	0.09
R	0.06	0.13	0.33	0.13	0.21	0.04	0.19	0.17	0.21	0.26	0.22	0.11	0.18	0.09	0.13	0.11	0.01	0.10	0.08	0.04
H	0.07	0.21	0.38	0.16	0.33	0.04	0.55	0.30	0.19	0.08	0.20	0.07	0.18	0.08	0.14	0.11	0.03	0.08	0.04	0.06
D	0.02	0.08	0.18	0.09	0.23	0.00	0.21	0.20	0.09	0.13	0.06	0.03	0.06	0.03	0.04	0.10	0.01	0.05	0.02	0.04
E	0.07	0.07	0.27	0.15	0.22	0.01	0.33	0.24	0.29	0.28	0.13	0.06	0.13	0.04	0.05	0.05	0.02	0.04	0.01	0.11
N	0.04	0.09	0.25	0.13	0.24	0.03	0.16	0.17	0.18	0.16	0.09	0.02	0.08	0.05	0.07	0.11	0.02	0.07	0.01	0.07
Q	0.06	0.06	0.18	0.18	0.19	0.01	0.19	0.27	0.13	0.22	0.08	0.06	0.17	0.10	0.14	0.23	0.01	0.07	0.05	0.13

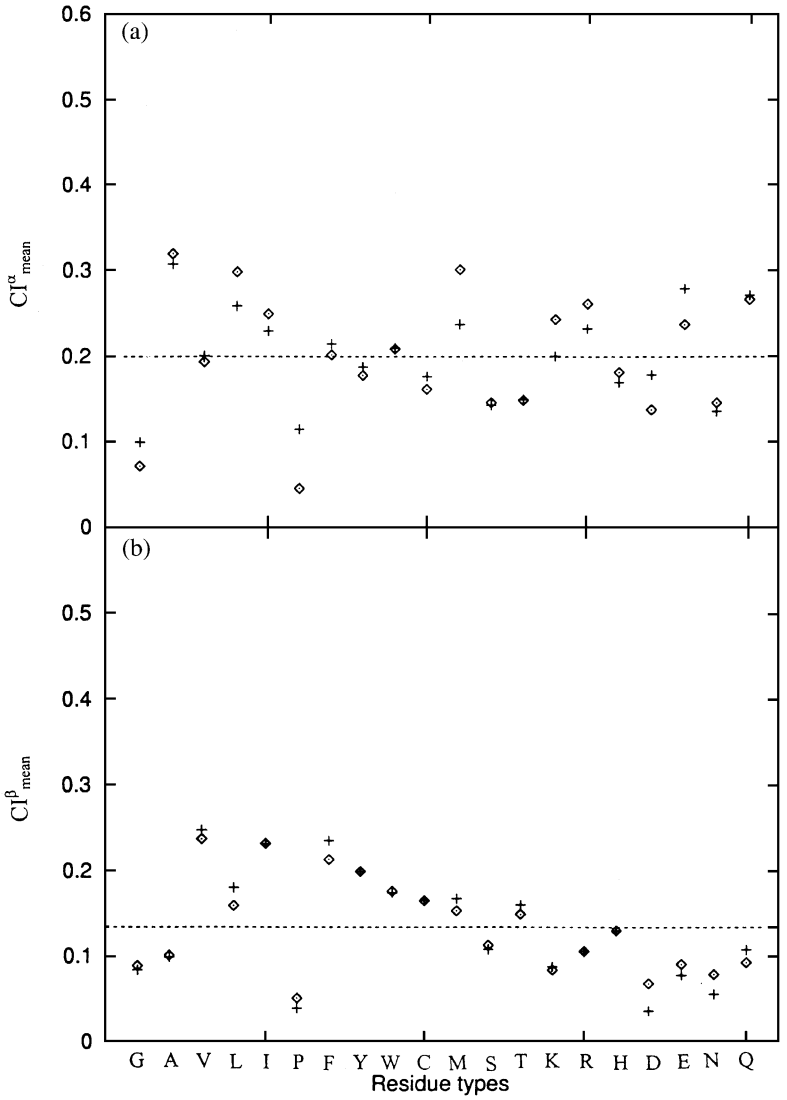

Fig. 2. Scatter plots showing (a) the relative CI^{α}_{mean} values for each residue type averaged over all possible nearest-neighbor partners for α-helix, and (b) CI^{β}_{mean} values for individual residue types averaged over all possible nearest-neighbor partners in the case of β-strand structures. The symbols ◇ and + represents the cases of analysis along downstream and upstream, respectively. The dashed lines represent the cut-off values in respective cases.

irrespective of the nearest neighbor partners indicating their intrinsic nature of structural preference. Similarly, the residues G and P are intrinsically incompatible for an α-helix. The residues V and I have intrinsic preferences for a β-strand. Such an intrinsic preference may reflect the perfect compatibility of the side chains with the backbone of the residues. There are works, pointing out such possibilities before [31–34]. Fig. 3a,b represent the average *CIA* value (CIA^{α}_{mean} and CIA^{β}_{mean}) of each residue type when its adjacent partner is varied over all possible ways in the cases of an α-

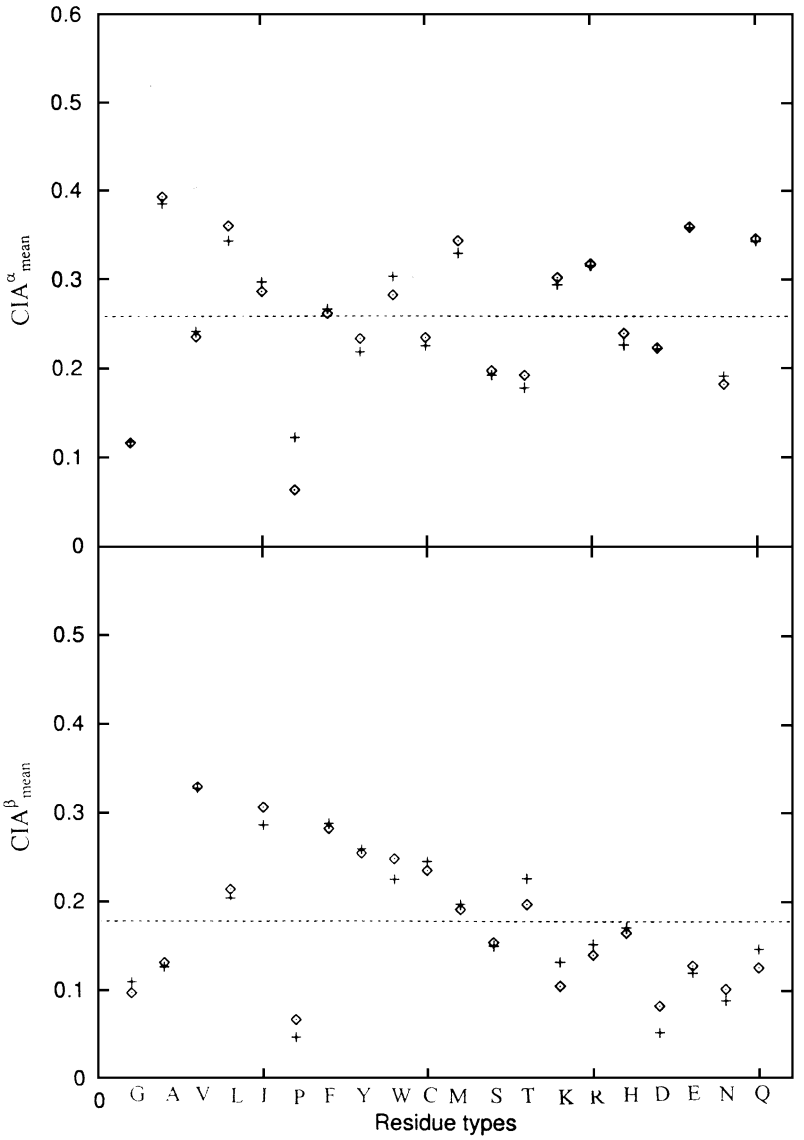

Fig. 3. Scatter plots showing (a) the relative CIA^{α}_{mean} values for each residue type averaged over all possible nearest-neighbor partners for α-helix, and (b) CIA^{β}_{mean} values for individual residue types averaged over all possible nearest-neighbor partners in the case of β-strand structures. The symbols ◇ and + represents the cases of analysis along downstream and upstream, respectively. The dashed lines represent the cut-off values in respective cases.

helix and β-strands, respectively. It is found that on average, the residues A, L, I, M, Q, R, K and E are found to have high CIA^{α}_{mean} values for an α-helix, while for a β-strand the corresponding residues are V, I, F, Y, W, M and T. It may further be pointed out that in Fig. 2a, even though for residues H, D and N CI^{α}_{mean} are lower than the CI^{α}_{ave} value, there are individual cases of nearest neighbor pairs involving any of these residues as a partner, has quite high value making the pair

compatible for an α-helix. Thus, for specific pairs the compatibility may be quite different from the average trend implying the importance of considering the compatibility at the level of residue pairs instead of individual residues.

3.2. The influence of tertiary contacts and intra-protein environment on compatibility features

The influence of tertiary contacts and intra-protein environment on the pairwise compatibility has not yet been addressed. In order to gain some insight in this issue, we have prepared two separate and independent training sequence sets. One set (S) contains sequences of monomolecular proteins with residue numbers (N) ≤ 200 and the other set (L) contains large sequences with $N \geq 600$. The ensembles S and L are of similar sizes having almost equal number of residues (approx. 9000). Using each of these ensembles, the compatibility indices (CI^{α} or CI^{β}) for an α-helix and β-strand have been computed. It is assumed that proteins with small N, have less possibility of tertiary contacts and hence, the influence of tertiary effects will be less compared to that for proteins with large N where a larger number of tertiary contacts are expected. As a control, we have prepared two general ensembles $G1$ and $G2$ of similar size as L (and S) and contain sequences of any size. To remove any artifact arising from inadequate sampling, we prepared a reduced data sets considering only those cases of pairs in CI^{α} (back) which were sampled larger than N_s times where N_s is the average number of occurrence if each pair was sampled equally. With these two reduced CI^{α} (back) data sets obtained from $G1$ and $G2$, correlation coefficient was computed as 0.85 as expected. Similarly, the reduced CI^{α} (back) data sets obtained from ensembles S and L yields a correlation coefficient of 0.83 indicating that the patterns of CI^{α} (back) sets are same as in $G1$ and $G2$, for α-helix. On the other hand, in the case of the β-strand structures, the correlation coefficients are found to be 0.56 (between S and L) and 0.79 (between $G1$ and $G2$) indicating that the patterns in the case of β-strands are considerably different in S and L ensembles. Thus, it demonstrates in a reasonable way that the influence of tertiary contacts on the pair-wise compatibility is significant for β-strands while, the influence is small for α-helices. Interestingly, the general trend of better accuracy in α-helix predictions compared to the accuracy for β-strands in statistical methods is in accord with this effect. Considering the complexity of the issue, the present demonstration appears significant.

3.3. Location-dependent local compatibility scores (LCI) of residues

Each of the 20 amino acid residues can appear in a protein sequence having a variety of nearest-neighbor partners as well as different adjacent local residue pairs. As a result, depending on the local sequence, the overall local compatibility index [$LCI^{\alpha}(I)$ or $LCI^{\beta}(I)$] of a particular type of residue can vary over a wide range as demonstrated in the Fig. 4a,b for an α-helix and β-strand, respectively. It is clearly seen that for some specific residues (G and P) the LCI^{α} and LCI^{β} scores are mostly below LCI^{α}_{r} and LCI^{β}_{r}, respectively, indicating again that these residues are intrinsically not compatible for any secondary structures. Moreover, it is seen that the LCI^{α} (or LCI^{β}) values for the residues A, V, L, K, R and E, vary over a wider range than the others do and thus their structural preferences are most sensitive to the local sequence.

Fig. 5 compares the relative preferences of different residue types towards α-helix and β-strand structures, averaged over a large number of different local sequences. It is evident from the plot that each of the residues Q, A, M, L, E, K, R, H and W has average equivalent $\overline{LCI}^{\beta}_{e}$ (corresponding to $\overline{LCI}^{\alpha}_{\mathrm{ave}}$) value above $\overline{LCI}^{\beta}_{r}$ and thus has potential to form an α-helix. Similarly, each of the residues V, I, Y and F is potentially capable of forming β-strand. However, since $LCI^{\beta}_{\mathrm{ave}}$ and $\overline{LCI}^{\beta}_{e}$ (corresponding to $LCI^{\alpha}_{\mathrm{ave}}$) are directly comparable, it is also seen that the residue Q, A, M, L, E, K, R, H and W are potentially more suitable for an α-helix and V, I, Y and F favors a β-strand on the average. It is also noticeable that most of the charged residues are highly compatible for an α-helix and do not support β-strand formation in general.

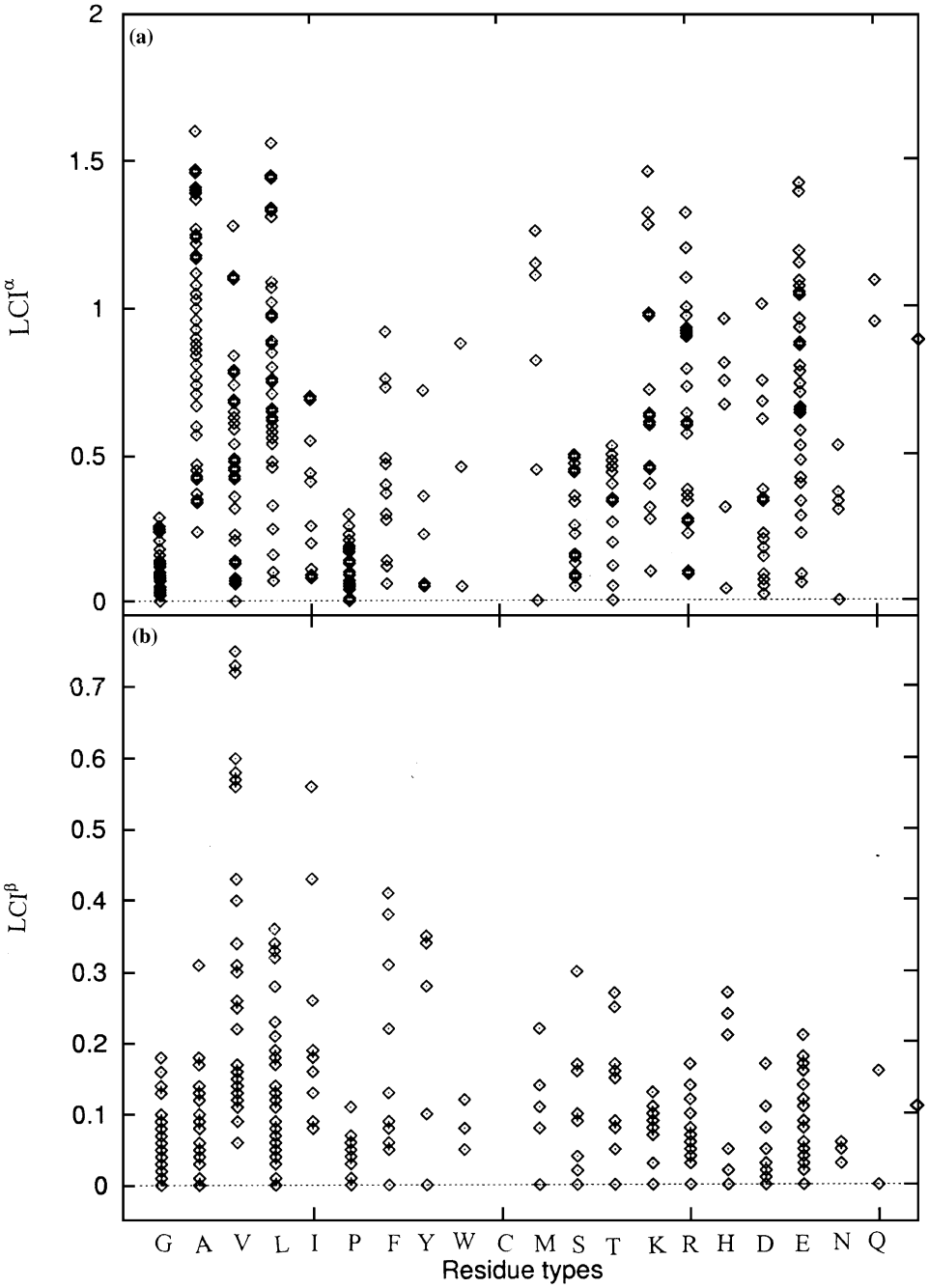

Fig. 4. Scatter plots of (a) LCI^{α} values and (b) LCI^{β} values, of different kinds of residues under different local sequences of a 345 residues protein (pdb-entry 1osj:a) as a typical case.

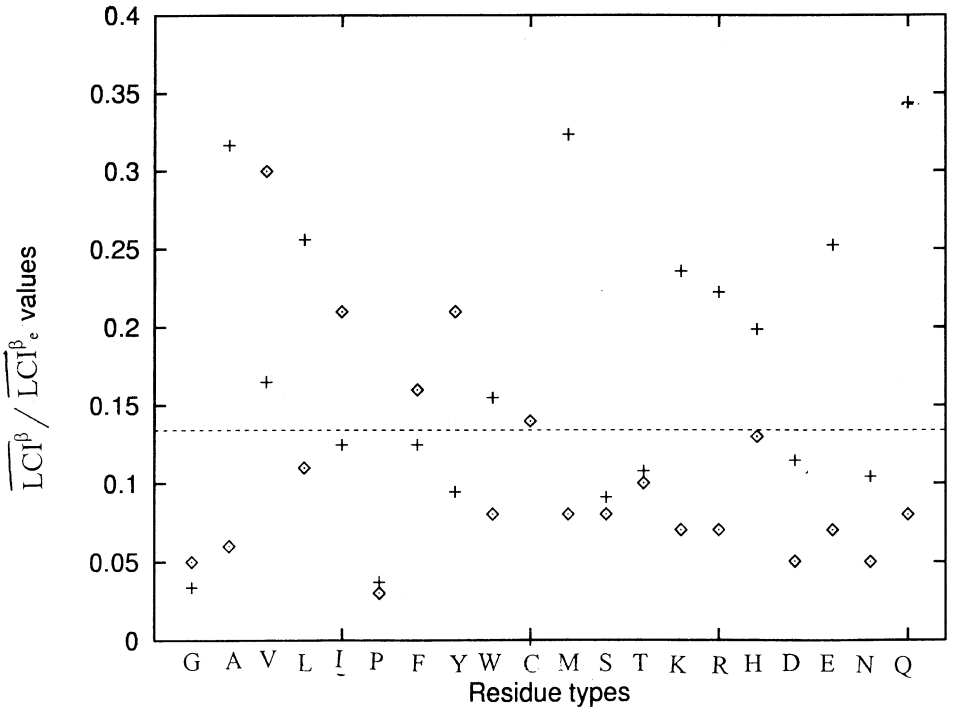

Fig. 5. Comparison of the relative average preference of the different residue types towards α-helix and β-strand structures. The symbol ◇ represents the LCI^{β} values averaged over different local sequences and + represents the equivalent LCI^{β}_{e} values corresponding to the LCI^{α} values for different residue types. Thus, these data sets are directly comparable. The dashed line represents the cut-off value.

3.4. Predictions of α-helices and β-strands in a sequence

In order to validate our method, we have predicted the secondary structures of 100 'test' sequences (not included in the 'training' set) and compared with their crystal structures. Some of the sequences are taken from other published test sequences [35] and some others are randomly selected. The *fg* scores for all of the 100 test sequences (a table showing these results can be obtained on request) indicate that such a minimal statistical method can provide an average prediction level of 59.6% (with a RMS fluctuation of ±6.2%) that is comparable to that of methods based on more sophisticated statistical schemes or neural network [11,12,35,36]. Fig. 6 represents the plot of *fg* values against the number of residues in each test sequence. An interesting observation is that replacing one or two odd assignments (say E) in a long stretch of an assigned specific secondary structure (say H) by the kind of the local trend (i.e. H in this example) increases the overall accuracy further to 63.3% (±6.6% r.m.s. fluctuation) and up to 15% in the cases of individual protein (see Fig. 6). This criterion is logical because in a long stretch of a structure a few residues with lower compatibility may be accommodated. It is clearly seen that in most of the cases the prediction is above 60% even without this correction. In individual case the highest degree of prediction accuracy for the selected test sequences was found to be 76% (without correction) and 79% (with correction).

4. Discussions

In the present work we have introduced the concept of compatibility of the adjacent residue pairs in determining the preferences for α-helix or

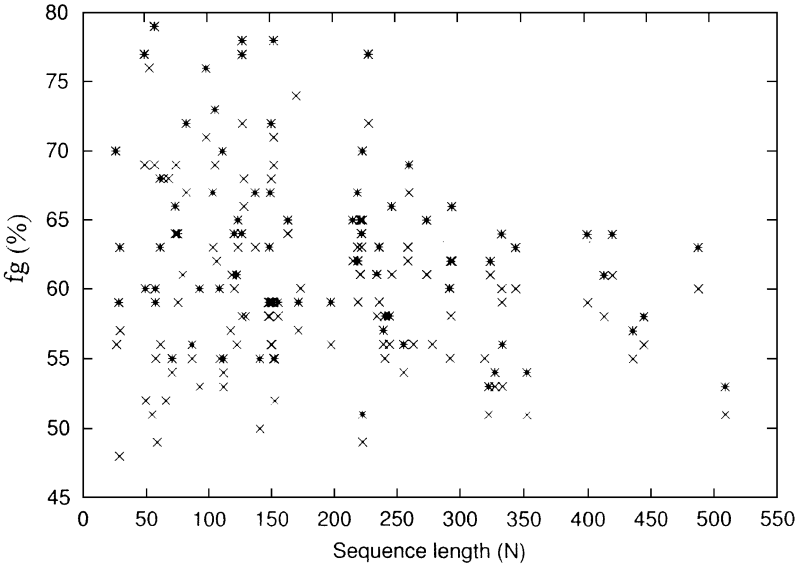

Fig. 6. Plot of the prediction accuracy (*fg*%) of the test sequences against the sequence lengths summarizing the prediction results. The symbols × and * represent the *fg* scores (%) with and without the correction for the continuous sequence stretch as discussed in the structure prediction section.

β-strands and have analyzed the relevant pairwise compatibility data from a large number of known structures. It may be noted that the statistical information on the compatibility of adjacent residue pairs has not been considered in earlier methods. These data have been used in a subsequent minimal statistical model for secondary structure prediction. However, such statistical methods use only compatibility information averaged over a large number of sequences and thus are incapable of describing individual sequences with full accuracy. Moreover, the effects of tertiary contacts are not considered which we have shown here to be important for β-strand predictions. In spite of these limitations, the present method gives a good overall picture of the secondary structural architecture of a given sequence indicating the validity of the consideration of structural features. Our approach is physically more realistic due to the inclusion of the intrinsic structural features of α-helices and β-strands in calculation. A sufficiently large ensemble of protein sequences and structures was used in computing the *CI* values to ensure adequate level of statistical confidence. Use of even larger ensemble may alter the individual *CI* values but the overall relative patterns are not expected to change. Moreover, as the same compatibility data sets are used to obtain the different critical values for different criteria in structure prediction, it is not expected that a larger ensemble will affect the generality of the results in a significant way. For further improvement of our insights in structure–sequence correlation it is essential to understand the physical basis of these observed compatibility properties at the atomic level. Work in this direction is in progress.

5. Summary and conclusions

The findings of the present work are summarized below.

We have presented the statistical data on the pair-wise compatibility of spatially nearest neighbors and the adjacent residues that has not been presented before. Statistical analysis of compatibility of different residue pairs as spatially nearest neighbor pairs for α-helices and β-strands shows quite different patterns. Interestingly, the difference in the patterns is more pronounced in the cases of the compatibility of adjacent residue pairs.

The compatibility of the same residue pair as the spatially nearest neighbors is different from that when the same pair is adjacent to each other indicating that the physical basis (or nature) of the compatibility in each type is different from the other.

Most interestingly it is found that the tertiary contacts influence the pairwise compatibility more in the case of β-strands than in the cases of an α-helix.

Depending on the local sequence, the LCI^{α} and the LCI^{β} values of the same residue type vary over a large range explicitly indicating the importance of the local context of the sequence in determining the preferred secondary structure. The structural preferences for the residues A, V, L, K, R and E are found to be most sensitive to the local sequence.

It has been shown that using the structural characteristics of the secondary structures, a minimal model can be constructed for predicting the secondary structures from the sequence and its predictive power is comparable to that of the more rigorous statistical methods.

Acknowledgments

This work was supported by the Council of Scientific and Industrial Research, India, (CSIR, scientists pool scheme). I would like to thank Dr Chittra Dutta of IICB for critically reading the manuscript and Mr Goutam Garai of the computer division of Saha Institute of Nuclear Physics for allowing to use their computers.

Appendix A:

List of PDB entry of protein sequences used in computing the compatibility indices.

1cbo:a, 1fug:a, 1lmk:a, 1quh:a, 1xlf:a, 9xim:a, 1auj, 1lnd:e, 1qyp, 1xpa, 1bab:a, 1bqx, 1bqz, 1bak, 1jan:a, 1jaw, 1jev:a, 1bmk:a, 1cfy:a, 1cmb:a, 1cax:a, 1caz, 1cdk:a, 1ceb:a, 1cne, 1ctn, 1cjw:a, 1czj, 1czm, 1bmr, 1bms:a, 1buz, 1bux, 1bzr:a, 1bzw:a, 1bzy:a, 1bao:a, 1bap, 1dbr:a, 1dbs, 1eac, 1efg:a, 1ena, 1eve, 1exp, 1eba, 1edk, 1edn, 1emf, 1hba:a, 1hdn, 1hgd:a, 1hjt, 1hng:a, 1hrh:a, 1htr:b, 1htt:a, 1hug, 1hui:a, 1hym:a, 1hyp, 1hea, 1hia:i, 1jgj:a, 1jae, 1jdc, 1jdo, 1jdy:a, 1jlx:a, 1jmf, 1jpc, 1jrh:h, 1jrs:a, 1jsa, 1jsg, 1jst, 1juk, 1jcr, 1jxp, 1kaa, 1kap:p, 1kay, 1kbc:a, 1kbg:h, 1kbp:a, 1kct, 1kda, 1kdu, 1kel:a, 1kev:a, 1kfs:a, 1kgg:a, 1kir:a, 1kmm:a, 1koa, 1kra:c, 1ksa,:a, 1ksi:a, 1kte, 1ktq, 1kvs, 1kwa:a, 1kxb, 1pba, 1bop:a, 1pgx, 1ubq, 1ptf, 1onc, 1tsc:a, 1fus, 1rev:a, 1and, 1cbn, 2ovo, 8rxn:a, 1gat:a, 3cla, 1tmc:a, 1apa, 2msb:a, 1aet, 1aax:a, 1alf, 1amp, 1ath:a, 1apx:a, 1awp:a, 1fax:a, 1fhi, 1ftg, 1frg, 1gal, 1gbe:a, 1gen, 1gma:a, 1gzi, 1gux:a, 1lzs:a, 1lsm, 1lic, 1lsn, 1lyf, 1mat, 1mam:h, 1mck:a, 1mri, 1mol:a, 1mlq, 1mnj:a, 1nir:a, 1noc:a, 1npk, 1nyz:a, 1nwp:a, 1ofg:a 1opd, 1oui, 1oya, 1ouu:a, 1pbo:a, 1pen, 1pjp:a, 1pot, 1bai:a, 1bah, 1beg, 1bmg; 1prx:a, 1psp:a, 1pxb, 1pyg:a, 1pyt:a, 1qhi:a, 1qrs:a, 1qtk:a, 1qul, 1ral, 1rby, 1rpm:a, 1rtd:a, 1rzl, 1djx:a, 1drh, 1dvr:a, 1ruh, 1rin:a, 1slg:b, 1sha:a, 1sbh, 1jen:a, 1jet:a, 1jfd:a, 1jhl:a, 1jia:a, 1jld:a, 1sue, 1sph:a, 1syn:a, 1szt, 1stf:e, 1tau:a, 1tas:a, 1tfr, 1taf:a, 1tsn, 1tuc, 1tyr:a, 1tyx, 1uky, 1uea:a, 1uro:a, 1upj, 1uvt:h, 1ice:a, 1ifd, 1irw, 1ixx:a, 1iro, 1iux, 1izb, 1vpn:a; 1vkx:a; 1vfa:a; 1vwg:b; 1vxg; 1vam:a; 1vwr:b; 1wej:f; 1wqq; 1wgt:a; 1wod; 1wkf; 1xis; 1xut; 1xys; 1xra; 1xif; 1xgs:a; 1yek:h; 1ydr:e; 1yma; 1ysc; 1yas:a; 1yyy:a; 1zbd:a; 1zwj:a; 1zqe:a; 1zsb; 1zxq; 1zin; 1zpr:a; 1aef, 1axp:a; 1azx:I; 1bcx; 1bvs:a; 1caj; 1cud:a; 1cyg; 1dea:a; 1dkt:a; 1djg:a; 1dut:a; 1elg; 1eni; 1gai; 1gky; 1ine:h, 1igc:a; 1mks; 1mlk; 1mng:a,; 1njd; 1nhr; 1ola:a; 1pop:a; 1sly; 1sox:a; 1qrx:a; 1quk; 1qtz:a; 1wet:a; 1wwc:a; 1ycr:a; 1ypa:I; 1yst:h, 1ytf:a.

References

[1] K.A. Dill, H.S. Chan, From Levinthal to pathway to funnel, Nat. Struct. Biol. 4 (1997) 10–19.
[2] V. Munoz, P.A. Thompson, J. Hfrichter, W.A. Eaton, Folding dynamics and mechanism of β-hairpin formation, Nature 390 (1997) 196–199.
[3] C.M. Dobson, M. Karplus, The fundamentals of protein folding: bridging together theory and experiment, Curr. Opin. Struct. Biol. 9 (1999) 92–101.
[4] F.B. Sheinermann, C.L. Brooks, Molecular picture of folding of a α/β protein, Proc. Natl. Acad. Sci. USA 95 (1999) 1562–1567.
[5] A.R. Dinner, T. Lazaridus, M. Karplus, Understanding β-hairpin formation, Proc. Natl. Acad. Sci. USA 96 (1999) 9068–9073.
[6] P. Luo, R.L. Baldwin, Interactions between water and the polar groups of the helix backbone: an important

determinant of helix propensity, Proc. Natl. Acad. Sci. USA 96 (1999) 4930–4935.
[7] P.Y. Chou, G.D. Fasman, Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from protein, Biochemistry 13 (1974) 211–222.
[8] P.Y. Chou, G.D. Fasman, Predictions of protein conformations, Biochemistry 13 (1974) 222–245.
[9] D.T. Jones, W.R. Taylor, J.M. Thornton, A new approach to protein folds recognition, Nature 358 (1992) 86–89.
[10] B. Rost, C. Sander, Prediction of protein secondary structure at better than 70% accuracy, J. Mol. Biol. 232 (1993) 584–599.
[11] F. Eisenhaberg, F. Imperiale, P. Argos, C. Frommel, Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods, Proteins 25 (1996) 157–168.
[12] F. Eisenhaberg, C. Frommel, P. Argos, Prediction of secondary structural content of proteins from their amino acid composition alone. Ii. The paradox with secondary structural class, Proteins 25 (1996) 169–179.
[13] R.B. Russel, R.R. Copley, G.J. Barton, Protein fold recognition by mapping predicted secondary structures, J. Mol. Biol. 259 (1996) 349–365.
[14] V. Solovyev, A. Salamov, A local secondary structure prediction using local alignments, J. Mol. Biol. 263 (1997) 31–36.
[15] L. Richlewski, A. Godzik, Secondary structure prediction using segment similarity, Prot. Eng. 10 (1997) 1143–1153.
[16] S.E. Brenner, C. Chothia, T.J.P. Hubbard, Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, Proc. Natl. Acad. Sci. USA 95 (1998) 6073–6078.
[17] J. Park, M. Karplus, C. Barrett, R. Hughey, D. Hanssler, T. Hubbard, Sequence comparisons using multiple sequences detect three times as many remote homologues as pair-wise method, J. Mol. Biol. 284 (1998) 1201–1210.
[18] L. Holm, C. Sander, Protein folds and families: sequence and structure alignments, Nucl. Acids Res. 27 (1999) 244–247.
[19] M.J.E. Sternberg, P.A. Bates, L.A. Kelley, R.M. MacCallum, Progress in protein structure prediction: assessment of CASP3, Curr. Opin. Struct. Biol. 9 (1999) 368–373.
[20] J.A. Cuff, G.J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, Proteins Struct. Funct. Genet. 40 (2000) 502–511.
[21] K. Nagano, Logical analysis of the mechanism of protein folding I. Prediction of helices, loops and β-sheets from primary structure, J. Mol. Biol. 75 (1973) 401–420.
[22] J. Garnnier, D.J. Ostguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structures of globular proteins, J. Mol. Biol. 120 (1978) 97–120.
[23] J. Garnier, Protein structure prediction, Biochimie 72 (1990) 513–524.
[24] S. Salzberg, S. Cost, Predicting protein secondary structure with nearest-neighbor algorithm, J. Mol. Biol. 227 (1992) 371–374.
[25] P. Stoloroz, A. Lapades, Y. Xia, Predicting protein secondary structure using neural net and statistical methods, J. Mol. Biol. 225 (1992) 277–363.
[26] T.M. Yi, E.S. Lander, Protein secondary structure prediction using nearest-neighbor methods, J. Mol. Biol. 232 (1993) 1117–1129.
[27] J.T. Yang, Prediction of protein secondary structure from amino acid sequence, J. Protein Chem. 15 (1996) 185–191.
[28] A.A. Salamov, V.V. Solovyev, Protein secondary structure prediction using local alignments, J Mol. Biol. 268 (1997) 31–36.
[29] L. Stryer, Biochemistry, Chapter 2, 3rd, W.H. Freeman and Co, New York, 1988.
[30] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.R. Flannery, Numerical Recipes in FORTRAN, 2nd, Cambridge University Press, 1992, p. 630.
[31] T.P. Creamer, G.D. Rose, Side chain entropy opposes α-helix formation propensities, Proc. Natl. Acad. Sci. USA 89 (1992) 5937–5941.
[32] J. Harman, A.G. Anderson, R.H. Yun, Differential helix propensity of small apolar side chains by molecular dynamics simulation, Biochemistry 31 (1992) 5646–5653.
[33] E.J. Spek, C.A. Olson, Z. Shi, N.R. Kallenbach, Alanine is an intrinsic α-helix stabilizing amino acid, J. Am. Chem. Soc. 121 (1999) 5571–5572.
[34] A.G. Street, S.L. Mayo, Intrinsic β-sheet propensities result from van der Waals interaction between side-chains and the local backbone, Proc. Natl. Acad. Sci. USA 96 (1999) 9074–9076.
[35] X. Zhang, J.P. Mesirov, D.L. Waltz, Hybrid system for protein secondary structure prediction, J. Mol. Biol. 225 (1992) 1049–1063.
[36] J.F. Gibrat, J. Garnier, B. Robson, Further developments of protein secondary structure prediction using information theory, new parameters and consideration of residue pairs, J. Mol. Biol. 198 (1987) 425–443.