Generation of deviation parameters for amino acid singlets, doublets and triplets from three-dimensional structures of proteins and its implications for secondary structure prediction from amino acid sequences

S A MUGILAN and K VELURAJA*

Department of Physics, Manonmaniam Sundaranar University, Tirunelveli 627 012, India

**Corresponding author (Fax, 91-462-322973; Email, bio@md3.vsnl.net.in).*

We present a new method, secondary structure prediction by deviation parameter (SSPDP) for predicting the secondary structure of proteins from amino acid sequence. Deviation parameters (DP) for amino acid singlets, doublets and triplets were computed with respect to secondary structural elements of proteins based on the dictionary of secondary structure prediction (DSSP)-generated secondary structure for 408 selected non-homologous proteins. To the amino acid triplets which are not found in the selected dataset, a DP value of zero is assigned with respect to the secondary structural elements of proteins. The total number of parameters generated is 15,432, in the possible parameters of 25,260. Deviation parameter is complete with respect to amino acid singlets, doublets, and partially complete with respect to amino acid triplets. These generated parameters were used to predict secondary structural elements from amino acid sequence. The secondary structure predicted by our method (SSPDP) was compared with that of single sequence (NNPREDICT) and multiple sequence (PHD) methods. The average value of the percentage of prediction accuracy for α-helix by SSPDP, NNPREDICT and PHD methods was found to be 57%, 44% and 69% respectively for the proteins in the selected dataset. For β-strand the prediction accuracy is found to be 69%, 21% and 53% respectively by SSPDP, NNPREDICT and PHD methods. This clearly indicates that the secondary structure prediction by our method is as good as PHD method but much better than NNPREDICT method.

1. Introduction

The sequential information of proteins has been increasing many fold than their three-dimensional counterpart. Any vital information obtained by the analysis of the three-dimensional structure in terms of sequence will have definite impact on the structure prediction methods and will have added value in this field of research, due to sequence automation and genome project. Earlier studies on the analysis of amino acid doublets and triplets in SWISSPROT sequence database have implications for the significance of the deviated doublets and triplets in the structural aspect of proteins (Veluraja and Mugilan 1997). Based on the information derived from the known three-dimensional structures methods were developed to predict the secondary structural elements of proteins, such as α-helix, β-strand and random structures (Chou and Fasman 1974a, b, 1978; Garnier *et al* 1978, Garnier 1990; Lim 1974). These methods suffered from a lack of data. Prediction was performed based on amino acid singlet information derived from relatively few known three-dimensional structures. The accuracy of prediction is between 56 to 60% (Kabsch and Sander 1984). The problem in these methods has been the inclusion of structures used to derive parameters in the set of structures used to assess the accuracy of the method. Amino acid doublets and triplets information were also used in the early works for secondary structure prediction (Nagano 1973, 1977a; Periti *et al* 1967; Ptitsyn and Finkelstein 1983; Kabat and Wu 1974; Wu and Kabat 1973) and the number of proteins used in deriving the parameters

Keywords. Amino acid singlets, doublets and triplets; deviation parameters; prediction accuracy; protein secondary structure prediction

were comparatively small due to the non-availability of enough three-dimensional structures. In the triplet parameter generation, Nagano (1977b) has grouped the 20 amino acids into 7 types leading to a total of 343 parameters. Kabsch and Sander (1983) have developed an algorithm to assign secondary structures for the structure solved proteins based on their X-ray crystal structure coordinates, which is commonly known as dictionary of secondary structure prediction (DSSP). The recent methods, ZPRED (Zvelebil *et al* 1987), NNPREDICT (Kneller *et al* 1990), SOPM (Geourjon and Deleage 1994), SSP (Solovyev and Salamov 1994), PHD (Rost and Sander 1993, 1994; Rost *et al* 1994), NNSSP (Salamov and Solovyev 1995) and SSPRED (Mehta *et al* 1995), use single or multiple sequence information for the secondary structure prediction. Nearly all of these now run via the world-wide web. But multiple sequence alignment remains a difficult task in molecular bioinformatics. Rigorous algorithms based on dynamic programming have the computational complexity of L^n (where L is the sequence length and n is the number of sequences) and can be impractical if many or long sequences are involved (Frishman and Argos 1997). Misaligned sequences can reduce prediction accuracy to a level lower than that achieved with mere single sequence information (Levin *et al* 1993; Di Francesco *et al* 1996). Any new method which can use single sequence information and intimately deliver secondary structural information will be of potential use in the field of bioinformatics.

In this present work the deviation parameters for amino acid singlets, doublets and triplets were computed and are subsequently used to predict protein secondary structure from amino acid sequence. A comparison of the results obtained by our method, secondary structure prediction by

Table 1. List of non-homologous proteins selected for analysis.

1ABR	2DLN	1CCR	1CMB	1DHR	1DRI	1FBA	1GKY	1GPB	1GPR
3GRS	3SDH	3TGL	4BLM	4BP2	4ENL	4FGF	4GCR	5P21	8ABP
8ATC	8CAT	8I1B	9LDT	9RNT	9RUB	1OVA	3RUB	1VSG	8ACN
1AAJ	1SNC	1TIE	1TTB	2AZA	2CDV	2CMD	2CPL	2DNJ	2END
2MAD	2PIA	2POR	2STV	1GST	1LPE	1LTS	1MIN	1MUP	1OMF
1OMP	1PDA	1PHG	1PRC	1PYA	1HLB	1HMT	1HMY	1HNG	1HPM
1HQA	1HRZ	1HSL	1HTM	1HTP	1HVK	1HUC	1HUL	1HXN	1IAE
1HUW	1OBP	1ONC	1ORO	1OXA	1OYC	1PBA	1PBE	1PBG	1PBN
1PCR	1PDG	1PDN	1PGS	1PI2	1PII	1PKM	1PKP	1PLQ	1PLS
1PNE	1POC	1PNK	1HVD	1ICE	1ILK	1INP	1IRK	1IRL	1ISC
1ISD	1ITG	1JAP	1JCV	1KBP	1KIF	1KNB	1KNY	1KPB	1KPT
1L17	1LFB	1LGR	1LIS	1LKI	1LKK	1LPB	1LPT	1LTD	1LXA
1LYL	1MAL	1CSE	1CSH	1CSM	1CTM	1CTN	1CTT	1CYU	1CYX
1DAA	1DAR	1DDT	1DEA	1DIH	1DIN	1POX	1POY	1PRE	1PRR
1PSD	1PRT	1PTX	1PTD	1PTV	1PUT	1PYP	1QRD	1QUK	1PXT
1QOR	1RCP	1RBU	1RCF	1RCI	1REC	1QPG	1PVD	1REG	1RFB
1RSY	1RTP	1RRG	1RVA	1RIB	1RPA	1DLC	1DLH	1DNP	1DOI
1DPB	1DPE	1DPG	1DSB	1DTR	1DTS	1DTX	1DUP	1DYN	1DYR
1ECA	1SLU	1SLT	1SMD	1SMN	1SRA	1SRI	1SRS	1STD	1SVA
1SVB	1SVC	1SVP	1TAG	1TAH	1TAM	1TBR	1TCA	1TFS	1THJ
1THT	1TII	1TIV	1TLK	1TML	1TPG	1TNR	1TPL	1TRK	1TRR
1TRY	1TSP	1TSS	1TUP	1TYS	1UBS	1UMU	1URN	1VCA	1VHH
1VHR	1VID	1VIL	1VPT	1VSD	1WAS	1WHT	1XAA	1XNB	1XYZ
1YPT	1YTB	1YUA	1ZAA	1SAC	1ECL	1ECP	1EDE	1EDG	1EFT
1ENY	1EPA	1ERI	1ERW	1ESC	4RHV	4SBV	5RXN	6TAA	7PCY
8RUC	8TLN	9PAP	2DKB	2CYP	2CWG	2CTC	2CBA	2CAS	2BTF
2BRD	2BOP	2BLT	2BGU	2BBV	2AYH	2ALP	2AK3	2ACQ	2ACG
2ABK	1ESF	1ESL	1ETC	1EXG	1SAT	1SBP	1SCH	1SCM	1SCU
1ADE	1ADN	1ADT	1AEP	1AER	1AFB	1AMG	1AMP	1AOR	1AOZ
1ARS	1ARV	1ASH	1AT1	2FAL	2FD2	2HBG	2HFT	2HMZ	2HTS
2KAI	2LIV	2MEV	2MNR	2MTA	2NAC	2ORA	2PII	2PLE	2POL
2PRD	2PRK	2PSP	2REB	2RSL	2SIL	2TCT	2TGI	2TMV	1ATP
1BAM	1BBT	1BCF	1BCO	1BDM	1BEC	1BER	1BGC	1BGW	1BIA
1BMC	1BMT	1BNC	1BND	1BNH	1BPB	1BPL	1BRI	1BRL	1BUC
1BVP	1BYB	1CDO	1CDR	1CEA	1CEL	1CEO	1CEW	1CFB	1CHD
1CHK	1CHM	1CID	1CKI	1CKS	2ABD	2ER7	2EBN	2DRP	2DLD
1CAU	1MAT	1MBB	1MDA	1MHC	1MHL	1MKA	1MLA	1MLS	1MMD
1MML	1MOL	1MPP	1MRJ	1MSA	1MSC	1MSE	1MUC	1MUT	1MXA
1CLC	1CNS	1CNV	1COL	1COM	1COW	1CPT	1CUS	1CYD	1CRL
1MAS	1HGX	1HJR	2BPA	1RCB	2SAS	3CD4	3CHY	3CLA	3COX
3DFR	1CPC	1GMF	2CCY	2SCP	4TS1	1ARB	1BBP		

deviation parameter (SSPDP), was made with the results predicted by multiple sequence alignment method PHD and single sequence method NNPREDICT.

2. Methods

The three-dimensional structure of proteins used for our computation were taken from Brookhaven Protein Data Bank (release #78, November 1996) using the PDB-SELECT sub-database (Hobohm *et al* 1992). The threshold value used for selecting the proteins is 25% at which a total of 408 non-homologous proteins were obtained.

The secondary structural information for the selected proteins were generated from the well known software package DSSP of Kabsch and Sander (1983). The α-helix and β-strand fragments were identified from the output of DSSP by imposing the constraint that for a β-strand or α-helix segment atleast three consecutive amino acids should contribute for each secondary structure. All the other amino acids which do not fall in this criteria are considered as amino acids of random structure. For random structure the constraint imposed is that a minimum of three consecutive amino acids should lie in this structure. This procedure was followed for all the selected proteins for identifying the secondary structures from the output of DSSP. The frequency of occurrence of amino acid singlets, doublets and triplets in the selected database was computed using the formula

$$P(X) = \frac{\Sigma N_i(X)}{\Sigma Y_i},$$

where, X = individual amino acid (A) for singlet, two consecutive amino acids (AB) for doublets, and three consecutive amino acids (ABC) for triplets in the selected dataset, $N_i(X)$ = number of counts for X in the ith protein, Y_i = T_i for singlets, T_i-1 for doublets, T_i-2 for triplets where T_i is the total number of amino acids in the ith protein, i varies from 1 to n and n = total number of proteins considered (in this case, 408).

Based on the frequency of occurrence of amino acid singlets, doublets and triplets, one can workout the expected count for these entities in the following way for the various structural elements:

Table 2. Number of generated deviation parameters for amino acid singlets, doublets and triplets.

Structure	Singlets	Doublets	Triplets
α-helix	20(20)	400(400)	4576(8000)
β-strand	20(20)	400(400)	3194(8000)
Random structure	20(20)	400(400)	6402(8000)

Numbers in parentheses indicate the possible theoretical parameters.

$$C_{\exp}(X) = P(X)\,\Sigma\, S_i,$$

X = amino acid singlet (A) or doublet (AB) or triplet (ABC), S_i = the total number of particular entity (X) in a particular structural element in protein i and $C_{\exp}(X)$ = expected counts.

Computed count (observed count) of singlets, doublets and triplets for the various secondary structural elements (α-helix, β-strand and random) are obtained from the output of DSSP by following the constraints mentioned earlier.

The deviation parameter (DP) for amino acid singlets, doublets and triplets in a particular secondary structural element is calculated as follows:

$$\mathrm{DP}(X) = \frac{C_{\mathrm{comp}}(X) - C_{\exp}(X)}{C_{\exp}(X)} \times 100,$$

X = amino acid singlet or doublet or triplet, $C_{\mathrm{comp}}(X)$ and $C_{\exp}(X)$ = computed and expected counts.

We generated the DP for the amino acid singlets, doublets and triplets with respect to the three secondary structural elements (α-helix, β-strand and random) using the above formula. These parameters are normalized with respect to α-helix but within the group. The normalized parameters are subsequently used for structure prediction from the amino acid sequences.

Table 3. Deviation parameters for amino acid singlets.

Amino acid	α-helix	β-strand	Random structure
L	31·3	9·7	– 27·4
A	43·3	– 16·4	– 18·5
S	– 21·8	– 11·4	22·1
G	– 51·6	– 21·4	48·8
V	– 9·6	59·8	– 32·5
E	35·9	– 17·7	– 12·7
K	13·0	– 14·2	0·4
T	– 22·5	15·8	4·9
I	7·4	50·4	– 37·8
D	– 15·3	– 33·0	31·8
R	22·2	– 6·5	– 10·8
P	– 59·3	– 39·4	65·7
N	– 23·0	– 27·6	33·5
Q	25·5	– 17·4	– 5·9
F	– 1·4	28·3	– 17·5
Y	– 3·5	29·3	– 16·7
M	30·6	9·0	– 26·6
H	– 14·0	1·0	8·8
C	– 7·7	16·5	– 5·5
W	5·4	19·8	– 16·5

Table 4. Deviation parameters for amino acid doublets for α-helix (**a**), β-strand (**b**), and random structure (**c**).

(a)

	L	A	S	G	V	E	K	T	I	D	R	P	N	Q	F	Y	M	H	C	W
L	56·1	79·1	2·8	– 30·3	3·8	61·0	56·2	– 6·0	27·2	12·2	70·0	– 72·6	8·5	63·1	24·7	21·9	71·9	17·9	48·6	35·6
A	87·6	86·2	11·1	– 35·4	36·0	72·5	62·8	12·4	56·3	5·5	87·8	– 74·4	– ·9	67·8	40·6	41·6	65·5	16·0	50·6	67·0
S	9·1	19·9	– 40·6	– 75·4	– 15·9	5·7	– 27·9	– 34·8	8·8	– 28·5	– 12·9	– 89·8	– 41·3	– 10·8	– 11·2	– 1·6	37·4	– 17·2	– 63·9	6·1
G	– 20·9	– 31·8	– 70·9	– 72·0	– 44·4	– 36·5	– 55·7	– 58·4	– 39·5	– 61·4	– 37·2	– 72·5	– 74·4	– 36·5	– 53·0	– 64·3	– 22·3	– 67·7	– 23·5	– 71·0
V	15·1	41·0	– 18·5	– 43·5	– 28·9	22·0	– ·3	– 43·2	– 19·6	– 28·6	28·1	– 76·9	4·9	14·9	– 12·4	– 36·1	20·1	– 30·1	– 39·8	– 10·5
E	70·9	88·1	7·1	– 36·3	24·6	67·7	54·1	15·0	32·0	8·7	48·5	– 72·6	– 5·1	85·7	27·8	39·4	71·2	18·9	18·1	33·9
K	39·1	66·4	– 12·1	– 62·4	– 14·7	71·2	32·5	– 3·9	10·4	·1	36·2	– 73·0	– 10·6	76·0	14·8	17·4	24·8	– 10·2	– 14·9	·4
T	11·6	22·3	– 44·2	– 66·9	– 29·9	4·6	– 6·4	– 35·5	– 6·4	– 38·5	– 2·2	– 85·5	– 38·8	– 4·4	– 16·9	– 25·4	25·3	1·4	– 22·8	– 26·3
I	28·4	56·2	– ·8	– 27·9	– 6·4	37·0	21·6	– 23·5	11·9	– 5·2	28·3	– 59·8	– 6·7	51·4	11·9	– 12·7	4·6	– 23·6	– 1·6	5·6
D	25·6	33·0	– 44·4	– 66·8	4·2	24·9	– 1·2	– 19·0	14·1	– 29·7	3·1	– 89·7	– 42·1	18·8	– 9·5	14·6	29·9	– 25·4	– 29·1	14·6
R	46·4	67·2	– ·8	– 52·8	– 7·5	67·1	35·8	– 2·4	20·7	35·4	41·9	– 76·5	4·0	69·5	– 1·6	20·8	84·7	33·5	– 14·9	29·8
P	– 39·4	– 25·6	– 65·2	– 79·1	– 55·2	– 14·1	– 58·8	– 67·9	– 28·6	– 64·0	– 46·1	– 98·3	– 83·1	– 62·3	– 52·6	– 64·7	– 38·6	– 46·2	– 57·0	– 67·4
N	4·3	30·9	– 32·6	– 64·1	– 31·5	3·3	– 14·1	– 37·1	– 7·7	– 61·6	– 1·1	– 92·8	– 56·4	– 12·3	– 8·0	– 33·6	19·9	– 21·4	– 48·1	– 29·1
Q	52·8	87·7	7·5	– 44·0	19·8	77·7	42·9	8·1	33·0	11·7	56·4	– 79·3	– 13·1	43·5	11·9	– ·4	50·6	1·6	– 37·2	27·8
F	40·4	31·1	– 24·1	– 43·9	– 1·2	31·0	– 8·7	– 16·8	– ·1	– 30·2	8·7	– 82·1	– 18·8	1·7	5·6	3·0	33·9	– 8·9	55·5	50·6
Y	48·2	41·4	– 44·2	– 56·3	– 13·2	25·1	12·4	– 48·6	– 1·9	– 23·3	– 14·7	– 74·4	– 18·1	11·0	5·0	24·5	50·6	– 20·7	– 7·3	– 40·8
M	70·3	80·7	– 4·3	– 31·5	6·7	54·6	60·1	50·6	41·6	7·6	49·2	– 70·5	17·0	35·6	23·8	9·5	28·3	42·7	17·9	15·9
H	18·5	9·7	– 18·8	– 59·7	– 24·2	42·6	– 5·4	– 33·9	– 6·3	– 45·0	27·7	– 92·9	– 3·8	– 20·3	– 20·2	– 21·3	22·7	– 10·6	– 27·7	– 15·9
C	21·5	12·3	– 51·1	– 57·0	4·1	47·3	7·6	– 22·5	– 3·4	– 17·8	35·9	– 93·5	– 21·9	50·6	– 2·3	– 21·4	– 4·9	11·0	– 66·5	67·4
W	52·0	60·7	8·3	– 33·1	– 12·1	·4	30·0	– 22·1	12·3	– 15·9	9·2	– 57·0	– 8·8	42·5	12·2	– 17·4	18·3	31·8	– 39·8	50·6

(b)

	L	A	S	G	V	E	K	T	I	D	R	P	N	Q	F	Y	M	H	C	W
L	27·9	– 12·0	5·1	– 9·0	90·8	– 6·9	– 14·7	41·7	66·6	– 16·3	– 10·3	– 24·8	– 15·1	– 7·8	34·3	60·8	17·9	14·5	18·4	43·7
A	– 14·9	– 31·8	– 17·3	– 31·5	35·0	– 33·7	– 29·6	– 3·0	41·2	– 40·0	– 32·4	– 43·0	– 30·7	– 32·6	12·3	3·0	– 13·4	– 1·0	– 5·9	– 18·2
S	10·9	– 22·4	– 39·4	– 31·1	55·2	– 41·1	– 22·6	2·9	26·3	– 42·3	– 30·8	– 51·7	– 40·0	– 26·1	25·2	20·1	7·9	– 26·0	11·8	30·0
G	– 19·4	– 18·7	– 34·0	– 38·2	5·7	– 30·7	– 41·0	– 18·2	9·8	– 46·3	– 20·4	– 40·9	– 41·2	– 33·1	8·2	·3	– 22·5	– 2·1	– 30·5	15·0
V	77·1	30·2	46·8	34·1	137·4	42·6	44·9	91·8	118·5	9·2	46·4	·5	13·7	30·3	98·2	123·3	82·5	80·0	77·4	85·7
E	– 8·7	– 37·1	– 22·1	– 46·3	33·8	– 41·4	– 43·4	– 14·2	41·5	– 47·9	– 18·0	– 32·7	– 46·3	– 40·0	22·4	·5	– 15·0	– 46·8	29·5	– 5·0
K	– 2·1	– 27·6	– 29·5	– 37·9	65·5	– 33·8	– 37·7	– 9·6	50·2	– 52·4	– 34·2	– 32·5	– 36·9	– 50·1	18·1	– 11·4	15·8	– 20·9	4·3	14·0
T	53·2	2·2	– 15·1	– 21·4	91·9	– 3·2	– 9·8	12·7	84·9	– 34·6	8·7	– 24·0	– 35·7	·7	66·9	55·7	30·5	30·1	50·8	37·7
I	74·4	40·5	42·0	60·9	110·9	28·5	25·7	70·1	93·6	12·9	38·0	– 17·5	22·4	2·8	79·2	96·3	85·0	67·7	113·8	52·3
D	– 32·4	– 44·0	– 37·7	– 48·9	– 4·0	– 54·9	– 55·5	– 39·6	·6	– 62·1	– 41·8	– 57·1	– 48·3	– 48·2	– 21·1	– 33·4	– 21·9	– 38·8	– 18·7	– 31·3
R	16·5	– 17·7	– 19·1	– 24·7	80·1	– 31·0	– 21·6	– 6·2	54·7	– 45·9	– 23·6	– 36·2	– 43·6	– 40·6	25·1	8·7	10·4	– 40·9	25·7	·9

Contd . . .

84 S A Mugilan and K Veluraja

P	– 42·6	– 55·4	– 56·3	– 58·3	13·2	– 62·2	– 51·5	– 43·6	– 2·7	– 57·9	– 52·8	– 65·1	– 62·4	– 58·8	– 30·1	– 28·5	– 24·4	– 35·7	– 20·1	– 31·5
N	– 9·7	– 44·5	– 42·0	– 52·8	25·9	– 48·0	– 34·3	– 30·8	6·9	– 50·9	– 45·4	– 56·5	– 55·1	– 44·1	– 13·4	– 1·7	– 28·2	– 52·7	– 21·7	– 4·2
Q	– 3·3	– 38·3	– 22·5	– 44·9	34·9	– 37·5	– 38·0	– 31·6	27·9	– 41·4	– 45·9	– 46·3	– 35·4	– 24·6	19·5	23·4	11·1	– 9·7	– 12·3	2·6
F	38·1	7·9	32·8	10·9	75·1	15·7	13·2	55·6	82·0	1·4	34·0	– 37·7	1·0	22·5	37·4	76·9	5·9	43·7	– 3·5	26·8
Y	26·2	7·5	40·5	3·8	87·9	17·1	– 2·4	66·7	101·7	10·7	21·7	– 25·1	7·1	10·7	54·8	55·0	8·7	28·0	59·2	73·6
M	14·5	– 8·4	6·6	– 3·1	83·8	30·0	– 7·3	– 19·6	36·8	– 26·0	– 15·8	– 13·5	– 39·9	– 17·8	54·4	62·3	30·2	10·7	61·4	46·6
H	6·5	– 13·1	– 4·2	– 33·5	55·8	– 34·7	– 21·2	16·5	43·3	– 22·7	– 24·2	– 49·0	– 39·6	·6	31·8	63·5	– ·1	– 26·0	11·8	17·0
C	12·4	– ·2	30·6	3·3	46·4	– 15·9	11·1	64·2	87·8	·1	– 34·8	– 24·2	– 6·2	– 27·9	30·6	47·1	36·6	– 41·1	5·9	– 30·5
W	9·9	1·8	– ·6	– 23·9	100·5	34·8	22·8	43·6	60·9	– 13·6	27·3	– 43·5	– 10·9	– 9·3	16·7	60·0	26·8	– 36·2	129·8	21·4

(c)

	L	A	S	G	V	E	K	T	I	D	R	P	N	Q	F	Y	M	H	C	W
L	– 55·3	– 46·0	– 5·0	26·0	– 58·9	– 36·9	– 28·8	– 21·8	– 59·7	1·9	– 40·9	64·4	3·7	– 37·8	– 38·0	– 52·5	– 59·7	– 21·1	– 44·2	– 51·2
A	– 50·0	– 38·5	3·2	43·5	– 46·0	– 28·1	– 24·1	– 6·6	– 63·6	21·1	– 39·2	76·9	19·7	– 25·5	– 35·1	– 30·0	– 36·0	– 10·2	– 30·5	– 34·0
S	– 12·9	·4	51·9	70·3	– 23·5	21·7	32·9	21·7	– 22·3	45·5	27·9	92·8	52·7	23·5	– 8·1	– 11·3	– 30·1	27·7	35·8	– 22·7
G	26·2	33·1	69·0	72·3	26·5	43·7	63·1	50·8	20·6	70·2	37·8	74·4	75·8	45·2	30·7	43·2	29·0	47·0	34·8	38·7
V	– 58·1	– 46·5	– 16·5	8·3	– 65·7	– 41·3	– 27·6	– 27·7	– 60·3	13·6	– 47·8	51·7	– 11·8	– 28·9	– 52·6	– 52·2	– 64·8	– 29·3	– 21·1	– 46·1
E	– 42·5	– 36·4	8·9	53·3	– 37·6	– 20·0	– 9·6	– 1·3	– 47·4	23·8	– 21·6	69·3	32·2	– 33·1	– 32·6	– 26·9	– 38·8	16·3	– 30·6	– 19·8
K	– 25·1	– 27·8	26·5	65·7	– 30·8	– 27·1	1·4	8·6	– 38·2	32·4	– 3·3	69·5	30·1	– 20·2	– 21·2	– 4·7	– 26·6	19·8	7·3	– 9·0
T	– 40·9	– 16·4	39·2	58·5	– 36·8	– 1·1	10·4	16·1	– 48·4	47·5	– 4·0	72·6	48·3	2·5	– 30·1	– 17·4	– 36·0	– 19·7	– 16·1	– 5·6
I	– 65·3	– 63·1	– 25·6	– 18·9	– 64·5	– 42·7	– 30·6	– 27·6	– 66·2	– 4·4	– 42·7	51·3	– 9·4	– 36·4	– 57·2	– 51·2	– 55·9	– 26·1	– 69·5	– 36·3
D	2·8	5·0	53·4	75·5	– ·3	17·2	35·2	37·4	– 9·9	58·6	23·8	96·0	58·5	17·2	19·5	10·8	– 6·6	41·3	31·3	9·6
R	– 41·6	– 34·4	12·4	51·0	– 44·7	– 26·1	– 10·8	5·5	– 47·9	4·6	– 13·7	74·1	24·4	– 21·8	– 14·5	– 19·5	– 63·7	2·7	– 5·9	– 20·7
P	53·0	51·6	79·0	89·6	29·1	48·1	71·7	72·9	21·0	79·1	63·9	106·8	94·8	78·6	54·2	61·4	41·3	53·4	51·0	65·1
N	3·1	6·7	48·1	76·1	5·2	27·5	30·8	44·2	·9	73·2	28·9	97·8	72·3	35·7	13·8	23·8	4·1	47·2	45·9	22·3
Q	– 33·7	– 35·5	8·9	57·6	– 35·1	– 29·2	– 5·4	14·1	– 39·7	17·8	– 9·6	82·4	30·8	– 14·1	– 20·2	– 14·3	– 41·1	5·0	32·8	– 20·4
F	– 50·9	– 25·9	– 4·1	22·9	– 45·8	– 30·7	– 2·3	– 23·1	– 50·8	19·5	– 27·0	78·9	12·1	– 15·1	– 27·0	– 49·8	– 26·6	– 21·1	– 35·3	– 50·8
Y	– 48·8	– 32·6	4·7	35·6	– 45·6	– 27·6	– 6·9	– 8·6	– 61·9	9·2	– 3·5	65·8	7·9	– 14·1	– 37·4	– 50·7	– 39·6	– 3·4	– 31·8	– 18·1
M	– 56·5	– 49·3	– 1·2	23·2	– 56·5	– 55·5	– 36·1	– 22·0	– 50·9	11·0	– 23·5	56·0	13·3	– 13·0	– 49·8	– 45·1	– 37·9	– 35·5	– 50·2	– 39·6
H	– 16·6	1·5	15·3	61·1	– 18·3	– 7·2	16·8	12·7	– 22·6	44·5	– 3·7	93·2	27·1	13·3	– 6·1	– 25·0	– 15·3	23·2	11·4	·2
C	– 22·2	– 8·2	15·6	36·4	– 31·5	– 22·1	– 12·0	– 24·6	– 52·2	12·0	– 2·6	78·1	18·7	– 16·9	– 17·4	– 14·8	– 19·4	18·1	41·3	– 26·6
W	– 41·2	– 42·1	– 5·2	37·2	– 54·1	– 21·9	– 34·4	– 12·1	– 46·1	19·2	– 23·2	65·5	12·7	– 22·9	– 18·6	– 25·5	– 29·0	1·0	– 53·7	– 47·5

3. Results and discussion

Table 1 lists the PDB files used for gathering information about the secondary structural elements using DSSP package. They were subsequently used to compute the DP for amino acid singlets, doublets and triplets with respect to secondary structural elements of proteins. Calculated parameters were normalized with respect to α-helix. Though normalization can also be carried out with respect to other structural elements, which will in turn affect the relative magnitude of the DP value but was not required as it would not have affected the final result. To ensure the absence of

Table 5. Model output showing secondary structure prediction from amino acid sequence by SSPDP for the protein 8ABP.

Sequence number	Amino acid	1st rank SSC	Summed value for 1st rank	2nd rank SSC	Summed value for 2nd rank	Difference in percentage between 1st and 2nd rank	Prediction by SSPDP
1	N	R	77	H	36	36	–
2	L	H	164	S	65	43	H
3	K	H	129	S	23	70	H
4	L	H	60	S	55	4	H
5	G	R	217	S	131	25	S
6	F	S	158	H	28	70	S
7	L	S	268	H	206	13	S
8	V	S	307	H	189	24	S
9	K	H	121	S	12	82	–
10	Q	R	157	H	129	10	R
11	P	R	302	H	15	91	R
12	E	H	153	R	122	11	R
13	E	R	118	H	108	4	R
14	P	R	371	S	66	70	R
15	W	R	108	S	90	9	R
16	F	S	113	R	1	98	–
17	Q	H	83	S	60	16	H
18	T	H	37	S	6	73	H
19	E	H	212	S	11	91	H
20	W	H	180	S	104	27	H
21	K	H	149	S	67	38	H
22	F	S	107	H	102	3	H
23	A	H	148	R	23	72	H
24	D	R	161	H	130	11	H
25	K	H	145	R	106	16	H
26	A	H	129	R	105	10	H
27	G	R	310	S	51	72	R
28	K	R	173	H	48	56	R
29	D	R	185	H	65	48	R
30	L	R	33	H	25	14	R
31	G	R	282	S	120	40	R
32	F	S	110	R	34	52	S
33	E	H	152	S	121	11	S
34	V	S	340	H	205	25	S
35	I	S	342	H	175	32	S
36	K	S	198	H	143	16	S
37	I	S	366	H	174	36	S
38	A	H	192	S	178	4	S
39	V	S	190	R	84	39	S
40	P	R	407	S	105	59	R
41	D	R	342	S	57	72	R
42	G	R	334	S	88	58	R
43	E	R	107	H	106	1	R
44	K	H	128	R	55	40	H
45	T	S	66	H	62	3	H
46	L	H	185	S	101	29	H
47	N	H	146	R	58	43	H
48	A	H	298	S	68	63	H
49	I	H	167	S	144	7	H
50	D	R	152	H	24	73	–

SSC, Secondary structural character; H, α-helix; S, β-strand; R, random structure.

relationship between sequences in the training set used to generate the deviation parameter and the protein sequence under prediction one at a time, i.e., Jack-Knife procedure, was followed. Each time one of the protein out of 408 proteins was removed from the dataset and the rest of the 407 proteins were used to generate the parameters. The generated parameters were used to predict the secondary structure of the removed protein. This procedure was repeated for all the selected proteins (table 1). The number of generated parameters for singlets, doublets and triplets in various secondary structural elements are shown in table 2 by indicating the possible number of theoretical parameters within parentheses. All the 20 amino acid singlets and 400 amino acid doublets possess DP value for α-helix, β-strand and random structures. For the amino acid triplets which have a computed count of zero (triplets not found in the selected dataset for a particular secondary structure), the DP values with respect to that secondary structure is assigned as zero. The reason for assigning the zero DP value is not to bias the structural prediction. The absence of triplets from the dataset does not mean that this will never occur in proteins. In our earlier analysis on amino acid sequences using the SWISSPROT protein sequence Release 28, March 94, having 36,000 proteins containing a total number of 1,24,96,420 amino acid entries, we found that all the triplets occurred in the protein sequence dataset with some of them having low occurrence and a few showing preferential selection (Veluraja and Mugilan 1997). Since the three-dimensional structure database is many fold smaller than sequence database, some of the triplets are not found in the selected dataset. Only 57% of amino acid triplets have DP value for α-helix, 40% for β-strand and 80% for random structures.

Table 3 gives average DP values for amino acid singlets in α-helix, β-strand and random structures. In the secondary structural elements, the DP value for singlets varies from 66 to – 59. In α-helix, alanine possesses the maximum DP value (43) and proline possesses the minimum DP value (– 59). Proline possesses the minimum DP value (– 39) in β-strand too. In β-strand the amino acid valine, possesses the maximum DP value (60). Though proline possesses a minimum parameter value for α-helix and β-strand, it has the highest parameter value (66) for random structures because of the inclusion of β-turns in random structures with proline being the amino acid of turn structures. In random structures isoleucine possesses the minimum DP value (– 38).

For the 400 amino acid doublets, the DP values for α-helix, β-strand and random structures are listed in table 4. For computing the maximum parameter, the average of 5% of those doublets (20 doublets) which have maximum DP value was taken and it worked out to be 90, 142 and 102 for α-helix, β-strand and random structures respectively. These values are roughly twice the maximum parameter value of the singlets for α-helix and β-strand. But for random structure, it is roughly 1·5 times the maximum parameter value of singlets. The minimum parameter value for doublets worked out to be – 90, – 82 and – 79 respectively for α-helix, β-strand and random structures, which were roughly 2 times higher in magnitude than the minimum parameter value of the amino acid singlets with respect to β-strand and random structures. But for α-helix structure, it was roughly 1·5 times the minimum parameter value of singlets.

In the case of triplets, the possible parameters were theoretically 24,000 (8000 each for α-helix, β-strand and random structures) and we could generate parameters only for 14172 triplets as the rest of the triplets were not found in the selected dataset. The maximum parameter of the 5% of the triplets (400 triplets) which had maximum DP values worked out to be 244, 372, and 224 for α-helix, β-strand and random structures respectively. These maximum parameters for α-helix and β-strand had a value 5 times more than that of the singlet parameters, and more than 2·5 times that of the doublet parameters. But for random structure, the triplet parameters were 3·4 times more than that of amino acid singlets, and 2 times than that of amino acid doublets. The minimum parameters (400 triplets) worked out to be – 101, – 112 and – 107 for α-helix, β-strand and random structures respectively. The generated deviation parameters (a total of 15,432) were used to predict the secondary structural elements from sequence by adopting the following procedure.

If we have sequence of a protein as – ALEGAML – the secondary structural information for G for a particular secondary structure can be a summed value calculated as, for α-helix

$$SUM_h = (G)_{sh} + \tfrac{1}{2}\,[(EG)_{dh} + (GA)_{dh}] + \tfrac{1}{3}\,[(LEG)_{th} + (EGA)_{th} + (GAM)_{th}],$$

where, sh = singlet deviation parameter for α-helix; dh = doublet deviation parameter for α-helix; th = triplet deviation parameter for α-helix.

Similar procedure is adopted to compute the summed value for β-strand (SUM_s) and random structures (SUM_r) and the rank order of these summed values was found out. The summed value of the third rank was normalized to zero and the subsequent rank orders were computed. For each amino acid in the sequence the summed value for α-helix, β-strand and random structures were calculated. Based on the rank order of the summed value secondary structural characters (H, S, R) were assigned to all amino acids in the sequence. If a minimum of three or more consecutive amino acids possessed the same secondary structural character, then that segment was assumed to belong to that secondary structure. If a long segment of the amino acid sequence bears the same secondary structural character (say H) with one or two secondary structural characters different, then the second rank order summed value was taken into account for its prediction. If the difference in percentage between the summed values of the first and

second rank order was less than 33·3 and if the second rank structural character (say H) was the same as that of the other amino acids of the selected long segment, these amino acids were also assigned to have the same secondary structural character (H) as that of the running sequence. A typical output for the arabinose binding protein (8 abp) is shown in table 5. Based on the above criteria the following secondary structures were assigned for the amino acid sequence of table 5: α-helix: 2–4, 17–26 and 44–49, β-strand: 5–8 and 32–39 and rest of the amino acids were assigned to the random structure.

To monitor the reliability of our method in secondary structure prediction, we generated secondary structure for all the proteins (408) from their sequence using the DP values. For the same sequences, secondary structures were also generated by NNPREDICT and PHD methods. Our classification of DSSP resulted into three states (α-helix, β-strand and random) as mentioned in the method of cal-

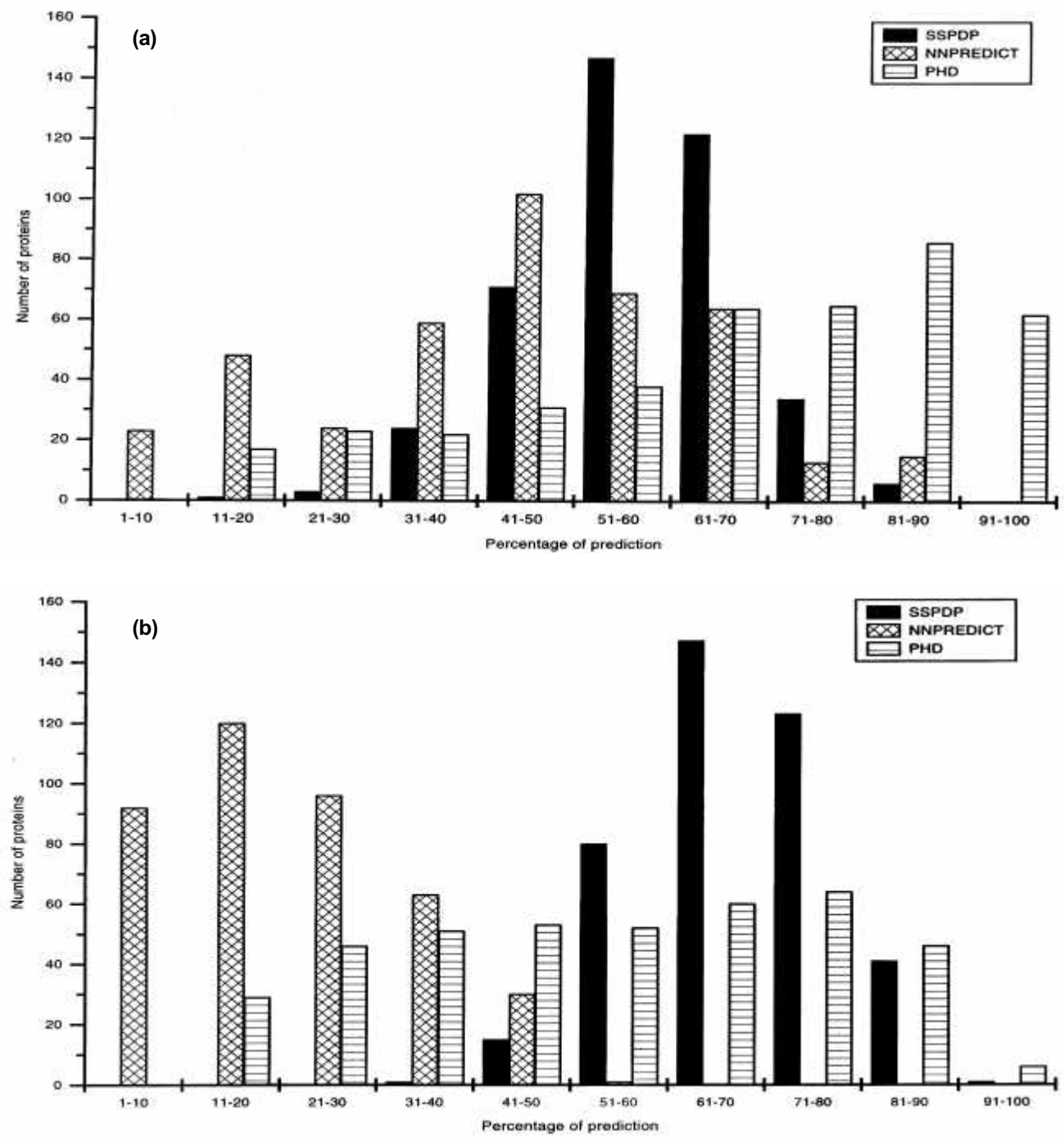

Figure 1. Percentage of prediction versus number of proteins by various methods (**a**) α-helix and (**b**) β-strand.

culation, analogous to that is used in the work of PHD and NNPREDICT.

The prediction accuracy in percentage was computed for the various methods as

$$\text{Prediction accuracy in \%} = \frac{\sum_{i=1}^{n} \text{Number of amino acids predicted in the segment of a particular secondary structure}}{\sum_{i=1}^{n} \text{Total number of amino acids in a particular secondary structure segments as generated by DSSP}} \times 100$$

where, n represents the total number of segments in the protein bearing the same secondary structure.

The prediction accuracy was computed for all the selected proteins. The prediction accuracy for α-helix varies from 11% to 90%, 0% to 91% and 11% to 100% respectively by SSPDP, NNPREDICT and PHD methods. The average value of the percentage of prediction accuracy for α-helix by SSPDP, NNPREDICT and PHD methods was found to be 57%, 44% and 69%. This clearly indicates that for helical prediction our method (SSPDP) has comparable prediction accuracy as that of PHD method but much better prediction than NNPREDICT. The prediction accuracy for β-strand varies from 36% to 96%, 0% to 56% and 11% to 97% by SSPDP, NNPREDICT and PHD methods. The average percentage value of prediction for β-strand is 69%, 21% and 53% respectively by SSPDP, NNPREDICT and PHD methods. The histogram with respect to helical prediction (figure 1a) indicates that the NNPREDICT method predicted the more number of proteins at a low accuracy scale (256 proteins have less than 50% prediction accuracy) whereas SSPDP and PHD methods predicted the more number of proteins at the higher accuracy scale (309 and 315 proteins have greater than 50% prediction accuracy for SSPDP and PHD methods respectively). The SSPDP method tends to dominate the prediction at 50% to 70% accuracy and the PHD method yields highest prediction accuracy with respect to α-helix. On the other hand in the β-strand prediction (figure 1b) NNPREDICT method predicts all the proteins with less than 50% accuracy, the SSPDP and PHD methods predict even more number of proteins with prediction accuracy higher than 50%.

In order to substantiate our prediction methods, we have randomly selected 10 proteins which are not included in the learning set. Using the average DP values, prediction was carried out and the results which were compared with NNPREDICT and PHD methods are shown in figure 2. It is seen from the figure that the average percentage of prediction for α-helix is 54%, 52% and 87% by SSPDP, NNPREDICT and PHD methods respectively. For β-strand the average percentage of prediction for SSPDP is about 61%, while it for NNPREDICT and PHD methods are 35% and 67% respectively. This also gives an indication that our method shows similar prediction accuracy for the proteins which are not part of the selected dataset.

Our method which is novel and does not use multiple sequence alignment, performs comparably well with the PHD method. This method is in the process of being improved with the addition of more number of proteins in the dataset. In conclusion, our method can also be used for protein secondary structure prediction from amino acid sequence.

Acknowledgements

The financial assistance received from the Department of Biotechnology (DBT), New Delhi for this work and the use of facilities available at Bioinformatics Centre, Madurai Kamaraj University, Madurai funded by DBT are gratefully acknowledged. The authors also wish to acknowledge Miss T Hema Thanka Christlet for her help and assistance with this work.

References

Chou P Y and Fasman G D 1974a Conformational parameters for amino acid in helical, β-sheet and random coil regions calculated from proteins; *Biochemistry* **13** 211–222

Chou P Y and Fasman G D 1974b Prediction of protein conformation; *Biochemistry* **13** 222–245

Chou P Y and Fasman G D 1978 Prediction of the secondary structure of protein from their amino acid sequence; *Adv. Enzymol.* **47** 45–148

Di Francesco V, Garnier J and Munson P J 1996 Improving protein secondary structure prediction with aligned homologous sequences; *Protein Sci.* **5** 106–113

Frishman D and Argos P 1997 Seventy-five percent accuracy in protein secondary structure prediction; *Proteins* **27** 329–335

Garnier J 1990 Protein Structure Prediction; *Biochimie* **72** 513–524

Garnier J, Osguthorpe D J and Robson B 1978 Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins; *J. Mol. Biol.* **120** 97–120

Geourjon C and Deleage G 1994 SOPM: a self optimised prediction method for protein secondary structure prediction; *Protein Eng.* **7** 157–160

Hobohm U, Scharif M, Schneider R and Sander C 1992 Selection of representative protein datasets; *Protein Sci.* **1** 409–417

Kabat E A and Wu T T 1974 Further comparison of predicted and experimentally determined structure of adenylate kinase; *Proc. Natl. Acad. Sci. USA* **71** 4217–4220

Kabsch W and Sander C 1983 Dictionary of protein secondary structure pattern – recognition of hydrogen-bonded and geometrical features; *Biopolymers* **22** 2577–2637

Kabsch W and Sander C 1984 On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations; *Proc. Natl. Acad. Sci.* **81** 1075–1078

Kneller D G, Cohen F E and Langridge R 1990 Improvements in protein secondary structure prediction by an enhanced neural network; *J. Mol. Biol.* **214** 171–182

Levin J, Pascarella S, Argos P and Garnier J 1993 Quantification of secondary structure prediction improvement using multiple alignment; *Protein Eng*. **6** 849–854

Lim V I 1974 Structural principles of the globular organization of proteins chains, A Stereochemical theory of globular protein secondary structure; *J. Mol. Biol.* **88** 857–872

Mehta P K, Heringa J and Argos P 1995 A simple and fast approach to prediction of protein secondary structure prediction from aligned sequence with accuracy above 70%; *Protein Sci*. **4** 2517–2525

Nagano K 1973 Logical analysis of the mechanism of protein folding; *J. Mol. Biol*. **75** 401–420

Nagano K 1977a Logical analysis of the mechanism of protein folding. IV. Super-secondary structures; *J. Mol. Biol*. **109** 235–250

Nagano K 1977b Triplet information in helix prediction applied to the analysis of super-secondary structures; *J. Mol. Biol*. **109** 251–274

Periti P F, Quagliarotti G and Liquori A M 1967 Recognition of α-helical segments in proteins of known primary structure; *J. Mol. Biol*. **24** 313–322

Ptitsyn O B and Finkelstein A V 1983 Theory of protein secondary structure and algorithm of its prediction; *Biopolymers* **22** 15–25

Rost B and Sander C 1993 Prediction of protein secondary structure at better than 70% accuracy; *J. Mol. Biol.* **232** 584–599

Rost B and Sander C 1994 Combining evolutionary information and neural networks to predict protein secondary structure; *Proteins* **19** 55–72

Rost B, Sander C and Schneider R 1994 PHD-an automatic mail server for protein secondary structure prediction; *Comput. Appl. Biosci*. **10** 53–60

Salamov A A and Solovyev V V 1995 Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiply sequence alignment; *J. Mol. Biol*. **247** 11–15

Solovyev V V and Salamov A A 1994 Predicting alpha-helix and beta-strand segments of globular proteins; *Comput. Appl. Biosci*. **10** 661–669

Veluraja K and Mugilan S A 1997 Amino acid doublets and triplets in protein sequences – A database analysis; *Curr. Sci*. **72** 572–577

Wu T T and Kabat E A 1973 An attempt to evaluate the influence of neighboring amino acids (n – 1) and (n + 1) on the backbone conformation of amino acid (n) in proteins. Use in predicting the three-dimensional structure of the polypeptide backbone of other proteins; *J. Mol. Biol*. **75** 13–31

Zvelebil M J, Barton G J, Taylor W R and Sternberg M J 1987 Prediction of protein secondary structure and active sites using the alignment of homologous sequences; *J. Mol. Biol*. **195** 957–961

MS received 9 August 1999; accepted 12 November 1999

Corresponding editor: DIPANKAR CHATTERJI