Theor Chem Acc (1999) 101:41–45
DOI 10.1007/s00214980m127

Theoretical
Chemistry Accounts

Regular article

A statistical analytical approach to predict the secondary structure of proteins from amino acid sequence information*

Shrish Tiwari, Boojala V.B. Reddy

Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, India

Received: 4 May 1998 / Accepted: 17 September 1998 / Published online: 10 December 1998

Abstract. A statistical analytical approach has been used to analyze the secondary structure (SS) of amino acids as a function of the sequence of amino acid residues. We have used 306 non-homologous best-resolved protein structures from the Protein Data Bank for the analysis. A sequence region of 32 amino acids on either side of the residue is considered in order to calculate single amino acid propensities, di-amino acid potentials and tri-amino acid potentials. A weighted sum of predictions obtained using these properties is used to suggest a final prediction method. Our method is as good as the best-known SS prediction methods, is the simplest of all the methods, and uses no homologous sequence/family alignment data, yet gives 72% SS prediction accuracy. Since the method did not use many other factors that may increase the prediction accuracy there is scope to achieve greater accuracy using this approach.

Key words: Helix – β-Strand and coil – Amino acid propensity – Statistical potentials – Sequence analysis – Secondary structure prediction

1 Introduction

The past few years have seen the development of several algorithms attempting to predict the secondary structure (SS) of proteins from a knowledge of the amino acid sequence alone [1–3]. The Chou-Fasman [4] and GOR [5] methods are the classical examples in the literature that have been extensively used, until recently, by experimentalists when they determine a sequence. These methods have been very useful for assessing the functional aspects of proteins by correlating the SS predictions with the experimental results. The present scenario has much greater expectations from the SS prediction methods. The most important rationale being to provide some information about the fold of the target protein from the predicted SSs [6–8]. Recent attempts using sequence similarity and alignment procedures from a set of homologous proteins and nearest-neighbor algorithms increased the prediction accuracy levels to 70% and above [9–18].

The most popular of all the methods is Rost and Sander's [10, 11] PHD program. The most recent methods which claim to be as good as PHD with conceptual simplicity and user-transparent algorithms are by Thompson and Goldstein [19] and Rychlewski and Godzik [20]. There are also quite a few reports on the possible limitations of SS prediction methods [21–23]. Frishman and Argos [3] argue that the accuracy of SS prediction cannot be greater than 80–85%, even with a tenfold increase in sequence data using sequence similarity and alignment procedures.

We have used a simple statistical analytical approach that was earlier used on DNA sequences to identify protein coding regions [24] and for splice-site predictions [25, 26]. We have successfully extended this approach to protein sequences to predict the intracellular stability of the protein from comparative sequence data analyses [27–29]. Here we describe a similar approach to predict the SS type of each amino acid from the given sequence of amino acids.

We describe our analysis of non-homologous best-resolved protein structures. We analyzed the SS of the amino acid as a function of the sequence of residues in the near-neighbor region. We calculated propensities and potentials of conditional occurrence of mono-, di- and tri-amino acids around the sequence region and used these values to predict the SS from the sequence.

2 Materials and methods

We have used 306 non-homologous (< 25% sequence identity), best-resolved (≤2.0 Å resolution) protein structures [30] from the Protein Data Bank (PDB) as a sample data set (data set 1) to

* Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: B.V.B. Reddy
e-mail: bvbreddy@ccmb.ap.nic.in

calculate the set of properties described later. We have prepared a test data set (data set 2) by taking the closest homologue of each of the proteins in data set 1 wherever it is available. The SS of each residue was evaluated using the SSTRUC program [31] from the coordinates of the atoms given in the PDB files. SSTRUC takes the PDB files and defines the SS type such as helix (α-helix or 3_{10}-helix), β-sheets, turns, coils and also ϕ, ψ and χ angles using Kabsch and Sander's [32] definition of SS. For analysis and prediction purposes we will be considering only three types of SSs, namely, helices (H), β-structures (B) and coils (C) (turns, loops and all non-regular conformations are included in coils). We have computed three different kinds of property using normalized occurrences in the neighborhood region of the sequence as defined below.

2.1 Amino acid propensity (AP)

The simple propensity of residues in the three classes of structures was calculated as follows:

$$P_s(x) = [N_s(x)/N(x)]/(N_s/N), \tag{1}$$

where $N_s(x)$ is the occurrence of amino acid x in the SS type s, $N(x)$ is the total number of x, N_s is the total number of residues in s and N is the total number of residues in the sample set. This equation was first defined by Chou and Fasman [4] in their popular SS prediction method.

2.2 Di-amino acid potentials (DP)

The conditional occurrence of near-neighbor residues and their SS type in each structure is used to calculate DP values, ${}_{iy}R_s(x)$.

$${}_{iy}R_s(x) = {}_{iy}N_s(x)/{}_{iy}N_{\mathrm{n-s}}(x), \tag{2}$$

where ${}_{iy}N_s(x)$ is the normalized occurrence of x in structure s with residue y at position i ($i = -n, -n+1,..., -1, 0, 1,...,n$; $i=0$ corresponds to the position of residue x) and ${}_{iy}N_{\mathrm{n-s}}(x)$is the same normalized occurrence in non-s. For example, a normalized occurrence is the occurrence with respect to one of the total di-amino acid occurrences, $\sum_{iy}M_s(x)/\sum_{iy}M_{\mathrm{n-s}}(x)$, whichever has the lowest value [*e.g.*, ${}_{iy}N_s(x) = {}_{iy}M_s(x)$ and $p_{iy}N_{\mathrm{n-s}}(x) = {}_{iy}M_{\mathrm{n-s}}(x)$ $[\sum_{iy}M_s(x)/\sum_{iy}\mathrm{M}_{\mathrm{n-s}}(x)]$ or vice versa].

2.3 Tri-amino acid potential (TP)

This is similar to DP, but in this case we do not take into account the type of the central residue, only the residue in the SS type is considered when calculating the TP value, ${}_{iz}T_s$.

$${}_{iz}T_s = {}_{iz}N_s/{}_{iz}N_{\mathrm{n-s}}, \tag{3}$$

where ${}_{iz}N_s$ is the normalized occurrence of residues in the SS type s with di-peptide z at a distance i (varies from -32 to $+32$). ${}_{iz}N_{\mathrm{n-s}}$ is the same normalized occurrence in non-s.

2.4 Matrices of amino acid propensities and potentials

The above-mentioned properties are computed for the 80,672 residues of the natural data sets in the form of separate matrices. In all our matrix definitions we followed the alphabetical order of the single residue code of amino acids and the order of SS type as C, H and B for $s = 1$, 2 and 3, respectively.

1. For the AP we have a 20×3 matrix, with each element expressing the propensity of a specific amino acid in a specific SS type. For example the element (1,1) of the matrix is the propensity of alanine in the coil-type structure.
2. The DP is a $20 \times 20 \times 3 \times m$ matrix, where $m = -n$ to $+n$ ($2n+1$ values) is the number of nearest neighbors considered and n is the maximum number of residues away from the central residue. Thus in this case, for example, the element (1,1,1,1) would represent the DP value of alanine in a coil-type structure with another alanine at a position $-n$ from the residue under consideration.
3. The TP matrix is of dimension $400 \times 3 \times m$. The element (1,1,1) here would represent the TP value of all amino acids in the coil-type structure with the di-peptide ($A_{i-n}\mathrm{A}_{\mathrm{i}-n+1}$) at $i - n$ from the central residue i.

2.5 SS prediction

The three calculations (Table 2) are used separately to predict the SS of a given sequence of amino acids.

1. The sum of the average individual residue propensities (AP) was calculated in the region m ($i - n$ to $i + n$) for each of the SS types. The highest value among the three scores (for three SS types) is used to predict the SS type of ith residue.
2. Using DP values the sum of the individual calculations for its nearest-neighbor residues is calculated for each of the three SS types as its score and is illustrated in Table 2. The class for which this score is highest is the predicted structure type for the central residue.
3. A similar procedure is followed for TP values to calculate SS scores for each ith amino acid. Here the SS of the ith residue is predicted based on its neighborhood dipeptide sequence and it is independent of the type of residue present at position i.
4. As a fourth calculation these individual scores were used to define an overall score to take into account the effect of all three properties in our final predictions. The overall score is a weighted sum (WS) of individual scores. Similarly, in this case the highest score among the three SS types is also used to predict the SS type of the ith residue.

3 Results

The percentage of SS prediction accuracy using each of the calculations is presented in Fig. 1 as a function of the number of nearest-neighbor residues in the sequence to calculate individual scores. We have systematically gone

Table 1. Average percentage of secondary structure predictions for the total amino acids in different data sets. The values given in parentheses are average percentages per protein

SMU[a]	Overall % prediction	Coil % prediction	Helix % prediction	β-strand % prediction
(a) Predictions on sample data set 1				
AP	49.4(49.5)	54.4(53.0)	46.7(44.5)	47.9(46.2)
DP	70.5(70.5)	65.9(65.1)	74.2(68.7)	70.0(67.8)
TP	65.8(66.0)	56.1(55.3)	73.1(65.6)	65.7(63.0)
WS	72.0(72.1)	65.4(64.2)	77.0(70.0)	71.7(69.2)
(b) Predictions on test data set 2				
AP	48.9(48.6)	52.5(50.5)	46.6(44.3)	48.2(46.5)
DP	64.5(63.3)	57.0(55.2)	70.6(63.3)	64.1(62.4)
TP	61.9(61.1)	49.3(48.2)	71.1(61.5)	62.6(60.2)
WS	66.6(65.5)	56.9(54.8)	74.0(64.7)	66.8(64.6)
(c) Predictions with jack-knife test on sample data set 1. The protein used for prediction was replaced with the closest homologue from test data set 2				
AP	49.1(49.0)	54.0(52.0)	46.1(43.1)	48.1(43.4)
DP	63.7(63.1)	58.2(57.1)	68.6(60.3)	62.7(57.5)
TP	61.3(60.6)	50.3(49.5)	69.5(59.2)	61.5(55.0)
WS	66.0(65.4)	58.1(56.8)	72.4(62.3)	65.6(59.0)
(d) Predictions with jack-knife test on sample data set 1				
AP	49.3(49.4)	54.4(53.0)	46.6(43.8)	47.6(42.8)
DP	52.9(52.9)	45.5(44.6)	61.0(54.4)	49.2(43.6)
TP	54.2(54.1)	41.4(41.0)	65.1(57.2)	65.1(57.2)
WS	56.6(56.5)	46.3(45.5)	66.6(58.2)	53.2(46.6)

[a]SMU = statistical measure used for predictions

Table 2. Illustrative example of the use of di-amino acid potentials (DP) to calculate scores to predict a possible secondary structure (SS) for residue alanine in its given neighborhood sequence. For example, (C, A) sub-matrix gives the coil (C) potential values of di-amino acids for alanine in the *i*th position and various residues in the $i-3$, $i-2$,..., $i+3$ positions, respectively. Similarly (H, A) for the helix (H) of alanine and (B, A) for the β-strand (B) of alanine potentials matrix. Matrices for each residue were computed separately. The DP values shown in bold correspond to the amino acid alanine and its near-neighbor sequence of amino acids is given in the first row. The scores to predict the secondary structure of alanine in its given sequence environment are 11.51, 3.84 and 6.89 for the C, H and B structures. The highest value is for the C structure. Therefore, the SS of alanine in that sequence neighborhood is predicted to be C. [The TP matrix was also calculated in a similar way, but with a slight difference. The row of the matrix is for 400 (20×20) combinations of amino acid pairs and the height of the column represents $i-32$ to $i+32$ rows. The TP matrix is not concerned with the residue type at the *i*th position but only its SS type. All residues at the *i*th position with each SS type will have one sub-matrix, i.e. a total of three sub-matrices for three SSs]

...PYV**A**PGP...

(C, A)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
−3	0.72	1.09	1.27	0.98	1.03	1.21	1.09	0.64	1.07	0.74	0.87	0.92	**1.43**	0.83	0.96	1.15	1.40	0.93	1.16	1.33
−2	0.65	2.03	0.99	0.54	0.92	1.47	1.92	0.79	0.97	0.84	0.67	1.80	2.45	0.83	0.74	1.10	1.14	0.61	1.13	**1.23**
−1	0.54	2.31	1.35	0.68	0.90	2.85	1.53	0.53	0.85	0.65	0.33	1.60	2.41	0.74	0.76	1.20	0.98	**0.68**	0.58	1.01
1	0.52	0.86	1.95	0.69	0.63	1.91	0.82	0.44	0.93	0.47	0.57	1.67	**4.18**	0.75	0.83	1.32	1.71	0.57	0.75	0.75
2	0.64	1.43	1.50	0.73	0.75	**1.56**	0.93	0.72	0.94	0.64	0.59	1.40	2.81	0.83	0.89	1.32	1.49	0.66	0.80	0.72
3	0.68	1.10	1.07	1.00	1.04	0.96	0.67	0.97	0.89	0.73	0.80	1.09	**2.43**	0.87	0.81	1.28	1.51	0.92	0.68	1.28

C score = **11.51**

(H, A)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
−3	1.66	0.65	1.03	1.28	1.02	0.65	0.79	1.16	0.89	1.32	1.37	0.92	**0.92**	1.37	1.05	0.77	0.73	0.78	0.91	0.76
−2	1.72	0.57	1.15	1.86	0.79	0.54	0.66	0.92	1.06	1.12	1.86	0.85	0.51	1.60	1.28	1.04	0.80	0.88	0.90	**0.74**
−1	2.06	0.39	1.02	1.86	0.91	0.33	0.62	0.81	1.25	1.48	1.84	0.76	0.55	1.66	1.11	0.98	0.86	**0.83**	1.13	0.80
1	2.08	0.88	0.69	1.60	0.98	0.62	0.84	0.93	1.32	1.84	1.36	0.74	**0.27**	1.66	1.61	0.79	0.47	0.82	1.19	0.84
2	1.84	0.74	0.69	1.78	0.97	**0.73**	0.83	0.98	1.53	1.33	1.44	0.66	0.34	1.49	1.71	0.65	0.59	0.88	0.83	1.05
3	1.74	0.76	0.82	1.27	1.01	0.77	1.39	1.02	1.41	1.43	1.28	0.99	**0.35**	1.67	1.13	0.61	0.69	0.80	0.82	0.94

H score = 3.84

(B, A)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
−3	0.63	1.62	0.70	0.68	0.94	1.47	1.26	1.22	1.10	0.90	0.70	1.23	**0.72**	0.74	0.96	1.24	1.09	1.50	0.98	1.08
−2	0.68	0.96	0.81	0.70	1.50	1.52	0.82	1.39	0.95	1.02	0.56	0.56	0.84	0.57	0.94	0.85	1.19	1.76	1.02	**1.23**
−1	0.60	1.29	0.65	0.55	1.26	1.20	1.20	2.09	0.84	0.86	1.00	0.84	0.75	0.61	1.12	0.82	1.25	**1.75**	1.35	1.34
1	0.61	1.36	0.73	0.71	1.52	0.91	1.51	1.98	0.69	0.81	1.07	0.84	**0.85**	0.61	0.56	1.04	1.51	1.98	1.04	1.61
2	0.61	1.03	1.07	0.55	1.34	**0.95**	1.39	1.39	0.51	1.02	0.96	1.25	1.17	0.65	0.44	1.34	1.33	1.66	1.56	1.25
3	0.63	1.33	1.24	0.68	0.95	1.47	0.91	1.01	0.66	0.80	0.87	0.92	**1.39**	0.48	1.04	1.51	1.07	1.45	1.77	0.82

B score = 6.89

up from 3 to 65 residue lengths (i.e. $m = i-32$ to $i+32$) to see what effect amino acid type in the nearest-neighbor region has on the predictions. This also helped us to set a length around the amino acid to be used to achieve the best possible predictions.

The percentage of correct predictions in a protein is defined as the ratio of the total number of correct predictions to the total number of amino acids in the protein sequence multiplied by 100. In the case of individual SS, for example in helices, it is the ratio of the total number of residues correctly predicted to be in the helix conformation and the total number of residues in the protein that are in the helix conformation that is multiplied by 100.

From the plots in Fig. 1 it is observed that predictions from simple propensity values are highest for six near-neighbor residues with a maximum predictability of only 49.39% (Fig. 1a). For the other two calculations the percentage of predictions increases with the increase in sequence length. The highest SS predictions of 70.47% and 65.80% are observed from DP and TP values, respectively, at $m=65$ (Fig. 1b, c). The WS of these three individual scores, defined as an overall score (Fig. 1d) gives the SS prediction as 71.96%, which surprisingly is the highest of the three individual characteristics.

The predictions (Table 1) for individual SS type using AP values gives the highest predictions of 54.43% ($m=5$), 48.64% ($m=17$) and 48.67% ($m=5$) for C, H and B, respectively. Using DP, the highest SS predictions are 65.92%, 74.22% and 69.96% for C, H, and B, respectively at $m=65$. Using TP at $m = 65$ the highest prediction of 73.13% for H and 65.69% for B are obtained while the highest for C is 57.43% at $m = 3$ residues. The overall predictions using WS values are 65.39% for C, 77.02% for H and 71.74% for B at $m = 65$. Among the three SS types the percentage prediction is the highest for H with DP as well as TP, and it is highest for C with AP values.

We have tested the WS predictions on data set 2 which has 190 proteins. The average prediction has come down to 66.6% (Table 1). We have also done a jack-knife test on predictions on data set 1 by replacing the protein that is being used for prediction with the closest homologue of that protein. This again gives a 66% prediction. If we do not include the closest homologue in the jack-knife test the average prediction drastically comes down to 56.6% (Table 1).

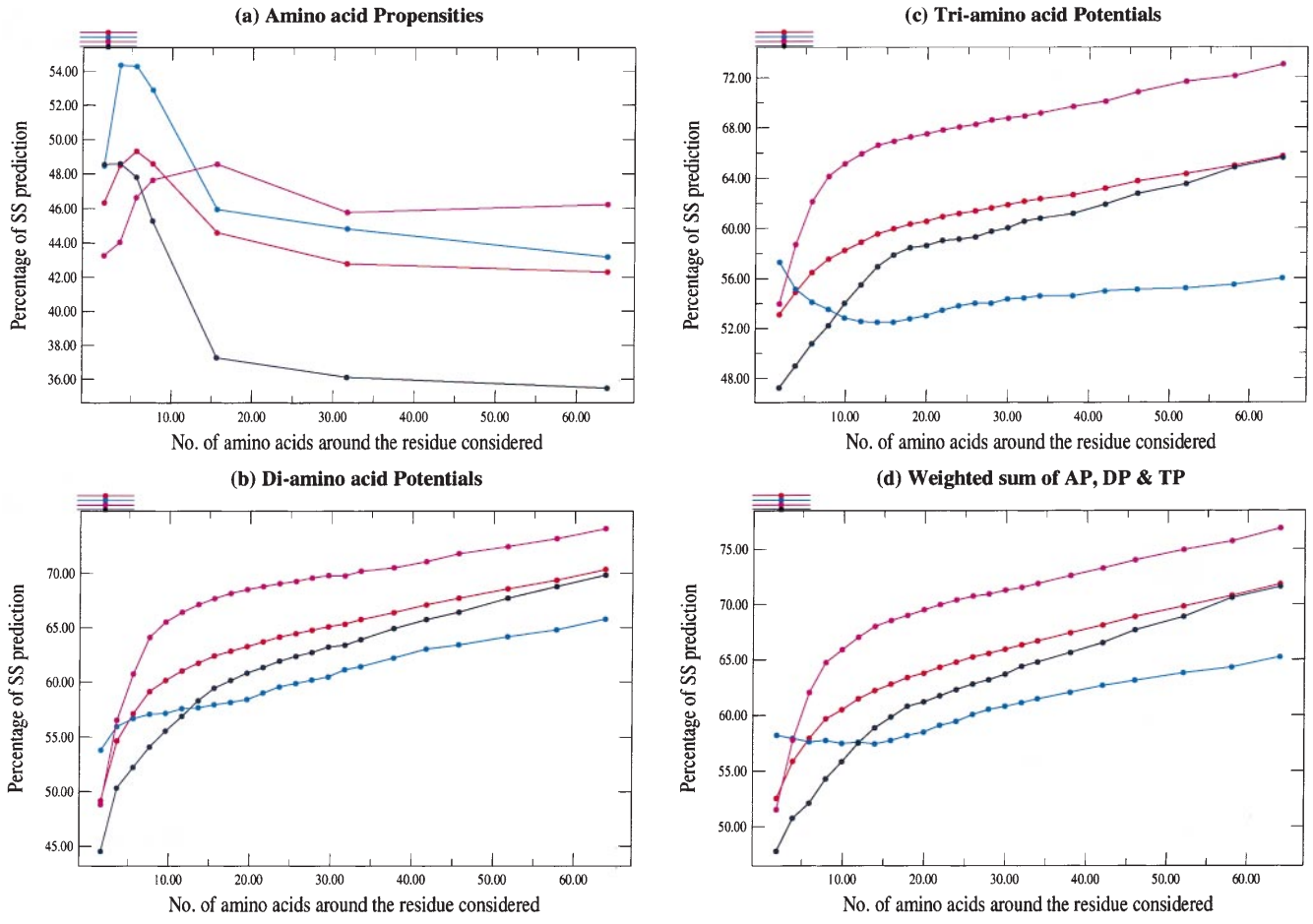

Fig. 1. Percentage of secondary structure (SS) predictions using all three types of measures, namely, **a** amino acid propensities, **b** di-amino acid potentials and **c** tri-amino acid potentials are given. Predictions from the weighted sum of these three characteristics is also shown **d**. The *xy*-plots for prediction of different SS types, namely, coil (*blue*), helix (*pink*), β-strand (*black*) and overall prediction (*red*) are shown for each characteristic

4 Discussion

From the method described it is clear that the SS taken by a residue at the *i*th position in a sequence is assessed based on the following factors:

1. The amino acid type at the *i*th position.
2. The effect of amino acids present at a distance of $i \pm n$ positions on the amino acid type at the *i*th position in its near-neighbor region.
3. The effect of di-peptides in the near-neighbor region at a distance $i \pm n$.

We observed that the percentage of SS prediction for individual cases, using AP, DP and TP, is lower than the WS of these predictions. This indicates that the contribution from each of the above-mentioned factors has a significantly high complementary effect in determining the SS of the amino acid.

In the final prediction approach, we took the WS of each score in a particular SS type taken at the highest overall prediction achieved by AP, DP and TP values. The WS gave 72% correct prediction accuracy. For some individual proteins the predictions are much better than this. In fact about 56% of proteins have a prediction level above 72% and 34 (11%) sequences have predictions even above 80%.

However, in cross-validation tests such as a blind test on data set 2 and the jack-knife test within data set 1, with and without replacing the protein subjected to test prediction by a closest homologue, the prediction accuracy decreases significantly (≥6%, see Table 1). This may be because of the non-homologous nature of the proteins in the sample data set. Each of the proteins in sample data set 1, that generates the weight matrices, has very important statistical information which is not supplemented by the other protein structures in the data set because of their non-homologous nature. We observe that the DP and TP matrices show larger prediction values than the AP matrices. Normalization of each individual protein sequence in the data set with all its available closest homologues should give rise to improved predictions in any cross-validation tests.

An analysis of proteins which have high percentages of prediction shows that these are structures dominated by one kind of SS such as all-α or all-β classes of proteins. The problem with proteins dominated by a particular structural motif, however, is that other structures seem to be almost completely wrongly predicted. For

example, if a protein is nearly all-H, few Cs and turns are necessary to stabilize the overall structure. In these proteins amino acids in some regions may have sufficient propensity to form rigid regular structures (H/B) but these are forced to form required turns/Cs for overall stability of the protein. In such cases nature may be choosing the sequence regions having weak propensity to continue as rigid SSs to form the Cs and turns. We need to improve the method which takes care of such regions of protein structures.

5 Concluding remarks

Our further interest is to use such a statistical analytical approach with respect to the sequence of amino acids and structural environment of the residues in known protein structures. We need to normalize each of the non-homologous proteins with total mono-, di- and tri- amino acid occurrences in all the corresponding closest homologues. We plan to use solvent accessibility type, hydrogen bonding, packing density and sequence variability as the amino acid residue environment-dependent parameters. The strategy to be followed is to define amino acid residue SS type as a function of these structural environment-dependent parameters. Thus, there is scope to increase the prediction accuracy by incorporating these structural environment-dependent parameters, sequence alignment data and nearest-neighbor algorithms.

In summary, we described a statistical analytical approach to predict SS from sequences of amino acids.

1. The database of 306 non-homologous best-resolved structures has been used to calculate properties of SSs.
2. These calculations take the effect of a near-neighbor sequence of amino acids into account in the form of propensities and potentials.
3. Overall prediction using the WS of individual predictions, from AP, DP and TP, gives about 72% prediction accuracy on average on the sample data. There is decrease of about 6% on cross-validation tests which may be improved by using more representative proteins with each of the proteins from the non-homologous data set. The individual characteristics, AP, DP and TP have significant complementary information to one another.
4. Since the present method does not use many other prediction-improving factors, such as residue structural environments, residue variability from sequence alignments and nearest-neighbor algorithms, there is a scope for improvement of prediction accuracy.

Acknowledgements. S.T. is the recipient of a fellowship under a DST grant. We acknowledge EBI for access to their databases and for a user account at EBI. We are grateful to the CCMB Bioinformatics facility. This work is supported by a grant (SP/SO/D-01/95) from the Department of Science and Technology, India.

References

1. Barton GJ (1995) Curr Opin Struct Biol 5: 372
2. Eisenhaber F, Persson B, Argos P (1995) Crit Rev Biochem Mol Biol 30: 1
3. Frishman D, Argos P (1997) Fold Des 2: 159
4. Chou PY, Fasman GD (1974) Biochemistry 13: 222
5. Garnier J, Osguthorpe DJ, Robson B (1978) J Mol Biol 62: 613
6. Rost B (1996) Methods Enzymol 266**:** 525
7. Benner SA, Jenny TF, Cohen MA, Gonnet GH (1994) Adv Enzyme Regul 34: 269
8. DiFrancesco V, Garnier J, Munson PJ (1997) J Mol Biol 267: 446
9. Salzberg S, Cost S (1992) J Mol Biol 227: 371
10. Rost B, Sander C (1993) J Mol Biol 232: 584
11. Rost B, Sander C (1994) Proteins 19: 55
12. Levin JM, Pascarella S, Argos P, Garnier J (1993) Protein Eng 6: 849
13. Wako H, Blundell TL (1994) J Mol Biol 238: 693
14. Mehta PK, Heringa J, Argos P (1995) Protein Sci 4: 2517
15. Salamov AA, Solovyev VV (1995) J Mol Biol 247: 11
16. Geourjon C, Delage G (1995) Comput Appl Biosci 11: 681
17. DiFrancesco V, Garnier J, Munson PJ (1996) Protein Sci 5: 106
18. Leszek R, Godzik A (1997) Protein Eng 10: 1143
19. Thompson MJ, Goldstein RA (1997) Protein Sci 6: 1963
20. Rychlewski L, Godzik A (1997) Protein Eng 10: 1143
21. Russell RB, Barton GJ (1993) J Mol Biol 234: 951
22. Colloch N, Etchebest C, Thoreau E, Henrissat B, Mornon JP (1993) Protein Eng 6: 377
23. Jenny TF, Benner SA (1994) Biochem Biophys Res Commun 200: 149
24. Kolaskar AS, Reddy BVB (1985) Nucleic Acids Res 13: 185
25. Reddy BVB, Deshpande M, Pandit MW (1991) In: Held KD, Brebbia CA, Ciskowski RD (eds) Computer prediction of splice sites in human genome. Computers in Bio-medicine. Computational Mechanics, Billerica, Mass, USA, pp 47–60
26. Reddy BVB, Pandit MW (1995) J Biomol Struct Dyn 12: 785
27. Guruprasad K, Reddy BVB, Pandit MW (1990) Protein Eng 4: 155
28. Reddy BVB (1996) J Biomol Struct Dyn 14: 201
29. Reddy BVB, Ramesh P, Tiwari S (1998) Bioinformatics 14: 225
30. Hobohm U, Scharf M, Schneider R, Sander C (1992) Protein Sci 1: 409
31. Smith D (1989) SSTRUC: a program to calculate secondary structural summary. Birkbeck College, University of London
32. Kabsch W, Sander C (1983) Biopolymers 22: 2577