PROTEINS: Structure, Function, and Bioinformatics 55:83–90 (2004)

Prediction of α-Turns in Proteins Using PSI-BLAST Profiles and Secondary Structure Information

Harpreet Kaur and G.P.S. Raghava*
Institute of Microbial Technology, Chandigarh, India

ABSTRACT **In this paper a systematic attempt has been made to develop a better method for predicting α-turns in proteins. Most of the commonly used approaches in the field of protein structure prediction have been tried in this study, which includes statistical approach "Sequence Coupled Model" and machine learning approaches; i) artificial neural network (ANN); ii) Weka (Waikato Environment for Knowledge Analysis) Classifiers and iii) Parallel Exemplar Based Learning (PEBLS). We have also used multiple sequence alignment obtained from PSIBLAST and secondary structure information predicted by PSIPRED. The training and testing of all methods has been performed on a data set of 193 non-homologous protein X-ray structures using five-fold cross-validation. It has been observed that ANN with multiple sequence alignment and predicted secondary structure information outperforms other methods. Based on our observations we have developed an ANN-based method for predicting α-turns in proteins. The main components of the method are two feed-forward back-propagation networks with a single hidden layer. The first sequence-structure network is trained with the multiple sequence alignment in the form of PSI-BLAST-generated position specific scoring matrices. The initial predictions obtained from the first network and PSIPRED predicted secondary structure are used as input to the second structure–structure network to refine the predictions obtained from the first net. The final network yields an overall prediction accuracy of 78.0% and MCC of 0.16. A web server AlphaPred (http://www.imtech.res.in/raghava/alphapred/) has been developed based on this approach. Proteins 2004;55:83–90.**
© 2004 Wiley-Liss, Inc.

Key words: neural networks; multiple alignment; tight turns; web server; weka; pebls

INTRODUCTION

For several decades, the protein tertiary structure prediction has been among the most challenging problem in the biological sciences. In the absence of homologous structures, an intermediate and useful step to solve protein tertiary structure prediction problem is to predict the protein secondary structure. The secondary structure of a protein is characterized by regular elements such as α-helices and β-sheets and irregular elements such as β-bulges, random coils and tight turns.[1] Tight turns provide a directional change for the polypeptide chain and act as a linking motif between different secondary structures. They often contain charged and hydrophilic residues and are thus mostly located on the surface of a protein. Depending on the number of residues forming the turn, the tight turns are classified as δ-turns, γ-turns, β-turns, α-turns and π-turns.[1] Among these tight turns, β- and γ-turns have been studied in detail and precisely classified in past. In comparison to β- and γ-turns, α-turns are little investigated due to their lower occurrence in proteins and peptides. The α-turn corresponds to a chain reversal involving five amino acids and may be stabilized by a hydrogen bond between the CO group of the first residue and the NH group of the fifth.

In 1996, Pavone et al. had undertaken a systematic search of isolated α-turns in a data set of 193 proteins and compared structures of different types of α-turns using a clustering procedure. This study has revealed that these structures are mainly characterized by hydrophilic amino acids. It has also been shown that these structures are not only exposed to solvent, but also protrudes outward from the protein surface with a hook-like shape and therefore these structural motifs can function in interaction mechanisms.[2]

It has been shown in past that α-turns have a functional role in molecular recognition and protein folding.[3–6] For instance, it was found that the residues in the α-turn in protein human lysozyme participate in a cluster of hydrogen bonds and they are located in the active site cleft suggesting the possibility of a functional role.[3] Cys residue in α-turn (residues 35–39) in protein Ferredoxin I and His and Met residues in α -turn (residues 117–121) in protein azurin are involved in the metal ion coordination.[4,5] In T-cell surface glycoprotein, residues 51–55 forming an α-turn are located in the putative binding region of HIV gp120 protein.[6] Moreover, α-turns are also relevant structural domains in small peptides, particularly in cyclopeptides containing 7–9 residues in their sequence.[7] Thus, these rarely occurring motifs whenever present on the protein surface might contain specific information about

This report has IMTECH communication No. 020/2003.

*Correspondence to: G.P.S. Raghava, Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India. Email: raghava@imtech.res.in

Received 25 April 2003; Accepted 14 June 2003

© 2004 WILEY-LISS, INC.

molecular recognition processes. Therefore, there is a need and also it will be useful to develop a prediction method for detecting α-turn residues in a given amino sequence.

One statistical method based on the Sequence Coupled Model has been reported in literature for prediction of α-turns.[8] In order to develop a highly accurate method for predicting α-turns, we have applied the different commonly used techniques in the field of protein secondary prediction. Firstly, we have implemented the Sequence Coupled Model as described by Chou (1997) using five-fold cross-validation. Further, three machine learning packages, ANN using SNNSv4.2 (Stuttgart Neural Network Simulator),[9] Weka3.2[10] and PEBLS[11] have been used. It has been demonstrated in past that secondary structure information and multiple sequence alignment (rather than single sequence) increases the performance of turn prediction methods significantly (e.g., β- and γ-turns prediction).[12,13] Thus, all the methods have been trained and tested using PSI-BLAST[14] profiles and PSIPRED[15] predicted secondary structure. It has been found that neural network method has the highest accuracy level in comparison to Sequence Coupled Model and the machine learning methods Weka and PEBLS. Based on this study, a neural network method AlphaPred for α-turn prediction has been developed which uses two steps. In the first step, a sequence-to-structure network is used to predict α-turns from multiple alignment of protein sequence. In the second step, it uses the structure-to-structure network where input is predicted α-turns obtained from first step and PSIPRED predicted secondary structure states.

MATERIALS AND METHODS

The Data Set

The method has been developed on the data set clustered by Pavone et al (1996).[2] It is comprised of 193 representative protein X-ray structures with resolution $\leqslant$ 2.5Å. A total of 356 isolated α-turns have been extracted. All secondary structures, other than α-turns have been picked up from the same 193 proteins and have been treated as non-α-turn data set.

Since the α-turns are often associated with β-turns, the dataset of α-turns have been analyzed for the presence of β-turns. The β-turns have been assigned using Promotif[16] program. Out of 356 α-turns, 100 have been found to contain β-turns and mostly there are two β-turns per α-turn with at least one overlapping residue.

5-Fold Cross Validation

A 5-fold cross-validation procedure has been used to develop the prediction method, where five subsets have been constructed randomly from the data set. Each set is an unbalanced set which retains the naturally occurring proportion of α-turns and non-α-turns. The datasets for Sequence Coupled Model, Weka and PEBLS consists of a training set and a testing set. The training set consists of four subsets and the testing is done on the remaining fifth set. The learning method is applied on the training set and its performance is tested on the testing set. To avoid over-training by neural network, the datasets for SNNS neural network learning consists of a training set, a validation set and a testing set. A training set is consisted of three subsets. The performance of the network has been monitored on validation set to avoid over-learning and the network is tested on the excluded set of proteins, the testing set. For all the methods, the prediction has been done five times to test for each subset and the final prediction results have been averaged over five testing sets.

Statistical Method: Sequence Coupled Model

Chou (1997)[8] proposed a residue-coupled model based on first order Markov chain to predict α-turns in proteins. The same approach has been applied here for α-turn prediction using 5-fold cross-validation. For detailed methodology, please refer to Chou (1997).

Machine Learning Method: Artificial Neural Network

In the present study, two feed forward back propagation networks with a single hidden layer have been used. The window size and the number of hidden units have been optimized. Both the networks have window eleven residues wide and have 10 units in a single hidden layer. The target output consists of a single binary number and is 1 or 0 for α-turn and non-α-turn residue respectively. The window is shifted residue by residue through the protein length, thus yielding N patterns for a protein with N residues. The neural network has been implemented using the publicly available free simulation package SNNS version 4.2.[9] It allows incorporation of the resulting networks into an ANSI C function for use in stand-alone code. A linear activation function is used. At the start of each simulation, the weights are initialized with the random values. The training is carried out using error-backpropagation with a sum of square error function (SSE).[17] The magnitude of the error sum in the test and training set is monitored in each cycle of the training. The ultimate number of cycles is determined, where the network converges. An overview of each network is given below.

First level—Sequence-to-Structure ANN

The input to the first network is either single sequence or multiple alignment profiles. Patterns are presented as window of eleven residues where a prediction is made for the central residue. In case of single sequence, binary encoding scheme has been used as input. In this scheme, each amino acid is coded by the binary vector (1, 0, 0,$\cdots$) or (0, 1, 0,$\cdots$), etc. The vector is 21-dimensional. Among the first twenty units of the vector, each unit stands for one type of amino acid. In order to allow a window to extend over the N terminus and the C terminus, the 21st unit has been added for each residue. In case of multiple alignment, the position specific scoring matrix generated by PSI-BLAST has been used as input to the net. The matrix has $21 \times M$ real number elements, where M is the length of the target sequence.

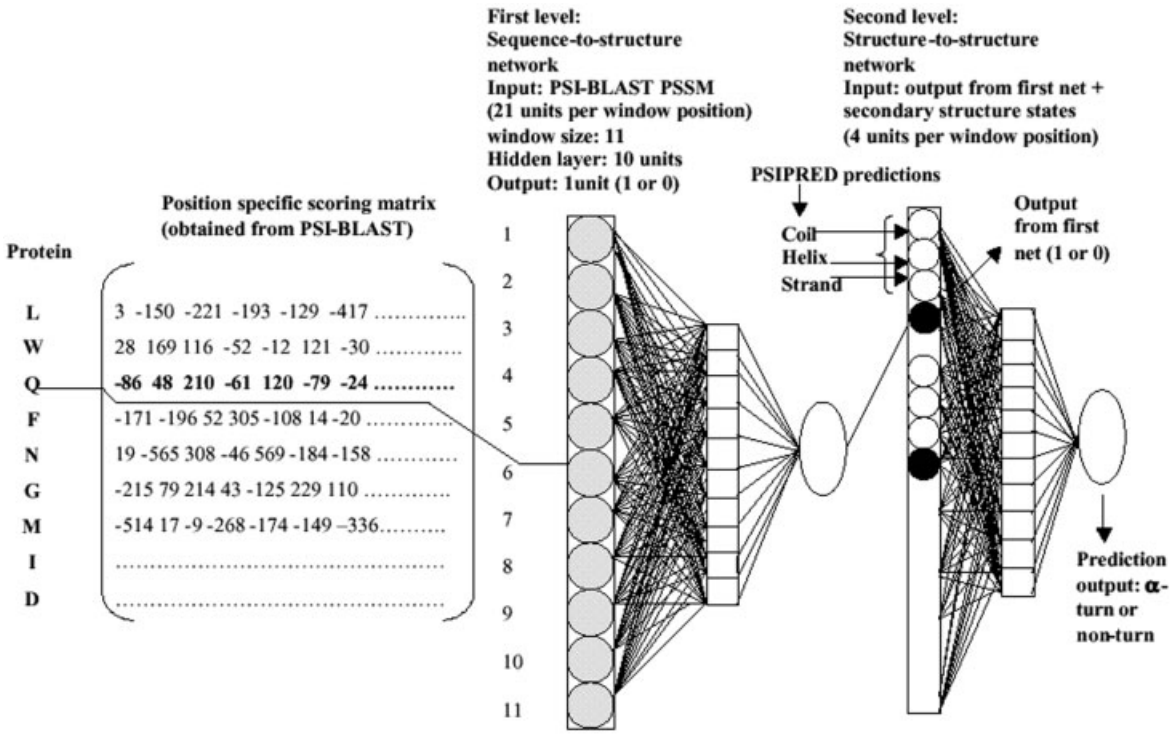

Fig. 1. The neural network system used for α-turn prediction. The network system consists of two networks: first level sequence-to-structure network and second level structure-to-structure network. ◍ Basic cell containing 20+1 units to code residues at that position in the window. Here, window size = 11. ▤ Hidden layer containing 10 units. In second level network, 4 units encode each residue. ● prediction obtained from first network ○ secondary structure state (helix, strand and coil) predicted byPSIPRED.

Second level—Structure-to-Structure ANN

The second network takes the outputs from first network and yields the final prediction based on these outputs. The input to second filtering network is prediction obtained from the first net and the secondary structure predicted by PSIPRED. Four units encode each residue where one unit codes for output from first net and remaining three units are the reliability indices of three PSIPRED predicted secondary structure states-helix, strand and coil (Fig. 1).

Machine Learning Method: Weka

The machine learning package Weka3.2 is a collection of machine learning algorithms such as ZeroR, OneR, Naïve Bayes, Logstic Regression, Linear regression, LWR, J48, decision table for classification and numeric prediction.[10] The data for Weka is represented in ARFF (attribute-relation function format) format that consists of a list of all instances, with the attribute values for each instance being separated by commas. The results from the Weka consists of a confusion matrix for both the training and testing set showing the number of instances of each class that have been assigned to each class. In this study the following three classifiers of Weka have been used: i) Logistic Regression[18] which is a variation of ordinary regression and particularly useful when the observed outcome is restricted to two values; ii) NaïveBayes[19] algorithm that implements Bayesian classification based on Bayes' theorem of conditional probability and iii) J4820 classifier based on C4.5 algorithm that generates a classification-decision tree for the given data set by recursive partitioning of data. As data in this study is highly unbalanced (non turns much more than α-turns), so we have used Weka's cost-sensitive classification option in which the data sets have been weighted according to the distribution of α-turns and non-α-turns and penalties have been assigned to each class (α-turn/non-α-turn) in the cost matrix. The penalties have been optimized by learning the classifier several times.

Machine Learning Method: Example-Based Learning

PEBLS is a nearest-neighbor learning system designed for applications where the instances have symbolic feature values.[11] It treats a set of training examples as points in a multi-dimensional feature space. Test instances are then classified by finding the closest exemplar currently contained in the feature space. The nearest neighbors or exemplars are determined by computing the distance to each object in the feature space using modified value distance metric (MVDM) based on the original value distance metric of Stanfill and Waltz.[21] These neighbors are then used to assign a classification to the test instance.

Multiple Sequence Alignment and Secondary Structure

PSIPRED uses PSI-BLAST to detect distant homologues of a query sequence and generate a position-specific scoring matrix as part of the prediction process.[15] These intermediate PSI-BLAST-generated position-specific scoring matrices are used as input in our method in cases where multiple sequence alignment is used. The matrix has 21× M elements, where M is the length of the target sequence and each element represents the frequency of occurrence of each of the 20 amino acids at one position in the alignment.[14] The secondary structure predicted by PSIPRED has been used in prediction. The predicted secondary structure states from PSIPRED are used to filter the α-turn prediction in the case of the Sequence Coupled Model and input for structure-to structure network, Weka classifiers, and PEBLS. In the final prediction, output has been filtered, where the α-turns are predicted only for those residues that are in the predicted coil region, i.e., eliminating the potential helix and strand-forming residues from α-turn prediction.

Filtering the Prediction

Since the prediction is performed for each residue separately, thus prediction includes a number of unusually short α-turns of 1, 2, 3, or 4 residues. To exclude such unrealistic turns in final prediction, we have retained only those turns that are five residues long.

Performance Measures

Threshold dependent measures

Four parameters have been used in present work to measure the performance of α-turn prediction method as described by Shepherd et al.[22] These four parameters can be derived from the four scalar quantities: p(number of correctly classified α-turn residues), n(number of correctly classified non-α-turn residues), o(number of non-α-turn residues incorrectly classified as α-turn residues) and u(number of α-turn residues incorrectly classified as non-α-turn residues). Another way to visualize and arrange these four quantities is to use a contingency or confusion matrix C

$$C = \begin{pmatrix} p & u \\ o & n \end{pmatrix}$$

The four parameters that can be derived from these four quantities are: i) Q*total* (or prediction accuracy), is the percentage of correctly classified residues; ii) Matthew's correlation coefficient (MCC), which is a more stringent measure of prediction accuracy accounts for both over and under-predictions; iii) Q*predicted* is the percentage of correctly predicted α-turns (or probability of correct prediction); and iv) Q*observed* is the percentage of observed α-turns that are correctly predicted (or percent coverage). The parameters can be calculated by following equations:

$$\mathrm{Q}total = \frac{p+n}{t}$$

$$\mathrm{MCC} = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}}$$

$$\mathrm{Q}predicted = \frac{p}{p+o} \times 100$$

$$\mathrm{Q}observed = \frac{p}{p+u} \times 100$$

where, t = p + n + o + u is the total number of residues. All these performance measures have been calculated at residue level, not at whole turn level.

Segment Overlap Measure (SOV)

The single-residue predictions do not completely reflect the quality of a prediction. It is important to estimate the number of α-turns and their lengths (number of residues in predicted α-turns). To address the overlapping between the observed and predicted α-turns, SOV[23] has been calculated as:

$$Sov = \frac{1}{N} \sum_{s} \frac{\min ov(S1;S2) + \delta}{\max ov(S1;S2)} \times \mathrm{len(S1)}$$

where S1 and S2 are the observed and predicted α-turns; len(S1) is the number of residues in the segment S1; minov(S1;S2) is the length of actual overlap of S1 and S2 or the extent for which both segments have residues in α-turn; maxov(S1;S2) is the length of the total extent for which either of the segments S1 or S2 has a residue in α-turn; δ is the integer value defined as equal to the min{(maxov(S1;S2)-minov(S1;S2));minov(S1;S2); int(len(S1)/2);int(len(S2)/2)}; N is the number of residues in α-turn and sum is taken over all the pairs of segments {S1;S2}, where S1 and S2 have at least one residue in α-turn in common.

Threshold independent measures

One problem with the threshold dependent measures is that they measure the performance on a given threshold. They fail to use all the information provided by a method. The Receiver Operating Characteristic (ROC) is a threshold independent measure that was developed as a signal processing technique. For a prediction method, ROC plot is obtained by plotting all sensitivity values (true positive fraction) on the y-axis against their equivalent (1-specificity) values (false positive fraction) for all available thresholds on the x-axis. The area under the ROC curve is taken as an important index because it provides a single measure of overall accuracy that is not dependent on a particular threshold.[24] It measures discrimination, the ability of a method to correctly classify α-turn and non-α-turn residues. Sensitivity (Sn) and specificity (Sp) are defined as

$$\mathrm{Sn} = \frac{p}{p+u} \text{ and } \mathrm{Sp} = \frac{n}{n+o}$$

TABLE I. Results of α-Turn Prediction Methods, When Single Sequence Has Been Used as Input. The Performance Is Averaged Over Five Test Sets†

Method	*Qtotal*	*Qpred.*	*Qobs.*	MCC
Sequence coupled model	52.6 ± 0.8	2.6 ± 0.4	40.4 ± 2.4	0.03 ± 0.01
	(57.8 ± 1.9)	(5.9 ± 0.6)	(43.2 ± 2.4)	(0.05 ± 0.01)
SNNS (Std. Backpropagation)	61.8 ± 4.0	6.5 ± 0.4	51.1 ± 6.7	0.06 ± 0.01
	(70.3 ± 2.5)	(7.4 ± 0.6)	(60.8 ± 5.2)	(0.13 ± 0.01)
Weka (Logistic Regression)	63.5 ± 1.8	7.4 ± 0.5	57.0 ± 2.0	0.09 ± 0.01
	(72.7 ± 0.8)	(7.6 ± 0.7)	(58.5 ± 2.3)	(0.13 ± 0.01)
Weka (Naïve Bayes)	67.9 ± 1.5	7.7 ± 0.6	51.8 ± 4.7	0.09 ± 0.01
	(64.7 ± 2.2)	(6.8 ± 0.5)	(69.0 ± 1.6)	(0.13 ± 0.01)
Weka (J48 classifier)	87.2 ± 0.5	7.8 ± 1.1	15.1 ± 2.2	0.04 ± 0.01
	(89.2 ± 0.4)	(7.8 ± 0.7)	(17.7 ± 1.4)	(0.07 ± 0.01)
PEBLS	90.9 ± 0.6	10.5 ± 1.3	11.1 ± 0.6	0.06 ± 0.01
	(93.3 ± 0.6)	(11.5 ± 2.0)	(11.4 ± 0.4)	(0.08 ± 0.01)

†Values in parentheses corresponds to the performance of α-turn prediction methods, when secondary structure information obtained from PSIPRED is also used.

RESULTS

All the methods have been trained and tested using five-fold cross-validation. The prediction performance measures have been averaged over five sets and are expressed as the mean ± standard deviation. In the case of the Sequence Coupled Model, the conditional probabilities have been calculated for α-turns and non-turns separately as described in Materials and Methods. In Weka, three classifiers have been used and the penalties for the cost matrix have been optimized. For PEBLS, we have varied the number of neighbors from single to multiple (1 to 30), and it has been found that single nearest neighbor gives best results. Hence, the result for PEBLS shown here is based on single nearest neighbor.

Prediction With Single Sequence

The prediction results of various methods with single sequence as input are shown in Table I. As can be seen that the probability of correct prediction, *Qpred* is in the range 2–10% and is indeed poor for all methods. Among all the methods, the best performance has been obtained with Weka classifiers Logistic Regression and Naïve Bayes. Surprisingly, these classifiers have also outperformed neural network method SNNS. The average MCC value with SNNS is 0.06, in comparison to Weka classifiers reaching a MCC of 0.09. The performance of PEBLS differ significantly from all other methods with respect to both the probability of correct prediction and coverage of α-turns and is much lower. When using the Sequence Coupled Model, a MCC of 0.03 has been obtained which is least among all other methods.

Prediction With Single Sequence and Secondary Structure

We next wanted to know whether including secondary structure information in prediction would be beneficial or not. We therefore used the secondary structure predicted from PSIPRED. In case of machine learning methods, secondary structure states predicted by PSIPRED have been used along with the prediction outputs (obtained from single sequence) as input to structure-to-structure net, Weka classifiers and PEBLS. The performance of all methods after incorporating secondary structure information has been improved (Table I). When applying a fivefold cross-validation test, the SNNS reached an overall accuracy of 70.3% and MCC value is increased from 0.06 with single sequence to 0.13. The best results have been achieved with SNNS and Weka classifiers Logistic regression and Naïve Bayes, having prediction accuracy around 70–72% and all having MCC value equal to 0.13. An improvement in all other measures can also be noticed after including secondary structure. As can be seen, SNNS and Weka classifiers have much better performance compared to PEBLS and Sequence Coupled Model. Even after including secondary structure, PEBLS and Sequence Coupled Model did not work well.

Prediction With Multiple Alignment

It is known that multiple sequence alignment rather than single sequence, improves prediction accuracy. Thus, all the machine learning algorithms have been trained and tested on PSI-BLAST-generated position-specific matrices. It is apparent from the results presented in Table II that there is a significant gain in prediction accuracy for all methods. The maximum MCC value of 0.13 has been obtained with Weka classifier Logistic Regression and is better than SNNS. With SNNS, the MCC improves from 0.06 with single sequence to 0.09. Thus, substantial improvements in prediction performance have come from the use of PSI-BLAST scoring matrices in preference to binary encoding of single sequence. Overall, the results of PEBLS and Weka classifier J48 are inferior in comparison to other methods.

Prediction With Multiple Alignment and Secondary Structure

Accuracy is further improved by using a second filtering network and secondary structure information. Output from the first step (trained on PSI-BLAST scoring matrices) and secondary structure predicted from PSIPRED is

TABLE II. Performance of SNNS, Weka and PEBLS Classifiers Using Multiple Alignment and Secondary Structure Information

	Multiple alignment					Multiple alignment and secondary structure				
		Weka classifiers					Weka classifiers			
	SNNS (first network)	Logistic Regression	Naïve Bayes	J48 Classifier	PEBLS	SNNS (second network)	Logistic Regression	Naïve Bayes	J48 Classifier	PEBLS
Qtotal	68.1	66.0	63.4	89.8	91.0	78.0	74.7	62.5	90.9	91.2
Qpred.	8.5	8.6	8.0	10.1	11.1	9.4	8.0	6.4	8.2	10.2
Qobs.	53.8	62.5	63.5	14.3	12.3	55.5	55.1	67.2	14.8	12.5
MCC	0.09	0.13	0.12	0.07	0.07	0.16	0.13	0.12	0.06	0.07

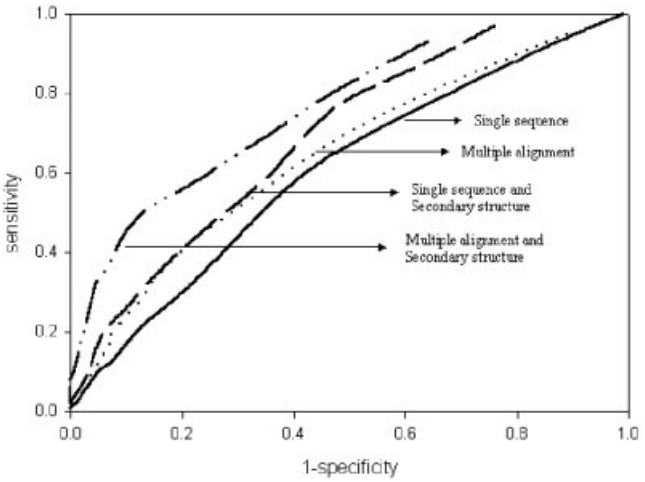

Fig. 2. ROC curves for four different neural network systems. Solid line indicates single sequence; dotted line indicates multiple alignment; dashed line indicates single sequence with secondary structure; and solid/dotted line indicates multiple alignment with secondary structure.

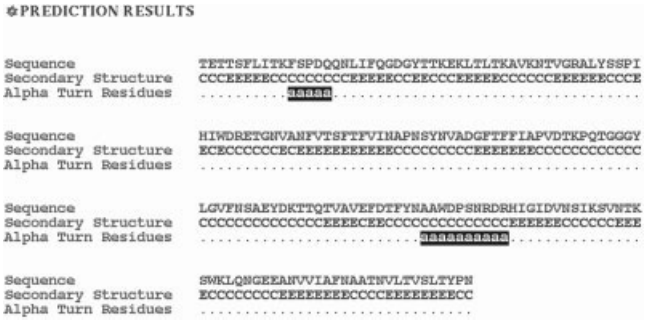

Fig. 3. Sample α-turn/non-turn predictions by AlphaPred server. Predictions are made using multiple alignment and secondary structure information. For each block, row 1 is the amino acid sequence, row 2 is the secondary structure predicted by PSIPRED (H = helix, E = strand, and C = coil) and row 3 is the predicted α-turns (denoted by 'a') and non-turns (denoted by '.').

further trained. A tremendous improvement can be noted especially for SNNS method (Table II). Use of secondary structure information improves the MCC to 0.16. The method shows segment overlap measure SOV = 42.4%, which is a more realistic assessment of prediction quality. The overall per-residue accuracy is 78% and is the best among all the methods tested. For Weka classifiers, a marginal increase in prediction accuracy has been observed but there is no change in MCC value. The same is true for PEBLS also, which shows negligible improvement after incorporating secondary structure information.

Receiver Operating Characteristic Results

By calculating the area under the receiver operating characteristic (ROC) curve, the performance of different networks has been assessed. Figure 2 shows the ROC curves for four different networks. The four curves have been compared by computing the area under the curves. The corresponding area under the curves is as follows: single sequence, 0.61; multiple alignment 0.70; single sequence with secondary structure 0.72; and multiple alignment with secondary structure, 0.80. These ROC values reflect the better classification by network system learned on PSI-BLAST profiles and secondary structure information.

AlphaPred Server

Based on current work, a web server AlphaPred has been developed to predict α-turns from a given primary amino acid sequence. Users can enter the one-letter amino acid sequence in fasta or plain text format. The output consists of secondary structure predicted by PSIPRED and predicted α-turn or non-α-turn residues. A sample of the prediction output has been shown in Figure 3.

DISCUSSION AND CONCLUSION

Irregular protein secondary structures are believed to be important structural domains involved in molecular recognition processes and protein folding. In this respect, tight turns are being studied in detail. Isolated α-turns, not participating in α-helical structures, have been studied little in comparison to other types of tight turns because α-turns present in proteins are sparse. In the past, a systematic study and classification of isolated α-turns into different types based on conformational similarity has been reported. In fact, on average, about two isolated α-turns were identified in each protein analyzed. The identification and classification of isolated α-turns has indicated the relevance of this irregular secondary structure. The α-turns are characterized by hydrophilic amino acids, thus are present on the protein surface and can function as keys in a lock-key interaction mechanism between proteins. Moreover, these turns provide a connection between extended peptide chains. In several cases, these turns are located at the tip of type 3 and 4 β-hair-

pins.[25] Also, α-turns are important structural domains in cyclo-peptides such as Ilamicinb1[26] and cyclolinopeptide A[27] that have important biological functions. Despite the importance of α-turns only one method based on sequence coupled model has been reported in past for predicting α-turn types in proteins. The average rates of correct prediction of this method by resubstitution test and jackknife test were 98.07% and 94.47% respectively.[1] To the authors best knowledge, this is a first attempt to develop a method using successful techniques and concepts in the field of protein structure prediction.

Initially, the training and testing of all the methods has been done with single sequences alone as well as with PSIPRED predicted secondary structure. Surprisingly, it has been found that the Weka classifiers Logistic regression and Naïve Bayes have better prediction performances than SNNS on single sequences. The MCC achieved by Weka classifiers is 0.09 in comparison to 0.06 of SNNS. PEBLS and Sequence Coupled Model perform poorly among all the methods. The results of Sequence Coupled Models are lower in this study than that reported by Chou (1997)[8] using the same method. This is due to the fact that we have selected a different threshold than Chou (1997). At a particular threshold value, one can achieve a high rate of correct prediction (the same as achieved by Chou, 1997) but at the cost of low MCC value and probability of correct prediction. Moreover, to assess the prediction performance, MCC is a more stringent criterion as compared to rate of correct prediction. The rate of correct prediction can be misleading owing to the disparity between α-turn and non-turn residues; hence it is possible to get high rate of correct prediction by the trivial strategy of predicting all residues to be non-turn residues. Thus, we have shown the results at a particular threshold value where there is a compromise between the rate of correct prediction and MCC value.

When secondary structure information is incorporated, the results of SNNS and Weka classifier Logistic Regression is comparable. One important point that can be noticed is that the Qpred., the probability of correct prediction, is significantly low in all the methods. All the machine learning methods have been trained and tested on PSI-BLAST obtained scoring matrices with and without secondary structure information. The inclusion of multiple sequence alignment information into the prediction scheme has given a significant boost in prediction accuracy of all methods. This is due to the fact that PSI-BLAST profiles have some basic advantages as more distant sequences are found; the probability of each sequence is properly weighted with respect to the amount of information it carries. For SNNS, MCC value increases from 0.06 with single sequence to 0.09 with PSI-BLAST whereas for Weka classifier Logistic Regression, it improves from 0.09 to 0.13 and is the best among all the method tested. Moreover, when secondary structure is used, the neural network and Weka classifiers have a final MCC of 0.16 and 0.13 respectively. Comparing the final results of neural network and Weka classifiers results in favor of the neural network. It is worth noting that the MCC so achieved is not so high and the probability of correct prediction is indeed poor.

One of the reasons for poor performance of the method described here is that the number of α-turns in proteins is very small in comparison to number of non α-turns (ratio of α-turns:non α-turns being 1:26). When a more robust measure of predictive performance is used such as MCC, α-turn prediction method appears far less successful in comparison to helix and sheet prediction methods. The fact that α-turns are very few resulted in poor Qpred. and MCC values in all the methods. The α-turn is a rare structure in contrast to regular secondary structures, helices and β-sheets. The number of helical and sheet residues in proteins are far greater than α-turn residues. Thus, it is possible to achieve high MCC values for helix (0.60) and β-sheet residue (0.52). However, for α-turn prediction, even after inclusion of multiple alignment and secondary structure, the maximum MCC value achieved is only 0.16. Also, when there is an abundance of data belonging to well-defined classes (helices, β-sheets), neural networks perform extremely well. However, neural nets perform poorly when the available data is sparse as the case of α-turn prediction. It is possible to achieve high accuracy by keeping equal number of α-turn and non α-turns for training, which actually reduces the number of false positives and false negatives. However if we move to a real situation where α-turns are present in very small proportion in proteins as compared to other secondary structures, the algorithms learned on balanced data sets would not work well.

In summary, this work is an attempt towards developing highly accurate method for α-turns in proteins. Further improvement of the suggested approach is possible with further elucidation of protein and peptide X-ray structures, which will probably clarify the biological role, and the occurrence of α-turns.

ACKNOWLEDGMENT

The authors are thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Govt. of India for financial assistance. We are also thankful to the developers of SNNS, Weka and PEBLS packages.

REFERENCES

1. Chou KC. Prediction of tight turns and their types in proteins. Analytical Biochem 2000;286:1–16.
2. Pavone V, Gaeta G, Lombardi A, Nastri F, Maglio O, Isernia C, Saviano M. Discovering protein secondary structures: classification and description of isolated α-turns. Biopolymers 1996;38:705–721.
3. Artymiuk PJ, Blake CCF. Refinement of human lysozyme at 1.5 Å resolution analysis of non-bonded and hydrogen-bond interactions. J Mol Biol 1981;152:737–762.
4. Stout CD. Refinement of the 7 Fe ferredoxin from Azotobacter vinelandii at 1.9 Å resolution. J Mol Biol 1989;205:545–555.
5. Baker EN. Structure of azurin from Alcaligenes denitrificans refinement at 1.8 Å resolution and comparison of the two crystallographically independent molecules. J Mol Biol 1988;203:1071–1095.
6. Wang J, Yan Y, Garrett TPJ, Liu J, Rodgers DW, Garlick RL, Tarr JE, Hosain Y, Reinherz EL, Harrison SC. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. Nature 1990;348:411–418.

7. Zanotti G, Wieland T, Benedetti E, Di Blasio B, Pavone V, Pedone C. Structure-toxicity relationships in the amatoxin series. Synthesis of S-deoxy[gamma(R)-hydroxy-Ile3]-amaninamide, its crystal and molecular structure and inhibitory efficiency. Int J Peptide Protein Res 1989;34:222–228.
8. Chou KC. Prediction and classification of α-turn types. Biopolymers 1997;42:837–853.
9. Zell A, Mamier G. Stuttgart Neural Network Simulator version 4.2. University of Stuttgart; 1997.
10. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco: Morgan Kaufmann; 1999.
11. Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning 1993;10:57–78.
12. Kaur H, Raghava GPS. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 2003;12:627–634.
13. Kaur H, Raghava GPS. A neural network based method for prediction of γ-turns in proteins from multiple sequence alignment. Protein Sci 2003;12:923–929.
14. Altschul SF, Madden TL, Alejandro AS, Zhang J, Zhang Z, Miller W, Lipman, DJ. Gapped blast and psi-blast: a new generation of protein databases and search programs. Nucleic Acids Res 1997;25: 3389–3402.
15. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
16. Hutchinson EG, Thornton JM. PROMOTIF: A program to identify and analyze structural motifs in proteins. Protein Sci 1996;5:212–220.
17. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagation errors. Nature 1986;323:533–536.
18. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 1989.
19. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 1997;29:103–130.
20. Quinlan JR. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.
21. Stanfill C, Waltz D. Toward memory-based reasoning. Communications of the ACM 1986;29:1213–1228.
22. Shepherd AJ, Gorse D, Thornton JM. Prediction of the location and type of β-turn types in proteins using neural networks. Protein Sci 1999;8:1045–1055.
23. Zemla A, Venclovas, C, Fidelis, K, Rost B. A modified definition of Sov, a segment based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
24. Deleo JM. Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, IEEE, Computer Society Press, College Park, MD; 1993. p 318–325.
25. Pavone V. On β-hairpin classification. Int J Biol Macromol 1988;10:238–240.
26. Iitaka Y, Nakamura H, Takada K, Takada T. An X-ray study of ilamycin B, a cyclic heptapeptide antibiotic. Acta Crystallogr B 1974; 30:2817–2825.
27. Di Blasio B, Rossi F, Benedetti E, Pavone V, Pedone C, Temussi PA, Zanotti G, Tancredi T. Bioactive peptides: x-ray and NMR conformational study of [Aib 5,6-D-Ala8] cyclolinopeptide A. J Am Chem Soc 1992;114:8277–8283.