W154–W159 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki588

BhairPred: prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques

Manish Kumar, Manoj Bhasin, Navjot K. Natt and G. P. S. Raghava*

Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India

Received December 15, 2004; Revised March 30, 2005; Accepted May 2, 2005

ABSTRACT

This paper describes a method for predicting a super-secondary structural motif, β-hairpins, in a protein sequence. The method was trained and tested on a set of 5102 hairpins and 5131 non-hairpins, obtained from a non-redundant dataset of 2880 proteins using the DSSP and PROMOTIF programs. Two machine-learning techniques, an artificial neural network (ANN) and a support vector machine (SVM), were used to predict β-hairpins. An accuracy of 65.5% was achieved using ANN when an amino acid sequence was used as the input. The accuracy improved from 65.5 to 69.1% when evolutionary information (PSI-BLAST profile), observed secondary structure and surface accessibility were used as the inputs. The accuracy of the method further improved from 69.1 to 79.2% when the SVM was used for classification instead of the ANN. The performances of the methods developed were assessed in a test case, where predicted secondary structure and surface accessibility were used instead of the observed structure. The highest accuracy achieved by the SVM based method in the test case was 77.9%. A maximum accuracy of 71.1% with Matthew's correlation coefficient of 0.41 in the test case was obtained on a dataset previously used by X. Cruz, E. G. Hutchinson, A. Shephard and J. M. Thornton (2002) *Proc. Natl Acad. Sci. USA*, 99, 11157–11162. The performance of the method was also evaluated on proteins used in the '6th community-wide experiment on the critical assessment of techniques for protein structure prediction (CASP6)'. Based on the algorithm described, a web server, BhairPred (http://www.imtech.res.in/raghava/bhairpred/), has been developed, which can be used to predict β-hairpins in a protein using the SVM approach.

INTRODUCTION

Currently available high-throughput sequencing facilities have generated a large amount of raw sequence data, making it possible to know the sequences of an increasing number of proteins. In contrast, there are a limited number of proteins whose structure is known at atomic level. To reduce the ever-widening gap between known sequences and known structure, there is a need to develop accurate methods for protein structure prediction. The available methods for structure prediction can be divided into three categories: (i) knowledge based methods, (ii) *ab initio* methods and (iii) a hierarchical approach. In the knowledge based approach, the rules are first derived from known structures and then these rules are used to predict the structure of proteins from their amino acid sequences (1). The *ab initio* methods attempt to derive the structure from first principles, i.e. without reference to any experimentally determined structure. In the hierarchical approach, first an intermediate structure (such as secondary structure) is predicted from the amino acid sequence of the protein and then this information is used to predict the tertiary structure of the protein. This paper is an attempt to improve the performance of hierarchal methods.

A large number of methods have been developed for predicting the regular secondary structures (α-helix, β-strand) and coils in proteins. These achieve an average accuracy of ∼80% (2–5). The prediction of tertiary structure from secondary structure is the most difficult step for two reasons. First, most of the methods can predict only the regular secondary structures, which cover on average ∼50% of the residues of a protein. Thus, these methods do not provide any information regarding the remaining 50% of the residues, which form irregular structures. Recently attempts have been made to predict irregular secondary structures in a protein that includes α-turns (6), β-turns (7–10) and γ-turns (11). Among β-turns, the type has been successfully predicted (12). Second, the secondary-to-tertiary structure prediction route has not been fully explored yet. In the past few years, methods have been developed to predict tertiary structures from supersecondary

*To whom correspondence should be addressed. Tel: +91 172 2690557/2690225; Fax: +91 172 2690632/2690585; Email: raghava@imtech.res.in

structures (predicted by taking the secondary structure as input) (13–15). Approaches for predicting supersecondary structural motifs fall into two main categories: the first is based on finding different supersecondary structural motifs in a protein sequence, all in one go, whereas in the other category attempts are being made to predict special structural motifs such as β-hairpins.

Supersecondary structure prediction is a field which is now being explored. Sun *et al*. (13) employed an artificial neural network (ANN) to predict 11 different commonly occurring supersecondary motifs and achieved an accuracy ranging between 70 and 80% and a Matthew's correlation coefficient (MCC) between 0.40 and 40.50. Cruz *et al*. (14) developed a method for predicting β-hairpins in a protein. They used a scoring scheme which utilized 14 scores based on alignment and such properties as secondary structure, accessibility, presence of turns, specific pair interactions and non-specific distance based contacts. Using this approach they attained an accuracy of 47.7% (±3.9). Kuhn *et al*. (15) attempted in 2004 to classify strand–loop–strand motifs by identifying local hairpins and non-local diverging turns using amino acid sequence as the input. This method, which attempted to predict the beginning and end of a hairpin and diverging turns or absence thereof, achieved an accuracy of 77.3% (±6.1) in predicting hairpins.

In the present paper an attempt has been made to develop a method for predicting β-hairpin motifs in protein sequences on similar lines to the approach of Cruz *et al*. (14). The main reasons behind selecting antiparallel β-hairpin structures were (i) that, after the alpha/beta (α/β) domain, they comprise the second largest group of protein domain structures, (ii) that functionally this group is very diverse and is found in enzymes, transporter proteins, antibodies and in viral coat proteins and (iii) that it is ubiquitous and exhibits simplicity of structure (16). In order to develop a classifier, we used two machine-learning techniques, a support vector machine (SVM) and an ANN, both of which are based on single sequence information, evolutionary profile, surface accessibility and secondary structure information (17,18). We trained and tested the method on a dataset of 2880 proteins using 5-fold cross-validation. In addition, we also trained and tested our approach on 534 proteins (referred to as Thornton's dataset below) used in the work of Cruz *et al*. (14). For a fair assessment of a method, one should evaluate its performance on protein structures not included in the dataset used for developing the method. To this end, we evaluated our method on all 63 proteins used recently in CASP6 (http://predictioncenter.llnl.gov/casp6/).

MATERIALS AND METHODS

Dataset of β-hairpins and non-hairpins

The dataset was generated from a large set of 2880 non-redundant (nr) protein chains of known structures, where no two protein chains have a percentage identity >33% (obtained from http://cubic.bioc.columbia.edu/eva/res/weeks.html#unique on November 25, 2002). The following steps were performed to generate the dataset.

(i) Secondary structure was assigned to each amino acid of all the 2880 proteins using DSSP (18).
(ii) From these proteins 12 653 unique amino acid patterns with secondary structure βcβ (minimum two consecutive amino acid residues in each state, later designated ECE patterns) were extracted (14).
(iii) PROMOTIF (19) was used to assign β-hairpin status in the 2880 proteins: 6675 β-hairpins, among which 6549 had unique amino acid patterns, were obtained.
(iv) A total of 5820 βcβ patterns (obtained from step ii), which were also assigned as hairpins by PROMOTIF, were finally considered as β-hairpins; the remaining 6833 βcβ patterns were considered as non-hairpins.
(v) Only 5548 hairpins and 6322 non-hairpins with length between 6 and 30 amino acid residues were kept.

The rationale behind selecting patterns of length 6–30 amino acids was the requirement of the machine-learning techniques for fixed-length patterns, as hairpin and non-hairpin patterns cannot be used without fixing the length. Therefore, fixed-length patterns of 17 amino acids were generated using the steps described below.

(i) If pattern length was <17, residues flanking the peptide in the primary amino acid sequence were appended at both the ends.
(ii) If pattern length was >17, only those patterns were kept for further study whose coil region was $\leqslant$10 residues long.
(iii) In the case of pattern length >17 and coil region $\leqslant$10 residues, the central coil residue was mapped and 9 residues from the left-hand side and 8 residues from the right-hand side were taken.

In this way, a total of 5102 hairpins and 5113 non-hairpins of length 17 were obtained.

Thornton's dataset

A Dataset of 534 proteins was obtained from Cruz *et al*. (14). To this dataset the same rules were applied to generate a library of hairpins and non-hairpins of length 17. DSSP generated 2229 ECE patterns, of which PROMOTIF classified 1169 as hairpins and 1060 as non-hairpins. Finally, after excluding the patterns with a chain break or heteroatom, 1076 β-hairpins and 878 non-hairpins of length 17 were obtained.

Secondary structure and surface accessibility

The secondary structure and surface accessibility were assigned using the DSSP program. The predictions of secondary structure and surface accessibility were made using the PSIPRED and NETASA (20) programs, respectively.

Multiple sequence alignment and position specific scoring matrices (PSSMs)

Multiple sequence alignment of each protein was performed using PSI-BLAST (21) at threshold 0.001 against a nr protein database (obtained from www.ncbi.nlm.nih.gov) with three iterations. Intermediate PSI-BLAST generated PSSMs were used as a direct input to the ANN and the SVM. The matrix had $21 \times M$ elements, where M is the length of a pattern. Each element in the matrix represents the frequency of occurrence of each of the 21 amino acids at a given position in the alignment.

Feature representation. For the sequence based model, each pattern was represented by 21×17 units, where 21 binary vectors were used to represent an amino acid (20 for the amino acid and one for terminal residues). In the case of multiple alignment or evolutionary profile, the PSSM matrix was used instead. In accessibility models, one unit was added to represent the accessibility of residues: 0 for buried and 1 for exposed residues. Thus, accessibility models have 17×22 units for each pattern. In the case of the secondary structure model, three units were used in the test case and one unit was added in the ideal case.

Artificial neural network

The artificial neural network was implemented using Stuttgart's neural network simulator (http://www-ra.informatik.uni-tuebingen.de/SNNS/). A feed-forward neural network with a back-propagation algorithm (22) was used to discriminate between hairpins and non-hairpins.

Support vector machine

In this study, SVM implementation was achieved using the SVM_light package (23), which provides a number of inbuilt parameters and kernels (e.g. linear, polynomial, radial basis function, sigmoid or any user-defined kernel). Three kernels, namely, linear, polynomial and sigmoid with different kernel parameters, were used for the training.

Consensus and combination prediction

In order to utilize the strength of both the ANN and SVM based approaches, we combined their results in the intersection and union modes, which are called the consensus and combined prediction modes, respectively. Consensus and combined approaches are analogous to the logical Boolean operators 'AND' and 'OR'. In consensus prediction, a pattern predicted as a hairpin by both methods was considered to be a hairpin; otherwise it was considered as a non-hairpin. In combined prediction, a hairpin predicted by either of the two methods was considered a hairpin.

Evaluation

Jack-knife test. The performance of any prediction algorithm is often checked by cross-validation or jack-knife tests. In this study the performance of all the methods and models was evaluated using 5-fold cross-validation in which the dataset was randomly divided into five equal sets, out of which four sets were used for training and the remaining one for testing. This procedure was repeated five times by changing the test dataset, so that each set was used for training as well as testing. The final performance was calculated by averaging over all five sets. During ANN architecture optimization, training was done on three sets; one set was used for validation to avoid overtraining the network.

Performance measures. We used the following standard measures to estimate the performance of our methods: (i) accuracy of prediction (Acc), (ii) MCC, (iii) sensitivity ($Q_{o(H)}$) or percentage coverage of hairpins, (iv) specificity ($Q_{o(Nh)}$) or percentage coverage of non-hairpins, (v) probability of positive (or hairpin) prediction ($Q_{p(H)}$) and (vi) probability of negative (or non-hairpin) prediction ($Q_{p(Nh)}$). See Supplementary Material for detailed equations (http://www.imtech.res.in/raghava/bhairpred/supli.html).

CASP6 proteins or independent dataset

We evaluated our method on all the 63 targets used in the recently concluded CASP6 competition. We submitted the amino acid sequences of these proteins to the BhairPred server and compared the predicted hairpins with PROMOTIF assigned hairpins. In addition, in order to evaluate the discriminatory capability of the method, all ECE patterns were sampled and the number of ECE patterns correctly classified as hairpin and wrongly classified as non-hairpin was computed.

RESULTS

The performance of the method was evaluated in two different cases, the ideal and the test case. In the ideal or true prediction case, only assigned values were used for training the algorithm, whether it was the secondary structure or the surface accessibility. The strategy adopted was thus similar to that of Cruz *et al*. (14). However, this approach does not reflect the true picture, because in real life the algorithm has to discriminate between a hairpin and a non-hairpin solely on the basis of sequence information; thus, the ideal case was used just to ascertain the upper limit of performance of the method when the highest quality of information (observed secondary structure and accessibility) was provided for training. We also performed a real-life test of our method (later designated as the test case), in which predicted information was used instead of the observed information. Our aim was to demonstrate the capability of our method to predict hairpins in real life, where the secondary structure of the protein is not known.

ANN approach

A two-layered neural network architecture was constructed, with a sequence-to-structure layer (PSI-BLAST profile used as input) and structure-to-structure layer. The input of the second layer or network is the output of first network and predicted secondary structure obtained from PSI-BLAST (Supplementary Figure S2). When only the amino acid sequence was used as input, the ANN was able to distinguish between hairpins and non-hairpins with an accuracy of 65.5%, with percentage coverage of 58.4% for hairpins and 78.5% for non-hairpins (Table 1). Accuracy and MCC were improved considerably (65.5 to 67.0% and 0.31 to 0.34%, respectively) when the PSI-BLAST profile was used as the input. Surface accessibility and secondary structure, when supplemented with amino acid sequence, increased the accuracy to 66.4 and 71.2%, respectively. In the ideal case, the maximum accuracy and MCC attained were 71.2% and 0.43, respectively. In order to train the method, in the test case we used predicted secondary structure and surface accessibility (real-life situation) with PSI-BLAST profile. As shown in Table 1, a maximum accuracy of 67.1% and MCC of 0.37 could be achieved using the ANN in the test case.

SVM approach

We achieved accuracies of 68.1 and 74.9% in the case of a single sequence and multiple alignments, respectively, using

Table 1. Prediction results with the 2880 protein dataset using the ANN

Approach	Coverage (%)		Probability (%)		Accuracy (%)	MCC
	$Q_{o(H)}$	$Q_{o(Nh)}$	$Q_{p(H)}$	$Q_{p(Nh)}$		
AA	58.4	78.5	68.1	63.7	65.5	0.31
MSA	66.7	67.3	66.7	53.4	67.0	0.34
AA + ACC_O	60.9	72.0	68.5	65.0	66.4	0.33
AA + SS_O	67.9	74.4	72.8	70.1	71.2	0.43
Seq–Str network (SS_O)	58.5	80.7	75.1	66.2	69.6	0.40
Seq–Str network (SS_P)	56.5	77.5	71.5	64.2	67.1	0.37

AA: amino acid sequence; MSA: multiple sequence alignment; ACC_O: observed accessibility (DSSP); SS_O: secondary structure observed (DSSP); seq: sequence; str: structure; SS_P: secondary structure predicted (PSIPRED).

Table 2. The performance of our SVM based modules on 2880 proteins using 5-fold cross-validation

Approach	Coverage (%)		Probability (%)		Accuracy (%)	MCC
	$Q_{o(H)}$	$Q_{o(Nh)}$	$Q_{p(H)}$	$Q_{p(Nh)}$		
AA	63.7	72.4	69.7	66.7	68.1	0.36
MSA (1)	77.3	72.4	73.7	76.3	74.9	0.49
AA + ACC_O (2)	69.1	70.6	70	69.7	69.9	0.39
AA + SS_O (3)	67.9	80.4	77.5	71.6	74.2	0.49
Hybrid (1 + 2 + 3)	82.6	75.7	77.2	81.4	79.2	0.59
AA + ACC_P (4)	64.1	71.9	69.5	66.9	68.0	0.36
AA+ SS_P (5)	68.9	72.3	71.2	70.0	70.6	0.41
Hybrid (1 + 4 + 5)	76.2	79.6	78.8	77.1	77.9	0.56

AA: amino acid; MSA: multiple sequence alignment; ACC_O: observed accessibility (DSSP); SS_O: secondary structure observed (DSSP); SS_P: secondary structure predicted (PSIPRED); ACCP: predicted accessibility.

the SVM technique (Table 2). This reflects the effect of evolutionary information on the performance of the classification method. Observed surface accessibility and secondary structure, when supplemented with amino acid sequence, increased the accuracies to 69.9 and 74.2%, respectively (in the ideal case). It is interesting to note that approximately the same performance was obtained with multiple sequence alignment as with the observed secondary structure. When all three sets of information—namely, evolutionary profile, surface accessibility and secondary structure—were combined, in the ideal case, an accuracy of 79.2% and MCC of 0.59 were achieved. In the test case an accuracy of 77.9% and MCC of 0.56 were obtained using the predicted secondary structure and accessibility (Supplementary Figure S3). These results clearly indicate the superior performance of the SVM over the ANN in the prediction of β-hairpins.

Combination of SVM and ANN

In order to utilize the capabilities of both approaches the prediction outputs of SVM and ANN were combined (Table 3). Supplementary Figure S4 shows the sensitivity and specificity of the consensus approach using the SVM (default threshold) and different thresholds of the ANN. As can be seen from the figure, it is possible to achieve high specificity of hairpin prediction but the sensitivity decreases drastically. However, the sensitivity may be increased at the cost of specificity or the probability of correct prediction. As shown in Supplementary Figure S5, the sensitivity increases but the specificity decreases in the case of the combined approach. The rationale behind using the consensus or combined approach lies in their inherent properties. The consensus approach would be an ideal choice if high specificity were desired during prediction, whereas the combined approach should be used if high sensitivity (detection of most of the hairpins) is desired.

Table 3. Performance of the consensus and combined approaches on the 2880 protein dataset

Consensus prediction		Threshold	Combined prediction	
Sensitivity (%)	Specificity (%)		Sensitivity (%)	Specificity (%)
76.1	79.6	0.1	98.9	10.6
75.1	79.8	0.2	94.8	27.5
72.2	80.7	0.3	89.3	43.4
67.6	82.2	0.4	84.9	55.7
61.5	84.6	0.5	81.6	64.7
53.1	87.4	0.6	79	71.9
41.3	91.2	0.7	77.4	76.4
25.7	95.7	0.8	76.4	78.9
5.8	99.8	0.9	76.2	79.5

Performance on Thornton's dataset

We also trained and tested our approach on datasets used in the past by Thornton's group (14). The performance of the SVM based method on this dataset is shown in Supplementary Table S1. The performance of our method decreased by 2.2, 4.1 and 6.8%, respectively when amino acid, the ideal case (combined information) and the test case (combined information) were used. With the ANN, performance dropped by 7.4 and 3.2% in using amino acid and sequence-to-structure network, respectively (Supplementary Table S2). These results clearly indicate that the performance of the method depends on the size and quality of the dataset. When the difference in performance of our method on our and Thornton's dataset was analyzed, it was observed that, although the performance decreased, the trend remained the same.

Performance on CASP6 proteins

To perform an impartial review of our methodology, we predicted hairpins in all the 63 CASP6 proteins using our web server (http://www.imtech.res.in/raghava/bhairpred/). To avoid any bias, during this prediction we did not optimize any prediction parameter (e.g. threshold) on the server. The following information was obtained for each protein: (i) number of ECE patterns predicted by PSIPRED, (ii) number of ECE patterns predicted as hairpins by BhairPred and (iii) number of predicted hairpins also assigned as hairpins by PROMOTIF. The mapping between ECE patterns, predicted hairpins and observed hairpins, along with the amino acid sequences of the proteins, is depicted diagrammatically in Supplementary Figure S1 (http://www.imtech.res.in/raghava/bhairpred/supli.html). A total of 201 ECE patterns were found by PSIPRED. The length of these patterns varied from 6 to >20 amino acid residues. After fixing the length at 17 residues, only 180 patterns remained. Out of these, 132 (73.33%) ECE patterns (47 hairpin and 85 non-hairpin; sensitivity 60.25% and specificity 83.33%) were correctly predicted by BhairPred (Table 4). Hairpins assigned by PROMOTIF were also

Table 4. Performance of BhairPred in predicting of ECE patterns as hairpins or non-hairpins

CASP6 categories	No. of proteins	No. of ECE patterns	No. of discarded ECE patterns	TP	TN	FP	FN
ALL	63	201	21	47	85	17	31
NF	4	7	0	0	4	2	1
FR(A)	6	12	1	4	4	1	2
FR(H)	10	23	1	6	6	4	6
CM	20	55	8	11	24	5	7

ALL: number of target proteins in CASP6; NF: new fold; FR(A): fold recognition (analogous); FR(H): fold recognition (homologous); CM: comparative modeling; TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

Table 5. Performance of Bhairpred on hairpins assigned by promotif in CASP6 proteins

CASP6 categories	No. of proteins	No. of hairpins (Promotif)	Exact matching ECE	Non-exact matching ECE	Non-exact at all
ALL	63	159	27 (22)[a]	51 (25)	61
NF	4	9	0	1 (1)	7
FR(A)	6	9	2 (2)	4 (2)	3
FR(H)	10	20	4 (3)	8 (3)	6
CM	20	46	5 (4)	13 (7)	20

[a]Correctly predicted hairpins by Bhairpred in parentheses.
ALL: number of target proteins in CASP6; NF: new fold; FR(A): fold recognition (analogous); FR(H): fold recognition (homologous); CM: comparative modeling.

examined (a total of 159 hairpins in all 63 proteins). Comparing PROMOTIF assigned hairpins with PSIPRED predicted ECE patterns, we found 27 exact matches the secondary structure of the hairpin and the ECE pattern were the same), 51 non-exact matches (the predicted and observed regions overlap, but the secondary structure of a few residues did not match) and 61 entirely misaligned secondary structure patterns (ECE pattern not predicted in the hairpin region; for detail see Supplementary Material). BhairPred was able to correctly predict 22 out of 27 hairpins (81.48% accuracy) in the case of exact matches, and 25 hairpins in the case of non-exact matches of secondary structure patterns. The performance of the BhairPred server in different categories of CASP6—namely comparative modeling (CM), fold recognition (FR) and new fold (NF) discovery—was also examined. Although the method performed reasonably well in the CM and FR categories, it was found to be unsuccessful in the NF category, which was due to the failure to correctly predict the ECE pattern by PSIPRED in the NF category (Table 5). These results unambiguously established the dependence of the BhairPred server on the performance of PSIPRED. In other words, if the ECE pattern is correct, BhairPred can predict hairpins with high accuracy.

Web server

A web server, BhairPred, was developed to predict β-hairpins in proteins using an SVM method, based on algorithms discussed in this paper. The server performs the following steps: (i) it accepts protein sequences from the user in any standard format (FASTA, PIR, EMBL); (ii) it predicts the secondary structure using PSIPRED; (iii) it identifies the βcβ regions in the protein; (iv) it generates an alignment profile for the protein using PSI-BLAST; (v) it predicts surface accessibility in the protein using NETASA and (vi) finally it predicts whether the identified βcβ regions are hairpins or non-hairpins from predicted secondary structure and PSI-BLAST profile. The BhairPred server provides various options to users, including selection of an appropriate threshold value for predicting β-hairpins (Supplementary Figure S6). It also allows output to be presented in tabular format with complete details of the predicted sheet–coil–sheet patterns in proteins and, among these patterns, specifies which are potential β-hairpins (Supplementary Figure S7). The server also provides advanced options such as assigning user-defined secondary structure to query the protein instead of using the predicted secondary structure obtained from PSIPRED.

DISCUSSION

In proteins, a few secondary structures are arranged in a definite, simple geometrical shape to form supersecondary structures. These supersecondary structures are the building blocks of the 3D structures of proteins, which are also called structural motifs. Many times these structural motifs are associated with specific functions. Accurate prediction of supersecondary structures can be one important step toward building a tertiary structure from the specified secondary structure. There is thus a need to understand supersecondary structures, particularly the linker regions which connect regular secondary structure α-helices and β-strands. There are a number of well-defined structural motifs, e.g. α–α and β–β motifs, α–β and β–α arches and α–α and β–β corners. One of the frequently occurring motifs in proteins is the β-hairpin, which connects two adjacent antiparallel hydrogen bonded β-strands. In proteins, several adjacent antiparallel β-sheets are found, but they need not be connected with each other by either covalent or non-covalent bonds. Thus, the prediction of β-hairpins is very important because it can reduce the number of possible folds available to that protein. So far as the prediction of protein structure is concerned, recently Cruz *et al*. (14) described an approach that predicts hairpins from predicted secondary structures. They demonstrated successfully that their approach could identify the hairpins with significant accuracy, much higher than that obtained by random prediction.

In this study, an attempt has been made to develop a better and quantitative method for discriminating hairpins from non-hairpins, on similar lines to Cruz *et al*. (14). Cruz *et al*. (14) compared the βcβ regions against a library of β-hairpins with the same number of residues and calculated 14 scoring terms from the alignment. These scoring terms were used as the input units of a neural network to discriminate between the potential hairpins and non-hairpins. They used the machine-learning method ANN, which is more qualitative than quantitative, and a major limitation was to get similar types of known hairpins. In the present method, we have worked with fixed-length hairpins and non-hairpin patterns, generated either by appending or by trimming at the pattern terminus (see Materials and Methods for the detail process). Patterns obtained in this way were used as input to the ANN and the SVM. In this study we have shown, for the first time, the importance of evolutionary information in predicting supersecondary structures (in this case β-hairpins), an importance

that known to hold for secondary structures. Besides this, from the result obtained, it was also evident that the evolutionary profile contains more information than the simple amino acid sequence. This conclusion is also supported by the improved performance of the SVM as well as the ANN when a PSSM is used for training (Tables 1 and 2; Supplementary Tables S1 and S2). We also trained, developed and evaluated the performance of the same algorithm on the dataset used by Cruz *et.al*. (14). Although the performance was inferior, the trend remained the same. It was thus inferred that the learning ability of any machine-learning technique is directly proportional to the size of the dataset used for training.

The real strength of any prediction method can be estimated only by evaluating its performance on an independent dataset. In our case we used the CASP6 proteins as an independent dataset to evaluate the performance of our method. As shown in Results, BhairPred performed reasonably well on the CASP6 proteins. It was observed that the performance of BhairPred depended to a great extent on the performance of PSIPRED, since it is the PSIPRED predicted secondary structure through which ECE patterns were extract.

The present method could discriminate between hairpins and non-hairpins with very high accuracy if the predicted secondary structure was correct. Because of this, we incorporated the option to give the secondary structure to the BhairPred server. One of the major things reported in this paper was that the method developed can predict with high accuracy non-hairpin ECE regions in proteins, thereby reducing the number of theoretical folds available to any protein, and thereby bringing down the effort and time for protein folding. Thus the method can be a good tool for protein structure prediction.

One of the limitations of the approach presented is that the length of patterns of a number of hairpins and non-hairpins had to be fixed, resulting in the removal of those which did not satisfy the criteria laid down. Thus, the result obtained should not be compared with that of Cruz *et al*. (14), because in that procedure all hairpins and non-hairpins were considered, whereas in the present case only those βcβ regions which satisfy the conditions described were considered. An average of 20% of hairpins and non-hairpins were excluded from our dataset. Therefore, the method given in this paper should be considered as a supplement to that of Cruz *et al*. (14). It is hoped that this study will be a useful addition to protein tertiary structure prediction.

ACKNOWLEDGEMENTS

We appreciate the suggestions of the anonymous referees to evaluate the performance of the method on an independent dataset (the CASP6 proteins). We are grateful to Dr J. M. Thornton (EBI, UK) for providing us the dataset used here and to Dr Amit Ghosh (IMTECH, India) for checking the manuscript. We are also gratified to the developers of PSIPRED, PSI-BLAST, SNNS, SVM^{light} and NETASA. We express thanks to the Council of Scientific and Industrial Research (CSIR) and the Department of Biotechnology (DBT), Government of India, for financial assistance. This manuscript carries IMTECH communication number 054/2004. Funding to pay the Open Access publication charges for this article was provided by DBT, Government of India.

Conflict of interest statement. None declared.

REFERENCES

1. Jones,D.T., Tress,M., Bryson,K. and Hadley,C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins*, **3** (Suppl.), 104–111.
2. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol*., **232**, 584–599.
3. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., **292**, 195–202.
4. Raghava,G.P.S. (2000) Protein secondary structure prediction using nearest neighbor and neural network approach. CASP4. 75.
5. Petersen,T.N., Lundegrad,C., Neilsen,M., Bohr,H., Bohr,J., Brunak,S., Gippert,G.P. and Lund,O. (2000) Prediction of protein secondary structure at 80% accuracy. *Proteins*, **41**, 17–20.
6. Kaur,H. and Raghava,G.P.S. (2004) Prediction of Alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins*, **55**, 83–90.
7. Kaur,H. and Raghava,G.P.S. (2002) BetaTPred: Prediction of β-turns in a protein using statistical algorithms. *Bioinformatics*, **18**, 498–499.
8. Kaur,H. and Raghava,G.P.S. (2002) An evaluation of β-turn prediction methods. *Bioinformatics*, **18**, 1508–1514.
9. Kaur,H. and Raghava,G.P.S. (2003) Prediction of β-turns in proteins from multiple alignment using neural network. *Protein Sci*, **12**, 627–634.
10. Kaur,H. and Raghava,G.P.S. (2003) BTEVAL: a server for evaluation of β-turn prediction methods. *J. Bioinform. Comput. Biol*., **1**, 495–504.
11. Kaur,H. and Raghava,G.P.S. (2003) A neural network based method for prediction of γ-turns in proteins from multiple sequence alignment. *Protein Sci*, **12**, 923–929.
12. Kaur,H. and Raghava,G.P.S. (2004) A neural network method for prediction of β-turn types in proteins using evolutionary information. *Bioinformatics*, **20**, 2751–2758.
13. Sun,Z., Rao,X., Peng,L. and Xu,D. (1997) Prediction of protein supersecondary structures based on artificial neural network method. *Protein Engg*., **10**, 763–769.
14. Cruz,X., Hutchinson,E.G., Shepherd,A. and Thornton,J.M. (2002) Toward predicting protein topology: an approach to identifying B hairpins. *Proc. Natl Acad. Sci. USA*, **99**, 11157–11162.
15. Kuhn,M., Meiler,J. and Baker,D. (2004) Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins*, **5**, 282–288.
16. Branden,C. and Tooze,J. (1999) *Introduction to Protein Structure,* 2nd edn. Garland Publishers, New York, pp. 67.
17. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., **292**, 195–202.
18. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
19. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci*, **5**, 212–220.
20. Ahmad,S. and Gromiha,M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, **18**, 819–824.
21. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
22. Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986) Learning representations by back-propagation errors. *Nature*, **323**, 533–534.
23. Joachims,T. (1999) Making large-Scale SVM learning practical. In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT-Press, Cambridge, MA, London, England, pp. 169–184.