Role of evolutionary information in prediction of aromatic-backbone NH interactions in proteins

Harpreet Kaur, G.P.S. Raghava*

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

Accepted 10 March 2004

First published online 9 March 2004

Edited by Robert B. Russell

Abstract In this study, an attempt has been made to develop a neural network-based method for predicting segments in proteins containing aromatic-backbone NH (Ar-NH) interactions using multiple sequence alignment. We have analyzed 3121 segments seven residues long containing Ar-NH interactions, extracted from 2298 non-redundant protein structures where no two proteins have more than 25% sequence identity. Two consecutive feed-forward neural networks with a single hidden layer have been trained with standard back-propagation as learning algorithm. The performance of the method improves from 0.12 to 0.15 in terms of Matthews correlation coefficient (MCC) value when evolutionary information (multiple alignment obtained from PSI-BLAST) is used as input instead of a single sequence. The performance of the method further improves from MCC 0.15 to 0.20 when secondary structure information predicted by PSIPRED is incorporated in the prediction. The final network yields an overall prediction accuracy of 70.1% and an MCC of 0.20 when tested by five-fold cross-validation. Overall the performance is 15.2% higher than the random prediction. The method consists of two neural networks: (i) a sequence-to-structure network which predicts the aromatic residues involved in Ar-NH interaction from multiple alignment of protein sequences and (ii) a structure-to structure network where the input consists of the output obtained from the first network and predicted secondary structure. Further, the actual position of the donor residue within the 'potential' predicted fragment has been predicted using a separate sequence-to-structure neural network. Based on the present study, a server Ar_NHPred has been developed which predicts Ar-NH interaction in a given amino acid sequence. The web server Ar_NHPred is available at http://www.imtech.res.in/raghava/ar_nhpred/ and http://bioinformatics.uams.edu/mirror/ar_nhpred/ (mirror site).

Key words: Aromatic; Backbone NH; Neural network; Multiple alignment; Secondary structure; Web server

1. Introduction

Hydrogen bonds are the key to many phenomena, including the formation/stabilization of secondary structures (e.g. α-helices and β-sheets), protein folding, molecular recognition and enzymatic reactions [1,2]. The conventional hydrogen bonds that involve electronegative atoms like oxygen and nitrogen have been thoroughly studied over the decades since their first introduction into the literature. In addition to conventional hydrogen bonding, there are non-conventional hydrogen bonding interactions, which are weak polar interactions in comparison to conventional hydrogen bonding [3]. Broadly, these interactions can be classified as C–H...π, N–H...π and C–H...O interactions [4]. It has been shown in recent studies that all these interactions play a crucial role in proteins, protein–protein, protein–ligand and drug binding interactions [5–7]. McPhail and Sim [8] reported the first example of an X–H...π hydrogen bond in a peptide crystal structure. Much later, N–H...π hydrogen bonds in proteins attracted greater attention following the observation of local non-random conformations formed by such interactions in bovine pancreatic trypsin inhibitor (BPTI) by Kemmink and Creighton [9,10]. These local structures are formed by interaction between the aromatic ring of Tyr10 and the backbone amide of Gly12 in BPTI.

An aromatic (Ar) NH interaction is one of the categories of non-conventional hydrogen bonding interaction. The aromatic residues Phe, Tyr, and Trp have a π ring system that can form a hydrogen bond with the NH moiety, thereby offering additional stability. Depending on the interactions with the main chain or side chain NH moiety, the interaction can be classified as Ar-NH (backbone) or Ar-NH (side chain) interaction respectively. Throughout this study, Ar-NH interaction denotes the interaction between the side chain aromatic ring and the backbone NH group.

In the past, several analyses of the characteristics of interactions between aromatic residues and main chain NH group have been performed, including structural and sequence features and relevance in different secondary structure elements. Studies by Burley and Petsko suggested the involvement of Ar-NH interactions in the stabilization of protein tertiary structures on the basis of their spatial distribution [11,12]. Further investigations by various research groups have established the role of these interactions in ligand recognition and stabilization of secondary structures, mainly β-sheets and helix termini [7]. Secondary structures of proteins may create three-dimensional space that could facilitate Ar-NH interactions, and these then further stabilize these structures [13]. Despite the relevance of Ar-NH interactions in proteins, to date not a single study has addressed the importance of predicting the residues involved in Ar-NH interaction. The prediction of the residues involved in Ar-NH interaction can be an interesting problem whose solution may be useful in protein folding and in de novo design.

*Corresponding author. Fax: (91)-172-690632;
http://imtech.res.in/raghava.
E-mail address: raghava@imtech.res.in (G.P.S. Raghava).

0014-5793/04/$30.00
doi:10.1016/S0014-5793(04)00305-9

1.1. An approach to prediction of Ar-NH interactions

In this study, a systematic attempt has been made to develop a method for predicting Ar-NH interactions in proteins from their amino acid sequence. It has been shown in the past that artificial neural networks (ANN) are a powerful tool in solving several kinds of problems, including predictions of both regular [14–16] and irregular [17–20] secondary structures, transmembrane helices [21], inter-residue contacts [22], and folding of initiation sites [23]. Thus, ANN has been used in the present study.

The preliminary analysis of Ar-NH interactions in the data indicates that a segment of seven residues provides sufficient information for prediction of segments having Ar-NH interaction. Beyond seven residues, the number of Ar-NH interactions decreases and moreover neural networks fail to capture long-range information [24]. For instance, in the present dataset nearly 88% of Ar-NH interactions occur with a separation of up to three residues between donor and acceptor pair. Thus, for prediction of Ar-NH interactions, an optimal window size of seven has been constructed by extracting fragments of length seven residues wide with an aromatic residue at the central position flanked by three residues on both sides. Depending on the position of the donor residue in the fragment, the Ar-NH interaction has been further categorized as Ar(*i*)-NH(*i*−3), Ar(*i*)-NH(*i*−2), Ar(*i*)-NH(*i*−1), Ar(*i*)-NH(*i*), Ar(*i*)-NH(*i*+1), Ar(*i*)-NH(*i*+2) and Ar(*i*)-NH(*i*+3) with the aromatic residue at the central *i*th position. For instance, Ar(*i*)-NH(*i*) is the interaction between the aromatic ring and NH of the *i*th residue and Ar(*i*)-NH(*i*+1) is the interaction between the aromatic ring of the *i*th residue and NH of the *i*th+1 residue and so forth.

The problems of prediction of Ar-NH interaction within the fragment and prediction of the actual location of the donor residue within the positively predicted fragment have been considered separately. First, we have developed a method for predicting the presence or absence of Ar-NH interactions

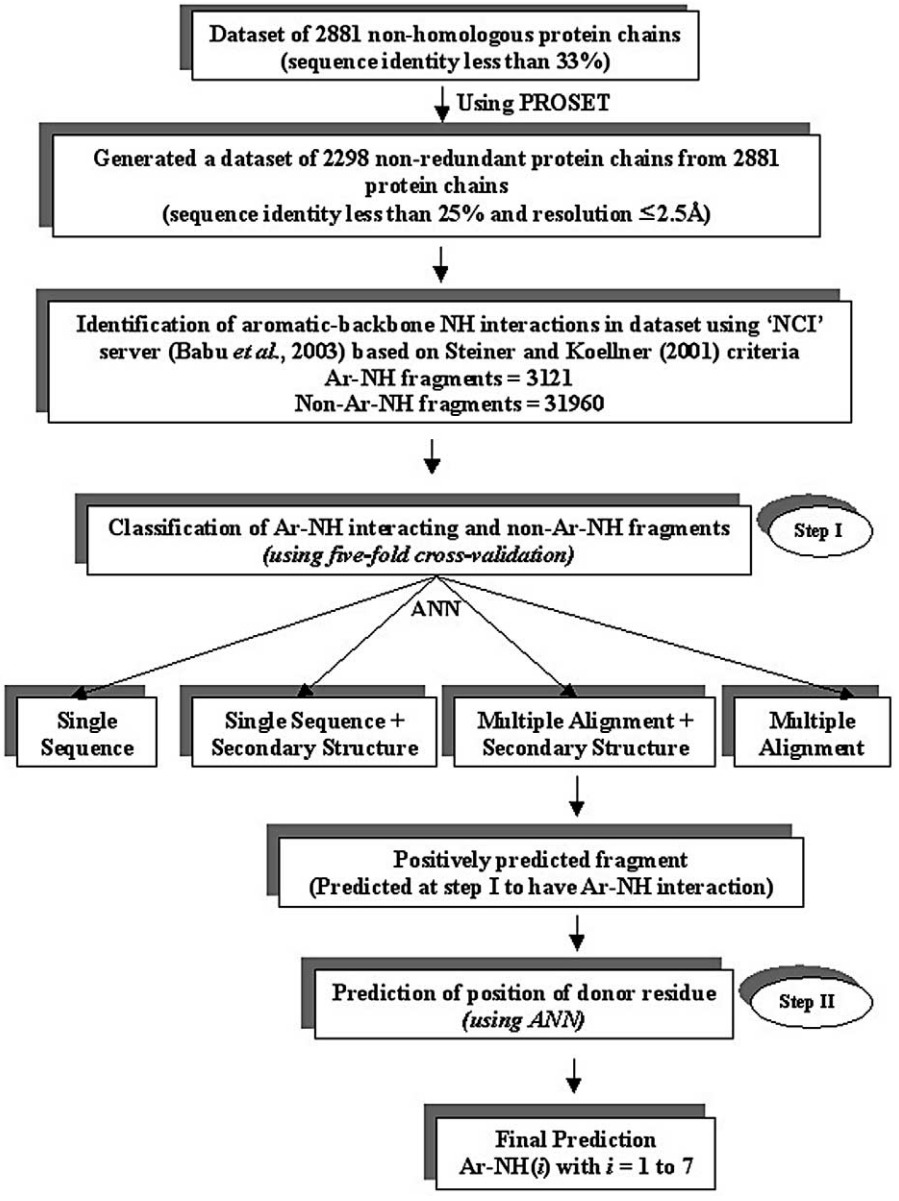

Fig. 1. Flowchart of the prediction method used to predict Ar-NH interactions in proteins.

in a given segment. For this, a first level 'sequence-to-structure network' has been trained on single amino acid sequences encoded as binary bits (0 or 1) or multiple sequence alignments (position-specific scoring matrices obtained from PSI-BLAST [25]). The accuracy of the method has been improved further by using a second structure-to-structure network where input to the network consists of the output obtained from the first network (sequence-to-structure network) and secondary structure information predicted using PSIPRED [26]. Finally, within the positively predicted fragment (fragment predicted to have Ar-NH interaction), the position of the donor residue has been predicted using a separate ANN trained with a single sequence on a dataset containing different types of interactions such as Ar(i)-NH(i), Ar(i)-NH(i+1) and so on (with the NH moiety at different positions). The predicted fragment is classified as Ar(i)-NH(i), Ar(i)-NH(i+1) and other types depending on the relative value of the network output. The outline of the prediction method is shown in Fig. 1.

2. Materials and methods

2.1. The dataset

In this study, we have used a non-redundant dataset of 2298 protein chains where no two protein chains have more than 25% sequence identity and have resolution 2.5 Å or better. The dataset has been generated from a large set of 2881 protein chains with sequence identity less than 33% (available at http://cubic.bioc.coulmbia.edu/eva/res/weeks.html#unique, retrieved on 25th November, 2002). Using the PROSET program [27] with default parameters, the dataset of 2881 protein chains was reduced by removing all proteins with more than 25% sequence identity.

All the protein structures have been extracted from the Protein Data Bank [28]. The overlapping fragments of length seven residues with Phe, Tyr or Trp at the central position, flanked with three residues on both sides have been generated from these proteins.

2.2. Identification of Ar-NH interaction in the dataset

The Ar-NH interactions in the dataset have been identified using the web server NCI (http://www.mrc-lmb.com.ac.uk/genome/nci/) [29] that is based purely on geometric criteria [30]. The default parameters ($N...\pi_m \leq 4.3$ Å; $H...\pi_m \leq 3.5$ Å, $N\text{-}H...\pi_m \geq 120°$ and $N...\pi_m...\pi_n \leq 30°$) have been used where π_m represents the mid-point of the π-ring and π_n represents the vector, normal to the plane of the ring. Further, Ar-NH interactions have been selected which have donor and acceptor sequential separation ($\Delta_{D-A} \leq \pm 3$) up to three residues. This has yielded a total of 3121 fragments having Ar-NH interaction with the NH (backbone) group as hydrogen donor and the π-aromatic ring as hydrogen acceptor. In these fragments, depending on the position of the donor residue, the interactions have been categorized as Ar(i)-NH(i), Ar(i)-NH(i+1) and so forth. The remaining fragments that do not satisfy the interaction criteria have been considered the negative dataset or the dataset with fragments having no Ar-NH interactions. Indeed, in all the fragments, the aromatic residue is present at the central position. The dataset contains much fewer Ar-NH interacting fragments than the number of non-Ar-NH fragments, the ratio being ~1:10.

2.3. Five-fold cross-validation

Assessment of a prediction method is often done by the jack-knife or cross-validation technique [14]. In a full jack-knife test of N proteins, one protein is removed from the set, the parameters are developed on the remaining $N-1$ proteins, then the accuracy of the method is tested on the removed protein. This process is repeated N times by removing each protein in turn. Since some training techniques are time-consuming, a more limited cross-validation is often performed. In the cross-validation technique, the set of N proteins are split into M subsets. Parameters are developed on $(M-1)N/M$ proteins, then tested on the remaining N/M proteins. This process is repeated M times, once for each subset.

In the present study, due to the size of the dataset and PSI-BLAST training, the jack-knife method was not feasible, so a five-fold cross-validation technique has been used by splitting the whole set of fragments into five subsets containing an approximately equal number of examples. At a given time three subsets have been used for training, one for validating and one for testing. The validation set is used to avoid over-training or over-learning of ANN. This process is repeated five times so that each subset is tested once. The final prediction results have been averaged over five testing sets.

2.4. ANN architecture

2.4.1. Software used. In this work, the SNNS version 4.2 neural network simulation package from Stuttgart University has been used (publicly available at http://www.informatik.uni-stuttgart.de/) to build ANN architecture [31]. It allows incorporation of the resulting networks into an ANSI C function for use in stand-alone code. A linear activation function has been used. At the start of each simulation, the weights are initialized with random values between −1.0 and 1.0. The training is carried out using error back-propagation with a sum of square error function [32]. The error is minimized for the validation subset, the parameters at this minimum error are used to compute the performance of ANN on the test set.

2.4.2. Classification of Ar-NH interacting and non-interacting fragments. To predict whether a given sequence segment contains Ar-NH interaction or not, two standard feed-forward ANNs have been used consecutively. Both have a single hidden layer with 10 units. An input window seven residues wide has been used. The target output consists of a single binary number and is 1 (having Ar-NH interaction) and 0 (having no Ar-NH interaction).

The input to the first 'sequence-to-structure' network is either a single sequence with amino acids as binary bits or multiple alignment profiles with PSI-BLAST-generated position-specific scoring matrices (PSSM). For single sequences, a binary encoding scheme has been used where each residue has been encoded as a vector of 21 elements corresponding to each residue, with one element set to 1 corresponding to the particular residue type and the remaining elements set to 0. Twenty elements encode 20 amino acids, and the one provides a signal when the input window overlaps the N- or C-terminus of the protein.

With multiple alignment input, PSSM generated by PSI-BLAST has been used as input to the neural network. The matrix has $21 \times M$ real number elements, where M is the length of the target sequence. Each element represents the likelihood of substitution of that particular residue at that position. Thus, 21 real numbers rather than binary bits encode each residue.

Using a second structure-to-structure network, the output obtained from the first network has been correlated. The input to the second filtering network is predictions obtained from the first network and PSIPRED-predicted secondary structure states. Four units encode each residue where one unit encodes interacting/non-interacting prediction output obtained from the first network and is the actual prediction score of the first network. The remaining three units correspond to the three secondary structure states (helix, extended and coil) obtained from PSIPRED. Secondary structure information is encoded by the actual probabilities of three states provided in the output of the PSIPRED prediction. The probabilities are just the strengths of the prediction for each of the three target states (helix, strand, coil) and are represented by a real number in the range 0–1.

2.4.3. Prediction of actual position of donor residue. The above prediction provides information whether a given fragment has Ar-NH interaction or not. It does not provide any information of the donor residue within the Ar-NH interacting fragment. Thus, the position or actual location of the donor residue in the fragment predicted positively (predicted to have Ar-NH interaction) has been further predicted using a separate 'sequence-to-structure' network. The input to the network is a single sequence with amino acids as binary bits. This network has window size seven and the target output has seven units, each representing one of the possible Ar-NH interactions (Ar(i)-NH($i-3$), Ar(i)-NH($i-2$), Ar(i)-NH($i-1$), Ar(i)-NH(i), Ar(i)-NH(i+1), Ar(i)-NH(i+2) and Ar(i)-NH(i+3)). For a given input and set of weights, the output of the network will be seven numbers between 0 and 1. The interaction type is the output unit having the highest activity level or value or actually corresponds to the position of the donor residue within the fragment. The architecture of the whole network system is shown in Fig. 2.

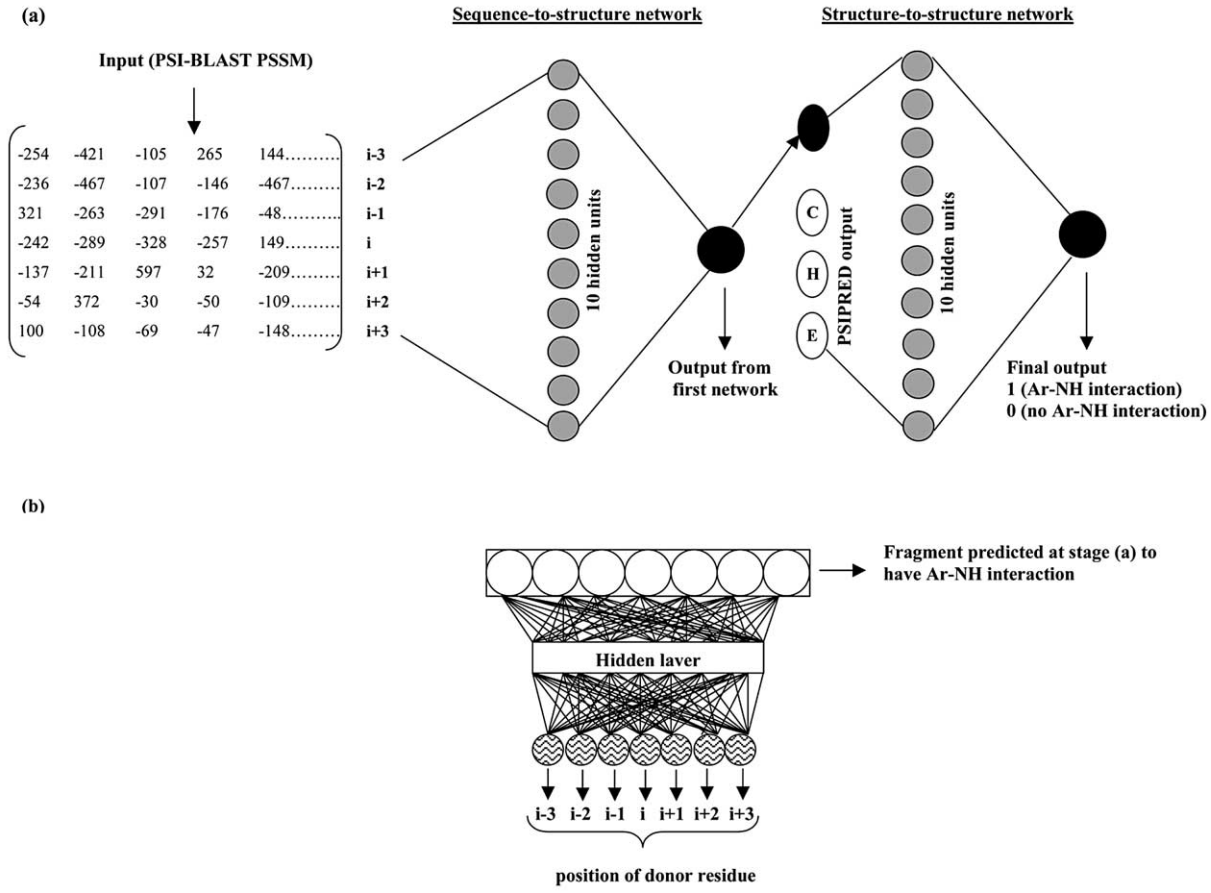

Fig. 2. The neural network system used for prediction of Ar-NH interactions. a: Network system used for prediction of Ar-NH interactions consisting of two networks: sequence-to-structure network and structure-to-structure network. The basic cell has 20+1 units. The second network has four units, one unit encodes the prediction obtained from the first network and the remaining three units encode three secondary structure states predicted by PSIPRED. b: Network system used for prediction of position/actual location of the donor residue within the fragment predicted positively by the first network.

2.5. Multiple alignment or position-specific scoring matrices

The secondary structure prediction method PSIPRED uses PSI-BLAST [25] to detect distant homologues of a query sequence and generate a position-specific scoring matrix as part of the prediction process, and here we have used these intermediate PSI-BLAST-generated position-specific scoring matrices as a direct input to the first level network. PSI-BLAST has been run against the standard NR (non-redundant) database. The position-specific scoring matrices have been obtained with three iterations of PSI-BLAST searches.

2.6. Secondary structure assignment and prediction

The protein secondary structure assignment by DSSP is used to establish an upper bound of predictive performance, i.e. the maximum performance that one can expect using secondary structure information. DSSP provides eight states assignment of secondary structure [33]. The eight states of DSSP have been decomposed into three states (G, H and I are taken as helices, B and E as strands and rest as coil). PSIPRED has been used to predict the secondary structure of proteins.

2.7. Performance measures

Both threshold-dependent and -independent measures have been used to assess the performance of the method.

2.7.1. Threshold-dependent measures. Five different parameters have been used to measure the performance of the prediction method. These five parameters can be derived from the four scalar quantities: TP (true positives: number of correctly classified Ar-NH interactions), TN (true negatives: number of correctly classified non-Ar-NH interactions), FP (false positives: number of non-Ar-NH interactions incorrectly classified as Ar-NH interactions) and FN (false negatives: number of Ar-NH interactions incorrectly classified as non-Ar-NH interactions). The following four parameters are calculated at different threshold or cut-off values.

1. Prediction accuracy: $((TN+TP)/t)\times 100$, where $t = TP+TN+FP+FN$ is the total number of examples.
2. Sensitivity: $(TP/(TP+FN))\times 100$ is the percentage of observed Ar-NH interacting fragments that are predicted correctly.
3. Specificity: $(TN/(TN+FP))\times 100$ is the percentage of observed non-interacting fragments that are predicted correctly.
4. Probability of correct prediction: the percentage of predicted examples that are predicted correctly. It has been calculated for interacting and non-interacting fragments separately as follows:
   Probability of correct prediction of positives: $(TP/(TP+FP))\times 100$
   Probability of correct prediction of negatives: $(TN/(TN+FN))\times 100$
5. Matthews correlation coefficient (MCC): the commonly used parameter prediction accuracy may be misleading due to disparity in the number of Ar-NH interacting fragments (10%) and non-Ar-NH fragments (90%); hence, it is possible to get an accuracy of about 90% by predicting all examples as non-Ar-NH fragments. Thus, there is a need to use more robust measures to evaluate a method. One of the best performance measures that accounts for unbalancing (both over- and under-prediction) is the Matthews correlation coefficient [34]. The correlation coefficient is defined by

$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

The MCC is a number between -1 and 1. If there is no relationship between the predicted values and the actual values the correlation

coefficient is 0 or very low (the predicted values are no better than random numbers). As the strength of the relationship between the predicted values and actual values increases so does the correlation coefficient. A perfect fit gives a coefficient of 1.0. Thus the higher the correlation coefficient the better is the prediction performance.

2.7.2. Performance with respect to random prediction. Another useful approach is to compare the accuracy of predictions with respect to predictions generated randomly. Here we have calculated the total number of patterns that are expected to be predicted correctly by randomly generated predictions. The requisite formula is

$$R_{\text{total}} = \frac{(\text{TP}+\text{FP})(\text{TP}+\text{FN}) + (\text{TN}+\text{FP})(\text{TN}+\text{FN})}{t}$$

To measure how well a method is performing compared with random (R_{total}), the normalized percentage better-than-random (S) has been calculated as

$$\text{S} = \frac{(\text{TP}+\text{TN}) - \text{R}_{\text{total}}}{\text{t} - \text{R}_{\text{total}}} \times 100$$

Perfect predictions score $S = 100\%$, predictions that are no better than random score $S = 0\%$ [17].

2.7.3. Threshold-independent measures. The performance measures described so far are threshold-dependent. One problem with the threshold-dependent measures is that they measure the performance at a given threshold. They fail to use all the information provided by a method. The receiver operating characteristic (ROC) is a threshold-independent measure, which is a trade-off between sensitivity and specificity. For a prediction method, an 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 curve always goes through two points (0,0 and 1,1). 0,0 is where the classifier finds no positives. In this case it always gets the negative cases right but it gets all positive cases wrong. The second point is 1,1 where everything is classified as positive. So the classifier gets all positive cases right but it gets all negative cases wrong. A classifier that randomly guesses has a ROC which lies somewhere along the diagonal line connecting 0,0 and 1,1. An important index of the ROC curve is its area. A random classifier has an area of 0.5, while an ideal one has an area of 1 [35].

2.7.4. A measure of statistical significance. When comparing different prediction approaches, we need to know whether the differences in performance measures (prediction accuracies or MCC values) among them are statistically significant or not. Statistics theory gives us a method to compute the 'significance interval' for the difference between two population proportions [36].

In this case, the 'proportion' is the percentage of cases in the test dataset which have been predicted correctly. If we assume that the prediction accuracies of two algorithms are p_1 and p_2 for two test datasets of r_1 and r_2 numbers of examples, respectively, and the test data are randomly selected, then we can say that we are $a \times 100\%$ confident that the two accuracies are really different if

$$|p_1 - p_2| > I$$

where

$$I = z(1 + a/2) \cdot \sqrt{p_1(1-p_1)/r_1 + p_2(1-p_2)/r_2}$$

z is the inverse cumulative normal distribution. The larger the difference between two prediction accuracies, the more significant it is. The above equation [37] has been used to determine whether the difference in the accuracies or other measures is statistically significant or not.

3. Results

3.1. Analysis of Ar-NH interactions

3.1.1. Distribution. The occurrence of different types of Ar-NH interactions found in the present dataset is presented in Fig. 3. A plot of the occurrence of Ar-NH interaction as a function of the NH position reveals prominent peaks at positions i and i+1. The peaks at position i+1, i+2 and i+3 are significantly higher than the corresponding ones at the negative side, implying that Ar(i)-NH(i+1, i+2, i+3) interactions are more common than the Ar(i)-NH(i−1, i−2, i−3) interactions. There are very few Ar(i)-NH(i+3, i−1, i−2, i−3) interactions. By far the most frequent Ar-NH interaction is the kind Ar(i)-NH(i) which occurs between the amino group and aromatic ring of the same residue. The second most frequent is the Ar(i)-NH(i+1) interaction which is three times more frequent than Ar(i)-NH(i+2) interactions. These findings suggest that the distance and number of separating residues between the backbone amide and the aromatic residue correlate with the occurrence of interaction. The difference in length seems to account for the predominance of Ar(i)-NH(i+1, i+2, i+3) interactions over Ar(i)-NH(i−1, i−2, i−3) interactions.

3.1.2. Acceptor and donor efficiencies. It has been found that among the aromatic residues, the most efficient π-acceptor is the indole group of the tryptophan residue followed by the phenol moiety of tyrosine and the benzene ring of phenylalanine. The higher acceptor efficiency of the Trp side chain compared to Tyr and Phe is due to the conjugate nature of two planar rings containing a heteroatom.

The protein fragments having Ar-NH interaction have further been analyzed to search for amino acid preferences or to determine the chance of the different amino acids to be involved in such interactions. From the absolute amino acid occurrences, the propensities of amino acids have been calculated and positional frequency histograms have been plotted for each aromatic amino acid for Ar(i)-NH(i+1, i+2 and i+3) interactions (Fig. 4). Ar(i)-NH(i−1, i−2 and i−3) interactions are very few, therefore, no statistical analysis is shown. The graph presented in Fig. 4 indicates that the NH groups of Asp, Cys, Ser, Thr, and Gln residues show a marked preference for Ar(i)-NH(i+1) interaction with the side chains of Phe, Tyr and Trp. On the other hand, amino acids such as Leu, Ala, Phe, and Tyr occur infrequently at position i+1 in Ar(i)-NH(i+1) interaction. This implies the involvement of polar residues in such interactions. This finding agrees well with the results of the database search by Toth et al. [13]. In particular, a striking feature that can be noted is that the donor residue in Ar(i)-NH(i+2, i+3) interaction is strongly preferred to be Gly while such a preference is not observed for Ar(i)-NH(i+1) interaction (Fig. 4b,c). Around 39% of Ar(i)-NH(i+2) interactions have Gly at the i+2 position.

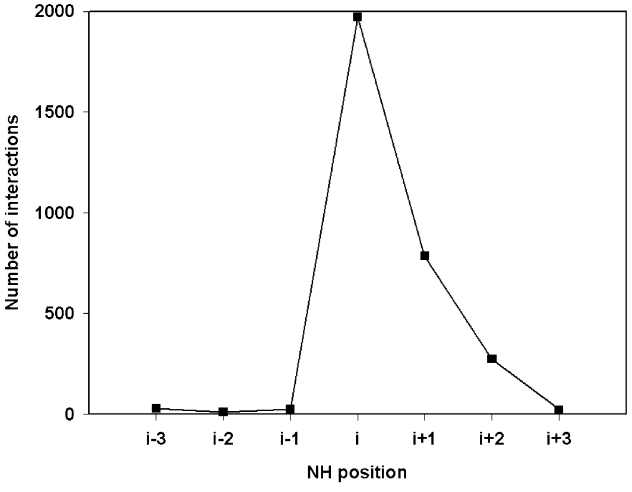

Fig. 3. A plot of the distribution of Ar-NH interactions as a function of distance between the aromatic ring and the backbone NH moiety.

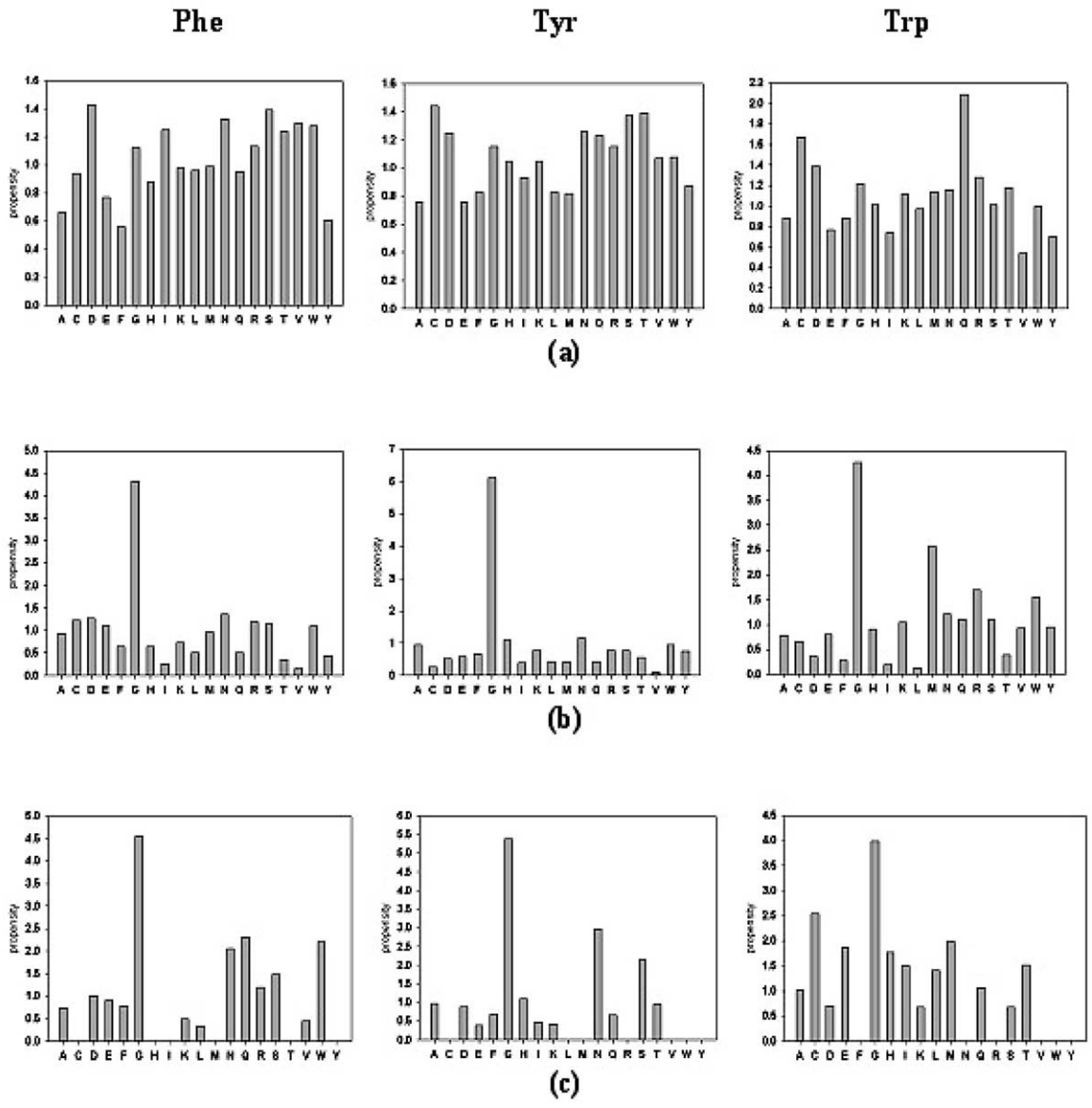

Fig. 4. Propensity values of 20 amino acids to occur in (a) Ar(*i*)-NH(*i*+1) interactions, (b) Ar(*i*)-NH(*i*+2) interactions, and (c) Ar(*i*)-NH(*i*+3) interactions.

3.1.3. Distribution in secondary structures. It has been found that Ar(*i*)-NH(*i*) interactions are equally present in turns and sheets. However, Ar(*i*)-NH(*i*+1) interactions are more predominant in β-sheets and are present at the edges of a parallel or antiparallel β-sheet, where half of the amide hydrogens are free to form an interaction with an aromatic side chain. These relatively free protons are generally present at the *i* or *i*+1 position. In contrast, in α-helices, most amide hydrogens are tied up in holding the structure together and consequently are less available for interaction with any aromatic rings.

3.2. Classification of Ar-NH interacting and non-interacting fragments

All these investigations of aromatic NH interactions concerning amino acid propensities and secondary structure preference can help in developing a method for predicting aromatic NH interactions from a given amino acid sequence. A neural network, which is a pattern recognition tool, can be used to learn these propensities/tendencies.

3.2.1. Architecture of ANN and evaluation. A number of architectures and parameters have been tried to search the best architecture and parameters for prediction. It has been observed that an ANN with seven input units and a single hidden layer with 10 units performs best, so in this study we have used this architecture. All the networks have been trained and tested using a five-fold cross-validation procedure. The prediction performance measures have been averaged over five sets and are expressed as mean ± S.D.

3.2.2. Single sequence. The ANN has been trained and tested on protein segments where amino acids in binary form (0 and 1) have been used as input, the performance of ANN is shown in Table 1. It has been found that Ar-NH interaction is predicted from sequence alone with an average accuracy of 58.3% and a MCC of 0.12. The averaged sensitivity and specificity of the network are 62.1% and 57.9%

Table 1
Performance of a network trained on a single sequence with and without secondary structure

	Network with single sequence	Network with single sequence and secondary structure	
		DSSP	PSIPRED
Accuracy	58.3 ± 0.8	61.8 ± 1.0	60.1 ± 1.0 (59.8 ± 1.0)
Sensitivity	62.1 ± 1.1	69.1 ± 2.0	66.5 ± 1.6 (66.0 ± 1.5)
Specificity	57.9 ± 2.0	63.1 ± 1.0	59.4 ± 2.2 (58.8 ± 2.0)
Probability of correct prediction of positives	13.0 ± 0.5	15.9 ± 0.8	14.2 ± 0.3 (14.0 ± 0.5)
Probability of correct prediction of negatives	93.8 ± 1.0	97.0 ± 0.4	94.6 ± 0.5 (94.0 ± 0.2)
MCC	0.12 ± 0.01	0.19 ± 0.01	0.15 ± 0.01 (0.15 ± 0.01)
Better-than-random (*S*)	7.4 ± 0.05	12.3 ± 0.05	9.8 ± 0.05 (9.8 ± 0.05)

respectively. The probabilities of correct prediction of positive examples (fragments having Ar-NH interactions) and negative examples (fragments having no Ar-NH interactions) are 13.0% and 93.8% respectively which indicates that the percentage of correctly predicted positive examples is comparatively low. Further, we have examined whether the result is better than random or not. It has been found that performance is 7.4% better than random prediction.

3.2.3. Single sequence and secondary structure information. In order to improve the performance of the method, we have incorporated secondary structure information. Here, two networks have been used. The first network classifies Ar-NH interacting and non-interacting fragments based on a single sequence. The second network utilizes the output obtained from the first network and secondary structure information. Both the observed (DSSP) and predicted (PSIPRED) secondary structure information has been used (Table 1). There is a gain of 2% in prediction accuracy using PSIPRED information. MCC is raised from 0.12 to 0.15 using predicted secondary structure information and to 0.19 using observed secondary structure information. An improvement of 4% in sensitivity and 2% in specificity has also been achieved. The probabilities of correct prediction of positives and negatives have also been increased by 1%. The performance with DSSP is higher than with PSIPRED, which is due to the accuracy of secondary structure prediction. To know whether the improvement due to incorporation of secondary structure information is real or the result of chance variation, statistical significance has been calculated. According to the statistical significance measure (described in Section 2), the improvement with secondary structure over sequence is found to be statistically significant at the 95% confidence level.

Although our dataset is non-homologous, it contains some of the protein chains used to train PSIPRED. As a consequence, we have cross-validated the results by removing those proteins from our dataset that were used to develop PSIPRED. The values are given in parentheses in Table 1. It is clear that the difference in prediction results is very small or almost negligible, thus the results are not biased by PSIPRED.

3.2.4. Evolutionary information in the form of multiple sequence alignment. We have employed evolutionary information (in the form of multiple sequence alignment) for prediction. In this case, the input to ANN is a multiple sequence alignment instead of a single sequence. For this purpose, multiple alignment in the form of position-specific scoring matrices (generated using PSI-BLAST) have been used. As shown in Table 2, the performance of the method is improved sig-

Table 2
Performance of network trained on multiple alignment with and without secondary structure

	Network with multiple alignment	Network with multiple alignment and secondary structure	
		DSSP	PSIPRED
Accuracy	64.1 ± 0.8	70.9 ± 1.0	70.1 ± 1.0 (69.4 ± 1.1)
Sensitivity	64.5 ± 1.1	69.0 ± 2.0	68.0 ± 1.6 (68.1 ± 1.5)
Specificity	64.5 ± 2.0	73.2 ± 1.0	71.0 ± 2.2 (71.0 ± 2.0)
Probability of correct prediction of positives	15.1 ± 0.5	19.5 ± 0.8	17.6 ± 0.3 (17.3 ± 0.4)
Probability of correct prediction of negatives	94.3 ± 1.0	95.6 ± 0.4	94.7 ± 0.5 (94.3 ± 0.5)
MCC	0.15 ± 0.01	0.22 ± 0.01	0.20 ± 0.01 (0.20 ± 0.01)
Better-than-random (*S*)	10.9 ± 0.05	17.2 ± 0.05	15.2 ± 0.05 (15.0 ± 0.05)

nificantly when multiple alignment is used. The prediction accuracy increases from 58.3% to 64.1% and MCC from 0.12 to 0.15. All other measures also increase. With this approach, the overall performance is improved as compared to that of a random prediction by 10.9%, which is better than that obtained with sequence only. Furthermore, the results are significant at the 95% confidence level, which indicates that the difference in performance so obtained using multiple sequence alignment is true and not spurious.

3.2.5. Multiple alignment and secondary structure information. In order to study the combined effect of multiple alignment and secondary structure information on prediction, we have used multiple alignment and secondary structure information from DSSP and PSIPRED as input to the ANN. The prediction results are shown in Table 2. This combination achieved prediction accuracy of 70.9%, 70.1% and MCC 0.22, 0.20 with DSSP and PSIPRED secondary structure respectively. The final sensitivity and specificity values are 68% and 71% with PSIPRED, which are respectively 2% and 13% higher than the performance obtained with sequence alone. These values indicate that use of secondary structure along with multiple alignment considerably increases the number of true positives and true negatives and decreases over- and under-predictions. Such improvements are also found to be statistically significant at the 0.95 confidence level. The network has a performance 15.2% higher than random prediction and is the best achieved so far in comparison to the other three above-mentioned approaches.

To check whether the better prediction performance with secondary structure information is due to PSIPRED or not, the results have been cross-validated by removing those proteins from the dataset that were used to develop PSIPRED. The results given in parentheses in Table 2 show negligible differences in performance measures.

3.2.6. ROC results. The results shown in Tables 1 and 2 depend on the threshold/cut-off value chosen according to the output of the network, so the results described so far are threshold-dependent. It means that one can achieve better performance for a particular method by varying the threshold. For instance, one can have a higher probability of correct prediction at the cost of low MCC. Thus, a better comparison between methods can be made using the single threshold-independent measure ROC. For all the methods, the ROC curve has been plotted between 1−specificity and sensitivity values. Fig. 5 shows the ROC curves of four different networks that have been used for prediction. By calculating the areas under

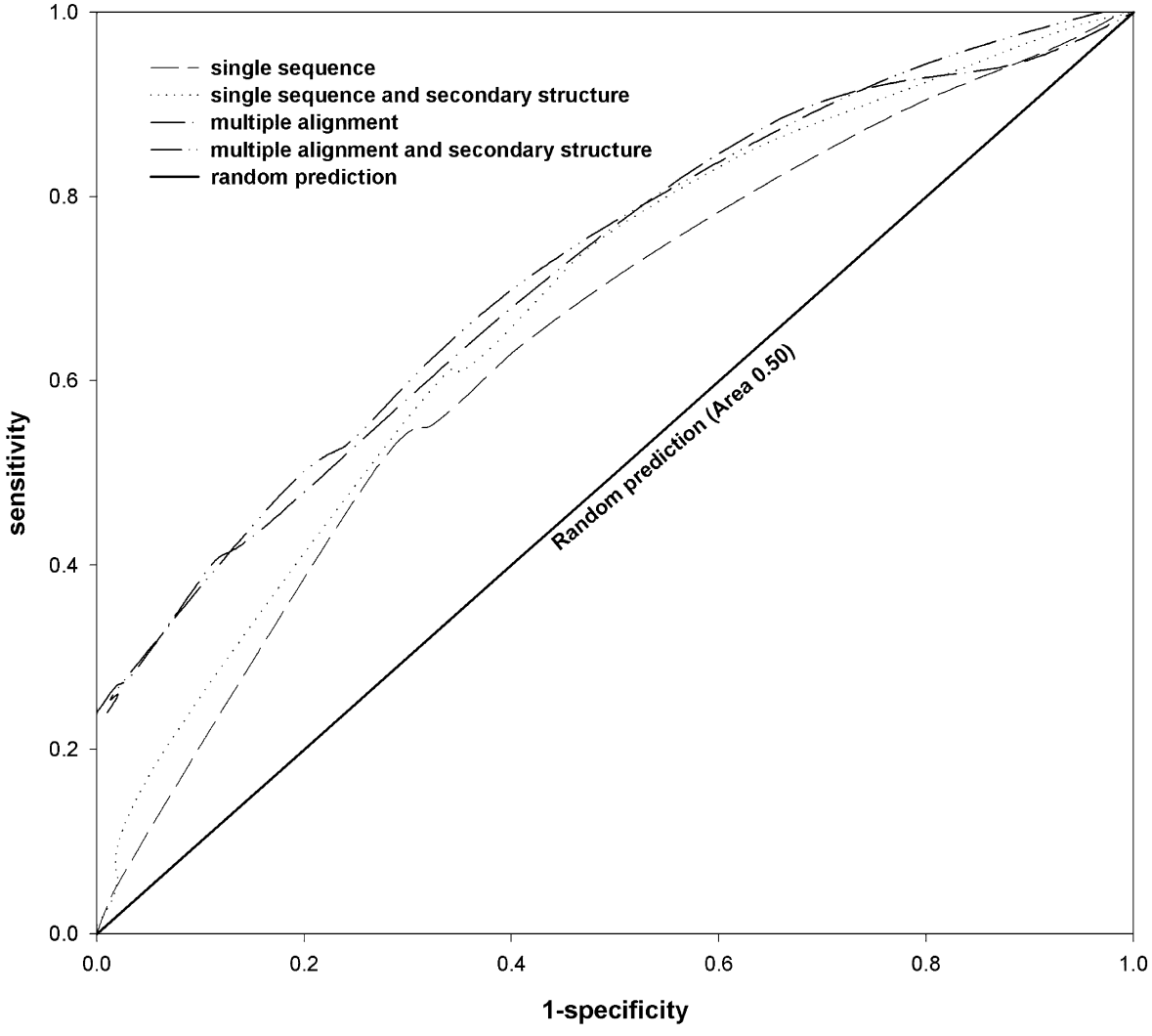

Fig. 5. ROC curves for four different network systems.

Table 3
Results of prediction of position of donor residue with sequence-to-structure network

	Interaction type		
	Ar(*i*)-NH(*i*)	Ar(*i*)-NH(*i*+1)	Ar(*i*)-NH(*i*+2)
Accuracy	59.2	74.3	86.0
Sensitivity	84.3	22.1	17.4
MCC	0.23	0.13	0.18
Better-than-random (*S*)	19.6	11.4	17.3

the curves, the performance of different networks has been compared. The corresponding areas under the curves are: single sequence, 0.59; multiple alignment, 0.65; single sequence with secondary structure, 0.68; and multiple alignment with secondary structure, 0.74. Therefore, the final, fourth network system consisting of a first network trained on PSI-BLAST PSSM and a second network trained with secondary structure tends to show better prediction performance. The ROC results are in agreement with threshold-dependent results.

3.3. Prediction of position of donor residues

In the first stage (described above), we have predicted whether a given fragment has Ar-NH interaction or not. Further, within the positively predicted fragment, the actual location/position of the donor residue has been predicted using a separate network trained with single sequences. A window size of seven residues has been used. The performance of the method obtained after five-fold cross-validation is shown in Table 3. We found that the network reached overall accuracies of 59.2%, 74.3% and 86.0% for Ar(*i*)-NH(*i*), Ar(*i*)-NH(*i*+1) and Ar(*i*)-NH(*i*+2) interactions respectively. The corresponding MCC values are 0.23, 0.18 and 0.13. Among all the interaction types, a higher percentage of Ar(*i*)-NH(*i*) is predicted correctly with sensitivity 84.3% and probability of correct prediction 54% followed by Ar(*i*)-NH(*i*+1) interactions with sensitivity 22.1%. The large difference in sensitivity values can be attributed to the fact that Ar(*i*)-NH(*i*) interactions are present in a large fraction, comprising more than 50% of the total interactions found. Thus, its performance is better than other interaction types. Overall the results are better than random prediction, by 20%, 11%, and 17% for interactions Ar(*i*)-NH(*i*), Ar(*i*)-NH(*i*+1) and Ar(*i*)-NH(*i*+2) respectively. Surprisingly, the network does not recognize even a single Ar(*i*)-NH(*i*+3), Ar(*i*)-NH(*i*−3), Ar(*i*)-NH(*i*−2) or Ar(*i*)-NH(*i*−1) interaction. This is due to the fact that there are very few occurrences of these interactions in the dataset (< 5%) and also neural networks perform poorly when the available data are sparse.

3.4. Ar_NHPred server

Based on the present study, a web server Ar_NHPred has been developed that allows the user to predict the Ar-NH interactions in a given amino acid sequence. The user can enter a single-letter amino acid sequence and the output consists of predicted donor and acceptor residues along with their respective position numbers in the sequence. A sample of the prediction output is shown in Fig. 6.

4. Discussion and conclusions

In the present study, we have made a systematic attempt towards the prediction of Ar-NH(backbone) interactions in proteins utilizing the successful approaches commonly used in the field of protein structure prediction. It has been shown in past that ANN is a powerful classification tool and can predict protein secondary structures with high accuracy [14,17–20,26]. Thus, we have used ANN for prediction. It is also well known that evolutionary information in the form of multiple alignments and profiles significantly improves the accuracy [16]. This is because the secondary structure of a fam-

Type of Interaction	Acceptor Residue[a]	Donor Residue[b]
Ar(i)-HN(i+1)	F^{4}	C^{5}
Ar(i)-HN(i+2)	Y^{10}	G^{12}
Ar(i)-HN(i+2)	Y^{23}	A^{25}
Ar(i)-HN(i+1)	F^{33}	V^{34}
Ar(i)-HN(i+1)	Y^{35}	G^{36}

[a] position number of acceptor residue
[b] position number of donor residue

Fig. 6. A sample of the prediction output of the Ar_NHPred server. The output consists of input sequence (sequence submitted by the user) where the donor and acceptor residues involved in Ar-NH interaction are marked in blue and red. The residue predicted both donor and acceptor or in which the NH moiety interacts with the aromatic ring of the same residue is marked in green. Below the input sequence, the results are presented in tabular format consisting of type of Ar-NH interaction, donor and acceptor residues and their respective positions.

ily is more conserved than the primary amino acid sequence. Thus, in this work, multiple alignments have been used as input for ANN instead of single sequence for classification. First, the method predicts whether a given protein fragment will have Ar-NH interaction or not. For this, two different inputs coding to the network have been used. One is based on single sequence with amino acids as binary bits and the other is multiple sequence alignment in the form of PSI-BLAST-generated position-specific scoring matrices. In both cases, a second 'structure-to-structure' network has further been trained on secondary structure information from PSIPRED. It has been found that the performance of the network with multiple sequence alignment is superior to that of sequence alone. With multiple sequence alignment, the method has a prediction accuracy of 64.1% and MCC is 0.15, which is 11% higher than the random prediction. The sensitivity and specificity are 64.5%. Even with multiple sequence alignment, the probability of correct prediction of Ar-NH interactions is very low, just 15%. This is due to the fact that in the dataset the number of examples having Ar-NH interactions is far lower than the number of negative examples, which in turn results in a large number of false positive predictions and thus a lower probability of correct prediction. Using secondary structure along with multiple sequence alignment improves the overall performance with a final accuracy of 70.1% and MCC is 0.20. There is 2% gain in probability of correct prediction of positive examples, which is found to be statistically significant. This network predicts whether the query sequence has Ar-NH interaction or not and does not provide the actual position of the residue whose NH group is involved in the interaction. Thus, the donor residues within the positively predicted fragments have been predicted using a separate network trained with single sequences as input. The results clearly shows the ability of the method to predict Ar(i)-NH(i, i+1 and i+2) interactions (MCC greater than 0.1) and Ar(i)-NH($i-1$, $i-2$, $i-3$) interactions (MCC < 0.05). The same distinction applies to how well the method performs compared to chance; the normalized better-than-random performance exceeds 10% for Ar(i)-NH(i, i+1 and i+2) interactions. These results are in line with our expectations; the most numerous Ar(i)-NH(i) (more than 50%) are predicted more accurately than the less numerous ones. In the dataset, Ar(i)-NH($i-1$, $i-2$, $i-3$) interactions are less than 5% whereas Ar(i)-NH(i), Ar(i)-NH(i+1) and Ar(i)-NH(i+2) interactions comprise 63.1%, 25.2% and 8.8% respectively.

This is the first method developed for prediction of Ar-NH interactions in proteins, so we cannot compare it with any other method. The overall performance of the method is poor in comparison to secondary structure prediction or tight turn prediction methods. This is expected because the Ar-NH interaction is not so specific and is very rare ($\sim$10%) in proteins, thus one can expect poor performance due to large unbalancing between positive and negative examples. Secondly, these interactions are not a consistent feature of proteins unlike helices and β-sheets. Moreover, the number of helical and sheet residues is far greater than the number of residues involved in Ar-NH interactions. However, a comparison can be made with prediction of tight turns such as β-, γ- and α-turns, which occur in small fractions in proteins. β-turns constitute 25% of the protein residues and a similar approach of multiple alignment and secondary structure gave an MCC of 0.43 for a β-turn study [18]. No doubt, the dataset used in the present study is definitely more unbalanced (1:10) than the β-turn study (1:4), so the performance of the present method is not as good as β-turns. However, if we consider β-turn types such as prediction of type I β-turns, MCC is 0.22 and the S score is 18% [17]. For type VIII β-turns, MCC and the S score are just 0.06 and 4.5%, which clearly indicates that due to its rare occurrence type VIII β-turn is predicted poorly in comparison to type II β-turns [17]. Even in the present study, the MCC and the S score obtained with multiple sequence alignment and secondary structure are only 0.20 and 15% respectively. Ar(i)-NH(i) interactions are predicted with an MCC of 0.23, 20% higher than a chance prediction. For γ- and α-turns, the approach of multiple alignment and neural network gives MCCs of 0.17 [19] and 0.16 [20] respectively. Moreover, the probabilities of correct prediction of γ- and α-turns are only 6.3% and 9.4%. These values are also not so impressive, and such poor performance is due to the fact that neural networks perform poorly when the data are highly unbalanced (as in the case of γ- and α-turns) and are sparse. So, in the present case also, the performance is not up to the mark, which is due to the unbalanced dataset (resulting in a large number of false positive predictions) and such an interaction is not a constant feature of proteins. It is possible to increase the prediction accuracy using a balanced dataset (equal number of positive and negative examples) but this network would fail in real life where the number of interacting residues is low in proteins. The present work can be extended towards the analysis of correlation between the number of Ar-NH interactions and solvent accessibility and its further integration along with other biological features associated with such interactions into the prediction method. It is possible to achieve high efficiency if more information is given in the input to the network system.

In conclusion, Ar_NHPred is the first approach for prediction of Ar-NH interactions from sequence. The method could be particularly useful in protein structure prediction, for instance the architecture of a neural network for secondary structure prediction that utilizes multiple sequence alignments can be extended to include the information of these interactions as additional input. In addition, Ar-NH interactions are responsible for shaping local structures, thus provided the interactions between residues are known for a protein sequence, the major features of its three-dimensional structures can be deduced by combining this knowledge with correctly predicted motifs of secondary structure. It would lead to a better understanding of the mechanism of protein folding as these interactions assist in folding by stabilizing intermediate structures along the folding pathway [9,10]. Since it is known that β-sheets have a higher content of Ar-NH($i-1$, i+1) interactions in comparison to α-helices [13], the prediction results can be used to relate a protein to its structural class. One can also take into account Ar-NH interactions in proteins especially in the modeling of β-sheet regions. The prediction of Ar-NH interactions in a protein sequence can be related to the upfield and downfield NH shifts in its nuclear magnetic resonance data.

Acknowledgements: We thank the Council of Scientific and Industrial Research (CSIR) and the Department of Biotechnology (DBT), Government of India, for financial assistance. We are also thankful to the developers of SNNS and PSIPRED and special thanks are due to M. Madan Babu, MRC Laboratory of Molecular Biology, Cambridge,

UK for assisting us in the identification of Ar-NH interactions in the dataset.

References

[1] Baker, E.N. and Hubbard, R.E. (1984) Prog. Biophys. Mol. Biol. 44, 97–179.
[2] Jeffrey, G.A. and Saenger, W. (1991) Hydrogen Bonding in Biological Structures, Springer-Verlag, New York.
[3] Desiraju, G.R. and Steiner, T. (1999) The Weak Hydrogen Bond in Structural Chemistry and Biology, Oxford University Press, Oxford.
[4] Weiss, M.S., Brandl, M., Sühnel, J., Pal, D. and Hilgenfeld, R. (2001) Trends Biochem. Sci. 26, 521–523.
[5] Brandl, M., Weiss, M.S., Jabs, A., Sühnel, J. and Hilgenfeld, R. (2001) J. Mol. Biol. 307, 357–377.
[6] Derewenda, Z.S., Lee, L. and Derewenda, U. (1995) J. Mol. Biol. 252, 248–262.
[7] Steiner, T. and Koellner, G. (2001) J. Mol. Biol. 305, 535–557.
[8] McPhail, A.T. and Sim, G.A. (1965) Chem. Commun. 124–125.
[9] Kemmink, J. and Creighton, T.E. (1993) J. Mol. Biol. 234, 861–878.
[10] Kemmink, J. and Creighton, T.E. (1995) J. Mol. Biol. 243, 251–260.
[11] Burley, S.K. and Petsko, G.A. (1986) FEBS Lett. 203, 139–143.
[12] Burley, S.K. and Petsko, G.A. (1988) Adv. Protein Chem. 39, 125–189.
[13] Toth, G., Watts, C.R., Murphy, R.F. and Lovas, S. (2001) Proteins 43, 373–381.
[14] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[15] Rost, B. (1996) Methods Enzymol. 266, 525–539.
[16] Przybylski, D. and Rost, B. (2002) Proteins 46, 197–205.
[17] Shepherd, A.J., Gorse, D. and Thornton, J.M. (1999) Protein Sci. 8, 1045–1055.
[18] Kaur, H. and Raghava, G.P.S. (2003) Protein Sci. 12, 627–634.
[19] Kaur, H. and Raghava, G.P.S. (2003) Protein Sci. 12, 923–929.
[20] Kaur, H. and Raghava, G.P.S. (2004) Proteins 55, 83–90.
[21] Rost, B., Fariselli, P. and Casadio, R. (1996) Protein Sci. 5, 1704–1718.
[22] Fariselli, P. and Casadio, R. (1999) Protein Eng. 12, 15–21.
[23] Compiani, M., Fariselli, P., Martelli, P. and Casadio, R. (1998) Proc. Natl. Acad. Sci. USA 95, 9290–9294.
[24] Pollastri, G., Baldi, P., Fariselli, P. and Casadio, R. (2001) Bioinformatics 17, S234–S242.
[25] Altschul, S.F., Madden, T.L., Alejandro, A.S., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[26] Jones, D.T. (1999) J. Mol. Biol. 292, 195–202.
[27] Brendel, V. (1992) Math. Comp. Model. 16, 37–43.
[28] Bernstein, F.C., Koetzle, T.F., Williams, G., Mayer, E.F., Bryce, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T. and Tasumi, M. (1997) J. Mol. Biol. 112, 535–542.
[29] Babu, M.M. (2003) Nucleic Acids Res. 31, 3345–3348.
[30] Steiner, T. and Koellner, G. (2001) J. Mol. Biol. 305, 535–557.
[31] Zell, A. and Mamier, G. (1997) Stuttgart Neural Network Simulator, version 4.2, University of Stuttgart, Stuttgart.
[32] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Nature 323, 533–536.
[33] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[34] Matthew, B.W. (1975) Biochim. Biophys. Acta 405, 442–451.
[35] Deleo, J.M. (1993) in: Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, pp. 318–325, IEEE, Computer Society Press, College Park, MD.
[36] Daniel, W.W. (1987) Biostatistics: A Foundation for Analysis in the Health Sciences, John Wiley and Sons, New York.
[37] Zhang, X., Mesirov, J.P. and Waltz, D.L. (1992) J. Mol. Biol. 225, 1049–1063.