Prediction of Super-
Secondary Structure
Motif
β-hairpin
Chapter 8: Prediction of super-secondary structure motif β-hairpin

Currently available high-throughput sequencing facilities have generated large amount of raw sequence data, making it possible to know the sequences of an increasing number of proteins. The ultimate goal of all genome projects is to understand all levels of functions that include biochemical, cellular, metabolic as well as phenotypic. Most often function annotation is accomplished via sequence similarity to other proteins with known function. Because protein structures are more strongly conserved during evolution than sequences hence structure information should enhance the function prediction (Skolnick et al., 2000). Though protein databank (PDB) have around 50,000 protein but unique structures are limited (Berman et al., 2000). To reduce the gap between known sequences and know structure, there is need to develop accurate methods for protein structure prediction. The available methods of structure prediction can be divided into three categories; (i) knowledge based methods, (ii) ab initio methods and (iii) hierarchical approach. In absence of sequence similarity, hierarchical approach is most successful for predicting structure of proteins. In the hierarchical approach first an intermediate structure (like 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 (Jones, 1999; Kaur and Raghava, 2002; Kaur and Raghava, 2003; Kaur and Raghava, 2004). This work was an attempt to improve the performance of hierarchichal methods.

In this chapter a method developed for predicting super-secondary structural motif β-hairpin has been described. It is a machine learning based method which uses PSI-BLAST generated PSSM, surface accessibility and secondary structure. The final method was evaluated on an independent protein dataset not used during development of method. A web-server BhairPred (http://www.imtech.res.in/raghava/bhairpred) is also described which was developed for predicting β-hairpin in a protein sequence.

8.1. Methodology

8.1.1. Dataset of β-hairpins and non-hairpins

The dataset for this study was generated from 2880 non-redundant protein chains of known structures (henceforth referred as main dataset) obtained from
No two protein chains of main dataset have percentage identity >33%. The final dataset of hairpins and non-hairpins was generated using following strategy.

a. Secondary structure was assigned to each amino acid of all 2880 proteins using DSSP (Kabsch and Sander, 1983).

b. From these proteins 12653 unique amino acid patterns with secondary structure \( \beta\)-c-\( \beta \) (minimum two consecutive amino acids in each state, later designated ECE patterns) were extracted.

c. Using PROMOTIF (Hutchinson and Thornton, 1996) 6,675 \( \beta \)-hairpins were obtained, among which 6549 had unique amino acid patterns.

d. A total of 5,820 \( \beta \)-c-\( \beta \) patterns (obtained from DSSP as described in step b), which were also assigned as hairpins by PROMOTIF, were finally considered as \( \beta \)-hairpins; the remaining 6,833 \( \beta \)-c-\( \beta \) patterns as non-hairpins.

e. Most of the patterns obtained during step d had 6-30 residues. Hence, only 5,548 hairpins and 6,322 non-hairpins with length between 6 and 30 amino acid residues were kept in the final dataset.

f. Requirement of fixed-length patterns by machine-learning methods compelled us to fix the pattern length to 17 residues using following the steps.

   o If pattern length was less than 17, flanking residues were appended at both ends in order to make for the required length.

   o If pattern length was more than 17, only those patterns were kept for further study whose coil region was up to 10 residues long.

   o In the case of pattern length more than 17 and coil region was up to 10 residues, central coil residue was mapped and 9 residues from the left-hand side and 8 residues from the right-hand side were taken to make the final input pattern.

Using the strategy described above we obtained a library of 5102 hairpins and 5113 non-hairpins of length 17 residues.
8.1.2. Thornton's dataset

This dataset was made up of 534 protein chains obtained from Cruz et al. (de la Cruz et al., 2002). Using the rules described above a library of hairpins and non-hairpins of length 17 was generated. DSSP find 2229 ECE patterns, of which PROMOTIF classified 1169 as hairpins and 1060 as non-hairpins. Finally, after excluding the patterns having chain break or heteroatoms, 1076 β-hairpins and 878 non-hairpins of length 17 residues were obtained.

8.1.3. Secondary structure and surface accessibility

The secondary structure and surface accessibility were assigned using the DSSP program. Predicted secondary structure and surface accessibility were obtained from PSIPRED (Jones, 1999) and NETASA (Ahmad and Gromiha, 2002) programs respectively.

8.1.4. Multiple sequence alignment and PSSM

Multiple sequence alignment of each protein was performed using PSI-BLAST (Altschul et al., 1997) at e-value threshold 0.001 against ‘nr’ protein database (obtained from www.ncbi.nlm.nih.gov) with three iterations. Intermediate PSSM generated after three iterations of PSI-BLAST search (inclusion E-value 0.001) were used as direct input to the ANN and the SVM. Input PSSM had $21 \times M$ elements, where $M$ is the length of a pattern. For detail description please see section 3.3.5.1.

8.1.5. Feature representation

For the sequence based model, each pattern was represented by $21 \times 17$ units, where 21 binary vectors were used to represent an amino acid (20 for the amino acid and one for terminal residues) (see section 3.3.3. for detail). In the case of multiple alignment or evolutionary profile, the PSSM matrix corresponding to the pattern was used. In accessibility models, one unit was added with binary amino acid notation to represent the accessibility of residues: 0 for buried and 1 for exposed residues. Thus, accessibility models have $17 \times 22$ units for each pattern. In the case of the secondary structure model, three units were used in the test/predicted case and one unit was added in the ideal/assigned case.

8.1.6. Machine learning methods
Two machine-learning methods ANN and SVM have been used to develop the predictor. The overall structure of the ANN used for prediction is shown in Figure 8.1. It contains two standard feed forward back propagation neural networks—'Sequence-to-Structure' and 'Structure-to-Structure', In sequence-to-structure layer PSSM was used as input while in structure-to-structure layer output of first network and PSIPRED (Jones, 1999) predicted secondary structure were used as input. The input window size and the number of hidden units have been optimized in order to get the best prediction. In both networks the best performance was achieved with input window of size 17 with a single hidden layer. But the number of hidden nodes was different in both networks. After optimization it was observed that in 1st 12 and in 2nd 23 hidden units was optimal. In case of SVM Three kernels, namely, linear, polynomial and sigmoid with different kernel parameters, were used for the training. The attributes for single sequence are binary numbers 0 and 1, and for multiple alignments real values of PSSM obtained from PSI-BLAST search against ‘nr’ protein database.

![Figure 8.1: The neural network system used for prediction of β-hairpins.](image)

8.1.7. Consensus and combination prediction

In order to utilize the strength of both the ANN and SVM based approaches, the results were combined in intersection and union modes, called as the consensus and combined prediction modes, respectively. Consensus and combined approaches are
analogous to the logical Boolean operators ‘AND’ and ‘OR’ respectively. 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.

8.1.8. Evaluation and performance measures

Performance of all methods and models was evaluated using five-fold cross-validation. For performance estimation following parameters were used (i) accuracy of prediction, (ii) MCC, (iii) sensitivity or percentage coverage of hairpins, (iv) specificity or percentage coverage of non-hairpins, (v) probability of positive (or hairpin) prediction and (vi) probability of negative (or non-hairpin) prediction. For detail procedure about evaluation and performance measure please see section 3.4.

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. However, this approach does not reflect the true picture, because in real life the algorithm has to discriminate between hairpin and 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. The aim was to demonstrate the capability to predict hairpins in real life, when secondary structure of protein is not known.

8.1.9. Independent dataset

All the 63 targets released for CASP6 competition were used as independent dataset. It was later used for independent evaluation of performance. The prediction results were compared with hairpin assignment of PROMOTIF. In addition 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 also computed.

8.2. Results
8.2.1. Performance on Main dataset

8.2.1.1. ANN approach

When only 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 (Figure 8.2). Accuracy and MCC improved considerably (65.5 to 67.0% and 0.31 to 0.34%, respectively) when PSSM was used as 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. During training of the method, predicted secondary structure and surface accessibility was used with PSI-BLAST profile (real-life situation). In the test case maximum accuracy of 67.1% and MCC of 0.37 was achieved using the ANN (Figure 8.2).

8.2.1.2. SVM approach

The accuracies of 68.1 and 74.9% in the case of single sequence and multiple alignments, respectively, have been using the SVM technique (Figure 8.3). 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. The schematic diagram of overall architecture of SVM used for hybrid mode of prediction is shown in Figure 8.4. In the test case an accuracy of 77.9% and MCC of 0.56 were obtained using the predicted secondary structure and accessibility (Figure 8.3). These results clearly indicate the superior performance of the SVM over the ANN in the prediction of β-hairpins.
Figure 8.2: Prediction results with main dataset protein using ANN.
Prediction of \( \beta \)-hairpins

Figure 8.3: Prediction results with main dataset protein using SVM.
Prediction of β-hairpins

Figure 8.4: Overall prediction system of hybrid mode of SVM prediction.

Consensus approach

Combined approach

Figure 8.5: Combined and consensus mode of prediction.

8.2.1.3. Combination of SVM and ANN
In order to utilize the capabilities of both approaches the prediction outputs of SVM and ANN were combined. Figure 8.5 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 8.5, 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 Figure 8.5, 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.

8.2.1.4. Performance on Thornton's dataset

As shown in Figure 8.6 the performance of our method on Thornton's dataset decreased by 2.2, 4.1 and 6.8%, respectively when amino acid, 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 (Figure 8.7). 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 the present method on main and Thornton's dataset was analyzed, it was observed that, although the performance decreased, the trend remained the same.

8.2.2. Performance on CASP6 proteins

To perform an impartial review of our methodology, we predicted hairpins in all the 63 CASP6 target proteins using our BhairPred web server. In order to avoid any bias during prediction we used default prediction parameter (e.g. threshold) of the server. The following information was compiled for each protein: (i) number of ECE patterns obtained by PSIPRED predicted secondary structure, (ii) number of ECE patterns predicted as hairpins by BhairPred and (iii) number of hairpins assigned by PROMOTIF. 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 8.1).

Figure 8.6: Performance of SVM on Thornton’s dataset.
Hairpins assigned by PROMOTIF were also examined (a total of 159 hairpins in all 63 proteins). Comparing PROMOTIF assigned hairpins with PSIPRED predicted ECE patterns, we found 27 exact matches in which 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). 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
unsuccesful in the NF category, which was due to the failure to correctly predict the ECE pattern by PSIPRED in the NF category (Table 8.2). 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.

This proves that BhairPred can 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 assign the secondary structure of query protein during submission. 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 considerably 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.

<table>
<thead>
<tr>
<th>CASP6 categories</th>
<th>#of proteins</th>
<th># of ECE patterns</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>#of discarded ECE patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>63</td>
<td>201</td>
<td>47</td>
<td>85</td>
<td>17</td>
<td>31</td>
<td>21</td>
</tr>
<tr>
<td>NF</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FR(A)</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FR(H)</td>
<td>10</td>
<td>23</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>CM</td>
<td>20</td>
<td>55</td>
<td>11</td>
<td>24</td>
<td>5</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 8.1:   Performance of BhairPred on CASP6 target proteins.

TP= Number of correct hairpin prediction
TN= Number of correct non-hairpin prediction
ALL= Number of target proteins in CASP6
NF= New Fold

8.3. Limitations

FP= Number of incorrect predicted hairpin
FN= Number of incorrect predicted non-hairpin
FR(A)= Fold Recognition (Analogous)
FR(H)= Fold Recognition (Homologous)
CM= Comparative Modeling
The major and obvious limitation of BhairPred is strict requirement of fixed length input pattern. It resulted in removal of patterns that didn’t qualify in the criteria laid down for appropriate input pattern. Thus, the result obtained should not be compared with that of Cruz et al. (de la Cruz et al., 2002), because in that procedure all hairpins and non-hairpins were considered. In the present case only those β-c-β regions which satisfy the conditions described earlier were considered. An average of 20% of hairpins and non-hairpins were excluded from our dataset. Therefore, the method described should be considered as a supplement of Cruz et al. (de la Cruz et al., 2002). It is hoped that this study will be a useful addition to protein tertiary structure prediction.

<table>
<thead>
<tr>
<th>CASP6 categories</th>
<th>Number of proteins</th>
<th>Number of hairpins</th>
<th>Exact matching</th>
<th>Non-exact matching</th>
<th>Non-exact at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>63</td>
<td>159</td>
<td>27(22)</td>
<td>51(25)</td>
<td>61</td>
</tr>
<tr>
<td>NF</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>1(1)</td>
<td>7</td>
</tr>
<tr>
<td>FR(A)</td>
<td>6</td>
<td>9</td>
<td>2(2)</td>
<td>4(2)</td>
<td>3</td>
</tr>
<tr>
<td>FR(H)</td>
<td>10</td>
<td>20</td>
<td>4(3)</td>
<td>8(3)</td>
<td>6</td>
</tr>
<tr>
<td>CM</td>
<td>20</td>
<td>46</td>
<td>5(4)</td>
<td>13(7)</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 8.2: Number of hairpins assigned by BhairPred in each category of CASP6. Values in parentheses show number of correctly predicted hairpins by BhairPred.

8.4. Web server

The results shown above clearly shows that performance of our approach is better than existing similar methods. Hence to make it publicly available a web-server, BhairPred, has been developed. It is publicly available at http://www.imtech.res.in/raghava/bhairpred. BhairPred predicts β-hairpins in proteins using an SVM model developed on hybrid input (HYRBID2 input mode of Figure 8.3). The server performs following steps: (i) accepts protein sequences from the user in any standard format (FASTA, PIR, EMBL); (ii) predicts the secondary structure using PSIPRED; (iii) identifies the β-c-β regions in the protein; (iv) generates an PSSM using PSI-BLAST; (v) predicts surface accessibility of each residue using NETASA and (vi) finally predicts whether identified β-c-β regions are hairpins or non-hairpins from predicted secondary structure and PSI-BLAST profile. The
BhairPred server provides various advance options also, including selection of an appropriate threshold value for predicting β-hairpins (Figure 8.8). The output is displayed in tabular format with complete details of the predicted sheet–coil–sheet patterns in proteins. The final prediction is the list of potential β-hairpins (Figure 8.9). The server also provides advanced options such as assignment of user-defined secondary structure of query protein instead of using the PSIPRED predicted secondary structure.

**Figure 8.8:** The sequence submission page of BhairPred web-server.

**Figure 8.9:** Prediction result of BhairPred web-server.