Prediction of β-Hairpins in Proteins Using Physicochemical Properties and Structure Information

Jun-Feng Xia1,2, Min Wu1,3,4, Zhu-Hong You1,3, Xing-Ming Zhao1,5 and Xue-Ling Li1,*

1Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China; 2School of Life Science, University of Science and Technology of China, Hefei, Anhui 230027, China; 3Department of Automation, University of Science and Technology of China, Hefei, Anhui 230027, China; 4School of Electrical Engineering and Automation, Hefei University of Technology of China, Hefei, Anhui 230009, China; 5Institute of Systems Biology, Shanghai University, Shanghai, 200444, China

Abstract: In this study, we propose a new method to predict β-Hairpins in proteins and its evaluation based on the support vector machine. Different from previous methods, new feature representation scheme based on auto covariance is adopted. We also investigate two structure properties of proteins (protein secondary structure and residue conformation propensity), and examine their effects on prediction. Moreover, we employ an ensemble classifier approach based on the majority voting to improve prediction accuracy on hairpins. Experimental results on a dataset of 1926 protein chains show that our approach outperforms those previously published in the literature, which demonstrates the effectiveness of the proposed method.

Keywords: β-Hairpin, support vector machine, majority voting, protein supersecondary structure prediction.

INTRODUCTION

The gap between the number of known amino acid sequences and the number of known three dimensional structures grows rapidly because of the progress of genome-sequencing projects [1]. Hence, there is a great need for accurate protein structure prediction methods to bridge the gap [2-7]. In proteins, protein secondary structure elements are observed to associate in a particular geometric arrangement to form supersecondary structures. Common supersecondary motifs include α-β-motif, β-α-motif, α-helix hairpin, β-link and β-hairpin [8]. Elements of secondary structure and supersecondary structure can then combine to form the full three-dimensional fold of a protein. As the information gained can be used in the tertiary structure prediction, protein supersecondary structure prediction is a key step in the hierarchical approach to derive the protein tertiary structure [9]. The protein supersecondary structure prediction methods developed thus far can be divided into two categories [1]: the first focuses on finding a number of different supersecondary structures at once [10], whereas the second is based on predicting special structures such as β-hairpins. Our study falls into the latter category.

β-hairpin is a simple supersecondary structure formed by two adjacent, antiparallel, hydrogen-bonded β-strands [11, 12]. As one of the most common types of supersecondary motifs in proteins, β-hairpin bears great significance in folding stability [13], molecular recognition [14] and structure assembly [15] of a protein. The identification of β-hairpin in proteins is very important because it can reduce the number of possible folds available to that protein and then improve tertiary structure prediction [1, 16]. Several methods based on different algorithms for predicting β-hairpin have been proposed in the past few years. Cruz et al. [9] tried a neural network method [17, 18] based on the decomposition of the problem into the two separate parts: secondary structure prediction and β-hairpin motif identification. According to Cruz’s research, the main contributors to β-hairpin motif recognition are predicted accessibility and turn propensities. Kuhn et al. [1] also used the neural network method to predict strand-loop-strand motif types in proteins. This approach built two separate neural networks for identifying the state of the beginning and end residue in strand-loop-strand motif, respectively. They achieved an accuracy of 77.3% in predicting hairpins. Kumar et al. [16] predicted hairpins using two machine learning methods (neural network and support vector machine (SVM)) based on the single sequence information, evolutionary profile, surface accessibility and secondary structure information. Their final SVM model yielded a highest accuracy of 79.2% in predicting hairpins in a protein. Recently, Hu et al. [19] used a composite vector method to predict hairpin motifs in a protein by combining the score of position weight matrix, increment of diversity, the value of distance and autocorrelation information. They obtained an overall accuracy of 83.1% in a protein dataset with 3088 non-homologous protein chains.

In this paper, we present a new method to predict hairpins in proteins and its evaluation based on the SVM [20]. The SVM is a machine learning method based on statistical learning theory, which projects the input data into a higher dimensional space and finds a hyperplane that best separates the data. SVM are used successfully in a wide area of biological domains. Different from previous methods, new feature representation scheme based on auto covariance (AC)
[21, 22] is adopted. AC describes the level of the correlation between amino acids within a certain number of amino acids apart throughout the whole sequence in terms of their specific physicochemical property. As a result, it can take into account the neighboring residues effect, which is important for representing the hairpin information. In addition, we also investigate two structure properties of proteins (protein secondary structure and residue conformation propensity), and examined their effects on prediction. Finally, we employ an ensemble classifier approach based on the simple majority voting to improve prediction accuracy of hairpins. We train and test the proposed method on a dataset of 1926 protein chains using 5-fold cross-validation. The results indicate that our method yields significantly better prediction accuracy than those previously published in the literature.

The rest of the paper is organized as follows. Section 2 describes the dataset, the feature representation of β-hairpin, and the models used in this paper. The experimental results for our approach and corresponding discussions are given in Section 3, and Section 4 presents the conclusions.

MATERIALS AND METHODS

Dataset of β-Hairpins and Non-Hairpins

The representative protein dataset was selected from the pdbselect [23] 25% list of November 2009, in which no two proteins have more than 25% sequence identity. Protein chains in this dataset solved by X-ray crystallography were further filtered with a resolution better than 2.0 Å, and refined with an R factor ≤ 0.23. After these filtering procedures, we obtained the final dataset of 1926 protein chains. Then we identified β-hairpins with the following rules [16].

(i) Secondary structure was defined to each amino acid of all proteins using DSSP [24].

(ii) The eight secondary structure classes were reduced to three states according to the following method: H, G and I to H; E, B to E; the rest to C.

(iii) From these proteins 3932 unique amino acid patterns with secondary structure ECE (minimum two consecutive amino acid residues in each state, ) were extracted.

(iv) The program PROMOTIF [25] was implemented to identify the observed β-hairpin in these proteins, and 2007 β-hairpins were obtained.

(v) A total of 1694 patterns, which were also assigned as hairpins by PROMOTIF, were finally considered as β-hairpins, the remaining 2238 patterns were considered as non-hairpins.

Machine learning methods such as SVM and the neural network require a fixed number of inputs for training. However, hairpin and non-hairpin patterns are often unequal-length vectors because of patterns sequences with different lengths. So we generated uniform hairpin and non-hairpin patterns of 17 amino acids using the steps described below [26].

(i) If length of coil region is odd, the central coil was mapped and eight residues from both sides are taken, respectively.

(ii) If length of coil region is even, the central coil was mapped and eight residues from the left hand side and seven residues from the right hand side are taken.

(iii) In the case of pattern length <17, residues flanking the peptides in the primary amino acid sequence were appended at both the ends.

(iv) Finally, a total of 1377 β-hairpins and 2198 non-hairpins of length 17 were obtained.

Feature Representation

In the present study, three different characteristics are employed for the hairpins classification, including physicochemical features, secondary structure information assigned by DSSP and conformation propensity calculated by position weight matrix [19].

Physicochemical Features

The five physicochemical properties of amino acids are hydrophobicity, hydrophilicity, polarity, polarizability and average accessible surface area (Table 1). The original values of the five physicochemical properties for each amino acid are obtained from the AAindex database [27]. Each of these properties was centralized and standardized according to:

\[ P_i' = \frac{(P_i - \overline{P})}{\sigma} \]  

where \( \overline{P} \) is the average of the property of the 20 amino acids, and \( \sigma \) is the corresponding standard deviation, given by:

\[ \overline{P} = \frac{\sum_{i=1}^{20} P_i}{20} \]  

and

\[ \sigma = \sqrt{\frac{1}{20} \sum_{i=1}^{20} (P_i - \overline{P})^2} \]

Secondary Structure

The secondary structure information assigned by DSSP was encode as follows: helix → (1, 0, 0), strand → (0, 1, 0), and coil → (0, 0, 1).

Conformation Propensity

To adequately take into account the amino acids position conservation effect in hairpin sequence segments, we used a 2-dimensional vector calculated by position weight matrix to encode an amino acid. In this vector, the first unit represents the conformation propensity factor of hairpin and the second is for non-hairpin. More details about the implementation of this algorithm can be referred to the original paper [19, 28, 29].

Thus, we can extract 5, 3, 2 features from physicochemical features, secondary structure and conformation propensity, respectively. For secondary structure and conformation propensity, we used the conventional orthogonal encoding method. The conventional orthogonal encoding uses input vector of 17 contiguous amino acid residues, corresponding to a sliding window containing the target residue and 8
neighboring residues on either side of the target residue. Each of the 17 residues in the window is represented by 3-bit and 2-bit vector for secondary structure and conformation propensity respectively.

For the physicochemical features, we adopted AC [21, 22] descriptor to infer hairpins. As a new feature representation, AC has been adopted by more and more investigators for protein classification. For each sequence, AC variables describe the average interactions between residues, in a certain range of $d$ along the sequence. As a result, they can take into account the local environment of residues in the sequences. The AC variables were calculated as:

$$AC_{d,j} = \frac{1}{N-d} \sum_{i=d}^{N} (P_{i,j} - \frac{1}{N} \sum_{i=1}^{N} P_{i,j}) (P_{i+(d-j),j} - \frac{1}{N} \sum_{i=1}^{N} P_{i,j})$$

where $j$ represents one property of the amino acid, $d=1,2,...,16$ is the distance between one residue and its neighbors, $N$ is the length of the pattern sequence, and $P_{i,j}$ and $P_{i+(d-j),j}$ are the properties $j$ of the amino acid at positions $i$ and $i+d$, respectively.

**Model Construction**

The classification model for predicting hairpin was based on SVMs [20], which are a class of effective supervised learning methods that demonstrate high prediction accuracy while efficiently avoiding the overfitting problem. In this study, the software LIBSVM [30] was employed and the radial basis kernel function was selected to build the SVM models. The SVM models were created with a set of default parameters. To further improve the prediction performance, we used a simple majority voting method to combine the results of individual SVM models.

**Evaluation Measures**

The performance of the proposed ensemble method is measured using 5-fold cross-validation. Each data set is randomly divided into five subsets with an approximately equal number of polypeptide chains. Each classifier is trained and tested five times with one dataset. And for each time, four subsets are used as training data and the remaining subset is used as test data.

### Table 1. The Original Values of the Five Physicochemical Properties for Each Amino Acid

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Hydrophobicity</th>
<th>Hydrophilicity</th>
<th>Polarity</th>
<th>Polarizability</th>
<th>Average accessible surface area</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>-0.5</td>
<td>8.1</td>
<td>0.046</td>
<td>27.8</td>
</tr>
<tr>
<td>C</td>
<td>0.04</td>
<td>-1</td>
<td>5.5</td>
<td>0.128</td>
<td>15.5</td>
</tr>
<tr>
<td>D</td>
<td>-0.72</td>
<td>3</td>
<td>13</td>
<td>0.105</td>
<td>60.6</td>
</tr>
<tr>
<td>E</td>
<td>-0.62</td>
<td>3</td>
<td>12.3</td>
<td>0.151</td>
<td>68.2</td>
</tr>
<tr>
<td>F</td>
<td>0.61</td>
<td>-2.5</td>
<td>5.2</td>
<td>0.29</td>
<td>25.5</td>
</tr>
<tr>
<td>G</td>
<td>0.16</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>24.5</td>
</tr>
<tr>
<td>H</td>
<td>-0.4</td>
<td>-0.5</td>
<td>10.4</td>
<td>0.23</td>
<td>50.7</td>
</tr>
<tr>
<td>I</td>
<td>0.73</td>
<td>-1.8</td>
<td>5.2</td>
<td>0.186</td>
<td>22.8</td>
</tr>
<tr>
<td>K</td>
<td>-1.1</td>
<td>3</td>
<td>11.3</td>
<td>0.219</td>
<td>103</td>
</tr>
<tr>
<td>L</td>
<td>0.53</td>
<td>-1.8</td>
<td>4.9</td>
<td>0.186</td>
<td>27.6</td>
</tr>
<tr>
<td>M</td>
<td>0.26</td>
<td>-1.3</td>
<td>5.7</td>
<td>0.221</td>
<td>33.5</td>
</tr>
<tr>
<td>N</td>
<td>-0.64</td>
<td>2</td>
<td>11.6</td>
<td>0.134</td>
<td>60.1</td>
</tr>
<tr>
<td>P</td>
<td>-0.07</td>
<td>0</td>
<td>8</td>
<td>0.131</td>
<td>51.5</td>
</tr>
<tr>
<td>Q</td>
<td>-0.69</td>
<td>0.2</td>
<td>10.5</td>
<td>0.18</td>
<td>68.7</td>
</tr>
<tr>
<td>R</td>
<td>-1.76</td>
<td>3</td>
<td>10.5</td>
<td>0.291</td>
<td>94.7</td>
</tr>
<tr>
<td>S</td>
<td>-0.26</td>
<td>0.3</td>
<td>9.2</td>
<td>0.062</td>
<td>42</td>
</tr>
<tr>
<td>T</td>
<td>-0.18</td>
<td>-0.4</td>
<td>8.6</td>
<td>0.108</td>
<td>45</td>
</tr>
<tr>
<td>V</td>
<td>0.54</td>
<td>-1.5</td>
<td>5.9</td>
<td>0.14</td>
<td>23.7</td>
</tr>
<tr>
<td>W</td>
<td>0.37</td>
<td>-3.4</td>
<td>5.4</td>
<td>0.409</td>
<td>34.7</td>
</tr>
<tr>
<td>Y</td>
<td>0.02</td>
<td>-2.3</td>
<td>6.2</td>
<td>0.298</td>
<td>55.2</td>
</tr>
</tbody>
</table>
To assess the performance of classification methods, we adopted a number of commonly used measures: Specificity, Recall, Precision, Accuracy and Matthews correlation coefficient (MCC). These measures are defined as follows:

\[
\text{Specificity} = \frac{TN}{(TN + FP)} \\
\text{Recall} = \frac{TP}{(TP + FN)} \\
\text{Precision} = \frac{TP}{(TP + FP)} \\
\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \\
\text{MCC} = \frac{(TP*TN - FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},
\]

where TP, FP, TN and FN represent true positive (correctly predicted \(\beta\)-hairpin), false positive (non-hairpin incorrectly predicted as \(\beta\)-hairpin), true negative (correctly predicted non-hairpin) and false negative (\(\beta\)-hairpin incorrectly predicted as non-hairpin), respectively.

RESULTS AND DISCUSSIONS

Selection of Optimal \(d\)

The use of AC with large \(d\) value will result in more variables that account for interactions of amino acids with large distances apart in the sequence. In this study, several \(d\) values were optimized in order to achieve the best characterization of the protein sequences. The prediction results for SVM classifier with AC transformation of different \(d\) values are shown in Fig. (1). As seen from the curve, there is a peak point with an average MCC of 10.51% and the corresponding \(d\) of 13 amino acids. The results illustrate that AC with \(d\) less than 13 amino acids would lose some useful classification information of protein sequence and larger \(d\) could introduce noise instead of improving the prediction power of the model. So the optimal \(d\) is 13 amino acids.

![Figure 1](image-url)

**Figure 1.** The average MCC of SVM classifier with AC of different \(d\) values.

Effect of Different Features on \(\beta\)-Hairpin Prediction

In previous studies, many features have been adopted to get improved predictions of \(\beta\)-hairpins in proteins. Based on these studies, we employed three features: auto covariance based on physicochemical features (ACP), secondary structure information (SS) and conformation propensity (CP). SVM classifiers were then built to discriminate between \(\beta\)-hairpins and non-hairpins based on these features. The predictive performances of these models are illustrated in Table 2, with the details of the SVM classifiers based on 5-fold cross-validation. Using conformation propensity, secondary structure and physicochemical features as the input of SVMs, the average prediction Accuracy is 0.5650, 0.5773 and 0.5580, respectively. Among the two structure properties of proteins, the use of protein secondary structure information improves the accuracy of prediction compared with residue conformation propensity such as MCC was improved from 0.1507 to 0.1739. When physicochemical features were used as the input, it can correctly predict \(\beta\)-hairpins from the data set with Recall = 0.5359 and Precision = 0.4395. This means that it correctly predicted 53.59% of the true \(\beta\)-hairpins for this data set (Recall), and 43.95% of the predicted \(\beta\)-hairpins are identified as true \(\beta\)-hairpins (Precision). The physicochemical features efficiently identified non-hairpins (Specificity = 0.5719), while the structure features based methods could correctly identify \(\beta\)-hairpins (Recall = 0.6304, 0.6405, respectively). As the ACP based method can predict with high accuracy non-hairpin regions in proteins, it reduces the number of theoretical folds available to proteins, and thereby decreasing the time and cost for protein fold recognition methods. So it can be a good supplement for protein structure prediction.

Ensemble Classifier for Hairpin Prediction

In order to utilize the capabilities of different features, the above three models were further combined by majority voting (MV) so as to improve the final prediction performance. The prediction results are listed in Table 2. We can see that the model based on MV gives good results with the average Specificity, Recall, Precision, Accuracy and MCC of 0.5692, 0.6383, 0.4814, 0.5958, and 0.2020, respectively. The MV model loses some Specificity relative to the ACP based model but increases the recall on the other hand. In other words, more positive \(\beta\)-hairpin predictions are made with a slight lower percentage of true negatives. Each of single models correctly predicts a different subset of \(\beta\)-hairpins. In summary, we conclude that our ensemble model for predicting \(\beta\)-hairpins achieved a satisfactory performance.

Comparison with Other Studies

Currently, the best \(\beta\)-hairpin prediction is obtained by the composite vector method [19], a quadratic discriminant method that takes the scores of sequence segment, the increments of diversity, the values of distance and the autocorrelation information as the input parameters. A direct comparison with this study is difficult due to the differences in choice of dataset. We request the dataset from the authors but have not received any response. So we applied the composite vector method to our dataset. Table 3 summarizes the performance comparison of different methods on the same dataset. The MCC value obtained by the present method is 0.0996 higher than obtained by composite vector method (QD [PWM (Ap), ID (Bo), ID (Co), D (Ad), Z]). It has to be noted that the enhancement of MCC mainly comes from the improvement of Recall, and that the composite vector method has achieved higher Specificity than our model. In-
terestingly, we found that the results of composite vector method containing five parameters were not better than the algorithm with single parameter (i.e. PWM(Ap), ID(Bo), ID(Co) and D(Ad)), for more details on these algorithms, please refer to the original paper [19]), while in Hu’s article [19] the composite vector method obtained the best predictive results. Compared with the algorithm with single parameter, the decrease of MCC is mainly due to lower percentage of true positives (lower Recall), although it efficiently identified non-hairpins than the algorithm with single parameter (higher Specificity). The reason for this flaw might be that we used a small dataset while quadratic determinants method is highly dependent on the quality of training data, which, in some sense, supports that MV is more insensitive to dataset than the composite vector method.

Three factors may account for the remarkable performance of the present method. First, AC was adopted to encode the physicochemical features. AC accounts for the interactions between residues a certain distance apart in the sequence, so it can effectively takes the neighboring effect into account. Second, the protein secondary structure information was used as input the classifier. In previous studies [9, 16], it has been shown that when the protein secondary structure was added to the input, the computational method can discriminate β-hairpins from non-hairpins with very high accuracy. Third, an ensemble classifier approach based on the majority voting was proposed to improve prediction accuracy of hairpins. It is pointed out that one single classification system cannot always provide high classification accuracy [17, 31]. Instead, an ensemble classifier system is proved to be more accurate and robust than an excellent single classifier in many fields [32-35]. As a result, it’s not surprising that our method obtained higher prediction accuracy.

CONCLUSION

In this work, we develop a novel method to predict hairpins in proteins. The results reported here demonstrate that approaches based on protein physicochemical and structural properties can predict hairpins. Furthermore, we employ an ensemble classifier approach based on the simple majority voting to improve prediction accuracy of hairpins. Our method outperforms the current best hairpin prediction method on the same dataset. In conclusion, the proposed method will be a powerful tool to infer hairpins and expedite the study of protein tertiary structure prediction.

ACKNOWLEDGMENTS

This work is supported by the grants from the National Science Foundation of China (60873012, 60805021), the National Basic Research Program of China (2007CB-311002), the National High Technology Research and Development Program of China (2006AA02Z309), the Guide Project of Innovative Base of Chinese Academy of Sciences (KSCX1-YW-R-30) and the Knowledge Innovation Program of the Chinese Academy of Sciences (0823A16121).

REFERENCES


