A Protein Secondary Structure Prediction Method Based on BP Neural Network

Ru-xi YIN, Li-zhen LIU*, Wei SONG, Xin-lei ZHAO and Chao DU

Information and Engineering Department, Capital Normal University, Beijing 100048, PR China
*Corresponding author

Keywords: Protein secondary structure, BP neural network, Profile encoding, 5-bit encoding.

Abstract. Protein structure prediction is a main task in the field of bioinformatics, and the prediction of protein secondary structure is the key point of this task. Extracting representative features and efficient classification methods are the basis of the prediction process. In this paper, a prediction method based on BP neural network is proposed. We use valid protein features extracted by variable-sized sliding window and two different encoding modes (5-bit encoding and Profile encoding) as input data, to make predictions for the secondary structure of proteins. The prediction accuracies are calculated by Jackknife test on three commonly used low-similarity protein datasets: 25PDB, 1189 and 640, and this method achieves a high overall accuracy upon these three datasets.

Introduction

Protein structure prediction plays an important role in understanding the connection between protein structure and protein function. Therefore, it becomes the key task in the field of drug design, mutant design and so on. Protein structure prediction, which means predicting the 3-dimentional structure (contains the secondary structure and the tertiary structure) of a sequence of amino acid, is now an urgent problem that remains to be solved in the field of Molecular biology researches. Protein secondary structure prediction is a hotspot issue in bioinformatics, it is not only the bond between the primary structure and the tertiary structure of proteins, but also key step from the prediction of protein primary structure prediction to the prediction of protein tertiary structure. Protein secondary structure prediction is helpful to understanding the function and mechanism of action, and it is also significant for making precise predictions of protein structures. [1] Protein secondary structure prediction is a computational problem in molecular biology and a multi-dimensional nonlinear mapping problem in mathematics. Currently, there are several ways to identify protein secondary structures, based on biochemical experiments, but this process is time-consuming and laborious, and there are some limitations. Compared with the biochemical experiment method, the bioinformatics-based method has the characteristics of time-saving and effort-saving.

Protein structure prediction can be divided into four broad categories based on the theoretical basis of the model: statistical-based model, biochemical-based model, machine learning model (including the nearest neighbor method, neural network, support vector machine, etc.), and mixed model. [2] The artificial neural network method, which is a machine learning model, has been widely studied and applied because of its diverse methods and suitability for computer calculation.

In this paper, the method is based on BP neural network. We use different coding methods and sliding windows with different sizes, and select quasi-Newton BP algorithm as the training function to predict the protein secondary structure and obtain precise predicting results.

Methods and Materials

BP (Back Propagation) neural network is a one-way propagation of multi-layer feed-forward neural network. BP learning algorithm is a neural network with three or more layers, including the input layer, the middle layer (hidden layer) and the output layer. The upper and lower layers are fully connected, and there is no connection between each layer of neurons. When the learning sample is provided to the network, the activation value of the neuron propagates from the input layer (through
the middle layer) to the output layer, and the output response of the network is obtained for each neuron of the output layer. Then, according to the direction of reducing the error between the target output and the actual value, the output layer is corrected by the middle layers. This algorithm is called error back propagation algorithm, that is, BP algorithm. As the propagation of this error continues being proceed, the correct rate of the network to the input pattern is also rising. [3]

BP neural network transfer function requirements must be differentiable, the commonly used transfer function are:

Log-sigmoid type function:

\[
\log \text{sig}(n) = \frac{1}{1 + \exp(-n)}
\]

Tan-sigmoid type function:

\[
\tan \text{sig}(n) = \frac{2}{1 + \exp(-2*n)} - 1
\]

Linear function:

\[
\text{purelin}(n) = n
\]

In this paper, we use tan-sigmoid type function as the transfer function.

Dataset

Sequence similarity plays an important role in prediction accuracy of protein structures. [4] We select three low-similarity protein datasets: the 25 PDB dataset (443all-α, 443 all-β, 346 α/β, 441 α+β), the 1189 dataset (223 all-α, 294 all-β, 334 α/β, 241 α+β), and the 640 dataset (138 all-α, 154 all-β, 177 α/β, 171 α+β). The first one is extracted from the Protein Data Bank, and the sequence similarity is lower than 25%, so it is named by 25 PDB. As for the 1189 dataset, the sequence similarity is no more than 40%. And the last one, 640 dataset, contains 640 protein sequences, whose similarities are lower than 25%. All of the above mentioned three datasets are low-similarity protein datasets.

The three datasets are commonly used in researches, and they all include high-resolution non-homologous protein sequences. In this paper, we are supposed to obtain a reliable predicting result, and make it comparable with other methods, so these three datasets are suitable choices.

On the basis of the above datasets, we extract effective protein features as experimental data by using DSSP [5] software to generate the dssp file of each protein strand. And then extract the sequence information and corresponding secondary structure information of the protein sequences from their dssp file.

Encoding Mode

The appropriate coding method is one of the important factors which determines the final performance of the system. Different amino acid encoding modes have different influence on the accuracy of protein secondary structure prediction. [6] We still need further researches to find out which kind of amino acid encoding mode can achieve a high prediction rate. Currently, the commonly used encoding modes include: orthogonal encoding, 5-bit encoding, and Profile encoding.

Each protein is composed of 20 amino acids, and each of them could be represented by a 20-position binary vector, and the inner product of the coding vector values between different amino acids is zero. The advantage of orthogonal coding representation is that no algebraic correlation is involved. For example, the orthogonal code of amino acid A is 100000000000000000000000000 and the orthogonal code of amino acid Y is: 0000000000000000000000000001. The all 20 kinds of amino acids are marked as: A, C, E, D, G, F, I, H, K, M, L, N, Q, P, S, R, T, W, V, Y, and their orthogonal code are shown in Table 1.
Table 1. The orthogonal code of amino acids.

| A    | 10000000000000000000 | L    | 00000000001000000000 |
| C    | 01000000000000000000 | N    | 00000000000100000000 |
| E    | 00100000000000000000 | Q    | 00000000000100000000 |
| D    | 00010000000000000000 | P    | 00000000000010000000 |
| G    | 00001000000000000000 | S    | 00000000000001000000 |
| F    | 00000100000000000000 | R    | 00000000000000100000 |
| I    | 00000010000000000000 | T    | 00000000000000010000 |
| H    | 00000001000000000000 | W    | 00000000000000001000 |
| K    | 00000000100000000000 | V    | 00000000000000000100 |
| M    | 00000000010000000000 | Y    | 00000000000000000010 |

As for 5-bit encoding, that is, turning each decimal integer code of the amino acids to into 5-bit binary code. For example, if amino acid A corresponds to a decimal value of 1, after the step of conversion, its 5-bit binary encoding is 00001, and if amino acid Y corresponds to a decimal value of 20, its converted 5-bit binary code is 10100.

Profile is a two-dimensional array, the first dimension corresponds to the sequence number, each row is a 20-dimensional vector (corresponds to 20 kinds of amino acids), each element in the vector (corresponds to the second dimension of the array) represents the appearing frequency of each amino acid at this position. Profile encoding contains a lot of biological evolution information. In this paper, we extract the corresponding Profile code of each protein sequence was extracted from its dssp file, and then use 5-bit encoding method and Profile encoding method to make predictions for protein secondary structures.

Construction of BP Neural Network

Protein secondary structure prediction is based on the protein sequence, by using prediction methods and techniques to achieve the classification of secondary structure prediction. As for BP neural network, the input of the protein is the primary structure, the output is the type of the secondary structure. In this paper, a three-layer BP neural network (Figure 1) is used as a classifier to predict the secondary structure of the protein. And the input layer of the BP network is designed as a window that slides along the amino acid sequence, and the position of the window is symmetrical. The input layer of the BP neural network respectively uses sliding windows of 3, 5, and 9, which are sliding windows along the amino acid sequence. Each time, it makes prediction for the amino acids which are in the middle of the sliding window. As shown in Figure 1, the input layer is 7 amino acids: CD A W K R V. And now it is time to predict for the middle amino acid W, and then turns to the next one K.

![Figure 1. The BP neural network used in this paper.](image)

By the method of 5-bit encoding, the number of neurons in the sliding window of the corresponding BP neural network is 3*5, 5*5, and 9*5, respectively. While by the method of Profile encoding, the number of neurons in the sliding window of the corresponding BP neural network is 3*20, 5*20, and 9*20, respectively. And the numbers of neurons in the hidden layers are 10, 20, and 20. The output layer is composed of three neurons, corresponding to three kinds of protein secondary structures: H-spiral (1 0 0), E-fold (0 1 0) and C-curl (0 0 1). The three neurons of the output layer are encoded as 3-dimensional binary vectors. Assuming that the output of the three
neurons in the output layer is (0.5 0.6 0.2), the amino acid residues in the middle of the input layer are determined to be E-fold (0 1 0). The sliding window continues to move to the next amino acid residues position, and then all the amino acid residues could be predicted.

**Prediction Assessment**

Currently, there are three kinds of widely used testing algorithms: independent dataset test, subsampling test, and jackknife test. And in these three methods, the last one is believed as the most objective one, in order to get a comprehensive assessment for the experimental results, we choose jackknife test to assess this method.

In the process of assessment, we select one protein sequence at each time for testing, and the rest of them remain to be the training group. We take the Overall Accuracy (OA) as the measurement for the accuracy of the prediction method. In order to make a comprehensive assessment, we add three parameters: Sensitivity (Sens), Specificity (Spec), and Matthew’s Correlation Coefficient (MCC). The value of MCC ranges from -1 to 1, if it equals to 0, it means random correlation. The parameters are defined as below:

\[
\begin{align*}
OA &= \frac{Num^+(C_i) - Num^-(C_i) + TN}{TP + TN + FP + FN} \\
Sens &= \frac{TP}{TP + FN} \\
Spec &= \frac{TN}{TN + FP} \\
MCC &= \frac{1 - \frac{FN}{TP + FN} - \frac{FP}{TN + FP}}{\sqrt{\left(1 + \frac{FP}{TN + FP}\right)\left(1 + \frac{FN}{TP + FN}\right)}}
\end{align*}
\]

where, \(TP\) means true positive, \(TN\) means true negative, \(FP\) means false positive, and \(FN\) means false negative. Each benchmark dataset contains two parts: proteins which belong to this class, and those who do not belong to this class. The first part of each dataset is marked as \(C_i(i = 1, 2, 3)\), the summation of \(C_i\) could be defined as:

\[
\begin{align*}
N_{num^+} (C_i) \\
N_{num^-} (C_i)
\end{align*}
\]

\(TP\), \(TN\), \(FP\) and \(FN\) could be defined as:

\[
\begin{align*}
TP &= Num^+(C_i) - Num^-(C_i) \\
TN &= Num^-(C_i) - Num^+(C_i) \\
FP &= Num_-(C_i) \\
FN &= Num_+(C_i)
\end{align*}
\]

So that equation (4) could be defined as:

\[
\begin{align*}
OA &= \frac{Num^+(C_i) - Num^-(C_i) + Num^+(C_i) - Num^-(C_i)}{Num^+(C_i) + Num^-(C_i)} \\
Sens &= \frac{Num^+(C_i) - Num^-(C_i)}{Num^+(C_i)} \\
Spec &= \frac{Num^+(C_i) - Num^-(C_i)}{Num^-(C_i)} \\
MCC &= \frac{1 - \frac{\frac{Num^+(C_i)}{Num^+(C_i)} + \frac{Num^-(C_i)}{Num^-(C_i)}}{\sqrt{\left(1 + \frac{Num^-(C_i) - Num^+(C_i)}{Num^+(C_i)}\right)\left(1 + \frac{Num^-(C_i) - Num^+(C_i)}{Num^-(C_i)}\right)}}}
\end{align*}
\]

For each experiment, the amount of \(Num^+(C_i)\), \(Num^-(C_i)\), \(Num^+_i(C_i)\) and \(Num^-_i(C_i)\) could be...
easily figured up. The overall assessment could be made according to the values of the Overall accuracy (\(OA\)), Sensitivity (\(Sens\)), Specificity (\(Spec\)), and Matthew’s Correlation Coefficient (\(MCC\)).

**Results and Analysis**

In this paper, we select pairs of two, four and eight amino acids which are adjacent to each other to build the input vector with the window size of 3, 5 and 9 for the BP neural network. After the training step, we make prediction tests, and finally calculate the prediction accuracy based on 5-bit encoding method and Profile encoding method respectively. The overall accuracies are listed below in Table 3.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Win 3</th>
<th>Win 5</th>
<th>Win 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-spiral</td>
<td>63.90</td>
<td>74.65</td>
<td>60.32</td>
</tr>
<tr>
<td>E-fold</td>
<td>30.67</td>
<td>59.20</td>
<td>39.43</td>
</tr>
<tr>
<td>C-curl</td>
<td>57.91</td>
<td>77.34</td>
<td>60.01</td>
</tr>
<tr>
<td>average</td>
<td>50.83</td>
<td>70.40</td>
<td>53.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Win 3</th>
<th>Win 5</th>
<th>Win 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-spiral</td>
<td>59.23</td>
<td>63.41</td>
<td>62.54</td>
</tr>
<tr>
<td>E-fold</td>
<td>48.96</td>
<td>50.77</td>
<td>55.80</td>
</tr>
<tr>
<td>C-curl</td>
<td>62.54</td>
<td>67.93</td>
<td>67.46</td>
</tr>
<tr>
<td>average</td>
<td>56.91</td>
<td>60.70</td>
<td>61.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Win 3</th>
<th>Win 5</th>
<th>Win 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-spiral</td>
<td>50.43</td>
<td>53.90</td>
<td>53.88</td>
</tr>
<tr>
<td>E-fold</td>
<td>20.55</td>
<td>22.57</td>
<td>30.37</td>
</tr>
<tr>
<td>C-curl</td>
<td>46.83</td>
<td>50.76</td>
<td>48.66</td>
</tr>
<tr>
<td>average</td>
<td>39.27</td>
<td>42.41</td>
<td>44.30</td>
</tr>
</tbody>
</table>

From Table 3, we could see that with the improvement of window size, the average overall accuracy improve a lot. For almost every dataset, when the window size equals to nine, the overall accuracies are the highest. And it is quite obvious that, the prediction accuracies of Profile encoding are higher than those of 5-bit encoding, which refers that although Profile encoding is more complicated than 5-bit coding, it is more accurate because it preserves the biological evolution information.

**Conclusion**

Proteins are the building blocks of life, and they play important roles in the body mechanism. The function of a protein depends on its structure, so that the prediction of protein secondary structures is now a highly discussed question in the field of Bioinformatics. In the past days, a lot researchers have been devoted to find better methods to make predictions of protein structures, but there still remains a lot space for improvement in low-similarity datasets.

In this paper, we propose a prediction method based on sliding window, encoding methods and BP neural network. We select representative low-similarity datasets, make tests on different window sizes and encoding methods, and use the BP neural network classifier to predict the secondary structure of proteins. The method shows a high prediction quality, especially the Profile encoding method obtains satisfactory experimental results. The feasibility and effectiveness of the experimental method are proved.

To further improve the accuracy of the forecast, a variety of other neural networks can be combined with the classifier in this paper. Or we could use a series of BP neural network together. The number of neurons in the hidden layer could also be adjusted. The use of other coding methods which contains biological evolutionary information may also improve the prediction accuracies.
Acknowledgement

This work was supported in part by National Science Foundation of China under Grants No. 61402304, the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046, the Beijing Natural Science Foundation under Grants No. 4154065, the Beijing Educational Committee Science and Technology Development Planned under Grants No. KM201610028015, Science and technology innovation platform, Teaching teacher, and Connotation Development of Colleges and Universities.

References


