Predict the Tertiary Structure of Protein with Binary Tree and Ensemble Strategy

Yiming Chen\(^1\), a, Yuehui Chen\(^2\), b

\(^1\)Department of Information Science and Engineering, University of Jinan, Jinan, 250022, China
\(^2\)Department of Information Science and Engineering, University of Jinan, Jinan, 250022, China

\(^{a}yiming198807@163.com,^{b}yhchen@ujn.edu.cn

**Keywords:** tertiary structure; binary tree; selective ensemble; FNT

**Abstract.** In this paper, we intend to apply a new method to predict tertiary structure. Several feature extraction methods adopted are physicochemical composition, recurrence quantification analysis (RQA), pseudo amino acid composition (PseAA) and Distance frequency. We construct the binary tree Classification model, and adopt flexible neural tree models as the classifiers. We will train a number of based classifiers through different features extraction methods for every node of binary tree, then employ the selective ensemble method to ensemble them. 640 dataset is selected to our experiment. The predict accuracy with our method on this data set is 63.58%, higher than some other methods on the 640 datasets. So, our method is feasible and effective in some extent.

**Introduction**

Protein plays an important role in basic life support; the study of protein tertiary structure contributes to protein function, and understands the essence of life phenomenon.

In recent years, the gap between the number of protein sequence data and structure data become more and more big, the protein structure prediction is gradually urgent and important. Efforts from many researchers have been done for many years on this field. We should find some new feature extract methods and new classifiers to improve the predict accuracy of protein tertiary structure. We apply FNT as the base classifier in this field instead of the traditional classification models.

We take several steps in our experiment: 1. Establish a data set; 2. Extract feature to obtain the information of protein sequence; 3. Design a classification model; 4. Ensemble these based classifiers.

**Dataset**

There are four benchmark datasets in the field of protein tertiary structure; they are C204 datasets, 1189 datasets, 640 datasets and 25PDB datasets. This paper we select 640 dataset to make the experiment. The sequence homology of this dataset is about 25%. It make our method more persuasive duo to the lower sequence homology.

**Feature extract methods**

**Physicochemical Composion**

The 20 amino acids are divided into three groups on the basis of their physicochemical properties, including seven types \(^7\) of hydrophobicity, normalized van der Vaals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility. For instance, we use hydrophobicity attribute to divide amino acids into three groups: polar, neutral and hydrophobic. Then a protein sequence is transformed into a sequence of hydrophobicity attribute. Thus, the composition descriptor contains three values: the global percent compositions of polar, neutral and hydrophobic residues in the new sequence. PCC consists of a total of \(3\times7=21\) descriptor values because of seven types of attributes.

**Recurrence quantification analysis**
Recurrence quantification analysis [2-6] is a powerful nonlinear method in analyzing time series; before we extract protein feature, a recurrent plot (RP) [4] of a protein sequence should be obtained. We should make a transition for protein sequence. Firstly, we convert amino acids sequence into nucleotide sequence. The encoding method is listed in the table 1 [5]. Secondly, we use Chaos game representation (CGR) [6] to describe a nucleotide sequence on a plot. CGR is defined as a [0, 1] square for a nucleotide sequence; please refer to reference [3] for the details of RQA.

DET, ENT, VMAX, LAM, REC and TT are adopted in this paper, so we can obtain twelve features for every protein sequence because of the two time series X and Y.

Subsection distance frequency

The 20 native amino acids are divided into 6 classes [22] according to the properties. We show the result of classification in table 2.

After the partition, protein sequence can be represented by combination of the six letters. For every kind of amino acids, we will separately calculate the distance-value’s occurrence number of two letters which belong to same type. One sequence thus gets one vector Vi [23] on the basis of distance frequency:

\[
V_i = \left[ v_i^L, v_i^L, \ldots, v_i^L, v_i^B, v_i^B, \ldots, v_i^B, v_i^C, \ldots, v_i^C \right]
\]

This paper the value of s is set to 11. In order to obtain the partial information we try to divide the protein sequence into three parts.

**Pseudo Amino Acid composition (PseAA)**

According to PseAA composition [8], the protein sequence can be described as:

\[
P = \left\{ p_1, p_2, \ldots, p_{20}, p_{21}, \ldots, p_{20+L} \right\} \lambda < L
\]

\[
s_i = \sum_{j=1}^{20} \frac{f_{ij}}{\sum_{j=1}^{20} f_{ij}} 1 \leq i \leq 20
\]

\[
s_i = \sum_{j=1}^{20+\lambda} \frac{w \mu_i}{\sum_{j=1}^{20+\lambda} f_{ij}} 21 \leq i \leq 20 + \lambda
\]

The first 20 components are the occurrence frequencies of 20 amino acids in sequence. \(P(21 \leq i \leq 20 + \lambda)\) are the additional factors that reflect some sort of sequence order information. In this paper the parameter w is set to 5, the parameter \(\lambda\) is set to 20; \(L\) is the length of protein sequence.

Classification model

**Binary tree model**

We design the Binary tree Classification model based on FNT classifier as follows:

![Binary Tree classification model](image)

For the first node (FNT0), four classes of protein tertiary structure are divided into two groups \(\alpha\beta\) class and \(\alpha+\beta\) \(\alpha/\beta\) class. The second node is used to classify \(\alpha\) and \(\beta\) from \(\alpha\beta\) class, finally. We can distinguish \(\alpha+\beta\) and \(\alpha/\beta\) class through FNT2 node.

**Flexible Neural Tree**

The base classifier is flexible neural tree (FNT). We use Probabilistic Incremental Program Evolution (PIPE) and Particle Swarm Optimization algorithms (PSO) to optimize the structure and parameters of FNT. FNT allows input features selection and the individuals of FNT tend to simplify...
structure of the similar model due to the evolutionary algorithm. The flexible neuron instructor and FNT model are composed of the function set $F$ and terminal instruction set $T$ described as follows:

$$S = F \cup T = \{12 + \lambda_1 + \lambda_2 + \cdots + \lambda_N \} \cup \{x_1, x_2, \ldots, x_n\}$$

(3)

The $F$ set $+i$ $(i = 2, 3, N)$ are non-leaf nodes’ instructions which has $i$ inputs. $\{x_1, x_2, \cdots, x_n\}$ are leaf nodes’ instructions i.e.,. The output of a flexible neuron operator is calculated by the activation function.

$$o_{out} = f(a_n, b_n, net_n) = e^{(net_n - c_n)/h_n}$$

(4)

$$net_n = \sum_{j=1}^{m} w_j x_j$$

(5)

Due to the limited space, please see references [12][13][14] for details of FNT.

Integration

Selective ensemble method[24] is a learning algorithm, it trains different kinds of based classifier and selects some of them to ensemble. Selecting a part of based classifier is effective than that select all based classifier.

This paper we will use five kinds of feature extraction methods to construct five different based classifier for every node of Binary tree Classification model, they are Physical and chemical composition, the fusion of Recurrence quantification analysis and Physical and chemical composition, the fusion of Pseudo Amino Acid composition and Recurrence quantification analysis, subsection Distance frequency, the fusion of Pseudo Amino Acid composition and subsection Distance frequency. Then, we apply the selective ensemble method to ensemble these based classifier.

Experimental results

We generally use the cross-validation method to evaluate the performance of classification method. The 10-jackknife cross-validation was adopted in this paper [16]. We calculate the overall success rate and accuracy of every class. We show the result obtained from different method in the table 3. From the table 3 we can conduct that the accuracy of our method is better than the result of some other experiments.

Conclusion

We construct a Binary tree Classification model based on flexible neural tree models as the classifiers. We adopt five different to construct five different based classifier and use the selective ensemble strategy to ensemble them. The results listed in Table III show that our method may make some contribution for protein structure prediction.

Acknowledgment

This research was supported by the Natural Science Foundation of China (61070130), the Key Project of Natural Science Foundation of Shandong Province (ZR2011FZ001), the Key Subject Research Foundation of Shandong Province and the Shandong Provincial Key Laboratory of Network Based Intelligent Computing.

References


Table 1 The reverse encoding for amino acids

<table>
<thead>
<tr>
<th>A</th>
<th>G</th>
<th>M</th>
<th>S</th>
<th>C</th>
<th>H</th>
<th>N</th>
<th>T</th>
<th>D</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCT</td>
<td>GGT</td>
<td>ATG</td>
<td>TCA</td>
<td>TGC</td>
<td>CAC</td>
<td>AAC</td>
<td>ACT</td>
<td>GAC</td>
<td>ATT</td>
</tr>
<tr>
<td>CCA</td>
<td>GTG</td>
<td>GAG</td>
<td>AAG</td>
<td>CAG</td>
<td>TGG</td>
<td>TCA</td>
<td>CGA</td>
<td>TAC</td>
<td>ATT</td>
</tr>
</tbody>
</table>

Table 2. Amino acids hydration property classification

<table>
<thead>
<tr>
<th>Classification</th>
<th>Abbreviation</th>
<th>Amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydrophily</td>
<td>L</td>
<td>R,D,E,N,Q,K,H</td>
</tr>
<tr>
<td>hydrophobicity</td>
<td>B</td>
<td>L,I,V,A,M,F</td>
</tr>
<tr>
<td>neutral</td>
<td>W</td>
<td>S,T,Y,W</td>
</tr>
<tr>
<td>proline</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>glycocoll</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>cysteine</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Table 3. The comparison of the results with prior research

<table>
<thead>
<tr>
<th>algorithms</th>
<th>accuracy rate</th>
<th>overall accuracy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α</td>
<td>β</td>
</tr>
<tr>
<td>IB1[18]</td>
<td>53.62</td>
<td>46.10</td>
</tr>
<tr>
<td>Naïve Bayes[18]</td>
<td>55.07</td>
<td>62.34</td>
</tr>
<tr>
<td>Logistic regression[18]</td>
<td>69.57</td>
<td>58.44</td>
</tr>
<tr>
<td>SVM[18]</td>
<td>73.91</td>
<td>61.04</td>
</tr>
<tr>
<td>This method</td>
<td>74.07</td>
<td>66.67</td>
</tr>
</tbody>
</table>