Prediction Protein Structural Classes with a Hybrid Feature

Guangting Shao, Yuehui Chen
Computational Intelligence Lab, University of Jinan, Jinan, 250022, China
shaoguangting@163.com

Abstract—Select the proper feature of protein sequence is a crucial step in protein structural class prediction. In this paper we intend to propose a novel hybrid feature to describe the protein. This hybrid feature is composed of two parts, one is physicochemical composition (PCC), and another is the recurrence quantification analysis (RQA). A new classifier is constructed with the Error Correcting Output Coding (ECOC) which incorporates three binary Artificial Neural Network (ANN) classifiers. We select 1189 data set to verify the efficiency of classify. The accuracy of our method on this data set is 57.3%, higher than some other methods on the same datasets. Furthermore only 33 parameters are used in our method, lower than many other methods. This indicates that the hybrid feature we proposed here is promising to the prediction of protein structural classes.

Keywords-protein structural classes; PCC; RQA; ECOC; ANN

I. INTRODUCTION

Knowledge of protein structure plays an important role in protein function analysis. Examination the three-dimensional structures of proteins through the X-ray diffraction or NMR has shown that these methods are very tedious for variety of proteins[1]. On the other hand protein sequence information grows significantly faster than information on protein structure, the need for predicting the folding pattern of a given protein sequence naturally arises.

Researchers have developed various schemes for protein structures classification. Many of these methods are reported very effective when they are tested on high homology data sets. However, when low-homology data sets were used, these methods were no longer effective any more. That is to say, some new methods of protein structure prediction are urgently needed.

Our experiments show that existing simplex traditional descriptor of the protein sequence, like amino acid composition, pseudo-amino acid composition etc, can’t get a rational result on 1189 data set, thus we explore to fusion these methods to get a higher accuracy. Through numerous tests, we found that parameters extracted from physicochemical composition and recurrence quantification analysis perform perfect.

II. MATERIALS AND METHODS

A. Dataset

This paper we select a low-homology data set, 1189 data set to evaluate the effective of proposed method. 1189 data set is from Wang and Yuan (2000), it contains 1092 protein domains, of which 223 are all-α class, 294 all-β, 241 α + β, and 334 α/β. The sequence homology of this data set is below 40%.

B. Feature extract methods

1) Physicochemical Composition

Physicochemical Composition has seven attributes, including hydrophobicity, normalized van der Waals volume, polarity, polarizibility, charge, secondary structures and solvent accessibility[2]. 20 native amino acids are divided into three groups for each of these physicochemical properties. For example, with hydrophobicity attribute all amino acids are divided into three groups: polar, neutral and hydrophobic. Therefore, the composition descriptor consists of three values: the global percent compositions of polar, neutral and hydrophobic residues in the new sequence. For seven types of attributes, consists of a total of 7 x 3 = 21 descriptor values.

2) Recurrence quantification analysis

RQA [3-4] is a powerful nonlinear technique in analyzing time series, it has no requirement on the length of time series. Before we use RQA to analysis the protein chain, we should get a recurrent plot (RP)[5] of a protein chain. The procedure of change a protein sequence into a RP is as follows.

Firstly, reverse amino acids sequence into nucleotide sequence. As we all know one amino acids sequence should have many possible nucleotide sequences. Here we use the encoding method used by Deschavanne and Tuffery (2008)[6] which is listed in the table 1.

Secondly, using Chaos game representation (CGR)[7] to describe a nucleotide sequence on a plot. Here, we briefly recapture the concept of CGR. For a nucleotide chain, a CGR is defined in a [0,1] x [0,1] square, where the four vertices correspond to four letters A, C, G and T. the first point of the plot is placed half way between the center of the square and the vertex corresponding to the first letter of the nucleotide sequence; the i-th point of the plot is then placed half way between the (i-1)-th point of plot and the vertex corresponding to the i-th letter[8]. Then write down the x and y coordinates of every point, we get two time series X and Y. In order to simplify the problem, we just take X as an example next.
Then for a time series $X = \{x_1, x_2, \cdots, x_N\}$ of length $N$, with a length parameter $m$ and a delay parameter $\tau$, we can get $N_m = N - (m - 1) \times \tau$ sub vectors, where sub vector $\tilde{X}_i = \{x_i, x_{i+\tau}, x_{i+2\tau}, \cdots, x_{i+(m-1)\tau}\}$. $m$ and $\tau$ are two very important parameters, reference to the work of previous investigators we select $m=8$. Due to the length of the nucleotide vary greatly, we set $\tau$ as a variable, $\tau=1$ when the length of the sequence is below 30, else $\tau=2$. Every vector $\tilde{X}_i$ is a point in space $R^m$, for $N_m$ points we can get a $N_m \times N_m$ distance matrix $DM$, the element $DM(i,j)$ is the distance between $\tilde{X}_i$ and $\tilde{X}_j$.

After got $DM$, we transform it into a recurrence matrix $RM$ with a parameter $\epsilon$, when the element of $DM$ is less than $\epsilon$, element is signed into 0, else signed to 1. Just as equation 1.

$$RM(i,j) = \begin{cases} 
0 & DM(i,j) > \epsilon \\
1 & DM(i,j) < \epsilon 
\end{cases} \quad (1)$$

RP is just a visualization of $RM$, that is plotting points on a $i-j$ plane for those elements in $RM$ with value is 1, just as Fig 1.

We can notice that there is a main diagonal line and points in the RP are symmetrical about this main diagonal line. It’s easy to understand these because the distance of two same point is zero and distance between point $i$ and point is equal to the distance between point $i$ and $j$.

Next we will use RQA to analysis the RP. Six RQA indexes, REC, DET, ENT, VMAX, LAM, and TT are used in this paper, the details for these indexes please refer to reference [4]. There are two time series for an amino acids sequence, so we can extract twelve attributes for one protein sequence.

C. ECOC classify model

ECOC[10-11] is a well-established method for solving multi-class problems by decompose multi-class problem into several complementary two-class problems. The performance of ECOC methods depends on base binary classifiers. This paper we use artificial neural network (ANN) as the base classifier of ECOC.

The procedure of ECOC prediction model are mainly divided into two steps: coding and decoding. The coding step is concern about how to decompose a multi-class problem to several binary ones. Main mission of this step is to establish a $r \times c$ code matrix $Z$. Every row of matrix is called the code word of each class. Each column of matrix is the label of each structure class in the process of training every base binary classifier. For example, in training classifier 1, samples in class 1 are labeled into 0, and when training the classifier 2, they are labeled into 1.

In the decoding procedure we concern about how to decide a final decision, we will get an output code for each test sample, it is composed by the output of three binary classifiers. The test pattern is assigned to the class that is represented by the closest code word, the distance of the p-th pattern to the i-th code word is defined as the hamming distance.

In this paper, in order to the balance number of training sets, we construct a code matrix as table 2.

It’s easy to notice that we don’t use an exhaustive code matrix, because through experiments we found it is helpless to promote the accuracy of prediction and is time exhaustive, though it can increase the hamming distance of code matrix.

III. RESULT ANALYSIS

Table 3 show the predict accuracy obtained by different algorithms on the 1189 data set. Remark that result of our method is obtained through 5 fold cross validation.

From the table 3 we can see that the over all accuracy of our method is higher than some work done by previous investigators. But notice that prediction accuracy of class is low, this mean that our method is less sensitive for this class.

IV. CONCLUSION

This paper we propose a novel multi-feature of protein sequence, adopt ECOC as the frame of the prediction engine, And the base classifier that of ECOC is ANN. Experiments show that this method may be a powerful method for protein structural prediction and is a complementarity of conventional methods. But just as mentioned above, our method is less sensitive for $\alpha + \beta$ class, so there still much space to promote the prediction accuracy. Next step we plan to combine features which can describe the $\alpha + \beta$ class well, this way we can get more information about protein sequence and promote the overall accuracy.

V. ACKNOWLEDGMENT

The authors thank to colleagues in Computational Intelligence Lab for their valuable advice and professor Jianyi Yang for providing the dataset. This paper was supported by the Natural Science Foundation of China (61070130, 60903176, 60873089), the Program for New Century Excellent Talents in university (NCET-10-0863), the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation of China (BS2009SW003).

REFERENCES


TABLE I. 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>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P</th>
<th>V</th>
<th>E</th>
<th>K</th>
<th>Q</th>
<th>W</th>
<th>F</th>
<th>L</th>
<th>R</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Figure 1. RP of a time series

TABLE II. CODE MATRIX OF ECOC

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Structure classes</th>
<th>ANN 1</th>
<th>ANN 2</th>
<th>ANN 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha + \beta$</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha/\beta$</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE III. RESULTS OF DIFFERENT METHODS

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\alpha + \beta$</th>
<th>$\alpha/\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
<td>F1 Score</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>Bayes classifier (Wang and Yuan, 2000)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Logistic regression (Kurgan and Homaeian, 2006)</td>
<td>57.0</td>
<td>62.9</td>
<td>25.3</td>
<td>64.6</td>
</tr>
<tr>
<td>StackingC ensemble (Kedarisetty et al., 2006)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Pseudo-amino acid composition (Zhang et al, 2008)</td>
<td>48.9</td>
<td>59.5</td>
<td>26.6</td>
<td>81.7</td>
</tr>
<tr>
<td>This method</td>
<td>62.4</td>
<td>55.3</td>
<td>15.3</td>
<td>82.1</td>
</tr>
</tbody>
</table>