Improved protein structural class prediction based on chaos game representation

Mohammad Olyaee
Department of computer engineering, Islamic azad university, Mashhad Branch, Iran
mh.olyaee@gmail.com

Mehdi yaghubi
Department of computer engineering, Islamic azad university, Mashhad Branch, Iran
yaghobi@mshdiau.ac.ir

Abstract—Determination of protein structural class from sequence information is a challenging task. In this paper, at first we apply chaos game representation to protein sequences and extract two time series then using phase space reconstruction theory and calculate phase space of all time series. Next, applying recurrence quantification analysis (RQA). For each protein sequence 16 characteristic parameters can be calculated with RQA. In order to classification we propose an ensemble classification method. The 10 fold cross validation test is used to test and compare our method with other existing methods. The overall accuracy for two datasets 1189 and 25PDB are 66.7% and 68.2% respectively that has much better performance toward compared methods

Keywords- Protein structural class, chaos game representation; phase space; recurrence plot; classifier ensemble

I. INTRODUCTION

Levitt and Chothia [1] defined the concept of protein structural classes according to this definition, proteins classified into four groups: (1) All a class, which includes proteins with only small strands, (2) All b which are formed by strands and with only small amount of helices, (3) a/b proteins which includes both helices and strands that strands are mostly parallel (4) a/b class which includes both helices and mostly anti parallel strands. The structural class is on of the important attributes of protein for example it can increase the accuracy of secondary structure prediction [2] or can help to reduce the searching scope of conformation in tertiary structure prediction[3, 4]. So it is useful to know the protein structural . There are many efforts to solve this problem [5, 6]. Most of them used Amino Acid composition (AA) however many important features about sequence order are missed that reduce the success rate of prediction. In view of this, various methods were presented that including the pair-coupled amino acid composition[3], polypeptide composition [7], Pseudo-amino acid composition[8], functional domain composition[9].Recently Yang et al [10] and Yu et al [14] proposed several other kinds of CGR of proteins. Yang et al [10] transform protein sequences into nucleotide sequences based on reserve encoding of amino acids [15] and use CGR for DNA sequence based on [16]. Since analysis of CGR is difficult, two time series are extracted from CGR. Secondly a new powerful nonlinear method Recurrence quantification analysis (RQA) is applied to analyze these time series. For each time series eight parameters are achieved then uses 16 (8x2) parameters to predict the structural classes. Yang et al gained to accuracies 65.8% and 64.2% for low homology 1189 (1092 domains) and 25PDB (1673 domains) datasets respectively. But before analyze time series with RQA, sets embedding dimension (m) 8 and delay time (τ) 2 for all time series. In this paper in order to improve this method we use GP algorithm and Auto correlation function and calculate phase space for all time series individually and applying an ensemble of classification algorithms. The 10 fold cross validation test that is much reliable and not time consuming [6] is used for evaluate and compare our method with other existing methods. Experimental results show that our method is much better than other and may play a complementary role to the existing methods.

II. MODELS AND METHODS

A. Amino Acid sequence to DNA sequence

There are several methods for transform protein sequences to nucleotide sequences. In order to do this we use encoding method used by [15] that is listed in Table1.

B. Chaos game representation

CGR of nucleotide sequences is defined in a square [0x1][0x1] where four vertices correspond to four letters A,T,C and G. the first point is placed halfway between the center of square and the vertex corresponding to the first letter of sequences; the ith point is then placed halfway between (i-1)th point and the vertex corresponding to ith
letter. The obtained plot is called CGR. The CGR of one sample is shown in Fig1. Since the analyze of the CGR plot isn’t easy; noticing that each point in CGR is determined by (x,y) coordinates; every CGR is decomposed into two time series. For example in Fig2 two time series are extracted from Fig1.

C. Phase space reconstruction

Packard and Takens firstly introduced phase space reconstruction theory to study the chaotic property of a system. Supposed that a time series is

\[ x(i) = x(t_i), \quad t_i = t_0 + i \Delta t, \quad i = 1, 2, \ldots, N \]

that \( \Delta t \) is a constant. After delay time is selected which be integral times of \( \Delta t \), k vectors with m-dimension can be constructed:

\[
X(1) = [x(1), x(1+\tau), \ldots, x(1+(m-1)\tau)] \\
X(2) = [x(2), x(2+\tau), \ldots, x(2+(m-1)\tau)] \\
\vdots \\
X(k) = [x(k), x(k+\tau), \ldots, x(k+(m-1)\tau)]
\]

(1)

Where \( N=k+(m-1) \tau \) is the length of time series, \( m \) is called embedding dimension, \( \tau \) is delay time. All \( k \) vectors form a reconstructed phase space. This kind of reconstruction has the same dynamical property of original.

D. Selection Delay time

The value of delay time \( \tau \) is very important. If \( \tau \) is too small, then all the coordinates are nearly constant on the other hand if \( \tau \) is too big, trajectory in phase space may be disconnected. Autocorrelation coefficient method determines delay time according to computing autocorrelation coefficient of the time series. That is defined as:

\[
C(\tau) = \frac{\sum_{i=1}^{N-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}
\]

(2)

[17] presents that a reasonably empirical value is \( \frac{1}{e} \), approximately 0.4. This paper determines delay time of CGRX of protein 1AK0 and the result is shown in Fig3. When \( \tau \) is equal to 2 autocorrelation coefficient attenuates to 0.4 nearly, so \( \tau = 2 \) is selected.

E. Selection Embedding dimension

Supposed that there are two points in the reconstructed phase space:

\[
X(i) = [x(i), x(i+\tau), \ldots, x(i+(m-1)\tau)] \\
X(j) = [x(j), x(j+\tau), \ldots, x(j+(m-1)\tau)]
\]

Let \( r_j(m) \) denote the distance between them, then we have

\[
r_j(m) = \|X_i - X_j\| \]

(3)

Given a certain distance \( r \), calculate the ratio of the number of the pairs of phase point, the distance between which is less than \( r \) to all the pairs of phase point. The ratio is denoted as:

\[
C_2(r, m) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \theta(r - \|X_i - X_j\|)
\]

(4)

Where \( \theta(z) \) is a Heavisible function:

\[
\theta(z) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 
\end{cases}
\]

(5)

\( C_2(r, m) \) is related to the \( r \) by the following relation:

\[
C_2(r, m) \propto r^{d_2}
\]

(6)
Where $D_2$ is the correlation exponent, given by equation 7.

$$D_2(m) = \frac{\ln C_2(r, m)}{\ln r} \quad (7)$$

Increase the embedding dimension until the estimated value of correlation dimension $D_2$ does not increase, the embedding dimension at the moment is called saturated embedding dimension. This method is proposed by Grassberger and Procaccia and called GP algorithm. \cite{17} We use this algorithm for protein 1AK0 CGRX and the result is shown in Fig 4.

The Fig 4 shows the curves of $\ln C_2(r, m)$ and $\ln r$. When $m$ increases from 5 to 10. From this figure when $m$ is 9 the slopes of the linear parts of the curve don’t change any more. So $m$ is 9.

F. Recurrence plot

Recurrence plot is a graphical tool that helps to detect patterns of recurrence in data and proposed by \cite{18} after $m$ and $r$. We can calculate distance matrix (DM) which is $N_m \times N_m$ and $N_m$ is the number of points in embedding space $R^m$. The elements of DM are distances between all possible combination of i-points and j-points. For simplicity DM can be rescaled by dividing DM by maximum distance of DM. Then we can transformed it to recurrence matrix (RM) which RM = $(R_{i,j}(\varepsilon))_{N_m \times N_m}$ and $R_{i,j}(\varepsilon) = \theta(\varepsilon - D_{i,j})$. RP is a visualization of RM such that if $R_{i,j}(\varepsilon) = 1$ we say $j$ point recure with reference to $i$ point. The RPs for two time series shown in Fig 5 and Fig 6.

$\varepsilon$ is a crucial parameter of RP. If $\varepsilon$ is too small there may be almost no recurrence points on the other hand if $\varepsilon$ is too large almost every point is neighbor of every other point \cite{19}.

G. Recurrence quantification analysis

RQA is a relatively new nonlinear technique based on RP. This method has been successfully applied to many different fields \cite{20, 21} and quantify the information supplied by RP. RQA provides us eight parameters from Rp those are:

$$%REC = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}(\varepsilon) \quad (8)$$

$$DET = \frac{\sum_{i=i_{min}}^{i} P(l)}{\sum_{l=1}^{N} P(l)} \quad (9)$$

$$Trend = \frac{\sum_{l=1}^{N} \left( \tau - \bar{N}/2 \right) (RR_t - <RR_t>)}{\sum_{l=1}^{N} (\tau - \bar{N}/2)^2} \quad (10)$$
\[ l_{\text{max}} = \text{Max}(\{l_i\}_{i=1}^{N_1}) \] (11)

\[ \text{ENTR} = - \sum_{l=l_{\text{min}}}^{N} p(l) \ln p(l) \] (12)

\[ TT = \sum_{v=v_{\text{min}}}^{N} p(v) \] (13)

\[ \text{LAM} = \frac{\sum_{v=v_{\text{min}}}^{N} p(v)}{\sum_{v=1}^{N} p(v)} \] (14)

\[ V_{\text{max}} = \text{Max}(\{v_i\}_{i=1}^{N}) \] (15)

For more information we recommend see [19].

III. DATA AND RESULTS

The two datasets analyzed here are used previously by Kurgan and Homaeian [6] and Yang et al [10] those are 25PDB and 1189 that have 25% and 40% homology respectively. The sequences of them can be downloaded from RCSB Data Bank (http://www.rcsb.org/pdb/home/home.do). For each protein sequence as mentioned above, two time series are extracted then calculate phase space for each time series individually next use RQA and extract 16 (8x2) features. In order to classification we use ensemble method such that features vectors are fed to four classifiers and the results then are combined by a voting module. The diagram of classifier is shown in Fig 7.

For examine the power of classification we use 10 fold cross validation test because it is not time consuming and is reliable [6]. The results of each classifier, are listed in Table 3.

![Diagram of classifier](image)

As seen in Table 3, the Logistic regression method has good results against other methods. All of methods have suitable accuracy in \( \alpha \) class while poor results in accuracy...
of $\alpha + \beta$ class. The corresponding accuracies of ensemble method are listed in Table 4.

**TABLE IV. Accuracy of our method compared with other methods**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reference</th>
<th>Accuracy (%)</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\alpha + \beta$</th>
<th>$\alpha/\beta$</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1189</td>
<td>This paper</td>
<td>99 62 31 75 66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[10]</td>
<td>62.3 67.7 63.1 66.5 65.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[5]</td>
<td>NA NA NA NA 58.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[6]</td>
<td>57 62.9 25.3 64.6 53.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25PDB</td>
<td>This paper</td>
<td>99 62 52 61 68.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[10]</td>
<td>64.3 65 61.7 65 64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[5]</td>
<td>NA NA NA NA 59.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[6]</td>
<td>69.1 61.6 60.1 38.3 57.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to compare with other methods the results of them are listed in Table 4 too. From this table, we can see that The accuracy in predicting $\alpha$ class is much more better than other existing methods but The accuracy in predicting $\alpha + \beta$ class is not good against others. However, the overall accuracy of our method is higher than other methods listed in Table 3. So the current method may play a complementary role.

**IV. CONCLUSION**

Prediction of protein structural class is a very important and challenging problem. Chaos game representation (CGR) of proteins provides a visualization of protein sequences and the protein sequence can be reconstructed uniquely from the CGR. Yang et al [10] proposed new method that decompose two time series from CGR of protein sequence. Toward improvement, we calculate phase space of each time series individually such that using GP algorithm for calculate embedding dimension and Auto correlation function to gain delay time and Next extract features based on Recurrence quantification analysis (RQA). RQA is a nonlinear method that can analyze time series without the requirement on the length of time series. The features get into an ensemble of four complementary classifiers to improve the accuracy of prediction.

The overall accuracy with such method are 66.7% and 68.2% for 1189 and 25PDB dataset respectively and compared with other method demonstrate that our results is better than methods compared here.

**REFERENCES**


