Predicting protein structural classes based on complex networks and recurrence analysis

Mohammad H. Olyaee \textsuperscript{a,\textast}, Ali Yaghoubi \textsuperscript{b}, Mahdi Yaghoobi \textsuperscript{c}

\textsuperscript{a} Department of Electrical and Computer Engineering, University of Zanjan, Zanjan, Iran
\textsuperscript{b} Engineering Department, Pars Razavi Institute of Higher Education, Gonabad, Iran
\textsuperscript{c} Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

HIGHLIGHTS

- Features are extracted based on CGR of predicted secondary structure contents.
- Two time series are extracted from the CGR which contain the coordinates of points.
- Recurrence matrices and RPs are acquired from time series.
- Complex network is used to extract some new features based on recurrence matrices.

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form 25 May 2016
Accepted 15 June 2016
Available online 18 June 2016

Keywords:
Prediction protein structural class
Chaos game representation
Recurrence matrix
Complex network

ABSTRACT

Protein sequences are divided into four structural classes. The determination of class is a challenging and beneficial task in the bioinformatics field. Several methods have been proposed to this end, but most utilize too many features and produce unsuitable results. In the present, features are extracted based on the predicted secondary structures. At first, predicted secondary structure sequences are mapped into two time series by the chaos game representation. Then, a recurrence matrix is calculated from each of the time series. The recurrence matrix is identified with the adjacency matrix of a complex network and measures are applied for the characterization of complex networks to these recurrence matrices. For a given protein sequence, a total of 24 characteristic features can be calculated and these are fed into Fisher’s discriminated analysis algorithm for classification. To examine the proposed method, two widely used low similarity benchmark datasets design and test its performance. A comparison with the results of existing methods shows that the current study’s approach provides a satisfactory performance for protein structural class prediction.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The prediction of the protein third structure is one of the most important and complicated issues in bioinformatics. Even so, their overall folding patterns are simple and regular (Chou, 2005; Zhang et al., 2008). Based on these patterns, proteins have been classified into (1) all-\(\alpha\) class which includes proteins with only small strands, (2) all-\(\beta\) class which is formed by strands and with only a small amount of helices, (3) \(\alpha/\beta\) proteins which includes both helices and mostly parallel strands, and (4) \(\alpha+\beta\)-class which includes both helices and mostly anti-parallel strands (Levitt and Chothia, 1976). The knowledge of the structural class is helpful to increase the accuracy of secondary structure prediction (Gromiha and Selvaraj, 1998) or to reduce the searching scope of conformation in tertiary structure prediction (Bahar et al., 1997; Chou and Zhang, 1995). There have been many efforts made to solve this problem (Bahar et al., 1997; Kedarisetti et al., 2006; Kurgan and Homaeian, 2006; Kong et al., 2014; Kong and Zhang, 2014; Liu et al., 2012; Zhang et al., 2011; Yang et al., 2009; Sahu and Panda, 2010; Zhou and Assa-Munt, 2001; Chou, 1995). Most of the early prediction methods have only focused on Amino Acid composition (AA) (Bahar et al., 1997; Zhou and Assa-Munt, 2001; Chou, 1995; Chou and Zhang, 1994; Zhou, 1998; Chou and Maggiora, 1998). The main development among these approaches was done by including the pair-coupled effect among different AA components. However, one of the major drawbacks to adopting these methods is that many important features related to sequence order have been missed which has reduced the success rate of prediction (Kurgan and Homaeian, 2006). In view of this, various methods have been presented, including polypeptide...
composition (Jin et al., 2003), pseudo-amino acid (PseAA) composition (Chou, 2001, 2011), functional domain composition (Chou and Cai, 2004) and recurrence quantification analysis (Yang et al., 2009).

It can be said that all these representations are different types of pseudo-amino acid composition (Liao et al., 2012). This concept was originally proposed by Chou (2001) and has been widely studied as a powerful feature representation technique (Sahu and Panda, 2010; Liao et al., 2012; Chen et al., 2012; Qin et al., 2012). Although these approaches are interesting and their results are encouraging, they suffer from low efficiency when encounter with low homology datasets (e.g., the 25PDB and 1189 datasets with sequence similarities lower than 25% and 40%, respectively).

In recent years, position-specific scoring matrix (PSSM) and predicted protein secondary structure content have been receiving much attention due to improve the prediction accuracy for low-homology proteins (Liu et al., 2012; Ding et al., 2012; Yang et al., 2010; Kurgan and Chen, 2007; Ding et al., 2014; Liu et al., 2010). Kong et al. (2014) developed 13 features which characterized general contents and spatial arrangements of the predicted secondary structure elements of a given protein sequence. In Althaus et al. (1993a) the authors proposed a sequence representation method based on position specific scoring matrix (PSSM). The PSSM is converted into 20 x 20 matrix by defined evolutionary difference equation and finally a 420 dimensional feature vector is extracted according to the gained matrix. In PSSS-PsePSSM (Zhang, 2015), the input of the classifier includes 111 features. Among them, 100 features are selected based on pseudo-position specific scoring matrix (PsePSSM) which contains evolutionary information and sequence order information. The other features are extracted according to predicted protein secondary structure elements.

The chaos game representation (CGR) initially was proposed by Jeffrey (1990) to visualize genomic sequences. CGR is a graphical tool which iteratively maps a genomic sequence into a 2-dimensional space. The map preserves all information of the input sequence and also provides an intuitive picture which helps to reveal hidden patterns and local structures more efficiently (Tanchotsrinon et al., 2015). Several studies, for example (Yang et al., 2009; Tanchotsrinon et al., 2015; Nair et al., 2010; Yang et al., 2008) have been showed that this graphical tool provides researchers with a number of novel features. These features are extracted based on the CGR of given sequences. Nair et al. (2010) developed a new method for classification of organisms. First, a variation of CGR named Frequency-CGR is used and 64 features are extracted directly from CGR image of DNA sequences. Next, the gained features are used by artificial neural network as a powerful machine learning approach for classification. In (Tanchotsrinon et al., 2015) the authors addressed the problem of predicting the Human Papillomavirus (HPV) genotypes from their genomes. For this purpose, two new feature extraction algorithms have been proposed based on CGR. These algorithms concentrate on the local information among nucleotides. Next, Singular value decomposition (SVD) as a matrix factorization technique, is deployed to decompose feature matrix and reduce information complexity. Yang et al. (2008) used CGR and multi-fractal analysis to predict protein structural classes. The main limitation of this method is that it is not appropriate for short amino acid sequences. To address this problem, in Yang et al. (2009) the same authors adopted recurrence quantification analysis (RQA) as a powerful nonlinear technique in analyzing time series without the requirement on the length of time series.

The current study, presents a new comprehensive feature set that was created from the predicted secondary structure as well. The experiments conducted on two benchmark datasets indicate that the prediction of protein structural classes can be supplementary enhanced for low-homology amino acid sequences.

In the first step, a protein structural class prediction is carried out based on the predicted secondary structure content. Next, the chaos game representation is used to represent a predicted secondary structure as a two time series, from which a comprehensive set of 24 features is generated using the complex network theory. Briefly, it can be said that the feature vector is indirectly constructed from the CGR image. Recently, the complex network theory has been introduced as an effective method for discovering the complexity in systems (Boccaletti et al., 2006). There have been some attempts to employ the concept of complex networks for studying the time series. The complex network theory is helpful in understanding complex interrelations and the information flow between secondary structure elements (Marwan et al., 2009). Using these new features can increase the possibility of representing all order information of the original data. Finally, to make a prediction of its protein structural class the resulting 24-dimensional feature vector is fed into the simple yet powerful Fisher’s discriminant algorithm. Jackknife tests on two low-similarity datasets show that our method is remarkable and may play a complementary role to existing approaches.

2. Materials and methods

As demonstrated by a series of recent publications (Jia et al., 2015, 2016a, 2016b; Chen et al., 2015a, 2016; Liu et al., 2015a, 2015b, 2015c) in compliance with Chou’s 5-step rule (Chou, 2011), to establish a really useful sequence-based statistical predictor for a biological system, we should follow the following five guidelines: (a) Select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce a powerful algorithm to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

2.1. Materials

The proposed method is tested on two benchmark datasets in low homology, namely 25PDB and 1189 dataset. The first is comprised of 1673 proteins with about 25% sequence identity, of which 443 are all-α class proteins, 443 are all-β class proteins, 346 are α/β class proteins and 441 are α+β class proteins (Kurgan and Hourmaian, 2006); The 1189 dataset is made up of 1092 proteins with about 40% sequence similarity consisting of 223 all-α class proteins, 294 all-β class proteins, 334 α/β class proteins, and 241 α+β class proteins (Wang and Yuan, 2006).

2.2. Secondary structure prediction

Every amino acid in a protein sequence can be predicted as one of the three secondary structural elements: H (helix), E (strand), and C (coil). There are many computational methods that predict the 3-state secondary structure from protein sequences. Among all of them, the present work prefers PSIPRED (Jones, 1999) to predict protein secondary structure because it has been shown to out-perform other competing prediction methods (Birznieks and Kramer, 2006) which employ the position specific scoring matrices generated by PSI-BLAST (Altschul et al., 1997).
2.3. Chaos game representation

Chaos game representation (CGR) is a popular technique for converting a one-dimensional sequence into a two-dimensional form that preserves subsequence structures and provides a visual representation (Jeffrey, 1992). This technique iteratively maps each residue in a sequence to a unique coordinates in a two-dimensional space. The uniqueness and preserving sequence order information are two main properties of CGR. Since CGR provides visual depiction of biological sequences and makes useful intuitive insights, it has been applied in many important biological topics (Niu et al., 2012).

Recently, this technique was applied to predict secondary structure contents (Yang et al., 2010). The concept of CGR for the secondary structure content is briefly reviewed here. An equilateral triangle is supposed with unit length and each related vertex is identified by a distinct letter, namely H, E and C. For each letter of the given secondary structure sequence, a point is plotted inside the triangle as follows. The first point is placed halfway between the center of the triangle and the vertex related to the first letter of the secondary structure sequence; the i-th point is then placed halfway between the (i-1)th point and the vertex matching the i-th letter.

The resulting plot is then called the CGR of the secondary structure sequence. The CGR of a typical secondary structure sequence is shown in Fig. 1.

2.4. Recurrence plot

Valuable intuition about the patterns of time-series can be achieved by the means of Recurrence Plot (RP) (Eckmann et al., 1987). Suppose given a time series \( x_1, x_2, \ldots, x_n \) of length \( N \). First, it should be embedded into the phase space \( \mathbb{R}^m \) with an embedding dimension \( m \) and a time delay \( \tau \). A point of time series in the phase space is as follows:

\[
X_i = (x_i, x_{i+\tau}, \ldots, x_{i+(m-1)\tau})
\]

Where \( i = 1, 2, \ldots, N_m \) and \( N_m = N - (m-1) \). Therefore, \( N_m \) vectors (i.e. points) in the embedding space \( \mathbb{R}^m \) are obtained.

Next, the distances between all pairs of embedded vectors are computed and are scaled down by the maximum distance. The gained measures generate a symmetric two-dimensional square matrix, called the Distance Matrix (DM).

According to Eq. (2), by applying threshold \( \varepsilon \) on the element values of DM, the Recurrence Matrix (RM) can be defined.

\[
RM_{ij}(\varepsilon) = \begin{cases} 1, & \text{if } |x_i - x_j| < \varepsilon \\ 0, & \text{otherwise} \end{cases}
\]

In the above equation, \( \theta(.) \) is the Heaviside function \( \theta(x) = 0 \), if \( x < 0 \) and \( \theta(x) = 1 \) otherwise and \( \|x\| \) is a norm.

For example, the RPs of the two x-time series (and y-time series) of Fig. 2 are presented in Fig. 3.

In the creation of an RP, the determination of \( \varepsilon \) is a fundamental task. If \( \varepsilon \) is too small, there may not be any recurrence points and so nothing can be learned about the recurrence structure of the underlying system. On the other hand, if \( \varepsilon \) is too large, almost every point is a neighbor of every other point, thus leading to numerous artifacts (Marwan et al., 2007).

2.5. Complex network theory

In recent years, Graphs or Complex network theory concepts have been applied to analyze the complicated systems of a wide range of problems, such as computer or neural networks, power grids, transportation networks, global climate systems, and so on (Boccaletti et al., 2006; Marwan et al., 2009; Donner et al., 2011; Tang et al., 2014; Antiqueira et al., 2009; Orouskhani et al., 2016). Moreover, introducing complex networks to biological systems can provide an intuitive vision and useful insights. Actually, various graph approaches have been successfully used to analyze many important biological problems, such as enzyme-catalyzed reactions (Chou et al., 1979; Chou and Forsén, 1980; Zhou and Deng, 1984; Chou, 1989), protein folding kinetics and folding rates (Chou, 1990; Chou and Shen, 2009), inhibition of HIV-1 reverse transcriptase (Althaus et al., 1993a, 1993b), non-steady drug metabolism systems (Chou, 2010), evolution of biological sequences (Wu et al., 2010), and using Wenxiang graphs (Chou et al., 2011) to analyze protein–protein interactions (Zhou, 2011; Zhou and Huang, 2013). This is mainly because the local and global properties of complex networks are useful for recognizing complicated interconnections and the information flow between different components in extended systems. The adjacency matrix, which describes the relations between the nodes of a network, is the basis of complex network analysis. Similar to the recurrence matrix, the adjacency matrix is also square, binary, and symmetric.

Marwan et al. (2009) suggested that the recurrence matrix and the adjacency matrix demonstrate a strong analogy in that a recurrence matrix represents neighbors in a phase space and an adjacency matrix represents links in a graph; both matrices include a pair-wise test of all components. Consequently, the recurrence matrix can be considered as the adjacency matrix of an undirected, unweighted network, thus allowing for the study of the extracted time series using a complex network approach. The present study employs these complex network measures (local and global properties) for recurrence matrices obtained from two time series. By applying these measures, further information is gathered from the recurrence plots.

3. Prediction protein structural class

A predicted secondary structure for each protein sequence is attained. Next, all sequences are mapped by the chaos game
representation in a 2D space. Since direct analysis of these plots is difficult, two time series (CGRX and CGRY) are extracted based on coordinates X and Y, the coordinates of each point is preserved. The next step is the determination of phase space, including the embedding dimension and time delay of each time series. Finally, as mentioned previously, recurrence matrices and RPs are acquired from CGRX and CGRY. From each time series, the total number of features to be extracted is 12 of which 8 are based on recurrence quantification analysis (RQA) and the remainder on complex network analysis. In view of the fact that RQA-based features have been introduced and used by (Yang et al., 2009; Yang et al., 2010; Marwan et al., 2007) in detail, they are referred for more study. The complex network-based features are presented below:

\[ K_{v_{\text{max}}} = \sum_{i=1}^{N} A_{v_{\text{max}},i} \]  

(3)

Degree centrality that gives the number of \( v_{\text{max}} \) neighbors

\[ K_{v_{\text{min}}} = \sum_{i=1}^{N} A_{v_{\text{min}},i} \]  

(4)

Degree centrality that gives the number of \( v_{\text{min}} \) neighbors

Average of the local clustering coefficient that provides the probability of neighboring two states:

\[ C_{ij} = \frac{\sum_{k=1}^{N} A_{ij,k}A_{ij,k}A_{ij}}{K_i(K_i-1)} \]  

(5)

The average length of the shortest paths between all pairs of nodes is given by:

\[ L = \frac{1}{N(N-1)} \sum_{ij=1}^{N} d_{ij} \]  

(6)

From each time series, 12 features were extracted; therefore, the total set has 12×2 features.

There are numerous presented prediction algorithms, of which most have recently been based on support vector machines (SVM).
In the current paper, the Fisher’s heterogeneous classifier method is used by using boundary and non-boundary patterns. The purpose of Fisher’s linear discriminant algorithm (FDA) is to perform the classification procedure in the mapping space. Suppose there are two classes, \( a_1 \) and \( a_2 \); one class as positive and the other is negative, the classes are represented by two normally distributed random vectors \( x_1 \sim N(m_1, S_1) \) and \( x_2 \sim N(m_2, S_2) \), respectively, with \( P_1 \) and \( P_2 \) as the priori probabilities.

After this, the converters of FDA are attained by two new random vectors \( y_1=Wx_1 \) and \( y_2=Wx_2 \) where \( x_1 \sim N(Wm_1,W_1W) \) and \( x_2 \sim N(Wm_2,W_2W) \) correspondingly represent a new class of mapping space. \( m_1 \) and \( S_1 \) are the mean vectors and covariance matrices in the original space, respectively the aim of FDA is to find a linear transformation matrix of \( W \) so that it may achieve maximum separation between the classes. The classification is performed by the Fisher method of classification procedure. Therefore by finding a \( d \times n \) matrix \( (W) \), the FDA criterion maximizes as follows:

\[
J_{FDA}=\text{tr}\left((WSW^T)^{-1}(WSW^T)\right)\]

(7)

\[S_W=P_1S_1+P_2S_2\]

(8)

\[S_I=(m_1-m_2)(m_1-m_2)^T\]

(9)

\[W=(m_1-m_2)S_W^{-1}\]

(10)

The classification procedure is done in such a way that \( x \) data belongs to \( a_1 \), if \( W^Tx > \frac{1}{2}(m_1+m_2) \) and otherwise it belongs to \( a_2 \). \( S_W \) and \( S_I \) are within-classes and between-classes scattering matrices, respectively.

FDA focuses on separating the means as best as possible. However, it ignores the difference between the covariance matrices of the classes. The effect of this property is specified when the classes have the same mean. In this case, the difference between the means for separating classes can be used because it causes a large overlap of classes in the project space. At the same time, in order to deal with this problem, class covariance is differentiated between classes covariance; in other words, two classes are completely separated. This project is proposed for handling many designs. In addition to this project is the use of distance distribution, in which discriminate information existing between classes covariance matrices is used. The heterogeneous Discriminant Analysis (HDA) algorithm (Duin and Loog, 2004) is employed. The HDA algorithm is proposed based on the distribution of Chernoff’s distance. An extension of the FDA, HDA aims to find the matrix \( W \) that maximizes the following formula:

\[
J_{HDA}=\text{tr}\left((WSW^T)^{-1}(WSW^T)\right)\]

\[\sum_{i=1}^{n} p_i \log \left( \frac{S^{-1}_i S^{-1}_i S_i^{-1}}{S_i^{-1} S^{-1}_i S_i^{-1}} \right)\]

\[P_1P_2\]

\[
S_{HDA}=S_W^{-1}
\]

(12)

After obtaining eigenvectors and eigenvalues, vector \( W \) is the eigenvector corresponding to the largest eigenvalue; therefore, the classification procedure is performed by FDA.

To increase the classification, \( S_W \) and \( S_I \) are designed in the standard scattering matrix HAD, based on boundary and non-boundary patterns. On the other hand, the use of the scattering matrix reduces the effect of class pairs that are far apart in the original space.

If \( X \) is to be considered as a sample set of boundary and non-boundary patterns by relevant pattern selection (RPS), which is the main idea of this technique, then the choice of Border Pattern is based on the criterion of proximity. The present study employs the technique expressed in (Na et al., 2010): \( X \) into two subsets \( \{X_1, X_2, \ldots, X_{nB}\} \) is a set of boundary patterns and \( \{X_{NB}\}\) is a set of non-boundary patterns. Therefore, \( n_{B} \) and \( n_{NB} \) are the number of boundary patterns and the number of non-boundary patterns, respectively. A new scattering matrix based on these two sets is defined in the following formula (Shin and Cho, 2007):

\[
S^{(B)}_{W(i)}=\sum_{i=1}^{n_{B}} \sum_{j=1}^{n_{B}} \left(x_{j}^{(B)} - m(i)\right)\left(x_{j}^{(B)} - m(i)\right)^T
\]

(13)

\[
S^{(W)}_{W(i)}=\sum_{i=1}^{n_{B}} \sum_{j=1}^{n_{B}} \left(x_{j}^{(NB)} - m(i)\right)\left(x_{j}^{(NB)} - m(i)\right)^T
\]

(14)

New scattering matrices \( S^{(B)}_{W(i)} \) and \( S^{(W)}_{W(i)} \) replace matrices \( S_W \) and \( S_I \) in the HDA algorithm. The classification procedure is performed by the same procedure of FDA. It must be taken into account that, by using scattering matrices, more features are obtained in comparison to the old criterion’s scattering matrices of HAD (Shin and Cho, 2007).

4. Results

4.1. Prediction assessment

Amongst the prediction assessment methods, the jackknife test has become increasingly prevalent and is generally used to examine the performance of various predictors. For this reason, the jackknife test was chosen to examine the power of the proposed method. In jackknife test, each protein sequence in the dataset is selected as a test sample, and the predictor is trained by the remaining protein sequences. The overall accuracy (OA) is computed for each dataset. Also, the following three standard performance measures have recently been widely used to estimate prediction accuracy: Sensitivity (Sens), Specificity (Spec) and Matthew’s Correlation Coefficient (MCC) (Liu et al., 2012; Ding et al., 2012, 2014; Zhang et al., 2013; Ding et al., 2013). The MCC value ranges between \(-1\) and \(1\), where \(0\) represents random correlation and larger positive (negative) values indicate a better (lower) prediction quality for a given class.

Clearly, these are defined by the following formula:

\[
OA=\frac{1}{N} \sum_{i=1}^{M} TP(i)
\]

(15)
that where \( \alpha \) is the constructed dataset and \( \beta \) is the number of samples that were incorrectly predicted \((\alpha \beta)\). \( \alpha \beta \) is the total number of the investigated samples in the subset which contains all samples of class \( i \). Let us represent TP, TN, FP and FN by the following equations:

\[
\text{TP}(i) = N^+ (i) - N^- (i)
\]

\[
\text{TN}(i) = N^- (i) - N_C^- (i)
\]

\[
\text{FP}(i) = N_C^+ (i)
\]

\[
\text{FN}(i) = N^- (i)
\]

Where \( N^+ (i) \) is the total number of the investigated samples in the subset \( S_i \), whereas \( N^- (i) \) is the number of samples in \( S_i \) that where incorrectly predicted belonging to the other subsets, and \( N^- (i) \) is the total number of samples in all of the other subsets, whereas \( N^- (i) \) is the number of samples that were incorrectly predicted belonging to \( S_i \). By substituting Eqs. (19)–(22) into Eq. (15)–(18), we obtain

\[
\text{OA} = \frac{1}{N} \sum_{i=1}^{M} [N^+ (i) - N^- (i)]
\]

\[
\text{Sens} (i) = 1 - \frac{N_C^+ (i)}{N^+ (i)}
\]

\[
\text{Spec} (i) = 1 - \frac{N_C^- (i)}{N^- (i)}
\]

\[
\text{MCC} (i) = \frac{\text{TP}(i) \times \text{TN}(i) - \text{FP}(i) \times \text{FN}(i)}{\sqrt{[\text{TP}(i) + \text{FP}(i)][\text{TP}(i) + \text{FN}(i)][\text{TN}(i) + \text{FP}(i)][\text{TN}(i) + \text{FN}(i)]}}
\]

Where TP is the true positive, TN is the true negative, FP is the false positive, FN is the false negative, \( M \) is the number of classes, and \( N \) is the number of protein sequences.

Since the above metrics are not easy to understand for most readers, particularly the MCC (the Matthews correlation coefficient), many recent studies have tried to make them in a more intuitive and easier-to-understand formulation (Liu et al., 2015a, 2015c, 2015d, 2015e; Xiao et al., 2015a; Ding et al., 2013, 2014; Chen et al., 2013, 2014a, 2014b, 2015b; Feng et al., 2013). Suppose \( S \) is the constructed dataset and \( S_i (i = 1, 2, 3, 4) \) is a subset which contains all samples of class \( i \). By substituting Eqs. (19)–(22) into Eq. (18), we obtain

\[
\text{OA} = \frac{1}{N} \sum_{i=1}^{M} [N^+ (i) - N^- (i)]
\]

\[
\text{Sens} (i) = 1 - \frac{N_C^+ (i)}{N^+ (i)}
\]

\[
\text{Spec} (i) = 1 - \frac{N_C^- (i)}{N^- (i)}
\]

\[
\text{MCC} (i) = \frac{1 - \left( \frac{N_C^+ (i)}{N^+ (i)} \right) \left( \frac{N_C^- (i)}{N^- (i)} \right)}{\sqrt{1 + \left( \frac{N_C^+ (i)}{N^+ (i)} \right) \left( \frac{N_C^- (i)}{N^- (i)} \right)}}
\]

It should be noted that the set of metrics as defined in Eqs. (15)–(18) or Eqs. (23)–(26) is valid only for the single-label systems. For the multi-label systems, whose emergence has become more frequent in system biology (Chou et al., 2012) and system medicine (Xiao et al., 2013), a completely different set of metrics should be used as given in (Chou, 2013).

All attempts are performed on two datasets using jackknife tests and, in addition to the overall accuracy (OA), the Sens, Spec, and MCC for each structural class are reported. The results are shown in Table 1. As seen, the overall accuracy for the two datasets is above 85%.

### 4.2. Comparison with existing methods

In this section, the performance of the proposed method is compared against other recently introduced methods, as shown in Table 2. The compared methods include (Kong et al., 2014; Kong and Zhang, 2014; Liu et al., 2012; Zhang et al., 2011, 2013; Yang et al., 2009, 2010; Ding et al., 2012, 2013, 2014). According to Table 2, the present study’s method resulted in acceptable overall prediction accuracy on both the 25PDB and the 1189 datasets. Specially, the overall prediction accuracy on 1189 and 25PDB datasets achieved 86.02% and 90.08%, respectively, which were the highest values among all of the results that were compared.

### 5. Conclusion

One of the critical issues in the bioinformatics field is the prediction of the protein structural class. The present paper proposes new features to represent a protein sequence. To make these features, a predicted secondary structure of each sequence is employed and a 2D visualization of sequences is created by the chaos game theory. For better analysis, two time series are extracted that preserve all properties of the CGR diagram and the phase space of each time series is constructed based on suitable parameters. Four new features based on the complex network theory combined with eight existing features based on RQA analysis individually are extracted from each time series and so a 24-feature set is prepared. These features are employed to represent protein sequences and are fed into FDA. A comparison with existing methods demonstrates that the proposed method can attain higher prediction accuracy and higher MCC values in most cases. In addition, it is shown that based on the complex network theory, the proposed features are vital to achieving satisfactory prediction accuracies.

As demonstrated in a series of recent publications (see, e.g., Liu et al., 2014, 2015a; Jia et al., 2015b; Ding et al., 2014; Chen et al., 2013, 2014a; Guo et al., 2014; Lin et al., 2014) in developing new prediction methods, user-friendly and publicly accessible web-servers will significantly enhance their impacts (Chou, 2015), we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper. Also, the codes used to prepare this article are available from the author upon request.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All α</td>
<td>All β</td>
<td>α/β</td>
<td>α + β</td>
</tr>
<tr>
<td>25PDB</td>
<td>Ding et al. (2012)</td>
<td>95.03</td>
<td>81.26</td>
</tr>
<tr>
<td>Ying et al. (2013)</td>
<td>95.07</td>
<td>80.8</td>
<td>82.4</td>
</tr>
<tr>
<td>Ding et al. (2014)</td>
<td>96.6</td>
<td>87.1</td>
<td>83</td>
</tr>
<tr>
<td>Ding et al. (2019)</td>
<td>92.8</td>
<td>83.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Kong and Zhang (2014)</td>
<td>94.1</td>
<td>87.1</td>
<td>81.4</td>
</tr>
<tr>
<td>Zhang et al. (2011)</td>
<td>95</td>
<td>85.6</td>
<td>81.5</td>
</tr>
<tr>
<td>Ding et al. (2013)</td>
<td>96.6</td>
<td>87.1</td>
<td>83.0</td>
</tr>
<tr>
<td>This paper</td>
<td>95.65</td>
<td>89.53</td>
<td>88.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Accuracy</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1189</td>
<td>Ding et al. (2012)</td>
<td>93.73</td>
<td>84.01</td>
</tr>
<tr>
<td>Ying et al. (2014)</td>
<td>94.0</td>
<td>84.3</td>
<td>84.4</td>
</tr>
<tr>
<td>Ding et al. (2014)</td>
<td>94.2</td>
<td>88.4</td>
<td>85.3</td>
</tr>
<tr>
<td>Yang et al. (2010)</td>
<td>89.2</td>
<td>86.7</td>
<td>85.6</td>
</tr>
<tr>
<td>Kong et al. (2014)</td>
<td>91.9</td>
<td>84.4</td>
<td>85.3</td>
</tr>
<tr>
<td>Kong and Zhang (2014)</td>
<td>83.7</td>
<td>86.1</td>
<td>86.8</td>
</tr>
<tr>
<td>Zhang et al. (2011)</td>
<td>92.4</td>
<td>87.4</td>
<td>82</td>
</tr>
<tr>
<td>Ding et al. (2013)</td>
<td>94.2</td>
<td>88.4</td>
<td>85.3</td>
</tr>
<tr>
<td>This paper</td>
<td>92.3</td>
<td>90.11</td>
<td>86.48</td>
</tr>
</tbody>
</table>

References


Feng, F.-M., Chen, W., Ding, S., Shi, Z., Yan, S., 2012. A protein structural classes prediction method based on PSI-BLAST pro


Table 2

Performance comparison of different methods on two datasets.


