Journal of Theoretical Biology 257 (2009) 618–626

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation

Jian-Yi Yang [a,b,*,1], Zhen-Ling Peng [c,1], Zu-Guo Yu [d,e], Rui-Jie Zhang [b], Vo Anh [e], Desheng Wang [a]

[a] Division of Mathematical Sciences, School of Physical & Mathematical Sciences, Nanyang Technological University, Singapore 637371
[b] School of Bioinformatics Science and Technology, Harbin Medical University, Heilongjiang 150081, China
[c] Department of Mathematics, Bijie University, Guizhou 551700, China
[d] School of Mathematics and Computational Science, Xiangtan University, Hunan 411105, China
[e] School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia

ARTICLE INFO

Article history:
Received 21 August 2008
Received in revised form
7 November 2008
Accepted 19 December 2008
Available online 8 January 2009

Keywords:
Sequence homology
Recurrence plots
Jackknife test
One-against-others

ABSTRACT

In this paper, we intend to predict protein structural classes (α, β, $\alpha+\beta$, or α/β) for low-homology data sets. Two data sets were used widely, *1189* (containing 1092 proteins) and *25PDB* (containing 1673 proteins) with sequence homology being 40% and 25%, respectively. We propose to decompose the chaos game representation of proteins into two kinds of time series. Then, a novel and powerful nonlinear analysis technique, recurrence quantification analysis (RQA), is applied to analyze these time series. For a given protein sequence, a total of 16 characteristic parameters can be calculated with RQA, which are treated as feature representation of protein sequences. Based on such feature representation, the structural class for each protein is predicted with Fisher's linear discriminant algorithm. The jackknife test is used to test and compare our method with other existing methods. The overall accuracies with *step-by-step* procedure are 65.8% and 64.2% for *1189* and *25PDB* data sets, respectively. With *one-against-others* procedure used widely, we compare our method with five other existing methods. Especially, the overall accuracies of our method are 6.3% and 4.1% higher for the two data sets, respectively. Furthermore, only 16 parameters are used in our method, which is less than that used by other methods. This suggests that the current method may play a complementary role to the existing methods and is promising to perform the prediction of protein structural classes.

1. Introduction

The tertiary structure of a protein is determined by its amino acid sequence via the process of protein folding (Anfinsen, 1973). In order to explore the mechanism of the protein folding process, some theoretical works have suggested the designability and other concepts based on lattice models (Li et al., 1996, 1998; Wang and Yu, 2000; Yang et al., 2007a). It is possible to predict the tertiary structure of a protein from its primary structure directly. But this is a challenging problem as there is no simple rule to map the primary sequence into the corresponding tertiary structure of a protein. Four main structural classes of proteins were recognized based on the types and arrangement of their secondary structural elements (Levitt and Chothia, 1976). They are the α class, the β class and those with a mixture of α and β shapes called the $\alpha+\beta$ class and the α/β class. It is especially important to predict protein structural classes (see, e.g., Chou and Zhang, 1995). Firstly, it is helpful for the prediction of protein secondary and tertiary structure. For example, the searching scope of conformation will be reduced if the structural class of the protein under study is known (Bahar et al., 1997). Secondly, the structural class is related to various properties of protein (e.g., biological function, and existence of disulfide bonds) (Nishkawa and Ooi, 1982).

Many methods have been proposed to predict structural classes of protein from their primary sequences (Anand et al., 2008; Chen et al., 2008a; Chou, 1995, 1999, 2000, 2005b; Chou and Zhang, 1994, 1995; Kedarisetti et al., 2006; Kurgan and Homaeian, 2006; Wang and Yuan, 2000; Zhang et al., 2008; Zhou, 1998; Zhou and Assa-Munt, 2001), some of which are based on amino acid composition (see, e.g., Chou and Zhang, 1995; Kedarisetti et al., 2006; Kurgan and Homaeian, 2006; Wang and Yuan, 2000). These methods were reported to be nearly perfect when high-homology data sets were used and the reason for their success in predicting structural class from amino acid composition was explored by Bahar et al. (1997) with lattice models. For example, the *359* data set (Chou and Maggiora, 1998) (over 95% homology) was used extensively in the past decade to test the effectiveness of various prediction methods (Kedarisetti et al.,

* Corresponding author at: Division of Mathematical Sciences, School of Physical & Mathematical Sciences, Nanyang Technological University, Singapore 637371 Tel.: +65 84372208
E-mail addresses: yangjianyiapple@163.com, yang0241@ntu.edu.sg (J.-Y. Yang).
[1] Joint first authors.

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.12.027

2006; Kurgan and Homaeian, 2006; Wang and Yuan, 2000). For this data set, the overall accuracy was higher than 90% (Kedarisetti et al., 2006; Kurgan and Homaeian, 2006). However, when low-homology data sets were used, these methods were not effective any more. For instance, for two low-homology data sets, *1189* and *25PDB* with sequence homology being 40% and 25%, respectively, the reported overall accuracy with these methods was less than 60% (Kedarisetti et al., 2006; Kurgan and Homaeian, 2006; Wang and Yuan, 2000). That is to say, the performance of existing methods was strongly affected by sequence homology. Kurgan and Homaeian (2006) discussed this problem in detail. Kedarisetti et al. (2006) calculated the accuracies for protein sequences with varying homologies and concluded that prediction of structural classes is more difficult for low-homology sequences than for higher-homology sequences. In order to predict protein structural classes for low-homology sequences, it is necessary to develop some new methods.

Recently, we have successfully used some different methods to predict structural classes of large proteins (i.e., protein sequences are long) (Yang et al., 2007b, 2008; Yu et al., 2006). Yu et al. (2006) used the hydrophobic free energy and solvent accessibility of amino acids to construct several parameter spaces. We found that some spaces could be used to distinguish and cluster the 43 selected large proteins from the four structural classes. With hydrophobicity scale of amino acids and a 6-letter model, we recently discussed the clustering of 49 large proteins via multifractal analysis (MFA) (Yang et al., 2007b).

Chaos game representation (CGR) of protein structures was first proposed by Fiser et al. (1994). We denote this CGR by 20-CGR as 20 kinds of letters are used to represent protein sequences. Later Basu et al. (1997) and Yu et al. (2004) proposed other kinds of CGRs for proteins, in which 12 and 4 kinds of letters were used for protein sequences, respectively. We denote them by 12-CGR and 4-CGR. We also applied a 6-letter model to cluster 49 large proteins (Yang et al., 2007b), hence we can discuss using 6 letters in CGR and it is denoted by 6-CGR. Recently, protein sequences are transformed into nucleotide sequences based on fixed reverse encoding of amino acids (Deschavanne and Tufféry, 2008) and the famous CGR for DNA sequence analysis (Jeffrey, 1990) can be used on protein sequences. Deschavanne and Tufféry (2008) showed that such method was able to classify functional families of proteins and protein structural classes. For convenience, we denote such method by AAD-CGR (Amino Acids to DNA) and it is the main method adopted here.

20-CGR have been used successfully by us to predict structural classes of 100 large proteins based on MFA recently (Yang et al., 2008). The protein sequences were transformed into two time series which were then analyzed by MFA. The disadvantage of this methods is that it requires that the length of protein sequences to be long enough (often >300, i.e., large proteins). However, most of proteins stored in RCSB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) are small proteins. In order to solve this problem, here we adopt *recurrence quantification analysis* (RQA) (Giuliani et al., 2002) to analyze the time series here. RQA is a powerful nonlinear technique in analyzing time series without the requirement on the length of time series.

In this paper, we intend to predict the protein structural classes of the two low-homology data sets, *1189* and *25PDB*, which contain 1092 and 1673 proteins, respectively. Firstly, the protein sequences are converted into two different time series with AAD-CGR. Secondly, RQA is applied to analyze these time series. For each time series, eight parameters are achieved. Thus, for a given protein sequence, a total of 16 (2×8) characteristic parameters can be calculated. These parameters are used to predict the structural classes of proteins by Fisher's linear discriminant algorithm. One *step-by-step* procedure is proposed to predict protein structural classes. The overall accuracies are 65.8% and 64.2% for *1189* and *25PDB* data sets, respectively. The jackknife test, which is the most rigorous and objective algorithm for evaluating the power of prediction methods, is used to evaluate and compare our method with five other existing methods. We found that the method proposed here has better performance than these methods, suggesting that the method proposed here may play a complementary role to the existing methods.

2. Models and methods

2.1. Reverse encoding for amino acids

It is known that there are 20 kinds of amino acids (AAs) and several kinds of coded methods for some AAs. As a result, there should have many possible nucleotide sequences for one given protein sequence. Here, we use the encoding method used by Deschavanne and Tufféry (2008) which is listed in Table 1. Deschavanne and Tufféry (2008) explained that the rationale for the choice of this fixed code is to keep a balance in base composition so as to maximize the difference between the amino acid codes. Two more arbitrarily selected reverse encoding methods are discussed in the Discussion section.

2.2. CGR and related time series

After one protein sequence is transformed into nucleotide sequences, we can use AAD-CGR of nucleotide sequences (Jeffrey, 1990) to analyze it. We recapture the concept of CGR briefly here. CGR for a nucleotide sequence is defined in a square $[0,1]\times[0,1]$, where the four vertices correspond to the 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 and the vertex corresponding to the i-th letter. The obtained plot is then called CGR of the nucleotide sequence, or AAD-CGR of the protein sequence. AAD-CGR of a typical protein is shown in Fig. 1 as an example.

It is not easy to analyze the obtained plot directly. Noticing that the AAD-CGR of proteins is determined by the (x,y) coordinates, we proposed to decompose the AAD-CGR into two time series and then analyzed them by MFA recently (Yang et al., 2008). Similarly, we decompose the AAD-CGR plot into two time series here. Any point in the AAD-CGR plot is determined by two coordinates, namely, x and y coordinates. Thus, two time series can be achieved from the AAD-CGR plot. Fig. 2 shows the two time series related to the AAD-CGR plot in Fig. 1. We denote them as CGRx and CGRy, respectively.

Because the AAD-CGR plot can be uniquely reconstructed from these two time series, all the information stored in the AAD-CGR plot is contained in the time series. And the information in the AAD-CGR plot comes from the primary sequence of proteins. Therefore, any analysis of the two time series is equivalent to indirect analysis of the protein primary sequence. We hope that such analysis provides better results than direct analysis of the protein primary sequences.

2.3. Recurrence plot

Recurrence plot (RP) is a purely graphical tool originally proposed by Eckmann et al. (1987) to detect patterns of recurrence in the data. For one time series $\{x_1,x_2,\ldots,x_N\}$ with length N, we can embed it into the space R^m with embedding

Table 1
The reverse encoding for amino acids.

A = GCT	G = GGT	M = ATG	S = TCA	C = TGC	H = CAC	N = AAC	T = ACT	D = GAC	I = ATT
P = CCA	V = GTG	E = GAG	K = AAG	Q = CAG	W = TGG	F = TTC	L = CTA	R = CGA	Y = TAC

This encoding method has been used in Deschavanne and Tufféry (2008).

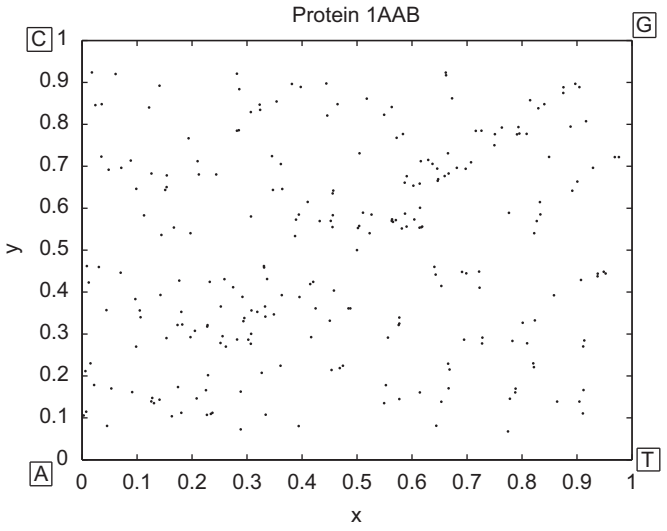

Fig. 1. AAD-CGR of protein 1AAB. One point in the figure represents one corresponding amino acid in the protein sequence. The order for the points (corresponding to the order in the protein sequence) is saved (not shown in the figure). See text for more details about how to get such plots.

dimension m and a time delay τ.

$$\vec{y}_i = (x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(m-1)\tau}), \quad i = 1, 2, \ldots, N_m, \tag{1}$$

where $N_m = N - (m-1)\tau$. In this way we get N_m vectors (points) in the embedding space R^m. Both embedding dimension m and a time delay τ have to be chosen appropriately from nonlinear dynamical theory (Riley and Van Orden, 2005). We will also give some numerical explanations for the selection of m and τ in the following.

From the N_m points, we can calculate the *distance matrix* (DM), which is a square $N_m \times N_m$ matrix. The elements of DM are the distances between all possible combinations of i-points and j-points. They are computed according to the norming function selected. Generally, Euclidean norm is used (Giuliani et al., 2002). DM can be rescaled by dividing down each element in the DM by a certain value as this allows systems operating on different scales to be statistically compared. For such value, the maximum distance of the entire matrix DM is the most commonly used (and recommended) rescaling option, which redefines the DM over the unit interval (0.0–100.0%) (Riley and Van Orden, 2005).

Once the rescaled DM $= (D_{i,j})_{N_m \times N_m}$ is calculated, it can be transformed into a *recurrence matrix* (RM) of distance elements within a *threshold* ε (namely radius). RM $= (R_{i,j}(\varepsilon))_{N_m \times N_m}$ and

$$R_{i,j}(\varepsilon) = H(\varepsilon - D_{i,j}), \quad i,j = 1, 2, \ldots, N_m, \tag{2}$$

where H is the Heaviside function

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geqslant 0. \end{cases} \tag{3}$$

RP is simply a visualization of RM by plotting points on i–j plane for those elements in RM with values equal to 1. If $R_{i,j}(\varepsilon) = 1$, we say j-points recur with reference to i-points. For any ε, since $R_{i,i}(\varepsilon) \equiv 1$ $(i = 1, 2, \ldots, N_m)$, the RP has always a black main diagonal line. Furthermore, the RP is symmetric with respect to the main diagonal as $R_{i,j}(\varepsilon) = R_{j,i}(\varepsilon)$ $(i,j = 1, 2, \ldots, N_m)$. For example, the RPs for the two time series shown in Fig. 2 are given in Fig. 3. ε is a crucial parameter of RP. If ε is chosen too small, there may be almost no recurrence points and we will not be able to learn about the recurrence structure of the underlying system. On the other hand, if ε is chosen too large, almost every point is a neighbor of every other point, which leads to a lot of artifacts (Marwan et al., 2007). The selection of ε will be discussed numerically below.

2.4. Recurrence quantification analysis

RQA is a relative new nonlinear technique based on RP. Webber and Zbilut (1994) and Zbilut and Webber (1992) quantify the information supplied by RP. RQA has been successfully applied to many different fields (Giuliani and Tomasi, 2002; Zaldívar et al., 2008). The ability of RQA to deal with protein sequences was investigated in Giuliani et al. (2000), Manetti et al. (1999), Webber et al. (2001), Zbilut et al. (2004), and Zhou et al. (2007). The works using signal analysis methods in elucidation of protein sequence–structure relationships were reviewed in Giuliani et al. (2002). In analyzing time series, unlike the MFA (Yu et al., 2006; Yang et al., 2007b, 2008), RQA does not have a strict requirement for the length of time series. Therefore, it is possible to apply RQA in the prediction of structural classes of small proteins. For convenience, we briefly describe RQA as follows. There are eight recurrence variables used to quantify RP (Marwan et al., 2007; Riley and Van Orden, 2005). We recommend (Marwan et al., 2007; Riley and Van Orden, 2005) as there are much more detailed descriptions of these variables in these two references. It should be pointed out that the recurrence points in the following definitions only consist of those in the upper triangle in RP (excluding the main diagonal line), which is similar to those in Riley and Van Orden (2005).

The first recurrence variable is *%recurrence* (*%REC*). *%REC* is a measure of the density of recurrence points in the RP. This variable can range from 0% (no recurrent points) to 100% (all points are recurrent):

$$\%REC = 100 \times (\#\text{ recurrent points in upper triangle})/(N_m(N_m - 1)/2)), \tag{4}$$

where # stands for counting the number of points.

The second recurrence variable is *%determinism* (*%DET*). *%DET* measures the proportion of recurrent points forming diagonal line structures. For this variable, we have to first decide at least how many adjacent recurrent points are needed to define a diagonal line segment. Obviously, the minimum number required (and commonly used) is 2 and such requirement is adopted in this paper:

$$\%DET = 100 \times (\#\text{ points in diagonal lines})/(\#\text{ recurrent points}). \tag{5}$$

The third recurrence variable is *linemax* (L_{max}), which is simply the length of the longest diagonal line segment in RP. This is a very important recurrence variable because it inversely scales with the largest positive Lyapunov exponent (Eckmann et al., 1987):

$$L_{max} = \text{length of longest diagonal line in RP.} \tag{6}$$

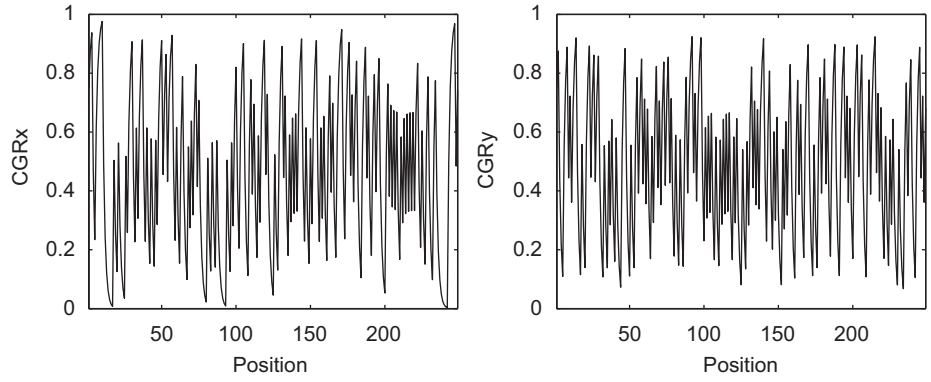

Fig. 2. Two time series related to Fig. 1. The first time series (left panel) is used to represent the *x*-coordinate of the points in Fig. 1 and the second (right panel) is used to represent the *y*-coordinate of the points in Fig. 1.

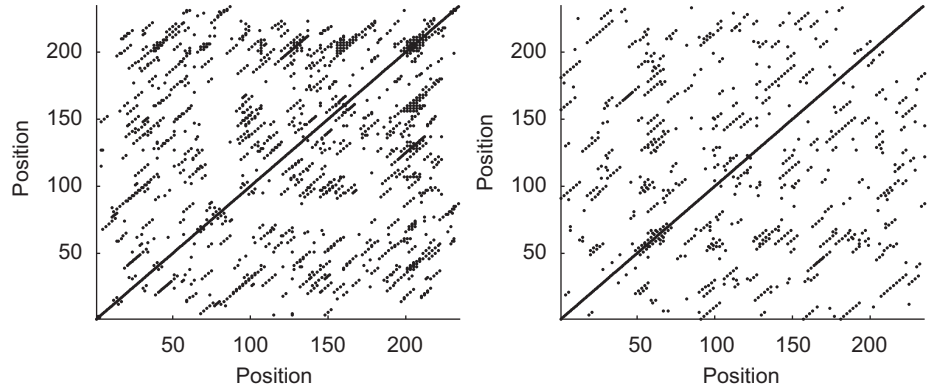

Fig. 3. The corresponding RPs for the two time series in Fig. 2. The parameters used are: $m = 8, \tau = 2, \varepsilon = 30\%$. It can be noticed that there is black main diagonal line in the plots as two identical points (in higher dimension, here it is 8-dimension point as $m = 8$) are always recurrent. The points in the RP are symmetric with respect to this main diagonal line.

The fourth recurrence variable is *entropy* (*ENT*), which is the Shannon information entropy of the distribution probability of the length of the diagonal lines:

$$ENT = -\sum_{k=L_{min},p(k)\neq 0}^{L_{max}} p(k)\log_2(p(k)), \tag{7}$$

where L_{min} is the minimum length of diagonal lines in RP and

$$p(k) = (\#\text{ diagonal lines of length}k\text{in RP})/(\#\text{ diagonal lines in RP}). \tag{8}$$

The fifth recurrence variable is *trend* (*TND*), which quantifies the degree of system stationarity. It is calculated as the slope of the least squares regression of *%local recurrence* as a function of the displacement from the main diagonal. It should be made clear the so-called %local recurrence is in fact the proportion of recurrent points on certain line parallel to the main diagonal over the length of this line. %recurrence is calculated on the whole upper triangle in RP while %local recurrence is computed on only certain line in RP, so it is termed as *local*. Multiplying by 1000 increases the gain of the *TND* variable (Riley and Van Orden, 2005):

$$TND = 1000 \times (\text{slope of \% local recurrence vs. displacement}). \tag{9}$$

The remaining three variables are defined based on the vertical line structure. The sixth recurrence variable is *%laminarity* (*%LAM*). *%LAM* is analogous to *%DET* but it is calculated with recurrent points comprising vertical line structures. Similarly, we also select 2 as the minimum number of adjacent recurrent points to form a vertical line segment:

$$\%LAM = 100 \times (\#\text{ points in vertical lines})/(\#\text{ recurrent points}). \tag{10}$$

The seventh variable, *trapping time* (*TT*), is the average length of vertical line structures. The eighth recurrence variable is *maximal length of the vertical lines* in RP (V_{max}), which is similar to L_{max}.

2.5. Determination of parameters

As mentioned in Section 2.3, we have to decide several parameters in RP: embedding dimension m, time delay τ, and radius ε. Because RP is analyzed by RQA, we use the quantity *%REC* to discuss the selections of these parameters. Similar to that done in Riley and Van Orden (2005), we decide to examine *%REC* output for embedding dimensions m between 6 and 9, and time delay τ from 1 to 4. A general guideline is that ε should be selected such that *%REC* remains low (often smaller than 5%) (Riley and Van Orden, 2005). We are looking for small (or smooth) changes in parameter settings yielding small (or smooth) changes in output measures, *%REC* values ranging between 0% and 5%.

The time series used for this analysis is the *x*-coordinate sequence (CGRx) in AAD-CGR of protein 1AAB. In order to decide the value of parameters in RQA, we present the surface plots for embedding dimensions m from 6 to 9, delays τ from 1 to 4 and radius ε from 10 to 39 in Fig. 4. From the figure, we set $m = 8$ as the changes for *%REC* are relatively small for $m = 7, 8, 9$ and the changes are smooth for $m = 8$ with the increase of radius ε and τ; we set $\tau = 2$ as the values of *%REC* almost do not change

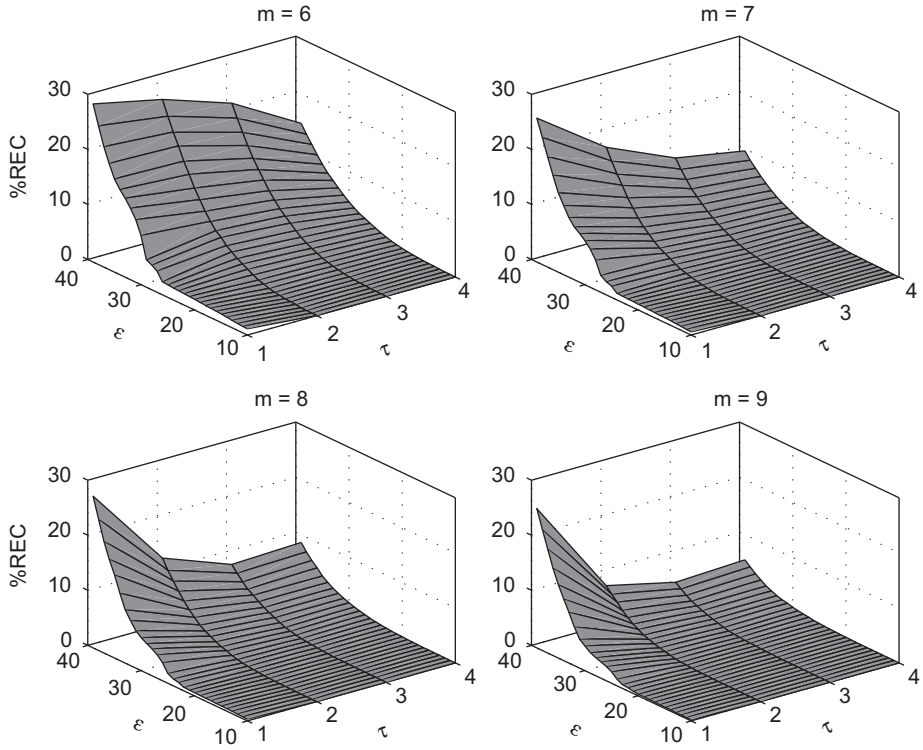

Fig. 4. Surface plots for embedding dimensions from 6 to 9 and delays from 1 to 4. *%REC* for different values of parameters in RQA is calculated for the time series (CGRx) in Fig. 2. From the figure, we set $m = 8, \tau = 2, \varepsilon = 30\%$. In the calculation, ε ranges between 10% and 49%. See text (Section 2.5) for details of such selection of parameters.

(or change smoothly) along with τ when $\tau \geqslant 2$. Finally, ε is set at 30% to make sure *%REC* is smaller than 5% (Riley and Van Orden, 2005).

3. Data and results

3.1. Data sets

For the test of one method in protein structural class prediction, many different data sets have been used in previous study. For example, a data set containing 359 proteins developed by Chou and Maggiora (1998) was used extensively. This data set includes highly homologous sequences (over 95% homology) and the prediction accuracy is nearly perfect (Kurgan and Homaeian, 2006; Kedarisetti et al., 2006). However, for the low-homology data sets, the prediction accuracy is notoriously low and we try to address this problem here by using two low-homology protein data sets.

The two protein data sets analyzed here are used previously (Wang and Yuan, 2000; Kurgan and Homaeian, 2006). The first one contains 1092 proteins/domains consisting of 223 α class proteins, 294 β class proteins, 334 α/β class proteins and 241 $\alpha+\beta$ class proteins. It is denoted by *1189*. Another is from Kurgan and Homaeian (2006) containing 1673 proteins/domains, in which 443 are from the α class, 443 from the β class, 441 from the α/β class and 346 from the $\alpha+\beta$ class. It is denoted by *25PDB*. The sequence homology of *1189* and *25PDB* data sets is 40% and 25%, respectively (Kurgan and Homaeian, 2006; Kedarisetti et al., 2006). One can download the protein/domain sequences of *1189* and *25PDB* from RCSB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) by the PDB ID listed in the Appendix A of Kurgan and Homaeian (2006).

3.2. Prediction of protein structure classes

For each protein sequence, two time series are derived from its AAD-CGR of amino acids sequence. Then we use RQA to analyze them which resulting in a total of 16 parameters (2×8). That is to say, each protein sequence is represented by 16 features which can be used to predict its structural class. There are a lot of existing prediction algorithms. For example, artificial neural network (Dubchak et al., 1995), Bayesian classification (Wang and Yuan, 2000), fuzzy clustering (Shen et al., 2005), LogitBoost (Feng et al., 2005) and support vector machines (Ding and Dubchak, 2001; Chen et al., 2006), etc. In the present paper, we use Fisher's linear discriminant algorithm.

Fisher's discriminant algorithm is used to find a classifier in the parameter space for a training set. The given training set $H = \{x_1, x_2, \ldots, x_n\}$ is partitioned into $n_1 \leqslant n$ training vectors in a subset H_1 and $n_2 \leqslant n$ training vectors in a subset H_2, where $n_1 + n_2 = n$ and each vector x_i is a point in the 16-D parameter space. Then $H = H_1 \cup H_2$. We need to find a parameter vector $\mathbf{w} = (w_1, w_2, \ldots, w_{16})^{\mathrm{T}}$ for the 16-D space such that $\{y_i = \mathbf{w}^{\mathrm{T}} x_i\}_{i=1}^{n}$ can be classified into two classes in the space of real numbers. If we denote

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x_i} \in H_j} \mathbf{x_i}, \quad j = 1, 2, \tag{11}$$

$$\mathbf{S}_j = \sum_{\mathbf{x_i} \in H_j} (\mathbf{x_i} - \mathbf{m_j})(\mathbf{x_i} - \mathbf{m_j})^{\mathrm{T}}, \quad j = 1, 2, \tag{12}$$

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2, \tag{13}$$

then the parameter vector $\mathbf{w}$ is estimated as $\mathbf{S}_w^{-1}(\mathbf{m_1} - \mathbf{m_2})$ (Duda et al., 2001). As a result, Fisher's discriminant rule becomes assign

$\mathbf{x}$ to H_1 if $(\mathbf{m_1}-\mathbf{m_2})^{\mathrm{T}}\mathbf{S}_w^{-1}[\mathbf{x}-\frac{1}{2}(\mathbf{m_1}+\mathbf{m_2})]>0$ and to H_2 otherwise (Duda et al., 2001).

For protein structural class classification problem, we have proposed a procedure to cluster protein structure in a 3D parameter space recently (Yang et al., 2007b; Yu et al., 2006). We name this procedure as a *step-by-step* method. The main idea of this method is classifying the proteins of certain class from other proteins by three steps. Correspondingly, one predictor is trained for each step. After testing all the possible steps (altogether 12 combinations), we find that the following steps are good (i.e., have high overall accuracy defined by formula (18)) to predict protein structural classes:

- *Step* 1: classify the proteins of the β class from the other proteins in the $\{\alpha,\alpha+\beta,\alpha/\beta\}$ classes;
- *Step* 2: classify the α/β class proteins from the other proteins in the $\{\alpha,\alpha+\beta\}$ classes;
- *Step* 3: classify the α class proteins from the proteins of the $\alpha+\beta$ class and the remaining proteins belong to the $\alpha+\beta$ class.

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent data set test, sub-sampling (such as fivefold or 10-fold sub-sampling) test, and jackknife test (Chou and Zhang, 1995). Of these three, however, the jackknife test is thought the most rigorous and objective that can always yield a unique result for a given benchmark data set, as elucidated in Chou and Shen (2008a) and demonstrated by Eq. (50) of Chou and Shen (2007d). Therefore, the jackknife test has been increasingly and widely adopted or recognized by many investigators to examine the power of various predictors (see, e.g., Chen et al., 2008a, b; Chen and Li, 2007a, b; Chou and Shen, 2006a, b, 2007a–c, 2008b; Du and Li, 2008; Jiang et al., 2008; Jin et al., 2008; Li and Li, 2008; Lin, 2008; Lin et al., 2008; Munteanu et al., 2008; Niu et al., 2008; Shen and Chou, 2007a, b; Shi et al., 2008; Wu and Yan, 2008; Zhang and Fang, 2008; Zhou et al., 2007). Hence, in this paper, we use the jackknife test to evaluate our method. In the jackknife test, each of the protein in the data set is in turn singled out as a tested protein, and the predictor is trained by the remaining proteins. Therefore, jackknife is also called a *leave-one-out* test.

Based on Fisher's discriminant algorithm, we calculate the prediction accuracies as follows (formula (14)–(18)). The prediction accuracies for the two data sets are given in Table 2. The overall accuracy for the data set *25PDB* is 1.6% lower than that for the data set *1189*, which might beso because the sequence homology for the former data set is lower than that for the latter. This is consistent with the conclusion in Kurgan and Homaeian (2006) (i.e., using highly homologous data sets results in higher accuracies).

$$\text{accuracy}_{\alpha} = \frac{\text{the number of proteins predicted correctly in the } \alpha \text{ class}}{\text{the number of proteins in the } \alpha \text{ class}}, \tag{14}$$

Table 2
Accuracies from AAD-CGR with step-by-step algorithm.

Data set	Prediction accuracy (%)				
	α	β	$\alpha+\beta$	α/β	Overall
1189	60.5	67.7	61.4	71.0	65.8
25PDB	58.0	65.0	65.1	69.9	64.2

$$\text{accuracy}_{\beta} = \frac{\text{the number of proteins predicted correctly in the } \beta \text{ class}}{\text{the number of proteins in the } \beta \text{ class}}, \tag{15}$$

$$\text{accuracy}_{\alpha+\beta} = \frac{\text{the number of proteins predicted correctly in the } \alpha+\beta \text{ class}}{\text{the number of proteins in the } \alpha+\beta \text{ class}}, \tag{16}$$

$$\text{accuracy}_{\alpha/\beta} = \frac{\text{the number of proteins predicted correctly in the } \alpha/\beta \text{ class}}{\text{the number of proteins in the } \alpha/\beta \text{ class}}, \tag{17}$$

$$\text{accuracy}_{\text{overall}} = \frac{\text{the number of proteins predicted correctly in all classes}}{\text{the number of proteins in all classes}}. \tag{18}$$

4. Discussions

4.1. Comparison with existing methods

For the four-class prediction problem, the *one-against-others* algorithm (Brown et al., 2000; Chen et al., 2006; Ding and Dubchak, 2001) is widely used to transfer it into a two-class problem. Therefore, in order to give a fair comparison with other existing methods, here we calculate the prediction accuracies of our method based on the one-against-others algorithm as well. The corresponding accuracies are listed in Table 3 in bold type. In order to compare with existing methods, other investigators' results for the two date sets are also listed in Table 3, which is taken directly from the original papers. From this table, we can see the following three points:

First, for *1189* data set, the overall prediction accuracy of our method is higher than other methods listed in Table 3 (from 6.3% to 11.4%). The accuracy in predicting α/β class proteins in Zhang et al. (2008) is higher than ours. Except for this, our prediction accuracies for three other protein classes are much more higher than other methods. For the available data in Table 3, this ranges between 4.8% and 37.8%.

Second, for *25PDB* data set, the overall prediction accuracy of our method is 4.1% and 6.9% higher than that in Kedarisetti et al. (2006) and Kurgan and Homaeian (2006), respectively. The accuracy in predicting α class proteins in Kurgan and Homaeian (2006) is higher than ours. However, for three other protein classes, our prediction accuracies are all higher than that in Kurgan and Homaeian (2006) (3.4%, 1.6% and 26.7% for β, $\alpha+\beta$ and α/β class proteins, respectively).

Third, only 16 parameters are used in our prediction, which is less than those used by other state-of-the-art methods. The method listed in Table 3 with the highest overall accuracy used 34 parameters (Kedarisetti et al., 2006), which is more than twice of the number of parameters used here. This clearly demonstrates that our method is promising in the classification of structural classes for proteins with low sequence homology.

Therefore, the current method proposed in this paper may play a complementaryrole to the existing methods.

Why the present method can get better results than other methods? It should be stressed that most methods in Kedarisetti et al. (2006), Kurgan and Homaeian (2006), and Wang and Yuan (2000) are based on amino acid composition. The important information, sequence order, is lost in these methods. For the method with pseudo amino acid (PseAA) composition (Chou, 2001, 2005a), some sort of sequence order information is incorporated but it is not enough. Sequence order is used in RQA and may play an important role in predicting protein structural classes (Riley and Van Orden, 2005). This can be proved from the fact that the results from PseAA (Zhang et al., 2008) are

Table 3
Accuracies of our method with the one-against-others algorithm and comparison with other results reported.

Data set	Reference	Parameters	Prediction accuracy (%)				
			α	β	$\alpha+\beta$	α/β	Overall
1189	**This paper**	**16**	**62.3**	**67.7**	**63.1**	**66.5**	**65.2**
	Wang and Yuan (2000)	19	NA	NA	NA	NA	53.8
	Kurgan and Homaeian (2006)	66	57.0	62.9	25.3	64.6	53.9
	Kedarisetti et al. (2006)	34	NA	NA	NA	NA	58.9
	Zhang et al. (2008)	27	48.9	59.5	26.6	81.7	56.9
	Anand et al. (2008)	50	NA	NA	NA	NA	54.7
25PDB	**This paper**	**16**	**64.3**	**65.0**	**61.7**	**65.0**	**64.0**
	Kurgan and Homaeian (2006)	66	69.1	61.6	60.1	38.3	57.1
	Kedarisetti et al. (2006)	34	NA	NA	NA	NA	59.9

The best overall results and the corresponding accuracies in predicting the four protein structural classes are highlighted in bold face.

Table 4
Two more ways of reverse encoding for amino acids.

A = GCA	G = GGA	M = ATG	S = TCG	C = TGT	H = CAT	N = AAT	T = ACA	D = GAT	I = ATA
P = CCC	V = GTA	E = GAA	K = AAA	Q = CAA	W = TGG	F = TTT	L = CTT	R = CGC	Y = TAT
A = GCC	G = GGG	M = ATG	S = TCC	C = TGT	H = CAT	N = AAT	T = ACC	D = GAC	I = ATC
P = CCT	V = GTC	E = GAG	K = AAG	Q = CAG	W = TGG	F = TTC	L = TTG	R = CGT	Y = TAT

The reverse encoding manner is selected randomly.

better than those from amino acid composition (Kurgan and Homaeian, 2006; Wang and Yuan, 2000). Detailed discussions were made about whether the coupling effect among different amino acid components can improve the prediction of protein structural classes (Cai, 2001; Chou et al., 1998; Eisenhaber et al., 1998). Chou et al. (1998) confirmed that the answer is yes. When it comes to the method proposed here, the protein sequence is converted into time series based on CGR and there is no information lost in this transition. Much more information of sequence order may be contained by analyzing these time series with RQA. Therefore, it is reasonable to expect better prediction results from RQA.

In order to predict protein structural classes more precisely, Chou and Cai (2004) proposed a higher level method (Chou and Shen, 2007d), the so-called "functional domain (FunD)" compositional approach. In this method, each protein is presented by a 7785-dimensional vector. It was used to predict protein structural classification among seven classes: α, β, α/β, $\alpha+\beta$, multi-domain, small protein, and peptide. A very high jackknife test rate (98%) was obtained on 2230 proteins in which none of protein has more than 20% pairwise sequence identity to any others. This suggests that a feature representation containing more information of a protein is helpful in predicting protein structural classes. Because the prediction for two data sets used here was not tested by this method, we cannot compare ours with it here. Anyway, this FunD compositional approach is highly valuablein that it directs us to explore those new methods able to incorporate as much information of proteins as possible when predicting protein structural classes.

4.2. Effect of reverse encoding of amino acids

For one amino acid, there are many possible ways to translate it into nucleotides. For example, the amino acid L can be translated by six different ways: TTA, TTG, CTT, CTA, CTC and CTG. Therefore, except for the encoding manner listed in Table 1, we need to test other kinds of encoding manners to see whether they affect the final results. However, it is not possible and practical to test all possible encoding manner. Here, we test the following two more manners listed in Table 4, which is randomly selected. The encoding manner in Table 1 is denoted by En1, and those in Table 2 by En2 and En3. Based on the one-against-others method, the prediction results are listed in Table 5. From the table, the overall accuracy from En1 is the best one, which is consistent with the conclusion in Deschavanne and Tufféry (2008) (i.e., the encoding manner in Table 1 resulted in a better result).

Table 5
Prediction accuracies from En1, En2 and En3 based on the one-against-others method.

Data set	Encoding	Prediction accuracy (%)				
		α	β	$\alpha+\beta$	α/β	Overall
1189	En1	**62.3**	**67.7**	**63.1**	**66.5**	**65.2**
	En2	46.6	54.1	59.0	65.9	57.2
	En3	53.1	57.5	51.5	63.5	56.7
25PDB	En1	**64.3**	**65.0**	**61.7**	**65.0**	**64.0**
	En2	49.2	58.2	58.3	72.5	58.8
	En3	52.6	59.6	57.4	69.4	59.2

The best overall results and the corresponding accuracies in predicting the four protein structural classes are highlighted in bold face.

4.3. Results with other kinds of CGRs

As mentioned in the Introduction, five kinds of CGRs for proteins will be discussed, i.e., 20-CGR, 12-CGR, 6-CGR, 4-CGR and AAD-CGR. The above results are calculated based on AAD-CGR (En1) and we should test the results from the other four CGRs. In the calculation, we firstly have to decide the parameters in RQA.

To this end, we use the similar method introduced in Section 2.5. Their parameters are listed in Table 6.

Based on the one-against-others procedure, the prediction accuracies with these methods are shown in Table 7, which suggests that the results are different for different CGRs. From the point of overall accuracy, the results from AAD-CGR (En1) are relatively better than other kinds of CGRs. As a result, we recommend AAD-CGR (En1) when using CGR of proteins to predict protein structural classes.

5. Conclusions

Identification of protein structural classes is important in the prediction of the 3D structures. CGR of proteins is a useful way to analyze proteins as it provides a visualization of protein sequences. With the reverse encoding of amino acids in Deschavanne and Tufféry (2008), AAD-CGR was introduced in this paper. In order to analyze AAD-CGR of proteins more conveniently, we decomposed the AAD-CGR of proteins into two time series.

RQA is a useful tool in many different fields. It has been used to deal with the problem relating to protein sequences (Giuliani et al., 2000; Manetti et al., 1999; Zbilut et al., 2004; Zhou et al., 2007). We used RQA in the prediction of structural classes from the primary sequences of proteins in this paper. Two kinds of time series were derived from AAD-CGR. RQA was used to analyze them and 16 (2×8) parameters were achieved. With these parameters altogether, we used the step-by-step procedure to predict protein structural classes based on Fisher's linear discriminant algorithm. The overall accuracies with such method are 65.8% and 64.2% for *1189* and *25PDB* data sets which are low-homologous data sets (sequence homology being 40% and 25%, respectively), respectively. At the same time, this suggests that the lower the sequence homology is, the more difficult it is to predict protein structural classes, which is consistent with the conclusion in Kurgan and Homaeian (2006)

Chou, K.C., 2005b. Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr. Protein Peptide Sci. 6, 423–436.
Chou, K.C., Cai, Y.D., 2004. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. 321, 1007–1009 (Corrigendum: Chou, K.C., Cai, Y.D., 2005. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. 329, 1362).
Chou, K.C., Liu, W.M., Maggiora, G.M., Zhang, C.T., 1998. Prediction and classification of domain structural classes. Proteins 31, 97–130.
Chou, K.C., Maggiora, G.M., 1998. Domain structural class prediction. Protein Eng. 11, 523–538.
Chou, K.C., Shen, H.B., 2006a. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.
Chou, K.C., Shen, H.B., 2006b. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.
Chou, K.C., Shen, H.B., 2007a. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734.
Chou, K.C., Shen, H.B., 2007b. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem. Biophys. Res. Commun. 357, 633–640.
Chou, K.C., Shen, H.B., 2007c. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Commun. 360, 339–345.
Chou, K.C., Shen, H.B., 2007d. Review: recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Chou, K.C., Shen, H.B., 2008a. Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Protocols 3, 153–162.
Chou, K.C., Shen, H.B., 2008b. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem. Biophys. Res. Commun. 376, 324–325.
Chou, K.C., Zhang, C.T., 1994. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J. Biol. Chem. 269, 22014–22020.
Chou, K.C., Zhang, C.T., 1995. Predicting of protein structural class. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Deschavanne, P., Tufféry, P., 2008. Exploring an alignment free approach for protein classification and structural class prediction. Biochimie 90, 615–625.
Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.
Du, P., Li, Y., 2008. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. J. Theor. Biol. 253, 579–589.
Dubchak, I., Muchnik, I., Holbrook, S.R., Kim, S.H., 1995. Prediction of protein-folding class using global description of amino-acid sequence. Proc. Natl. Acad. Sci. 92, 8700–8704.
Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley, New York.
Eckmann, J.P., Kamphorst, S.O., Ruelle, D., 1987. Recurrence plots of dynamical systems. Europhys. Lett. 4, 973–977.
Eisenhaber, F., Frömmel, C., Argos, P., 1998. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. Proteins 25, 169–179.
Feng, K.Y., Cai, Y.D., Chou, K.C., 2005. Boosting classifier for predicting protein domain structural class. Biochem. Biophys. Res. Commun. 334, 213–217.
Fiser, A., Tusnády, G.E., Simon, I., 1994. Chaos game representation of protein structures. J. Mol. Graphics 12, 302–304.
Giuliani, A., Benigni, R., Zbilut, J.P., Webber Jr., C.L., Sirabella, P., Colosimo, A., 2002. Nonlinear signal analysis methods in the elucidation of protein sequence–structure relationships. Chem. Rev. 102, 1471–1491.
Giuliani, A., Sirabella, P., Benigni, R., Colosimo, A., 2000. Mapping protein sequence spaces by recurrence: a case study on chimeric structures. Protein Eng. 13, 671–678.
Giuliani, A., Tomasi, M., 2002. Recurrence quantification analysis reveals interaction partners in paramyxoviridae envelope glycoproteins. Proteins 46, 171–176.
Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.
Jiang, X., Wei, R., Zhang, T.L., Gu, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept. Lett. 15, 392–396.
Jin, Y., Niu, B., Feng, K.Y., Lu, W.C., Cai, Y.D., Li, G.Z., 2008. Predicting subcellular localization with AdaBoost learner. Protein Pept. Lett. 15, 286–289.
Kedarisetti, K.D., Kurgan, L.A., Dick, S., 2006. Classifier ensembles for protein structural class prediction with varying homology. Biochem. Biophys. Res. Commun. 348, 981–988.
Kurgan, L.A., Homaeian, L., 2006. Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognition 39, 2323–2343.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Li, F.M., Li, Q.Z., 2008. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept. Lett. 15, 612–616.
Li, H., Helling, R., Tang, C., Wingreen, N.S., 1996. Emergence of preferred structures in a simple model of protein folding. Science 273, 666–669.
Li, H., Tang, C., Wingreen, N.S., 1998. Are protein folds atypical? Proc. Natl. Acad. Sci. 95, 4987–4990.
Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.
Lin, H., Ding, H., Guo, F.B., Zhang, A.Y., Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept. Lett. 15, 739–744.
Manetti, C., Ceruso, M.A., Giuliani, A., Webber Jr., C.L., Zbilut, J.P., 1999. Recurrence quantification analysis as a tool for the characterization of molecular dynamics simulations. Phys. Rev. E 59, 992–998.
Marwan, N., Romano, M.C., Thiel, M., Kurths, J., 2007. Recurrence plots for the analysis of complex systems. Phys. Rep. 438, 237–329.
Munteanu, C.B., Gonzalez-Diaz, H., Magalhaes, A.L., 2008. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J. Theor. Biol. 254, 476–482.
Niu, B., Jin, Y.H., Feng, K.Y., Liu, L., Lu, W.C., Cai, Y.D., Li, G.Z., 2008. Predicting membrane protein types with bagging learner. Protein Pept. Lett. 15, 590–594.
Nishkawa, K., Ooi, T., 1982. Correlation of the amino acid composition of a protein to its structural and biological characters. J. Biochem. 91, 1821–1824.
Riley, M.A., Van Orden, G.C., 2005. Tutorials in contemporary nonlinear methods for the behavioral sciences. Retrieved March 1, 2005, from ⟨http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp⟩.
Shen, H.B., Chou, K.C., 2007a. Signal-3L: a 3-layer approach for predicting signal peptide. Biochem. Biophys. Res. Commun. 363, 297–303.
Shen, H.B., Chou, K.C., 2007b. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem. Biophys. Res. Commun. 364, 53–59.
Shen, H.B., Yang, J., Liu, X.J., Chou, K.C., 2005. Using supervised fuzzy clustering to predict protein structural classes. Biochem. Biophys. Res. Commun. 334, 577–581.
Shi, M.G., Huang, D.S., Li, X.L., 2008. A protein interaction network analysis for yeast integral membrane protein. Protein Pept. Lett. 15, 692–699.
Wang, B., Yu, Z.G., 2000. One way to characterize the compact structures of lattice protein model. J. Chem. Phys. 112, 6084–6088.
Wang, Z.X., Yuan, Z., 2000. How good is the prediction of protein structural class by the component-coupled method? Proteins 38, 165–175.
Webber Jr., C.L., Giuliani, A., Zbilut, J.P., Colosimo, A., 2001. Elucidating protein secondary structures using alpha-carbon recurrence quantifications. Proteins 3, 292–303.
Webber Jr., C.L., Zbilut, J.P., 1994. Dynamical assessment of physiological systems and states using recurrence plot strategies. J. Appl. Physiol. 76, 965–973.
Wu, G., Yan, S., 2008. Prediction of mutations in H3N2 hemagglutinins of influenza a virus from North America based on different datasets. Protein Pept. Lett. 15, 144–152.
Yang, J.Y., Yu, Z.G., Anh, V., 2007a. Correlations between designability and various structural characteristics of protein lattice models. J. Chem. Phys. 126, 195101.
Yang, J.Y., Yu, Z.G., Anh, V., 2007b. Clustering structures of large proteins using multifractal analyses based on a 6-letter model and hydrophobicity scale of amino acids. Chaos Solitons Fractals, in press, doi:10.1016/j.chaos.2007.08.014.
Yang, J.Y., Yu, Z.G., Anh, V., 2008. Protein structure classification based on chaos game representation and multifractal analysis. in: Proceedings of the 4nd International Conference on Natural Computation (ICNC2008), 18–20 October 2008, Jinan, China, vol. 4, pp. 665–669.
Yu, Z.G., Anh, V., Lau, K.S., 2004. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J. Theor. Biol. 226, 341–348.
Yu, Z.G., Anh, V., Lau, K.S., Zhou, L.Q., 2006. Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins. Phys. Rev. E 73, 031920.
Zaldívar, J.M., Strozzi, F., Dueri, S., Marinov, D., Zbilut, J.P., 2008. Characterization of regime shifts in environmental time series with recurrence quantification analysis. Ecol. Modelling 210, 58–70.
Zbilut, J.P., Mitchell, J.C., Giuliani, A., Colosimo, A., Marwan, N., Webber Jr., C.L., 2004. Singular hydrophobicity patterns and net charge: a mesoscopic principle for protein aggregation/folding. Physica A 343, 348–358.
Zbilut, J.P., Webber Jr., C.L., 1992. Embeddings and delays as derived from quantification of recurrence plots. Phys. Lett. A 171, 199–203.
Zhang, G.Y., Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J. Theor. Biol. 253, 310–315.
Zhang, T.L., Ding, Y.S., Chou, K.C., 2008. Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J. Theor. Biol. 250, 186–193.
Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.
Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins 44, 57–59.
Zhou, X.B., Chen, C., Li, Z.C., Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248, 546–551.
Zhou, Y., Yu, Z.G., Anh, V., 2007. Cluster protein structures using recurrence quantification analysis on coordinates of alpha-carbon atoms of proteins. Phys. Lett. A 368, 314–319.