American Journal of Biochemistry and Biotechnology

Original Research Paper

Low-Homology Protein Structural Class Prediction from Secondary Structure Based on Visibility and Horizontal Visibility Network

[1]Zhi-Qin Zhao, [2]Liang Luo and [1]Xiao-Yan Liu

[1]College of Science, Xi'an Shiyou University, Xi'an, 18 Second Dianzi Rd, 710065, Shaanxi, China
[2]Department of Mathematics, School of Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China

Article history
Received: 19-01-2018
Revised: 01-03-2018
Accepted: 10-03-2018

Corresponding Author:
Zhi-Qin Zhao
College of Science, Xi'an Shiyou University, Xi'an, 18 Second Dianzi Rd, 710065, Shaanxi, China

Email: zhiqin_zhao2014@163.com

Abstract: In this study, based on the predicted secondary structures of proteins, we propose a new approach to predict protein structural classes (α,β,α/β,$\alpha+\beta$) for three widely used low-homology data sets. Fist, we obtain two time siries from the chaos game representation of each predicted secondary structure; second, based on two time series, we construct visibility and horizontal visibility network, respectively and generate a set of features using 17 network features; finaly, we predict each protein structure class using support vector machine and Fisher's linear discriminant algorithm, respectively. In order to evaluate our method, the leave one out cross-validating test is employed on three data sets. Results show that our approach has been provided as a effective tool for the prediction of low-homology protein structural classes.

Keywords: Protein Structure Class, Secondary Structure, Chaos Game Representation, Visibility and Horizontal Visibility Network, Support Vector Machine, Fisher's Linear Discriminate

Introduction

The roles of proteins are varied and complex. Levitt and Chothia (1976) first propose the protein structural classes. In their pioneering work, four structural classes of protein, namely all-α, all-β, α/β and $\alpha+\beta$ can be obtained. The all-α and all-β classes represent structures that consist of mainly α-helices and β-strands, respectively. The α/β and $\alpha+\beta$ classes contain both α-helices and β-strands which are mainly interspersed and segregated, respectively (Murzin *et al.*, 1995).

A knowledge of protein structure class is very important in both theoretical and experimental studies in protein science. The information of structure classes has been employed to improve the prediction accuracy of the protein secondary structure (Gromiha and Selvaraj, 1998), to reduce the search space of various possible conformations of the tertiary structure (Carlacci *et al.*, 1991; Bahar *et al.*, 1997). However, for newly-found proteins, the structural class prediction method of automated and accurate are urgently needed. Therefore, the problem of protein structural class prediction is very important towards the protein structure prediction problem. Despite the significance of this problem, when the sequence similarity rate is low, finding the most precise computational method to solve this problem still remains an unsolved problem.

To predict the protein structural class, the current classification methods mainly focus on two aspects: Feature extraction and classification algorithms. The method of feature extraction contains several aspects. Such as physicochemical based information (Dehzangi *et al.*, 2013a; Sharma *et al.*, 2013), structural based information (Yang *et al.*, 2009; 2010; Zhang *et al.*, 2013; Liu and Jia, 2010; Zhang *et al.*, 2011; Ding *et al.*, 2012; Han *et al.*, 2014; Dehzangi *et al.*, 2013b; 2014; Wang *et al.*, 2014). Yu *et al*. (2017) use Chous pseudo amino acid composition and wavelet denoising to prediction structural class. From 2014 to now, several papers (Dehzangi *et al.*, 2014; Wang *et al.*, 2014; Jones, 1999; Faraggi *et al.*, 2012) show that the protein secondary structure is significanc to predict protein structural classes. Firstly the features are extracted, secondly all kinds of algorithms can be used to implement the classification prediction, such as Fisher's linear discriminant algorithm (Yang *et al.*, 2009), Support Vector Machine (SVM) (Cai *et al.*, 2003) and so on.

In this study, based on the predicted protein secondary structure, we attempt to predict the protein structural classes of the three low-homology data sets. Fist, we obtain two time siries from the chaos game representation of each predicted secondary structure, based on two time series, we generate a set of features using 17 network features of visibility or horizontal visibility network. The structure class for each protein is predicted with support vector machine and Fisher's linear discriminant algorithm, respectively. In order to evaluate our approach, the leave one out cross-validating test is employed on three data sets. The result shows that network features are valid features.

Materials and Methods

Data Sets

To evaluate our proposed approach, we employe three benchmarks with low sequence identity including 25PDB(the homology-range between 22 and 45%) (Yang *et al.*, 2009), 1189 (less than 40% sequence similarities) (Yang *et al.*, 2009) and 640 (with 25% sequence identity) (Yang *et al.*, 2010), respectively. The data sets in this study and the number of proteins belonging to four structural classes are shown in Table 1.

Secondary Structure Prediction

First, we can predict each amino acid in a protein sequence into one of the three secondary structural elements, C (coil), E (strand) and H (helix). For instance, the amino acid sequence of protein 1A1W as follows: MDPFLVLLHSVSSSLSSSELTELKYLCLGRVGKRKL ERVQSGLDLFSMLLEQNDLEPGHTELLRELLASLR RHDLLRRVDDFELEHHHHHH. In this study, if we submit this amino acid sequence to the web server of PSIPRED (http: //globin.bio.warwick.ac.uk/psipred. or http://bioinf.cs.ucl.ac.uk/psipred/) (Jones, 1999), the predicted secondary structure to be returned will be CCHHHHHHHHHHHHCCHHHHHHHHHHHHHHCC CHHHHHCCCHHHHHHHHHHCCCCCCCCHHHHH HHHHHHCHHHHHHHHHHHHHHHCCCCC.

Chaos Game Representation of Predicted Secondary

Fiser *et al*. (1994) firstly propose the concept of Chaos Game Representation (CGR) of protein structures. Yang and co-workers proposed CGR of predicted protein secondary structure sequence (Yang *et al.*, 2010) to predict protein structure class.

In this study, based on the method of Yang *et al*. (2010), the CGR of four proteins secondary structure sequence as shown in Fig. 1. The blue points represent the CGR points，the blue edge represents the sides of equilateral triangles. corresponding to the order in the predicted secondary structure, the order of the blue points is saved, but not shown in the figure. We can see that the plotted points tend to be distributed around the sides HC and EC, respectively, for proteins in the α and β classes. However, the points lie around both sides HC and EC without preference for proteins in the mixture classes.

Each secondary structure sequence generates a distinct (x, y)-coordinate sequence of the plotted points. Hence we model a CGR plot as two time series, one composed of the x-coordinates，namely x-time series and the other of the y-coordinates, namely y-time series, as shown in Fig. 2.

Recent research showed that the theory of complex network was an effective approach to analyze time series (Lacasa *et al.*, 2008; Luque *et al.*, 2009; Liu *et al.*, 2014). In this study, we hope to reveal some information in the above time series from the perspective of the visibility network (Lacasa *et al.*, 2008) and the horizontal visibility network (Luque *et al.*, 2009).

Visibility Network (VN): Let $\{x_i\}_{i=1,2,\cdots,N}$ be a time series of length N. We can obtain a visibility graph from the mapping of a time series of n data into a network of n nodes (where each datum is associated to a specific node and where temporal order is preserved in the node labelling) according to the following visibility criterion: Two arbitrary data (t_i, x_i) and (t_j, x_j) in the time series have visibility and consequently become two nodes in the associated graph, if any other data (t_n, x_n) such that $t_j < t_n < t_i$ fulfills (Lacasa *et al.*, 2008):

$$x_n < x_i + \left(x_j - x_i\right)\frac{t_n - t_i}{t_j - t_i}$$

Some basic properties of the mapping include undirectedness, connectedness (the visibility graph is always connected by definition) and invariance under affine transformations.

Horizontal Visibility Network (HVN): Let $\{x_i\}_{i=1,2,\cdots,N}$ be a time series of length N. The algorithm assigns each datum of the series to a node in the network. Two nodes i and j in the network are connected if one can draw a horizontal line in the time series joining x_i and x_j that does not intersect any intermediate data height. Hence, i and j are two connected nodes if the following geometrical criterion is fulfilled within the time series (Luque *et al.*, 2009):

$$x_i, x_j > x_n$$

For all n such that $i < n < j$. As a result, given each time series, its HVN is unweighed, undirected and connected (each node sees at least its nearest neighbors (left-hand side and right-hand side).

Network features: Here, we briefly introduce the considered features, namely network characteristics, that we extract from the visibility network and the horizontal visibility network. The network can be represented by graph.

Table 1: The number of proteins belonging to four structural classes in the datasets

Data set	all-α	all-β	α/β	$\alpha+\beta$	Total
25PDB	443	443	346	441	1673
1189	223	294	334	241	1092
640	138	154	177	171	640

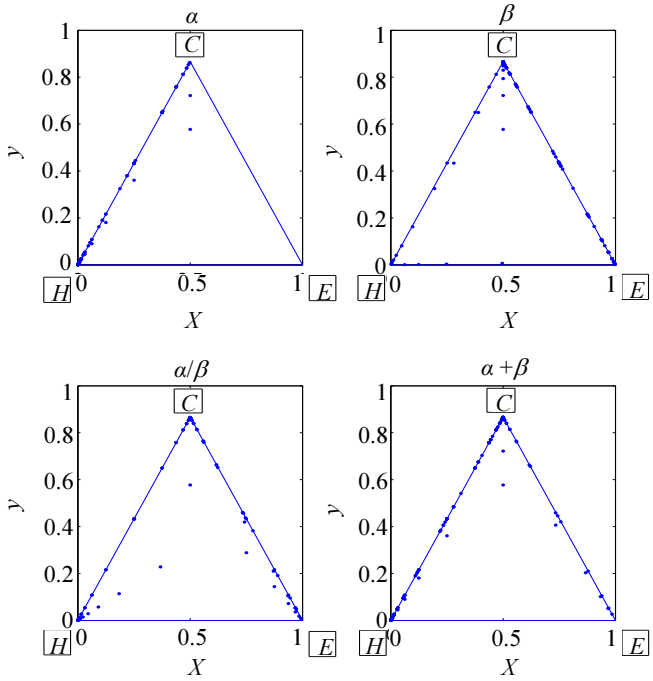

Fig. 1: The CGRs of predicted secondary structure for proteins from four different structural classes. The PDB IDs for four different proteins are 1A1W (α), 1A1X (β), 1ABA (α/β), and 169LA ($\alpha+\beta$) (Yang *et al*., 2010)

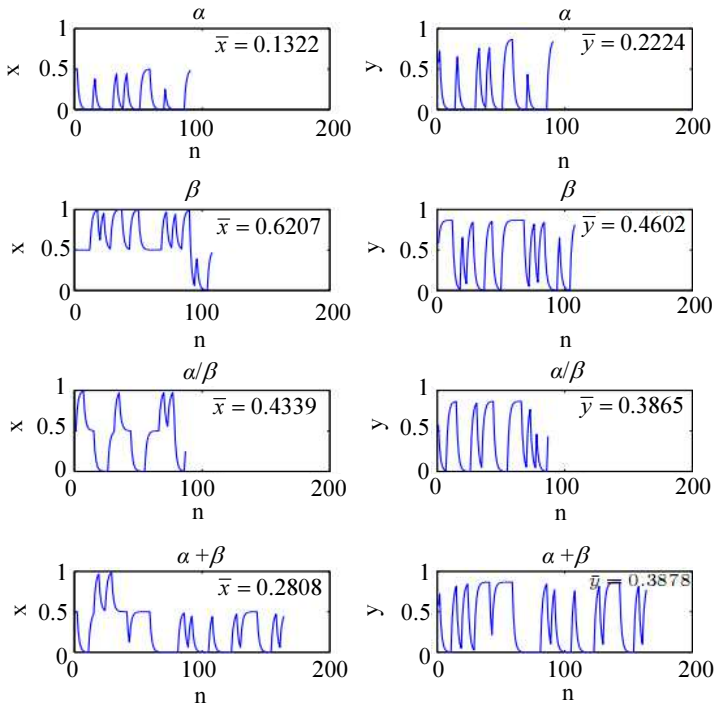

Fig. 2: Eight time series that represent the four CGRs in Fig. 1. Each panel in Fig. 1 gives rise to two time series (x- and y-coordinates, respectively). As a result, we obtain eight time series for four CGRs

Let $G = (V, E)$, $N = |V|$, $M = |E|$ be an unweighed and undirected graph, where N and M are the number of nodes and the number of edges, respectively. Let A be the adjacency matrix of the graph G.

The number of nodes (N) is an important feature of network.

Average degree $\left(\overline{K}\right)$: The degree of any vertex i is given by $K_i = \sum_{j=1}^{N} A_{ij}$. The average degree of the network can be written as (Chang *et al*., 2008) $\overline{K} = \frac{1}{N}\sum_i K_i$.

Characteristic path length (L): It is calculated as:

$$L = \frac{1}{N_P}\sum_{j>i} d_{ij} = \frac{1}{N(N-1)}\sum_{i=1}\sum_{j=i+1} d_{ij}$$

where, N_p represents the number of pairs of nodes of the network and d_{ij} is the shortest path (Floyd, 1962) between nodes i and j (Chang *et al*., 2008). The characteristic path length L is the average of the shortest path lengths.

Diameter (D): The diameter D is defined as the largest value of all the shortest path lengths in a network. Diameter is a measure of the compactness in a network and is computed by (Emerson and Gothandam, 2012): $D = Max\{d_{ij}\}$, $\forall$ *i-j* pairs of shortest paths.

Clustering coefficient of the network (C): The clustering coefficient of any node i is the ratio between the total number of links actually connecting its neighbors and the total number of all possible links between these neighbors. It is given by $C_i = \frac{e_i}{k_i(k_i-1)/2}$, where e_i is the actual number of edges between the neighbors of node j. The clustering coefficient of the network is the average of C_i overall nodes. It is calculated as (Chang *et al*., 2008): $C = \frac{1}{N}\sum_i C_i$.

Pearson correlation coefficient (r): To understand whether our unweighed undirected networks are of assortive or disassortive type, we calculate the Pearson correlation coefficient r of the degrees at either ends of an edge. For this, we use the expression suggested by Newman (2002):

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i 0.5(j_i + k_i)\right]^2}{M^{-1}\sum_i 0.5(j_i^2 + k_i^2) - \left[M^{-1}\sum_i 0.5(j_i + k_i)\right]^2}$$

Here, j_i and k_i are the degrees of the nodes at the two ends of the i-th edge, with $i = 1,2,\ldots,M$.

Average closeness centrality (ACC): Network centrality measures were developed by Freeman (1978; Beauchamp, 1965; Sabidussi, 1966). Basically "closeness centrality" of node i is calculated as:

$$CC(i) = (N-1)\Big/\sum_{j\in V, j\neq i} d_{ij}$$

The closeness value is therefore the inverse of the average distance between node i and the other nodes. The average closeness centrality is calculated as:

$$ACC = \sum_i CC(i)/N$$

Energy (E): The energy (Gutman and Zhou, 2006) of the graph is defined as $E = \sum_{i=1}^{n} |\lambda_i|$, where λ_i is the *ith* eigenvalue of the adjacency matrix A.

Laplacian Energy (LE): Let us define the Laplacian matrix as $L = D - A$, where D is a diagonal matrix containing the vertex degrees. The Laplacian energy, LE (Gutman and Zhou, 2006), is defined as:

$$LE = \sum_{i=1}^{n}\left|\mu_i - \frac{2m}{n}\right|$$

where, μ is the i-th eigenvalue of the Lappacian.

In this subsection, given a secondary structure sequence, we can convert a protein into two series: x time series and $y$time series. Each time series can construct corresponding visibility and horizontal visibility network, respectively. Nine network features can be obtained from a network. The features are the number of nodes (N), average degree (K), characteristic path length (L), network diameter (D), clustering coefficient of network (C), Pearson correlation coefficient (r), average closeness centrality (ACC), Energy (E) and Laplacian Energy (LE). Different time series for the same protein, under the same constructing of network, the number of nodes is the same. Hence we can obtain 1+8×2 = 17 features in total for each protein. So, each protein is described as a real-valued vector of 17 features.

Feature Space of Proteins

As mentioned above, In this study, suppose we use n features to represent a protein sample. Thus, the i-th protein sample P^i should be a real-valued vector in a n-D (dimensional) space, i.e.:

$$P^i = \left[p_1^i \quad p_2^i \quad \cdots \quad p_n^i\right]^T \tag{1}$$

Here p_j^i is the *j*-th (j = 1,2,…n) feature of the P^i and can be derived by following the setps.

Before prediction, each of the *n* features in Equation (1) should be normalized by:

$$p_j^i \Leftarrow \left(p_j^i - \mu_j\right)/\sigma_j \quad \left(i = 1,2,\cdots,m; j = 1,2,\cdots n\right) \tag{2}$$

where, m is the number of the total proteins in the data set, $\sigma_j = \sqrt{\sum_{i=1}^{m}\left(p_j^i - \mu_j\right)^2/(m-1)}$ and $\mu_j = \sum_{i=1}^{m} p_j^i/m$ are the standard and mean deviation of the *j*-th feature over the m protein samples. The normalized values obtained by Equation (2) will have a zero mean value over the m protein samples (Huang *et al.*, 2010).

Support Vector Machine

Vapnik (1995) introduced a machine learning method of Support Vector Machine (SVM). In our study, we choose Gaussian kernel function. The kernel width parameter γ and the regularization parameter c are optimized using a grid search strategy within a limited range, where $\gamma = 2^i$, i = –15,–14,–13,…,4,5 and $c = 2^i$, i = –5,–4,–3,…,14,15. We find the optimal SVM parameters c and using 10-folding cross validation on the training set for each turn in the leave-one-out cross validation process. The publicly available LIBSVM software (Chang and Lin, 2001) is used to implement the SVM classifier in our paper. The software toolbox can be freely downloaded from http://www.csie.ntu.edu.tw/cjlin/libsvm. Version 3.22 released on December 22, 2016.

Fisher's Discriminant Algorithm

Fisher's discriminant algorithm (Duda *et al.*, 2001) is used to find a classifier in the parameter space for a training set. A training set $H = \{x_1\ x_2\ \ldots\ x_n\}$ contains training vectors from two classes. There are n_1 training vectors from one class forming a subset H_1 and n_2 training vectors from another class forming a subset H_2. Hence, $H = H_1 \cup H_2$ and $n_1 + n_2 = n$. Suppose that each x_i is a m-dimension vector. Then, a parameter vector $\omega = (\omega_1\ \omega_2\ \ldots\ \omega_m)^T$ is estimated such that it allows as many training vectors as possible to be accurately predicted. Specifically:

$$m_j = \frac{1}{n_j}\sum_{x_i \in H_j} x_i, j = 1,2$$

$$S_j = \sum_{x_i \in H_j}\left(x_i - m_j\right)\left(x_i - m_j\right)^T, j = 1,2$$

$$S_\omega = S_1 + S_2$$

Then the parameter vector ω is estimated as $S_\omega^{-1}(m_1 - m_2)$ (Duda *et al.*, 2001). By Fisher's discriminant rule, x is assigned to the class of H_1 if $dist = \left(m_1 - m_2\right)^T S_\omega^{-1}\left[x - \frac{1}{2}\left(m_1 + m_2\right)\right] > 0$ and to the class of H_2 otherwise (Duda *et al.*, 2001).

The above algorithm is designed for a two-class problem. In this study, we transform a four-class problem of protein structural classes prediction into six two-class problems, namely, α-vs-β, α -vs-α/β, α-vs-$\alpha + \beta$, β-vs-α/β, β-vs-$\alpha + \beta$ and α/β-vs- $\alpha + \beta$ (Yang *et al*., 2010).

Performance Evaluation

The jackknife test (*leave-one-out* test) (Chou, 1995) is employed in our study.

The individual sensitivity S_n, the individual specificity S_p and the overall accuracy OA over the entire data set, as well as Matthew's correlation coefficient *MCC* (Xu *et al.*, 2013) are used to evaluate performance.

Results and Discussion

Prediction Performances of our Method

The prediction approach is examined with three benchmark data sets in low similarity by *leave-one-out* test and report the Sensitivity, Specificity and *MCC* for each structural class, as well as the OA.

By constructing of visibility network, a protein is described as a real-valued vector of 17 features. The results are shown in Table 2. From Table 2, we can see that the overall accuracies for the three data sets are close to or above 80%. Specifically, when SVM is used to implement the classification prediction, the overall accuracies of 82.07, 79.03 and 80% are achieved for the data sets 25PDB, 1189 and 640, respectively; when Fisher's linear discriminant algorithm is used to implement the classification prediction, the overall accuracies of 80.69, 79.40 and 80% are achieved for the data sets 25PDB, 1189 and 640, respectively. If comparing the four protein structural classes to each other, the predictions of proteins in the α classes are always the best (with accuracies higher than 90% for all the data sets.

By constructing of horizontal visibility network, a protein is described as a real-valued vector of 17 features. The results are shown in Table 3. From Table 3, we can see that the overall accuracies for the three data sets are close to or above 80%. Specifically, when SVM is used to implement the classification prediction, the overall accuracies of 82.85%, 79.21% and 81.25% are achieved for the data sets 25PDB,

1189 and 640, respectively; when Fisher's linear discriminant algorithm is used to implement the classification prediction, the overall accuracies of 82.19, 79.30 and 81.41% are achieved for the data sets 25PDB, 1189 and 640, respectively. If comparing the four protein structural classes to each other, the predictions of proteins in the α classes are always the best (with accuracies higher than 90% for all the data sets).

From Table 2 and 3, referring to the classes, our method also performs satisfactorily with prediction accuracies of about 80%. However, it seems very challenging to predict the α/β class and $\alpha + \beta$ class as their prediction accuracies are relatively low when compared with the other classes.

Comparison with Existing Methods

In this section, the proposed approach is further compared with other recently reported prediction approachs on the same three data sets. The results are shown in Table 4.

As can be seen from Table 4, our methods obtain the high prediction accuracies for all-α ,and all-β classes among all the tested methods. But our methods obtain the low prediction accuracies for α/β and $\alpha + \beta$ classes among all tested methods. But our method shows that network features are useful for prediction of protein structure class.

Table 2: 17 features (VN): The prediction quality of our method on the three data sets with SVM and Fisher algorithms

	SVM				Fisher		
Data set	Class	Sens	Spec	MCC	Sens	Spec	MCC
25PDB	all-α	0.9413	0.9579	0.8906	0.9300	0.9581	0.8828
	all-β	0.8352	0.9516	0.7988	0.8420	0.9358	0.7793
	α/β	0.7572	0.9455	0.7179	0.7688	0.9402	0.7171
	$\alpha+\beta$	0.7347	0.8800	0.6040	0.6780	0.8780	0.5545
	OA	0.8207			0.8069		
1189	all- α	0.9148	0.9565	0.8574	0.9148	0.9595	0.8628
	all-β	0.8673	0.9530	0.8274	0.8741	0.9472	0.8237
	α/β	0.8144	0.8704	0.6726	0.7455	0.9169	0.6788
	$\alpha+\beta$	0.5477	0.9002	0.4685	0.6515	0.8690	0.5044
	OA	0.7903			0.7940		
640	all- α	0.9493	0.9744	0.9173	0.9275	0.9821	0.9162
	all-β	0.7987	0.9534	0.7720	0.8312	0.9389	0.7716
	α/β	0.8531	0.8848	0.7164	0.8362	0.9055	0.7317
	$\alpha+\beta$	0.6257	0.8862	0.5241	0.6316	0.8745	0.5108
	OA	0.8000			0.8000		

Table 3: About 17 features (HVN): The prediction quality of our method on the three data sets with SVM and Fisher algorithms

	SVM				Fisher		
Data set	Class	Sens	Spec	MCC	Sens	Spec	MCC
25PDB	all-α	0.9549	0.9611	0.9055	0.9436	0.9608	0.8968
	all-β	0.8014	0.9600	0.7869	0.8397	0.9400	0.7834
	α/β	0.7890	0.9545	0.7595	0.8064	0.9456	0.7551
	$\alpha+\beta$	0.7596	0.8736	0.6145	0.6939	0.8901	0.5852
	OA	0.8285			0.8219		
1189	all- α	0.9058	0.9485	0.8360	0.9148	0.9484	0.8424
	all-β	0.8401	0.9611	0.8191	0.8673	0.9488	0.8207
	α/β	0.7964	0.8848	0.6762	0.7455	0.9182	0.6806
	$\alpha+\beta$	0.6224	0.8904	0.5151	0.6556	0.8741	0.5160
	OA	0.7921			0.7930		
640	all-α	0.9348	0.9631	0.8855	0.9203	0.9801	0.9070
	all-β	0.8052	0.9612	0.7904	0.8312	0.9493	0.7887
	α/β	0.8475	0.9113	0.7485	0.8701	0.9017	0.7531
	$\alpha+\beta$	0.6842	0.8838	0.5690	0.6550	0.8911	0.5556
	OA	0.8125			0.8141		

Table 4: Performance comparison of different methods on three data sets

		Prediction accuracy (%)				
Data set	Method	all-α	all-β	α/β	$\alpha+\beta$	OA
25PDB	Yang *et al*. (2009)	0.5800	0.6500	0.6990	0.6510	0.6420
	Yang *et al.* (2010)	0.9280	0.8330	0.8580	0.7010	0.8290
	Zhang *et al.* (2013)	0.9570	0.8080	0.8240	0.7550	0.8370
	Liu and Jia (2010)	0.9260	0.8130	0.8150	0.7600	0.8290
	Zhang *et al.* (2011)	0.9500	0.8560	0.8150	0.7320	0.8390
	Ding *et al*. (2012)	0.9503	0.8126	0.8324	0.7755	0.8434
	Han *et al.* (2014)	0.9460	0.8760	0.8410	0.7820	0.8630
	Dehzangi *et al.* (2014)	0.9680	0.9370	0.9010	0.8700	0.9220
	Wang *et al.* (2014)	0.9500	0.9140	0.7750	0.8870	0.8880
	This paper VN (SVM)	0.9413	0.8352	0.7572	0.7347	0.8207
	This paper VN (Fisher)	0.9300	0.8420	0.7688	0.6780	0.8069
	This paper HVN (SVM)	0.9549	0.8014	0.7890	0.7596	0.8285
	This paper HVN (Fisher)	0.9436	0.8397	0.8064	0.6939	0.8219
1189	Yang *et al.* (2009)	0.6050	0.6770	0.7100	0.6140	0.6580
	Yang *et al.* (2010)	0.8920	0.8670	0.8260	0.6560	0.8130
	Zhang *et al.* (2013)	0.9240	0.8440	0.8440	0.7340	0.8360
	Liu and Jia (2010)					
	Zhang *et al.* (2011)	0.9240	0.8740	0.8200	0.7100	0.8320
	Ding *et al*. (2012)	0.9372	0.8401	0.8353	0.6639	0.8196
	Han *et al.* (2014)	0.9100	0.8880	0.8740	0.6930	0.8450
	Dehzangi *et al.* (2014)	0.9820	0.9150	0.8380	0.7220	0.8630
	Wang *et al.* (2014)	0.9640	0.9290	0.8200	0.7840	0.8710
	This paper VN (SVM)	0.9148	0.8673	0.8144	0.5477	0.7903
	This paper VN (Fisher)	0.9148	0.8741	0.7455	0.6515	0.7940
	This paper HVN (SVM)	0.9058	0.8401	0.7964	0.6224	0.7921
	This paper HVN (Fisher)	0.9148	0.8673	0.7455	0.6556	0.7930
640	Yang *et al.* (2009)	-	-	-	-	-
	Yang *et al.* (2010)	0.8910	0.8510	0.8810	0.7140	0.8310
	Zhang *et al.* (2013)	-	-	-	-	-
	Liu and Jia (2010)	-	-	-	-	-
	Zhang *et al.* (2011)	-	-	-	-	-
	Ding *et al*. (2012)	0.9493	0.7662	0.8927	0.7427	0.8344
	Han *et al.* (2014)	-	-	-	-	-
	Dehzangi *et al.* (2014)	-	-	-	-	-
	Dehzangi *et al.* (2014)	0.9570	0.8960	0.8930	0.9010	0.9090
	This paper VN (SVM)	0.9493	0.7987	0.8531	0.6257	0.8000
	This paper VN (Fisher)	0.9275	0.8312	0.8362	0.6316	0.8000
	This paper HVN (SVM)	0.9348	0.8052	0.8475	0.6842	0.8125
	This paper HVN (Fisher)	0.9203	0.8312	0.8710	0.6550	0.8141

Conclusion

The problem of protein structural class prediction is still a challenge problem. Though some of approachs have shown the state-of-the-art performance, there is always room for improvement. In this study, we used matlab software to write programs. 17 network features are utilized to predict low-homology protein structural class. By comparisons with other existing approachs, we may attribute the high prediction accuracy. Three widely used data sets, 25PDB, 1189 and 640, with low sequence similarity, are adopted to evaluate the performance of our approach. Results by *leave-one-out* test show that our proposed method provides an effective tool for the accurate prediction of low-homology protein structural classes.

Acknowledgements

This work was Supported by National Natural Science Foundation of China (Grant No. 11626187), the Doctoral Scientific Research Foundation of Shaanxi Province (2016BS14), Foundation of Shaanxi Education Committee (No:16JK1603; No:16JK1708) and Shaanxi young science and technology nova (2017KJXX-60).

Author's Contributions

Zhi-Qin Zhao: Designed and developed the method, performed the numerical experiments, analyzed the data and wrote the paper.

Liang Luo: Performed the numerical experiments and revised the manuscript.

Xiao-Yan Liu: Revised the manuscript.

Ethics

The authors declare their responsibility for any ethical issues that may arise after the publication of this manuscript.

Conflict of Interest

The authors declare that they have no competing interests. The corresponding author affirms that all of the authors have read and approved the manuscript.

References

Bahar, I., A.R. Atilgan, R.L. Jernigan and B. Erman, 1997. Understanding the recognition of protein structural classes by amino acid composition. Proteins, 29: 172-185. DOI: 10.1002/(SICI)1097-0134(199710)29:2<172::AID-PROT5>3.0.CO;2-F

Beauchamp, M.A., 1965. An improved index of centrality. Behav. Sci., 10: 161-163. DOI: 10.1002/bs.3830100205

Cai, Y.D., X.J. Liu, X.B. Xu and K.C. Chou, 2003. Support vector machines for prediction of protein domain structural class. J. Theor. Biol., 221: 115-120. DOI: 10.1006/jtbi.2003.3179

Carlacci, L., K.C. Chou and G.M. Maggiora, 1991. A heuristic approach to predicting the tertiary structure of bovine somatotropin. Biochemistry, 30: 4389-4398. DOI: 10.1021/bi00232a004

Chang, C.C. and C.J. Lin, 2001. LIBSVM: A library for support vector machines.

Chang, S., X. Jiao, C. Li, X. Gong and W. Chen *et al*., 2008. Amino acid network and its scoring application in protein-protein docking. Biophy. Chem., 134: 111-118. DOI: 10.1016/j.bpc.2007.12.005

Chou, K.C., 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins, 21: 319-344. DOI: 10.1002/prot.340210406

Dehzangi, A., K. Paliwal, J. Lyons, A. Sharma and A. Sattar, 2014. Proposing a highly accurate protein structural class predictor using segmentation-based features. BMC Genomics 15: 1-13. DOI: 10.1186/1471-2164-15-S1-S2

Dehzangi, A., K.K. Paliwal, A. Sharma, O. Dehzangi and A. Sattar, 2013a. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. IEEE/ACM Trans. Comput. Biol. Bioinformat., 10: 564-575. DOI: 10.1109/TCBB.2013.65

Dehzangi, A., K.K. Paliwal, J. Lyons, A. Sharma and A. Sattar, 2013b. Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy. Proceedings of the 8th IAPR International Conference on Pattern Recognition in Bioinformatics, RIB'13, Springer, Berlin, Heidelberg, pp: 208-219. DOI: 10.1007/978-3-642-39159-0_19

Ding, S.Y., S.L. Zhang, Y. Li and T.M. Wang, 2012. A novel protein structural classes prediction method based on predicted secondary structure. Biochim., 94: 1166-1171. DOI: 10.1016/j.biochi.2012.01.022

Duda, R.O., P.E. Hart and D.G. Stork, 2001. Pattern Classification. 2nd Edon., John Wiley and Sons, New York, ISBN-10: 0471056693, pp: 654.

Emerson, A. and K.M. Gothandam, 2012. Network analysis of transmembrane protein structures. Phy. A: Statistical Mechan. Applic., 391: 905-916. DOI: 10.1016/j.physa.2011.08.065

Faraggi, E., T. Zhang, Y. Yang, L. Kurgan and Y. Zhou, 2012. Spine x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J. Comput. Chem., 33: 259-267. DOI: 10.1002/jcc.21968

Fiser, A., G.E. Tusnady and L. Simon, 1994. Chaos game representation of protein structures. J. Mol. Graph., 12: 302-304. DOI: 10.1016/0263-7855(94)80109-6

Floyd, R.W., 1962. Algorithm 97: Shortest path. Commun. ACM, 5: 345-345. DOI: 10.1145/367766.368168

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. Social Netw, 1: 215-239. DOI: 10.1016/0378-8733(78)90021-7

Gromiha, M. and S. Selvaraj, 1998. Protein secondary structure prediction in different structural classes. Protein Eng. Design Select., 11: 249-251. DOI: 10.1093/protein/11.4.249

Gutman, I. and B. Zhou, 2006. Laplacian energy of a graph. Linear Algebra Applic., 414: 29-37. DOI: 10.1016/j.laa.2005.09.008

Han, G.S., Z.G. Yu and V. Anh, 2014. Secondary structure element alignment kernel method for prediction of protein structural classes. Curr. Bioinformat., 9: 253-257. DOI: 10.2174/1574893609999140523124847

Huang, T., X.H. Shi, P. Wang, Z.S. He and K.Y. Feng *et al*., 2010. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One, 5: 1-9. DOI: 10.1371/journal.pone.0010972

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol., 292: 195-202. DOI: 10.1006/jmbi.1999.3091

Lacasa, L., B. Luque, F. Ballesteros, J. Luque and J. Carlos-Nuno, 2008. From time series to complex networks: The visibility graph. Proc. Natl. Acad. Sci. USA, 105: 4972-4975. DOI: 10.1073/pnas.0709247105

Levitt, M. and C. Chothia, 1976. Structural patterns in globular proteins. Nature, 261: 552-558. DOI: 10.1038/261552a0

Liu, J.L., Z.G. Yu and V. Ahh, 2014. Topological properties and fractal analysis of recurrence network constructed from fractional Brownian motions. Phys. Rev. E, 89: 1-25. DOI: 10.1103/PhysRevE.89.032814

Liu, T. and C.Z. Jia, 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol., 267: 272-275. DOI: 10.1016/j.jtbi.2010.09.007

Luque, B., L. Lacasa, F. Ballesteros and J. Luque, 2009. Horizontal visibility graphs: Exact results for random time series. Phys. Rev. E, 80: 1-17.

Murzin, A.G., S.E. Brenner, T. Hubbard and C. Chothia, 1995. Scop: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247: 536-540. DOI: 10.1016/S0022-2836(05)80134-2

Newman, M.E.J., 2002. Assortative mixing in networks. Phys. Rev. Lett., 89: 1-5. DOI: 10.1103/PhysRevLett.89.208701

Sabidussi, G., 1966. The centrality index of a graph. Psychometrika, 31: 581-603. DOI: 10.1007/BF02289527

Sharma, A., K.K. Paliwal, A. Dehzangi, J. Lyons and S. Imoto *et al*., 2013. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinformat., 14: 1-11. DOI: 10.1186/1471-2105-14-233

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. 1st Edn., Springer-Verlag New York, ISBN-10: 978-1-4757-2440-0, pp: 188.

Wang, J.R., Y. Li, X. Q. Liu, Q. Dai and Y.H. Yao *et al*., 2014. High-accuracy prediction of protein structural classes using PseAA structural properties and secondary structural patterns. Biochime, 101: 104-112. DOI: 10.1016/j.biochi.2013.12.021

Xu, Y., J. Ding, L.Y. Wu and K.C. Chou, 2013. iSNO-PseAAC: Predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. Plos One, 8: 1-7. DOI: 10.1371/journal.pone.0055844

Yang, J.Y., Z.L. Peng and X. Chen, 2010. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformat., 11: 1-9. DOI: 10.1186/1471-2105-11-S1-S9

Yang, J.Y., Z.L. Peng, Z.G. Yu, R.J. Zhang and V. Anh *et al*., 2009. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J Theor. Biol., 257: 618-626. DOI: 10.1016/j.jtbi.2008.12.027

Yu, B., L.F. Lou and S. Li, 2017. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J. Mol. Graph. Modell., 76: 260-273. DOI: 10.1016/j.jmgm.2017.07.012

Zhang, L.C., X.Q. Zhao and L. Kong, 2013. A protein structural class prediction method based on novel features. Biochimie, 95: 1741-1744. DOI: 10.1016/j.biochi.2013.05.017

Zhang, S., S. Ding and T. Wang, 2011. High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. Biochimie, 93: 710-714. DOI: 10.1016/j.biochi.2011.01.001