Journal of Molecular Graphics and Modelling 76 (2017) 260–273

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Topical Perspectives

Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising

Bin Yu[a,b,c,*,1], Lifeng Lou[a,1], Shan Li[a,c,1], Yusen Zhang[d], Wenying Qiu[a,c], Xue Wu[a,c], Minghui Wang[a,c], Baoguang Tian[a,c,*]

[a] *College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China*
[b] *CAS Key Laboratory of Geospace Environment, Department of Geophysics and Planetary Science, University of Science and Technology of China, Hefei 230026, China*
[c] *Bioinformatics and Systems Biology Research Center, Qingdao University of Science and Technology, Qingdao 266061, China*
[d] *School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China*

ARTICLE INFO

Article history:
Received 21 April 2017
Received in revised form 11 July 2017
Accepted 12 July 2017
Available online 14 July 2017

Keywords:
Protein structural class prediction
Pseudo amino acid composition
Two-dimensional wavelet denoising
Support vector machine

ABSTRACT

Prediction of protein structural class plays an important role in protein structure and function analysis, drug design and many other biological applications. Prediction of protein structural class for low-similarity sequences is still a challenging task. Based on the theory of wavelet denoising, this paper presents a novel method of prediction of protein structural class for the first time. Firstly, the features of the protein sequence are extracted by using Chou's pseudo amino acid composition (PseAAC). Then the extracted feature information is denoised by two-dimensional (2D) wavelet. Finally, the optimal feature vectors are input to support vector machine (SVM) classifier to predict protein structural classes. We obtained significant predictive results using jackknife test on three low-similarity protein structural class datasets 25PDB, 1189 and 640, and compared our method with previous methods The results indicate that the method proposed in this paper can effectively improve the prediction accuracy of protein structural class, which will be a reliable tool for prediction of protein structural class, especially for low-similarity sequences.

1. Introduction

Protein structural class plays a key role in protein function analysis, protein fold recognition, protein folding rate analysis and many other related fields. Using machine learning methods to predict the structure of protein can make up for shortcomings such as long cycle, high cost and difficult technology to adapt to the rapid expansion of protein sequence data in bioinformatics database. For any protein with unknown structure, if we can use the protein sequences information to obtain its structural class information, we can narrow the predicted protein conformation search range, and then accelerate the speed of structure identification. At present, protein structural class prediction is one of the most important research fields of bioinformatics.

The concept of the protein structural class was originally proposed by Levitt and Chothia [1] in 1976. They divided proteins into four major structural classes according to their different topological structures of the secondary structure fragment: all-α, all-β, α/β and $\alpha+\beta$ classes. The all-α and all-β proteins consist of only α-helices and β-strands, respectively. The α/β and $\alpha+\beta$ proteins contain both α-helices and β-strands, where the former consist of parallel β-sheets and the latter anti-parallel β-sheets. Subsequent studies on protein structure prediction mostly use this classification method.

During the past two decades, many feature extraction methods and classification algorithms have been proposed for protein structural class prediction. At present, the research on protein structural prediction mainly focuses on two aspects: One is the construction of feature vector. The feature vector, constructed from protein sequence or its derivatives, is used to extract protein spatial structure property. The other is choosing classification algorithms. There are a wide range of classification algorithms based on machine learning, which can be used to analyze prediction issues of protein structural class.

* Corresponding authors at: College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China.
E-mail addresses: yubin@qust.edu.cn (B. Yu), tianbaoguangqd@163.com (B. Tian).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jmgm.2017.07.012
1093-3263/

In the process of protein structure class prediction, protein feature vector construction plays a very important role. In previous studies, the extracting of amino acid composition from protein sequence to predict protein types is a widely used classical method [2–5]. This method uses a simple 20-dimensional vector to calculate the protein sequence which contains the percentage of 20 common amino acid content characteristics. Amino acid composition feature extraction methods consider a similar structure type for similar protein sequences, but fail to list the order of amino acids in the protein primary structure, which reduce prediction accuracy. In order to improve the prediction accuracy of protein structural class, Chou's pseudo amino acid composition (PseAAC) [6–9] and polypeptide composition [10,11] have been proposed, which can reflect the sequence-order information of amino acids in protein primary structure. The first 20 features of the feature vectors are the percentage of the 20 amino acids commonly found in protein sequences. The rest of the features are used to reflect the sequence-order information.

This method performs better on datasets with high-similarity sequences than low-similarity sequences. As low-similarity sequences may have very similar local spatial folding patterns, many methods, based on secondary structure information, have been proposed to improve prediction accuracy. Kurgan et al. [12] proposed the SCPRED method, which used protein secondary structure information for the first time to represent proteins with linear secondary structural sequences. In the SCPRED method, the author designed more than 2300 features and got 9 features through the feature selection algorithm, one feature is based on collocation of Leucine and Glycine in the amino acid sequence, and the other 8 features are extracted based on protein secondary structure with PSIPRED. SCPRED improved highly prediction accuracy of low-similarity datasets, reaching about 80%. SCPRED improved highly prediction accuracy of low-similarity datasets, reaching about 80% [13]. Yang et al. [14] proposed the RKS-PPSC method to predict 24 features based on protein secondary structure fragment information, recurrence quantification analysis and k-string information entropy. The accuracy of prediction can reach about 82%. Liu et al. [15] found that the α-helices and β-strands in α/β proteins alternate more frequently than in α + β proteins, and count their alternating frequency as well as the content of parallel β-sheets and anti-parallel β-sheets. Ding et al. [16] removed the random coil in the predicted secondary structure sequence, resulting in a new E-H sequence, and extracted a series of secondary structural features based on the E-H sequence, to better reflects the permutation distribution of α-helices and β-strands in the secondary structure sequence. Zhang et al. [17] proposed a method, which incorporates alternating word frequency and normalized Lempel-Ziv complexity, and obtains good prediction results. Wang et al. [18] extracted feature information based on PseAA structural properties and secondary structure patterns, using support vector machine in the protein structural class datasets. Prediction accuracy was more than 87%. The feature extraction method based on structure-driven was proposed by Kong et al. [19], and prediction accuracy reached above 86% using SVM. Zhang et al. [20] proposed a chaos game representation based on the predicted secondary structure and applied it to the protein structural class datasets 25PDB, 640 and 1189. The prediction accuracy rates of the three datasets are all more than 88%. Olyaee et al. [21] proposed predicting protein structure method based on complex networks and relapse analysis. 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. The prediction accuracy of the 25PDB and 1189 datasets were 90.08% and 86.02%, respectively. Marrero-Ponce et al. [22] proposed 3D protein descriptors based on the bilinear algebraic form in the R^n space on the coulombic matrix to obtain the characteristic information of the protein sequence. They used linear discriminant analysis (LDA) algorithm for classification, and achieved a good prediction effect. Hayat et al. [23] proposed a promising model employing hybrid descriptor space in conjunction with optimized evidence-theoretic K-nearest neighbor algorithm. Using the particle swarm optimization technique, high discriminative descriptors were selected from the mixed space. Using jackknife test, the average accuracy of the three low similarity benchmark datasets (including 25PDB, 1189 and 640) were 87.0%, 86.6% and 88.4%, respectively. Sahu et al. [24] introduced a feature representation method based on Chou's pseudo amino acid composition, which was composed of the amino acid composition information, the amphiphilic correlation factors and the spectral characteristics of the protein. Thus the sample of a protein was represented by incorporating both the sequence order and the length effect. They used radial basis function network based classifier to predict protein structural class. Hayat et al. [25] used pseudo average chemical shift to extract the protein sequences. Using SVM as a classifier, the overall prediction accuracy of the 25PDB dataset, 1189 dataset and the 640 dataset were 90.0%, 90.0% and 85.8%, respectively. Dehzangi et al. [26] proposed segment distribution and segment auto covariance feature extraction methods to capture local and global discriminatory information from evolutionary profiles and predicted secondary structure of the proteins. Using SVM as a classifier, the overall prediction accuracy of the 25PDB and 1189 datasets were 92.2% and 86.3%.

The evolutionary knowledge of protein sequences derived from the primary structure contains more. Many researchers proposed several protein structural class prediction models based on evolutionary information [27–39]. Liu et al. [27] fused the evolutionary information and amino acid sequence information in PSSM to obtain high prediction accuracy on low-similarity protein datasets. In the two low-similarity protein datasets 25PDB, 1189, the prediction accuracy was more than 70%. AAC-PSSM-AC method was proposed by Liu et al. [28]. This method combined PSSM with autocovarince transformation to extract features, and the prediction accuracy reached about 74% in both datasets 25PDB and 1189. Chen et al. [29] proposed the SCEC method. This method extracted amino acid pairs, and achieved 61%–96% accuracy. Zhang et al. [30] proposed the AATP method, which extracted the feature information based on the transition probability matrix of PSSM, and obtained the prediction accuracy of 70% or more on the datasets. Liang et al. [31] proposed a feature extraction method. Firstly, the 3600-dimensional feature vector is extracted by correlation analysis on the PSSM. Then the NMF algorithm is used to reduce the dimension to 200-dimensional. Finally, the SVM predicts two datasets 25PDB and 1189, and the overall prediction accuracy of the two datasets was more than 75%. Xia et al. [32] extracted feature information using PSSM and the square root of the occurrences of 20 amino residues in the protein sequence, and achieved overall accuracy of 83.1%. Ding et al. [33] first used the physicochemical properties of amino acids to obtain simplified PSSM, and then extended the protein sequence representation of amino acid composition and dipeptide composition to PSSM. Ding et al. [34] further extracted the long-range correlation information and linear correlation information of the protein sequence from the PSSM to obtain the feature vectors, and obtained higher prediction accuracy. Wang et al. [35] proposed a feature extraction method based on fusion of simplified PSSM and position-based secondary structure features for low-similarity sequences. The prediction precision obtained by this method has been improved by 3%–5% compared with other methods. Li et al. [36] proposed a PSSP-RFE method, which integrates PSSM, GO and PROFEAT for feature extraction to obtain 1080-dimensional vector, then, the author used SVM-RFE method and SVM as classifier to reduce dimensionality and got higher prediction accuracy. Zhang [37] proposed the PSSS-PsePSSM method, which selected 111-dimensional features. While 100-dimensional features, selected to describe evolutionary and sequence-order information, are based on PsePSSM, other 11-dimensional features are based on predicted protein secondary structure sequences (PSSS). The SVM was used to obtain the high prediction accuracy, and the overall prediction accuracy of the datasets 1189 and 25PDB was above 86%. Tao et al. [38] proposed a feature extraction technique which was based on tri-grams computed directly from position-specifc scoring matrix (PSSM). Then support vector machine-recursive feature elimination (SVM-RFE) was feature selection, and reduced features were input SVM classifier to predict the structure of a given protein. The results showed that their approach had achieved good results. Nanni et al. [39] combined position-specific scoring matrix (PSSM), amino acid sequence and secondary structure sequence to extract the protein sequences. Using SVM as classifier, the overall prediction accuracy of FC699, 1189, 640, 25PDB datasets were 94.5%, 91.1%, 91.4% and 83.0%, respectively.

Another important factor in protein structural class prediction is the classification algorithm. Successful classification algorithms should be able to separate different types of proteins efficiently and correctly. In recent years, machine learning methods have been widely used in prediction algorithms. Commonly used algorithms include neural network [40], fuzzy clustering [41], Bayesian classification [42], rough sets [43], support vector machine (SVM) [35–37] and so on. Neural network has a strong nonlinear fitting ability, which can map any complex nonlinear relationship, and its learning rules are easy for computer implementation. It has strong robustness, memory ability, nonlinear mapping ability and strong self-learning ability. However, the most serious problem of neural network is the inability to explain their reasoning process and reasoning basis, and if the data is not sufficient, the neural network cannot work. Changing all the features of the problem into numbers and changing all the reasoning into a numerical calculation will eventually lead to information loss. The advantage of fuzzy clustering is intuitive, and its conclusion is concise. However, if the sample size is large, it is difficult to obtain the clustering results. The Bayesian classifier has the lowest error rate compared with other classification methods. However, this is not always the case, because the Bayesian classifier assumes that the attributes are independent of each other. This assumption is often not true in practical applications, which has a certain impact on the correct classification of Bayesian classifier. Rough sets can get a good classification result in the case of less attribute set, and reduce the storage cost. But if the attribute set is relatively large, it will get a lower prediction accuracy rate. Support vector machine (SVM) is widely used. SVM can solve the problem of machine learning in the case of small sample, improve the generalization performance, solve the high-dimensional problem and nonlinear problem, and avoid neural network structure selection and local minimum point problem. However, SVM is particularly sensitive for missing data. There is no general solution for the nonlinear problem, and the kernel function must be carefully chosen for processing. Although each classification algorithm has its own advantages and disadvantages, it has been found in recent years that SVM has been widely used because of its excellent predictive performance in predicting protein structural class [38,44], protein–protein interaction [45–47], G protein-coupled receptors [48–50] and protein subcellular localization [51–55] and other bioinformatics field.

In this paper, we propose a novel protein structural class prediction method. Firstly, the features are extracted from protein sequences by Chou's PseAAC, and the feature vectors are denoised by two-dimensional (2D) wavelet. The features of each class of proteins are more prominent after wavelet denoising. Finally, the optimal feature vectors after wavelet denoising are input to the SVM classifier for prediction. In order to evaluate the proposed method, we selected the different λ values, wavelet functions, different wavelet decomposition scales and kernel functions on the results with the jackknife test on three widely used low-similarity benchmark datasets: 25PDB, 1189 and 640. Through the comparative analysis, the optimal parameters of the prediction model are determined. Our method obtained overall accuracies of 93.1%, 90.8% and 89.5% on 25PDB, 1189 and 640 datasets, respectively. According to the comparision with other existing methods, the experimental results show that our method can remarkably improve the prediction accuracy of protein structural class, particularly for low-similarity sequences.

Table 1
The compositions of three datasets adopted in this study.

Dataset	Number of proteins				
	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

2. Materials and methods

2.1. Datasets

In order to facilitate the comparison with the previous works, three widely used low-similarity benchmark datasets are adopted in our work: 25PDB dataset [56], 1189 dataset [42], 640 dataset [29], which include 1673, 1092 and 640 protein domains with sequence similarity lower than 25%, 40%, and 25%, respectively. The detailed information about the three datasets are shown in Table 1.

2.2. Feature extraction

In this study, Chou's pseudo amino acid (PseAA) composition method is used to extract feature vectors in protein sequences [6]. PseAA composition method is the protein primary structure coding method. The descriptor uses a 20 + λ dimensional vector to represent the protein sequence. The former 20 dimensions are the amino acid composition, and the posterior λ dimensions are the sequence correlation factor reflecting different levels of amino acid sequence information. According to the PseAA composition discrete model, the protein P can be represented as

$$P = [p_1, p_2, p_3, \cdots, p_{20}, p_{20+1}, \cdots p_{20+\lambda}]^T \tag{2.1}$$

where the (20 + λ) components are given by

$$p_u = \begin{cases} \dfrac{f_\mu}{\sum\limits_{\mu=1}^{20} f_\mu + \omega \sum\limits_{k=1}^{\lambda} \tau_k}, 1 \le \mu \le 20 \\[2ex] \dfrac{\omega \tau_{\mu-20}}{\sum\limits_{\mu=1}^{20} f_\mu + \omega \sum\limits_{k=1}^{\lambda} \tau_k}, 21 \le \mu \le 20 + \lambda \end{cases} \tag{2.2}$$

where P is the feature vector, ω is the weight factor, which was set at 0.05 in Ref. [6]. f_μ is the occurrence frequency of 20 amino acids in the protein P. τ_k is the k-tier sequence correlation factor, which can be calculated by the sequence correlation function

$$\tau_k = \frac{1}{L-k} \sum_{i=1}^{L-k} J_{i,i+k} (k < L) \tag{2.3}$$

$$J_{i,i+k} = \frac{1}{3} \left\{ [H_1(R_i) - H_1(R_{i+k})]^2 + [H_2((R_i) - H_2(R_{i+k}))]^2 + [M(R_i) - M(R_{i+k})]^2 \right\} \tag{2.4}$$

where $H_1(R_i)$, $H_2(R_i)$, $M(R_i)$ are the hydrophobicity value, hydrophilicity value and side chain mass of the amino acid residue R_i, respectively. The choice of parameter λ for the pseudo amino acid composition should be special attention, and the maximum value of λ must be less than the length of the shortest sequence in all protein structural class datasets. The length of the shortest sequence in 25PDB dataset is 13, in 1189 dataset is 10, and in 640 dataset is 37, so correlation factor λ is in the range of 1–9. PseAA composition has excellent properties, so it has attracted the attention of researchers, and developed a series of different forms of PseAA composition to predict various proteins [57–62].

2.3. *Wavelet denoising*

Protein structural class prediction information becomes worse because of a lot of noise in the process of generation and transmission, so purposefully obtaining useful information from the obtained feature vector, that is, to reduce noise, is an important step for prediction. Wavelet analysis is a combination of functional analysis, Fourier analysis and harmonic analysis of the product. The wavelet analysis for signal processing has a good time-frequency localization ability in the field of bioinformatics [63–66]. Wavelet denoising (WD) is one of the most important applications of wavelet analysis. For a large class of signals, their energy is highly concentrated in a small amount of wavelet coefficients or wavelet packet coefficients, and due to its randomness of the noise, the energy is more evenly distributed in the wavelet transform domain. So threshold method is used to suppress the noise to the maximum when preserving the signal in the transform domain [67,68]. In order to eliminate the noise genes effectively and select more accurate classification informative genes, we used the multi-scale wavelet threshold method to denoise the data in the gene expression profiles, and got a satisfying tumor classification accuracy [69]. The decomposition of protein amino acids signal by wavelet analysis can effectively remove redundant information and extract characteristic signals of each kind of proteins, which is very important to ensure the high accuracy of prediction [70].

2.3.1. *One-dimensional wavelet denoising*

A one-dimensional (1-D) model with noise [67] can be expressed as:

$$s(i) = f(i) + \sigma e(i) \quad i = 1, 2, \cdots n-1 \tag{2.5}$$

where $f(i)$ is the real signals, $e(i)$ is the noise, $s(i)$ is the signals with noise, and σ is the noise intensity. The purpose of wavelet denoising is to suppress noise $e(i)$ and recover real signals $f(i)$.

In general, 1-D wavelet denoising process can be divided into three steps:

(1) Wavelet or wavelet packet decomposition. Select a wavelet and determine the level of decomposition N, and then decompose the wavelet coefficients or wavelet packet coefficients of the calculated signal $s(i)$ at the N layer.
(2) Threshold quantization of the high-frequency coefficients based on wavelet decomposition. The appropriate thresholds are selected to quantify the high-frequency coefficients at each decomposition scale.
(3) Reconstruct the 1-D signals. For the 1-D wavelet, the signals are reconstructed on the basis of the approximate wavelet coefficients of the N layer and the detail wavelet coefficients from 1 to the N layer after the threshold quantization. For the wavelet packet, the information is reconstructed using the wavelet packet coefficients of each node in the N layer.

Wavelet or wavelet packet denoising has two key points: how to choose the threshold and how to use the threshold quantization wavelet coefficients or wavelet packet coefficients. Wavelet or wavelet packet denoising uses the threshold to quantize the wavelet coefficients, i.e., the absolute values of the wavelet coefficients are compared with the threshold values. The wavelet coefficients or wavelet packet coefficients less than or equal to the threshold are set to zero, while the wavelet coefficients or wavelet packet coefficients larger than the threshold or wavelet packet coefficient become the difference between the two. The default thresholds, the thresholds obtained by Birge-Massart strategy, and the thresholds obtained by pealty strategy are used to select the 1-D wavelet denoising thresholds. In the course of the experiment, this paper aims at improving the overall accuracy of the selection of appropriate methods.

2.3.2. *Two-dimensional wavelet denoising*

A two-dimensional (2-D) model with noise [68,71] can be expressed as:

$$s(i,j) = f(i,j) + \sigma e(i,j) \quad i,j = 1, 2, 3, \cdots m-1 \tag{2.6}$$

where (i, j) is the size of the signal, $f(i, j)$ is the real signals, $e(i, j)$ is the noise, $s(i, j)$ is the signals with noise, and σ is the noise intensity. The purpose of wavelet denoising is to suppress noise $e(i, j)$ and recover real signals $f(i, j)$.

2-D wavelet denoising's steps are as follows:

(1) Wavelet decomposition of 2-D signal. Under the 2-D wavelet transform, decompose the 2-D data to four scale spatial frequency band signal, that is, low-frequency scale space frequency band signals f^1, horizontal scale spatial frequency band signals d^H, vertical scale spatial frequency band signals d^V, the signals on the diagonal scale spatial frequency band d^D. It is the orthogonal decomposition process on the different scales, that is, $f = f^1 \oplus d^H \oplus d^V \oplus d^D$. The low-frequency signals f^1 can also continue to be decomposed into four directions: the signals of protein structural class on the low-frequency scale space f^2, the signals on the horizontal scale spatial band $d^{1,H}$, the signals on the vertical scale spatial frequency band $d^{1,V}$, and the signals on the diagonal scale spatial frequency band $d^{1,D}$. The decomposition process is still orthogonal which is $f^1 = f^2 \oplus d^{1,H} \oplus d^{1,V} \oplus d^{1,D}$. The signal energy of the protein structural class is limited, so the number of decomposition of the 2-D wavelet is also limited. The process of the n wavelet decomposition is $f^{n-1} = f^n \oplus d^{n-1,H} \oplus d^{n-1,V} \oplus d^{n-1,D}$.
(2) Thresholding quantization of the decomposed high-frequency coefficients. Threshold quantization is performed by selecting the appropriate thresholds for the high-frequency coefficients from 1 to N horizontal, vertical and diagonal directions of each layer.
(3) Reconstruct the 2-D signals. The 2-D signals are reconstructed on the basis of the low-frequency coefficients of the wavelet decomposition and the high-frequency coefficients after the threshold quantization.

2-D signals commonly used threshold models: default threshold determination model, and threshold model determined by Birge-Massart strategy. We can choose the wavelet functions or wavelet packets to confirm the default threshold. In the experiment, this study aims at select appropriate methods to improve the overall accuracy.

2.4. *Support vector machine*

SVM is a supervised machine learning algorithm based on statistical learning theory proposed by Vapnik et al. [72]. SVM usually has outstanding classification performance, and it has been widely applied in prediction of protein structural classes [25,73] and other bioinformatics fields. SVM is used to solve convex optimization problems. Compared with other machine learning algorithms, it can solve the problems with small sample, high dimension, nonlinearity and local minimal value. The basic idea of SVM is to transform the space of the input sample into a high-dimensional Hilbert space by nonlinear transformation, and obtain the optimal linear separating hyperplane in this space. The nonlinear transformation is defined by the appropriate kernel function to achieve. The decision function is given by:

$$f(x) = sign\left\{\sum_{i=1}^{n} y_i\alpha_i K\left(x_i, x_j\right) + b^*\right\} \tag{2.7}$$

where α_i is Lagrange multiplier, b^* is classification threshold value, and $K(x_i, x_j)$ is the kernel function. The selection of kernel function is very important in the process of training the support vector. Generally, four kinds of kernel functions, including linear kernel function, polynomial kernel function, radial basis function (RBF), and sigmoid kernel function, can be available to perform prediction.

As protein structural class prediction is a multi-class classification problem, SVM usually adopts One-Versus-One (OVO) or One-Versus-Rest (OVR) strategy to solve multi-class problems [53,66]. OVR strategy is the earliest method for SVM to apply to multi-class problem. In this strategy, each class of samples is used as a positive sample in proper order, and the remaining samples as negative samples to construct a classifier for classification. This method is simple, effective and efficient. If the dataset is large, the sample size may be much less than others, leading to the low prediction accuracy. Also, this method may result in a false positive. OVO strategy uses any two types of samples in datasets to construct classifiers. Considering the existence of false positives in OVR strategy, this paper uses OVO strategy to predict protein structural classes. This study used LIBSVM software developed by Chang and Lin [74], which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.5. Performance evaluation and model building

In statistical theory, the following three methods are often used to evaluate the effectiveness of the models [36,37]: independent dataset test, jackknife test, and self-consistency test. In this study, jackknife test is used to evaluate the performance of the prediction model. In addition, we used precision ratio (P), sensitivity (Sens), specificity (Spec), overall accuracy (OA), F-measure and Matthew's correlation coefficient (MCC) to evaluate the prediction model. These measures are defined as follows:

$$P_j = \frac{TP_j}{|C_j|} \tag{2.8}$$

$$Sens_j = \frac{TP_j}{TP_j + FN_j} = \frac{TP_j}{|C_j|} \tag{2.9}$$

$$Spec_j = \frac{TN_j}{FP_j + TN_j} = \frac{TN_j}{\sum\limits_{k\neq j}|C_k|} \tag{2.10}$$

$$OA = \frac{\sum\limits_{j} TP_j}{\sum\limits_{j}|C_j|} \tag{2.11}$$

$$F - measure = \frac{2 \times P_j \times Sens_j}{P_j + Sens_j} \tag{2.12}$$

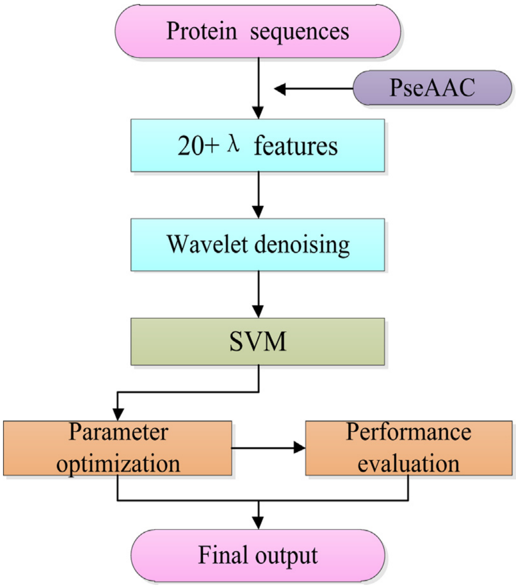

Fig. 1. Flow chart of protein structural class prediction based on WD-PseAAC method.

$$MCC_j = \frac{TP_j \times TN_j - FP_j \times FN_j}{\sqrt{\left(TP_j + FP_j\right)\left(TP_j + FN_j\right)\left(TN_j + FP_j\right)\left(TN_j + FN_j\right)}} \tag{2.13}$$

where TP_j represents the number of true positives, FP_j represents the number of false positives, TN_j represents the number of true negatives, FN_j represents the number of false negatives, and $|C_j|$ represents the number of proteins in each structural class C_j (all-α, all-β, α/β and $\alpha+\beta$ classes). F-value measure is the harmonic mean of P and Sens. The higher are P and Sens values, the better. But in fact these two are in conflicts under certain circumstances. F-value measure synthesizes the results of P and Sens. The higher is F-value measure, the better is model performance.

In addition, The ROC curve can test the robustness of the model. The area under the ROC curve (AUC) is used as a quantitative indicator of the robustness of the model. AUC is the area calculated under receiver operating characteristic (ROC) curve plotted by FP rate vs TP rate. Its values range from 0 to 1.

For convenience, the method proposed in this paper is called WD-PseAAC, and the general framework of our method is shown in Fig. 1. We have implemented it in MATLAB R2014a in windows 10 running on a PC with system configuration Intel (R) Core (TM) i7-4510U CPU @ 2.00 GHz 2.60 GHz with 8.00 GB of RAM.

The steps of the WD-PseAAC prediction method are described as follows:

(1) Input protein sequences of the benchmark datasets, and the class label corresponding to 4 kinds of proteins;
(2) Using PseAAC online service system developed by Chou et al. [8] to carry out feaure extraction of the protein sequences in the datasets and the protein sequences transform into numerical sequences, and the 20 + λ dimensional vector is obtained, where 1 ≤ λ ≤ 9;
(3) Wavelet denoising is performed for each feature vector extracted in (2), and the redundant information in the sequence is eliminated to obtain the vector after wavelet denoising, and the dimensionailty does not change
(4) The optimal feature vectors of noise reduction are input to the SVM classifier for protein structural class prediction;
(5) According to the accuracy prediction, the optimal parameters are determined. The parameters include the λ values,

Table 2
Prediction results of protein structural class by selecting different values of λ.

Dataset	λ								
	Jackknife test (%)								
	1	2	3	4	5	6	7	8	9
25PDB	80.04	84.28	84.40	85.77	87.99	89.60	88.64	88.94	88.64
1189	80.77	82.78	83.24	84.07	85.81	88.37	86.81	86.72	85.71
640	78.28	78.75	81.87	80.31	84.53	88.28	86.25	86.72	86.88

wavelet basis function, wavelet decomposition scales, threshold method and kernel function in SVM;

(6) According to the optimal parameters of the model got in (5), the performance of the model is evaluated using the given evaluation indexes P, Sens, Spec, OA, F-measure and MCC.

3. Results and discussion

3.1. Selection of optimal parameter λ

In this paper, we used PseAAC to extract features in protein amino acid sequences. In feature extraction process, the selection of λ value plays an important role. If the λ value is set too small, the sequence information contained in the feature vectors will be very small, and the features of the protein sequences cannot be extracted completely. If the λ value is set too large, it will bring more redundant information, which may have negative effect on prediction. And when the value λ exceeds the length of a protein sequence in the dataset, it will not make sense to use PseAAC to feature extraction of protein sequences. Since the length of the shortest protein sequence in the three benchmark protein structural class datasets is 10, the maximum λ value for this paper is 9. To find the optimal λ value in the model, we set λ values from 1 to 9 in turn. The results are tested by jackknife cross-validation method, and the sym8 wavelet basis function is selected. The wavelet decomposition scale is 5, the threshold is obtained by 2-D wavelet denoising method, and RBF kernel function is used in SVM. The overall prediction accuracy of the protein structural class obtained in the datasets 25PDB, 1189 and 640 are shown in Table 2.

As shown from Table 2, when other conditions in WD-PseAAC model are certain, we can get great results by changing λ value. With the constant change of λ value, the accuracy of protein structural class prediction is also constantly changing. The highest predictive accuracy of 89.60%, 88.37%, and 88.28% was achieved respectively when λ = 6 on three datasets 25PDB, 1189 and 640. Therefore, this paper selects λ = 6 as the optimal parameter of the model.

3.2. Selection of wavelet function and optimal decomposition scale

When using PseAAC to extract features in protein sequence, λ = 6 is chosen and 2-D wavelet denoising is used to obtain thresholds. RBF kernel function is used to classify in SVM and jackknife cross-validation are used to test the results. Different wavelet basis functions and wavelet decomposition scales are used to classify the protein structural. Different wavelet basis functions produce different wavelet families, each family with different signal processing capabilities, and the results are different [60,66]. Wavelet functions have many excellent properties, such as compactly support, orthogonality, symmetry, stationary and high order vanishing moments. There are some conflicting constraints in the choice of wavelet function, and none of wavelet function possesses all these properties. In this study, we chose 11 representative wavelet bases db3, db5, db8, sym4, sym6, sym8, coif2, coif4, bior2.2, bior2.4 and bior2.6, and select decomposition scle from 2 to 5. The overall prediction accuracy of the protein structural classes is obtained on 25PDB, 1189 and 640 datasets, as shown in Tables 3–5. In order to find the optimal wavelet basis function and decomposition scale, a lot of work has been done. And several wavelet basis functions are chosen to obtain the overall prediction precision of protein structural class of three datasets and draw the line chart (the specific results see supplementary materials Tables S1–S3, Figs. S1–S3).

Table 3
Prediction results of protein structural class on 25PDB dataset by different wavelet function and wavelet decomposition scales.

Function	Scale			
	Jackknife test (%)			
	2	3	4	5
db3	74.90	76.51	82.55	87.09
db5	76.33	78.84	83.08	87.69
db8	75.55	75.91	81.29	82.37
sym4	76.33	79.80	88.22	93.07
sym6	75.85	77.88	86.49	89.00
sym8	76.87	81.29	87.99	89.60
coif2	75.67	79.26	83.80	87.09
coif4	76.93	81.11	86.37	84.04
bior2.2	76.21	81.71	87.57	88.88
bior2.4	76.57	80.75	86.61	89.42
bior2.6	78.24	80.16	85.53	87.81

Table 4
Prediction results of protein structural class on 1189 dataset by different wavelet function and wavelet decomposition scales.

Function	Scale			
	Jackknife test (%)			
	2	3	4	5
db3	72.25	72.80	80.77	87.18
db5	78.57	82.51	80.04	80.77
db8	78.39	77.11	80.13	82.88
sym4	79.85	82.51	87.64	90.75
sym6	76.74	78.39	83.88	83.79
sym8	80.13	81.78	81.50	88.37
coif2	78.30	80.49	81.23	86.45
coif4	79.49	82.14	81.96	77.66
bior2.2	78.30	83.33	87.00	88.74
bior2.4	80.86	80.59	85.26	86.45
bior2.6	82.97	82.97	86.17	89.10

It can be seen from Tables 3–5 that different wavelet functions and different decomposition scales will affect the prediction accuracy of the model. When the sym4 wavelet function is chosen and the decomposition scale is 5, the highest overall prediction accuracy is 93.07% on dataset 25PDB, 90.75% on dataset 1189, and 89.53% on

Table 5
Prediction results of protein structural class on 640 dataset by different wavelet function and wavelet decomposition scales.

Function	Scale			
	Jackknife test (%)			
	2	3	4	5
db3	78.13	82.19	88.44	88.59
db5	80.16	81.72	81.09	84.84
db8	82.97	81.87	82.97	79.84
sym4	77.50	81.09	89.53	89.53
sym6	82.34	79.22	85.31	86.88
sym8	78.91	80.78	84.22	88.28
coif2	77.03	77.66	84.22	85.16
coif4	78.28	79.69	88.44	75.47
bior2.2	77.03	82.03	85.94	87.50
bior2.4	80.00	81.87	86.88	87.81
bior2.6	80.94	81.41	86.72	87.50

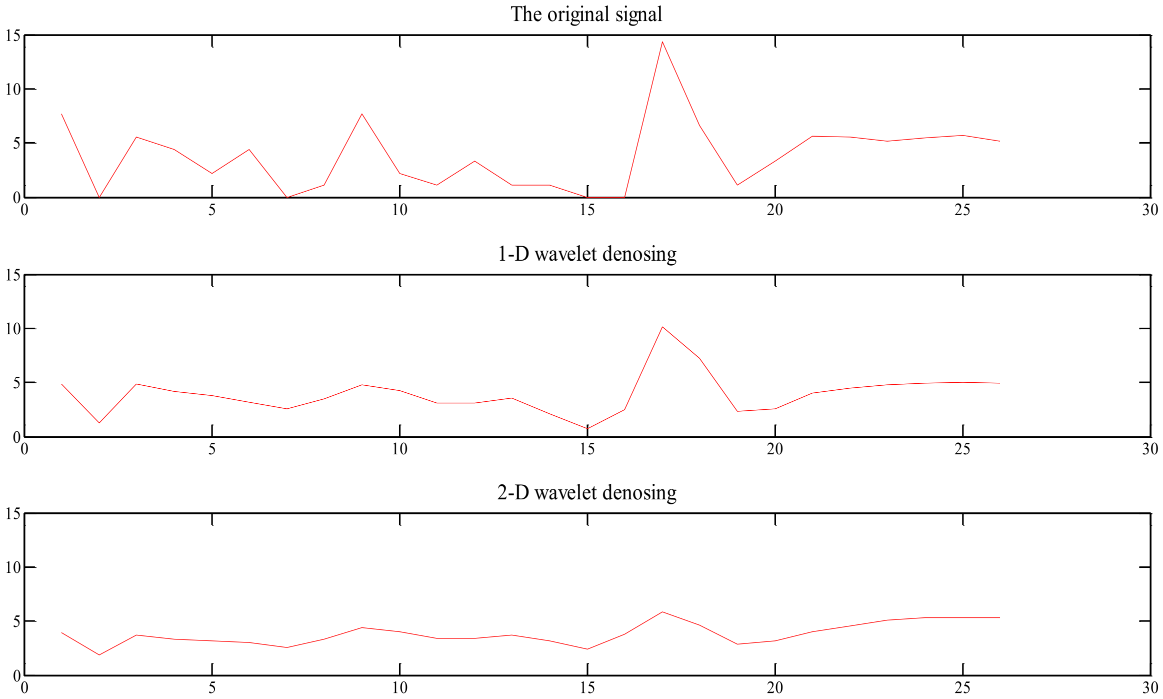

Fig. 2. Wavelet denoising using sym4 wavelet function and decomposition scale is $j = 5$.The y axis indicates the intensity of signal; the x axis indicates the residue position along the sequence.

dataset 640. On these three datasets, the sym4 wavelet function is superior to other wavelet functions, because the sym4 wavelet function has orthogonality and biorthonality, compact support. Through the discrete wavelet transform, at the same time, time and frequency domain have a good locality characteristics, excluding protein sequence redundancy information, making the extracted protein feature information more prominent. Therefore, we use sym4 as the optimal wavelet function, and 5 as the optimal decomposition scle.

3.3. Selection of denoising threshold method

In this paper, we use PseAAC to carry out feature extraction on protein sequences, and then extracted the feature vectors of the proteins by 1-D and 2-D wavelet denoising respectively. The specific process is as follows: Firstly, the wavelet decomposition function of the two methods is chosen to be sym4 wavelet, and the decomposition scale level is 5. Then the high-frequency coefficients are processed by threshold, and the thresholds are selected by wavelet function to get the default threshold. Finally, the signals are reconstructed on the basis of the low-frequency coefficients of the wavelet decomposition and the high-frequency coefficients after the threshold quantization. In order to more intuitively compare the effect of 1-D and 2-D wavelet denoising on protein structural class datasets, we pick PDB ID 1IGD protein of $\alpha+\beta$ class from the 640 dataset, and the results are shown in Fig. 2. The protein is an immunoglobulin binding protein, which has 61 amino acid residues. The 3-D structure of the protein 1IGD is determined by X-ray crystallography [75], and the 3-D structure of the protein is shown in Fig. 3.

It can see from Fig. 2, the 1IGD protein by 1-D and 2-D wavelet denoising, using 2-D wavelet denoising can effectively remove redundant information in protein sequence, so as to extract the characteristic signals of the protein itself. Moreover, from the principle of 1-D and 2-D wavelet denoising, we can see that the decomposition frequency band of 1-D wavelet denoising is not detailed enough, and the overall prediction accuracy of the protein structural class by 1-D wavelet denoising threshold is low (the specific results see supplementary material Table S4). 2-D wavelet denoising can deal with the whole dataset. The data is denoised from the low-frequency, horizontal, vertical and the diagonal scale space. After several experiments, we found that 2-D wavelet denoising method can effectively improve the feature of each class of protein structural classes, so we chose 2-D wavelet denoising method.

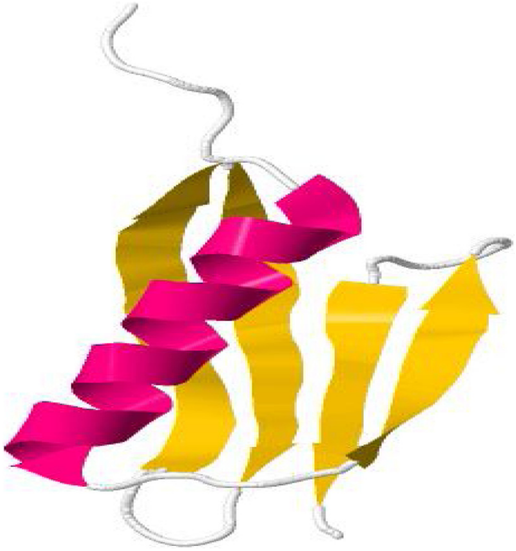

Fig. 3. The three-dimensional structure of protein 1IGD.

2-D wavelet denoising method to obtain the threshold value includes the default threshold and Birge-Massart strategy, which the default threshold can be obtained in two different ways. One is to use the 2-D wavelet to obtain the threshold, and the other is to use 2-D wavelet packet. The method of obtaining the threshold value by 2-D wavelet is called WV method, and the method of

Table 6
Prediction results of protein structural class datasets by selecting different threshold methods.

Dataset	Method		
	Jackknife test (%)		
	WV	WP	B-M
25PDB	93.07	92.89	87.87
1189	90.75	90.75	84.62
640	89.53	89.38	80.78

Table 7
Prediction results of protein structural class datasets by selecting different kernel functions.

Dataset	Function			
	Jackknife test (%)			
	1	2	3	4
25PDB	91.51	85.06	93.07	74.12
1189	89.93	78.48	90.75	73.99
640	81.09	82.19	89.53	66.09

obtaining the threshold value by 2-D wavelet packet is called WP method. The method of obtaining threshold by Birge-Massart strategy is called B-M method. We chose $\lambda = 6$ using PseAAC to extract feature information, used the sym4 wavelet function, the decomposition scale is 5, and selected RBF kernel function as the kernel function of the SVM. The thresholds are obtained by different methods, and then the protein structural class datasets are classified. The overall prediction accuracy of protein structure is shown in Table 6.

As shown from Table 6, the overall prediction accuracy using the B-M method on the three protein structural class datasets is lower than that using the WV and WP methods. In the dataset 1189, both WV method and WP method have the same overall prediction accuracy of 90.75%, and WV method performed better than WP on datasets 25PDB and 640. Therefore, we use the 2-D wavelet to obtain the default threshold method WV to reduce the noise of the datasets.

3.4. Effect of kernel function on results

The kernel function of SVM has important influence on the prediction results. In this paper, the feature information was extracted by PseAAC, choosing $\lambda = 6$. The sym4 wavelet function is used in the 2-D wavelet denoising, and the decomposition scale is 5.The different kernel functions are used to predict the protein datasets, and the overall prediction accuracy of protein structural class, as shown in Table 7. In Table 7, 1 represents a linear kernel function, 2 represents a cubic polynomial kernel function, 3 represents an RBF kernel function, and 4 represents a sigmoid kernel function. Fig. 4 shows the histogram of the overall prediction results of protein structural class by different kernel functions.

As is shown in Table 7 and Fig. 4, the RBF kernel function on the 25PDB dataset obtains the highest overall prediction accuracy of 93.07%, which is 1.56%-18.95% higher than the overall prediction accuracy obtained by other kernel functions. The highest overall prediction accuracy obtained by using the RBF kernel function on dataset 1189 is 90.75%, which is 0.82%–16.76% higher than the overall prediction accuracy obtained by other kernel functions. Using the RBF kernel function on the dataset 640, the highest overall prediction accuracy is 89.53%, which is 7.34%–23.44% higher than the overall prediction accuracy obtained by other kernel functions. It can be seen that the RBF kernel functions are used to obtain the highest overall prediction accuracy on the three low-similarity datasets because the RBF kernel function is excellent in dealing with nonlinear problems [36,37]. By the comparison and analysis, the RBF kernel function is selected as the kernel function of SVM algorithm.

Table 8
Prediction results of protein structural class datasets by two different feature extraction methods on 25PDB dataset.

Algorithm	Locations				
	Jackknife test (%)				
	All-α	All-β	$\alpha + \beta$	α/β	OA
PseAAC	72.69	61.63	32.20	58.38	56.13
WD-PseAAC	95.71	97.74	84.35	94.80	93.07

Table 9
Prediction results of protein structural class datasets by two different feature extraction methods on 1189 dataset.

Algorithm	Locations				
	Jackknife test (%)				
	All-α	All-β	$\alpha + \beta$	α/β	OA
PseAAC	56.05	63.61	14.52	72.75	54.03
WD-PseAAC	98.65	98.98	68.88	94.01	90.75

Table 10
Prediction results of protein structural class datasets by two different feature extraction methods on 640 dataset.

Algorithm	Locations				
	Jackknife test (%)				
	All-α	All-β	$\alpha + \beta$	α/β	OA
PseAAC	54.35	51.30	33.92	71.19	52.81
WD-PseAAC	92.75	95.45	78.95	92.09	89.53

3.5. Effect of feature extraction on results

One of the key aspects of protein structural class prediction is the extraction of protein sequences. The selection of appropriate protein sequence feature extraction methods tends to make our model more excellent. In this paper, we compare the effects of PseAAC algorithm and WD-PseAAC algorithm on the prediction results of three different datasets, and choose the best feature extraction algorithm. We choose $\lambda = 6$ using PseAAC to extract feature information and 2-D wavelet noise reduction using sym4 wavelet function, decomposition scale of 5. PseAAC and WD-PseAAC feature extraction methods are classified by the RBF kernel function of the support vector machine, and three different prediction results of the 25PDB dataset, the 1189 dataset and the 640 dataset are obtained, as shown in Tables 8–10. In addition, we use the ROC (receiver operating characteristic) curve to compare the robustness of the model under different feature extraction algorithms. In general, ROC curve is applicable to evaluate the prediction performance of a binary classifier, but structural class prediction is a four-class prediction problem. We first use the one-versus-rest (OVR) strategy to transform the multi-classification problem into multiple two-classification problem. And then for these two-classification ture positive rate and false positive rate, the average of them was taken as the final result. Figs. 5–7 are the ROC curves obtained by the two different feature extraction methods for the 25PDB dataset, 1189 dataset, and 640 dataset, respectively.

As show from Table 8, different feature extraction algorithms have different prediction accuracy for 25PDB dataset. The overall prediction accuracy of the PseAAC algorithm is 56.13%, which is 36.94% lower than the overall prediction accuracy using the WD-PseAAC algorithm. In addition, The prediction accuracy of WD-PseAAC algorithm are 23.02%, 36.11%, 52.15% and 36.42% higher

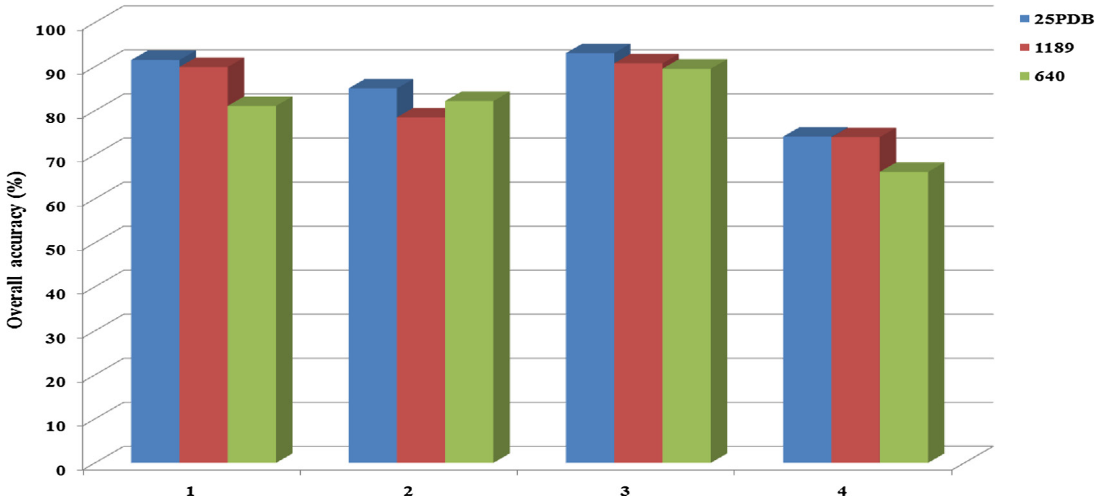

Fig. 4. The overall prediction accuracy of three protein structural datasets 25PDB, 1189 and 640 under different kernel functions.

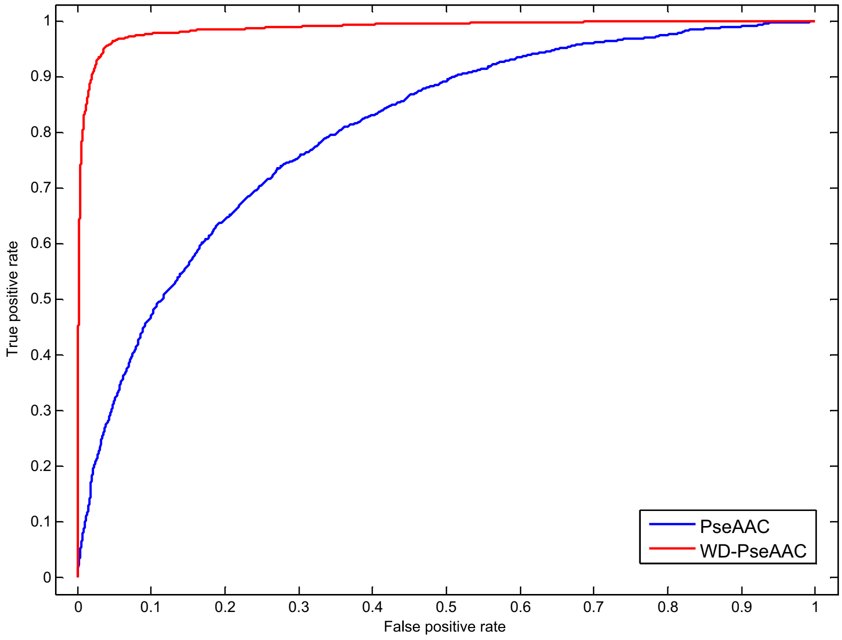

Fig. 5. This graph shows the ROC curves of 25PDB dataset.

than PseAAC algorithm for the All-α class, All-β class, α + β class, α/β class, which significantly improves the prediction accuracy of the four protein structural classes in the dataset. As show from Fig. 5, the coverage area of the ROC curve using the WD-PseAAC feature extraction algorithm is the largest for the 25PDB dataset, and AUC is 0.9874, which is significantly higher than the PseAAC algorithm, the AUC value of PseAAC is 0.8036.

As show from Table 9, the different feature extraction algorithms are different for the prediction accuracy of each of the protein structural class in the 1189 dataset. Using the WD-PseAAC algorithm, the overall prediction accuracy is 90.75%, which is 36.72% higher than the overall prediction accuracy of the PseAAC algorithm. In addition, prediction accuracy by using the WD-PseAAC algorithm are 42.60%, 35.37%, 54.36%, 21.26% higher than the PseAAC algorithm for All-α class, All-β class, α + β class, α/β class, which significantly improves the prediction accuracy of the four protein structural classes in the dataset. As show from Fig. 6, the coverage area of the ROC curve using the WD-PseAAC feature extraction algorithm is the largest for the 1189 dataset, AUC is 0.9785, which is significantly higher than the PseAAC algorithm, the AUC value of PseAAC is 0.7815.

As show from Table 10, the different feature extraction algorithms are different for the prediction accuracy of each of the protein structural class in the 640 dataset. Using the WD-PseAAC algorithm, the overall prediction accuracy is 89.53%, which is 36.72% higher than the overall prediction accuracy of the PseAAC algorithm. In addition, prediction accuracy by using the WD-PseAAC are 38.4%, 44.15%, 45.03%, 20.9% higher than the PseAAC algorithm for All-α class, All-β class, α + β class, α/β class, which significantly improves the prediction accuracy of the four protein structural classes in the data set. As show from Fig. 7, the coverage area of the ROC curve using the WD-PseAAC feature extraction algorithm is the largest for the 640 dataset, AUC is 0.9770, which is significantly higher than the PseAAC algorithm, the AUC value of PseAAC is 0.7843.

In conclusion, comparing the effects of different feature extraction algorithms on the results and the robustness of two different feature extraction algorithms from the ROC curve, we determine

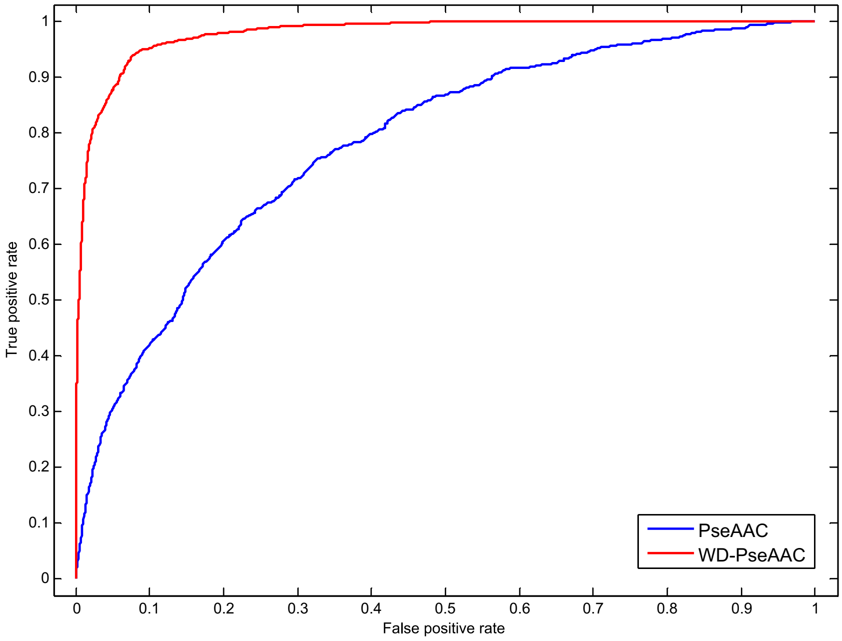

Fig. 6. This graph shows the ROC curves of 1189 dataset.

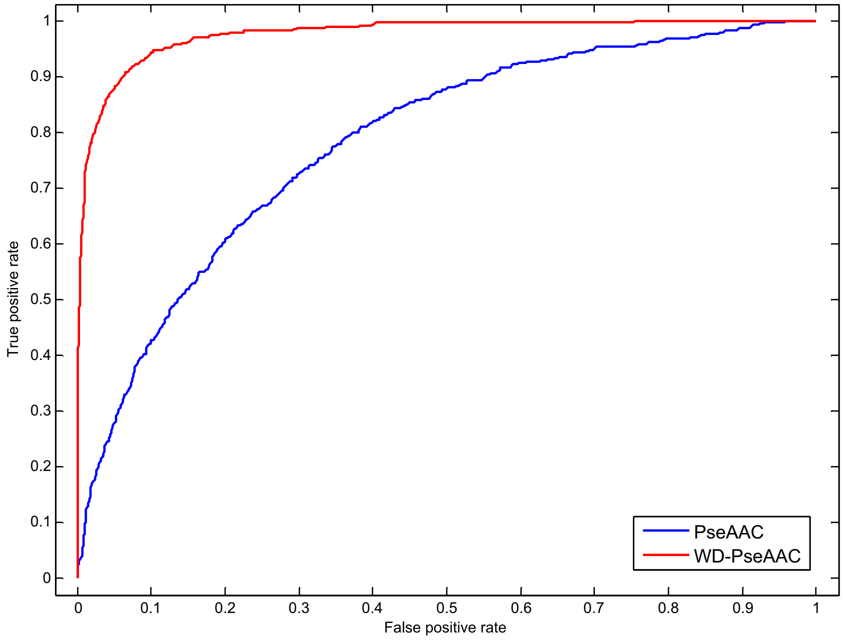

Fig. 7. This graph shows the ROC curves of 640 dataset.

that WD-PseAAC is the best feature extraction algorithm for this study.

3.6. *Prediction performance of our method*

In this study, three datasets 25PDB, 1189 and 640 are extracted features by PseAAC, then SVM algorithm based on 2-D wavelet denoising is used to predict the structural classes of proteins. According to the above analysis, we chose $\lambda = 6$ using PseAAC to extract feature information and used the sym4 wavelet function, the decomposition scale is 5. The 2-D wavelet denoising is applied to obtain feature data, and the RBF kernel function is used as the kernel function of SVM. The results of protein structural class prediction on the datasets 25PDB, 1189 and 640 are obtained by jackknife test and performance evaluation indicators P, Sens, Spec, F-measure, MCC and OA are shown in Table 11.

As shown from Table 11, the overall prediction accuracy of the datasets25PDB, 1189 and 640 are 93.1%, 90.8% and 89.5%, respectively. Comparing the prediction accuracy of four protein structural classes, the performance evaluation indexes of all-α class are about 91.0%–96.8% on dataset 25PDB, 94.0%–98.7% on dataset 1189 and 92.8%–99.3% on dataset 640. The performance indexes of all-β class are about 89.6%-99.0% on the three protein structural class datasets. The main reason for the good performance of all-α class and all-β class is that these protein sequences are helices or stands, and single highly repetitive structures are easier to predict.

At the same time, the predition results of α/β class protein are also satisfactory. The performance indexes of the three datasets are

Table 11
Prediction quality of our method on three benchmark datasets.

Dataset	Structure class	P (%)	Sens (%)	Spec (%)	F-measure	MCC	OA (%)
25PDB	All-α	92.0	95.7	96.8	0.94	0.91	93.1
	All-β	96.4	97.7	98.6	0.97	0.96	
	α/β	89.6	94.8	97.0	0.92	0.90	
	$\alpha+\beta$	93.7	84.4	97.9	0.89	0.85	
1189	All-α	94.0	98.7	98.2	0.96	0.95	90.8
	All-β	90.4	99.0	95.8	0.95	0.92	
	α/β	90.0	94.0	95.1	0.92	0.88	
	$\alpha+\beta$	88.8	68.9	97.5	0.78	0.73	
640	All-α	97.7	92.8	99.3	0.95	0.94	89.5
	All-β	89.6	95.5	96.2	0.93	0.90	
	α/β	83.6	92.1	92.8	0.88	0.83	
	$\alpha+\beta$	90.0	79.0	96.7	0.84	0.79	

all above 83.0%, and the specificity index on the 25PDB dataset is 97.0%. In contrast, the individual indicators of the predictive results of the $\alpha+\beta$ class is low. For example, the sensitivity on 1189 dataset is 68.9%, and MCC is 0.73. Only the sensitivity, F-measure and MCC of the $\alpha+\beta$ class on three datasets are low, and the results of the other two classes are above 88.8%, and the F-measue of the all-β class on 25PDB dataset even reaches 0.97. The prediction accuracy of the $\alpha+\beta$ class on the three low-similarity datasets is low, and we speculate that it may be due to its non-negligible overlap with the other classes resulting in low accuracy. The fact indicates that there are still many challenges in the future research to improve prediction accuracy of $\alpha+\beta$ class.

3.7. Performance comparison with existing methods

In this section, to demonstrate the effectiveness of the proposed method WD-PseAAC, we compared with other recently reported prediction methods on the same datasets. All the methods are performed using jackknife cross-validation test. We selected the accuracy of each class and overall accuracy as evaluation indexes that are shown in Table 12. The compared methods include the famous methods SCPRED [12] and MODAS [13]. SCPRED is based on the extraction of information from the predicted protein secondary structure sequence. MODAS combines evolutionary profiles and predicted secondary structure information. SCEC [29] incorporates evolutionary information encoded using PSI-BLAST profile-based collocation of amino acids pairs. RKS-PPSC [14] extracts feature vectors combining recurrence quantification analysis, K-string based information entropy and segment-based analysis. The comparison method also includes other PSSM-based methods such as MEDP [73], LCC-PSSM [34] and PSSS-PsePSSM [37]. Kong et al. [19] proposed feature extraction method based on structure-driven. Hayat et al. [25] proposed feature extraction method using pseudo average chemical shift. Comparison of our method with other prediction methods on the datasets 25PDB, 1189 and 640 are shown in Figs. 8–10.

As shown from Table 12 and Fig. 8, the WD-PseAAC method achieves the highest overall prediction accuracy of 93.1% on the 25PDB dataset, which is 3.6%–18.3% higher than the other prediction methods. For all-β class, α/β class, and $\alpha+\beta$ class, the highest accuracy reached 97.7%, 94.8% and 84.4%, respectively. The prediction accuracy of all-α is not the highest, which is 0.7% lower than that of PSSS-PsePSSM.

As shown from Table 12 and Fig. 9, the WD-PseAAC method achieves the highest overall prediction accuracy of 90.8% on the 1189 dataset, which is 4.2%–23.2% higher than the other prediction methods. For all-α class, all-β class, α/β class, the highest accuracy reached 98.7%, 99.0% and 94.0%, respectively. The prediction accuracy of $\alpha+\beta$ is not the highest, which is 5.0% lower than that of PSSS-PsePSSM.

Table 12
Performance comparison of different methods on three benchmark datasets.

	Prediction accuracy (%)					
Dataset	Method	All-α	All-β	α/β	$\alpha+\beta$	Overall
25PDB	SCPRED [12]	92.6	80.1	74	71	79.7
	MODAS [13]	92.3	83.7	81.2	68.3	81.4
	MEDP [73]	87.8	78.3	76	57.4	74.8
	RKS-PPSC [14]	92.8	83.3	85.8	70.1	82.9
	LCC-PSSM [34]	91.7	80.8	79.8	64	79
	Kong et al. [19]	94.1	87.1	84.1	74.4	85
	PSSS-PsePSSM [37]	96.4	90.5	90.2	81.2	89.5
	Hayat et al. [25]	92.2	85.4	82.2	76.9	84.2
	WD-PseAAC	95.7	97.7	94.8	84.4	93.1
1189	SCPRED [12]	89.1	86.7	89.6	53.8	80.6
	MODAS [13]	92.3	87.1	87.9	65.4	83.5
	SCEC [29]	75.8	75.2	82.6	31.8	67.6
	RKS-PPSC [14]	89.2	86.7	82.6	65.6	81.3
	MEDP [73]	85.2	84	84.4	45.2	75.8
	LCC-PSSM [34]	89.2	88.8	85.6	58.5	81.2
	Kong et al. [19]	93.7	86.1	86.8	73.9	85.2
	PSSS-PsePSSM [37]	91.9	91.8	87.7	73.9	86.6
	Hayat et al. [25]	92.2	84	88.6	74.6	85
	WD-PseAAC	98.7	99	94	68.9	90.8
640	SCPRED [12]	90.6	81.8	85.9	66.7	80.8
	SCEC [29]	73.9	61	81.9	33.9	62.3
	RKS-PPSC [14]	89.1	85.1	88.1	71.4	83.1
	MEDP [73]	84.8	75.3	86.4	53.8	74.7
	LCC-PSSM [34]	92.8	88.3	85.9	66.1	82.7
	Kong et al. [19]	91.3	77.3	91.5	76	83.9
	PSSS-PsePSSM [37]	87	81.2	84.7	70.8	81
	Hayat et al. [25]	85	83	86.3	83.2	86.4
	WD-PseAAC	92.8	95.5	92.1	78.9	89.5

As shown from Table 12 and Fig. 10, the WD-PseAAC method achieves the highest overall prediction accuracy of 89.5% on the 640 dataset, which improves the accuracy by 3.1%–27.2% compared to other prediction methods. For the all-α class, the all-β class, the α/β class, the highest accuracy reached 92.8%, 95.5% and 92.1%, respectively. The prediction accuracy of $\alpha+\beta$ is not the highest, which is 4.3% lower than that of Hayat et al. [25], and 12.2% and 8.1% higher than SCPRED and PSSS-PsePSSM, respectively. In this paper, the most contribution of our proposed method is the improvement of the accuracy using 2-D wavelet denoising. Wavelet denoising can well describe the non-stationary characteristics of the signal, such as edges, spikes, breakpoints and so on, and decorreduce the signal and whiten the noise after transformation, which makes the obtained feature information of each class of proteins more prominent. In summary, the above comparison results show that our method is very promising for protein structural class prediction, especially for low-similarity datasets.

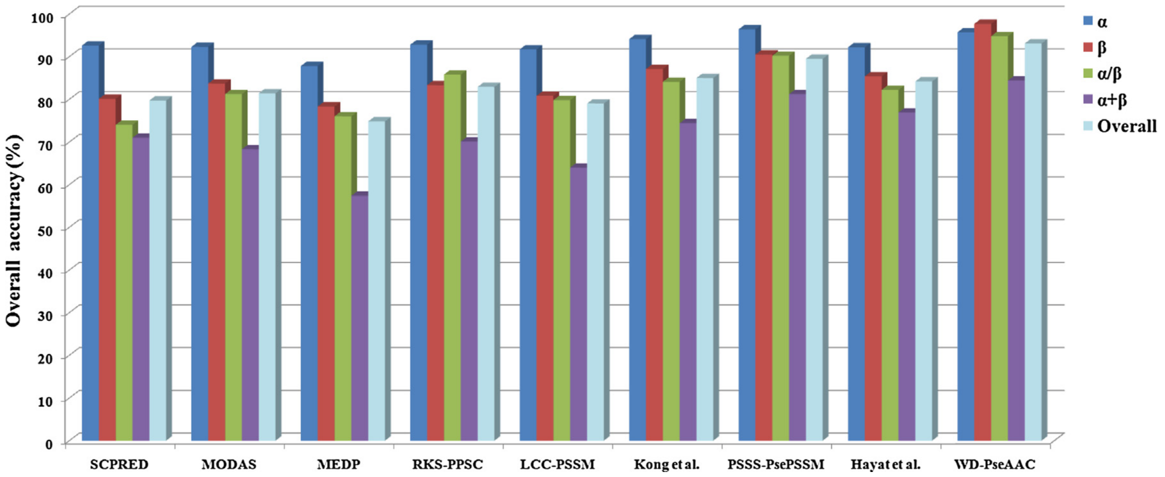

Fig. 8. Comparison of our method with other prediction methods on the 25PDB dataset.

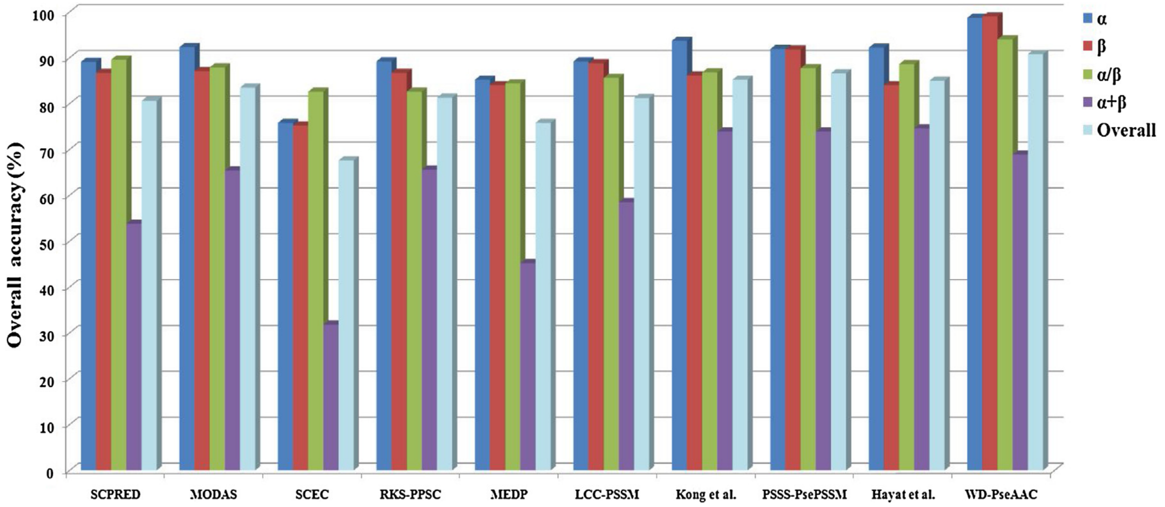

Fig. 9. Comparison of our method with other prediction methods on the 1189 dataset.

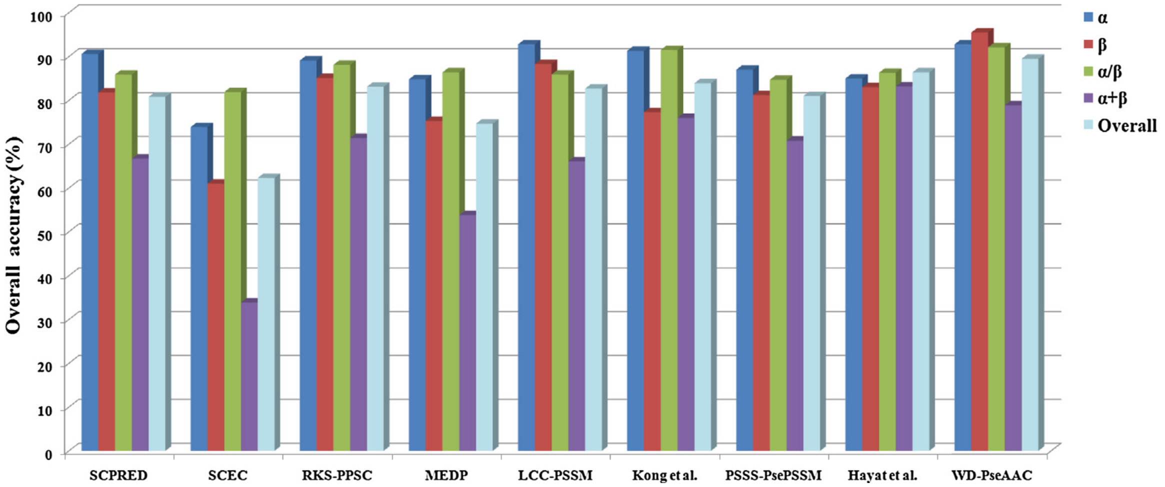

Fig. 10. Comparison of our method with other prediction methods on the 640 dataset.

4. Conclusions

Accurate prediction of protein structural classes for low-similarity sequences is a complicated and challenging task. In this paper, a novel protein structural class prediction method WD-PseAAC is proposed. Chou's PseAAC is used to extract the feature of protein sequences, and protein structural class is predicted by SVM algorithm based on 2-D wavelet denoising. PseAAC feature extraction can avoid the loss of protein sequence-order information, and obtain more detailed protein sequence information. 2-D wavelet denoising can effectively remove redundant information in protein sequences and extract the characteristic signal of each class of protein, which is very important for guaranteeing high prediction accuracy. SVM classification algorithm can deal with high-dimensional data, to avoid over-fitting and effectively removal of non-support vector. The performance of the three benchmark low-similarity datasets 1189, 25PDB and 640 is evaluated by jackknife test. The overall prediction accuracies of the three datasets reach 93.1%, 90.8% and 89.5%, respectively. The experimental results show that our proposed method not only can effectively improve the prediction accuracy of protein structural class in particular for low-similarity datasets, but also is expected to be used for the prediction of other attributes of proteins. We shall make efforts in our future work to provide a web-server for the method presented in this paper. The source code for implementing in this study is available from the author upon request.

Acknowledgements

The authors sincerely thank the anonymous referees for their many valuable comments that have improved this manuscript. This work is supported by the National Natural Science Foundation of China (Nos. 51372125 and 21571112), the Natural Science Foundation of Shandong Province of China (Nos. ZR2014FL021 and ZR2016BQ28), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), and the Applied Basic Research Project of Qingdao of China (No. 15-9-1-83-JCH).

References

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, Nature 261 (1976) 552–557.
[2] K.C. Chou, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins 21 (1995) 319–344.
[3] G.P. Zhou, An intriguing controversy over protein structural class prediction, J. Protein Chem. 17 (1998) 729–738.
[4] K.C. Chou, A key driving force in determination of protein structural classes, Biochem. Biophys. Res. Commun. 264 (1999) 216–224.
[5] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Prediction of protein structural classes by support vector machines, J. Comput. Chem. 26 (2002) 293–296.
[6] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Struct. Funct. Bioinform. 43 (2001) 246–255.
[7] T.L. Zhang, Y.S. Ding, K.C. Chou, Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern, J. Theor. Biol. 250 (2008) 186–193.
[8] H.B. Shen, K.C. Chou, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[9] X. Xiao, S.H. Shao, Z.D. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, J. Comput. Chem. 27 (2006) 478–482.
[10] X.D. Sun, R.B. Huang, Prediction of protein structural classes using support vector machines, Amino Acids 30 (2006) 469–475.
[11] R.Y. Luo, Z.P. Feng, J.K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, Eur. J. Biochem. 269 (2002) 4219–4225.
[12] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, BMC Bioinf. 9 (2008) 226.
[13] M.J. Mizianty, L. Kurgan, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, BMC Bioinf. 10 (2009) 414.
[14] J.Y. Yang, Z.L. Peng, X. Chen, Prediction of protein structural classes for low homology sequences based on predicted secondary structure, BMC Bioinf. 11 (2010) S9.
[15] T. Liu, C.Z. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, J. Theor. Biol. 267 (2010) 272–275.
[16] S.Y. Ding, S.L. Zhang, Y. Li, T.M. Wang, A novel protein structural classes prediction method based on predicted secondary structure, Biochimie 94 (2012) 1166–1171.
[17] S.L. Zhang, Y.Y. Liang, X.G. Yuan, Improving the prediction accuracy of protein structural class: approached with alternating world frequency and normalized Lempel-Ziv complexity, J. Theor. Biol. 341 (2014) 71–77.
[18] J.R. Wang, Y. Li, X.Q. Liu, Q. Dai, Y.H. Yao, P.A. He, High-accuracy prediction of protein structural classes using PseAA structural properties and secondary structural patterns, Biochimie 101 (2014) 104–112.
[19] L. Kong, L.C. Zhang, Novel structure-driven features for accurate prediction of protein structural class, Genomics 103 (2014) 292–297.
[20] L.C. Zhang, L. Kong, X.D. Han, J.F. Lv, Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure, J. Theor. Biol. 400 (2016) 1–10.
[21] M.H. Olyaee, A. Yaghoubi, M. Yaghoobi, Predicting protein structural classes based on complex networks and recurrence analysis, J. Theor. Biol. 404 (2016) 375–382.
[22] Y. Marrero-Ponce, E. Contreras-Torres, C.R. García-Jacas, S.J. Barigye, N. Cubillán, Y.J. Alvarado, Novel 3D bio-macromolecular bilinear descriptors for protein science: predicting protein structural classes, J. Theor. Biol. 374 (2015) 125–137.
[23] M. Hayat, M. Tahir, S.A. Khan, Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces, J. Theor. Biol. 346 (2014) 8–15.
[24] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.
[25] M. Hayat, N. Iqbal, Discriminating protein structure classes by incorporating pseudo average chemical shift to Chou's general PseAAC and support vector machine, Comput. Meth. Prog. Biomed. 116 (2014) 184–192.
[26] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, A. Sattar, Proposing a highly accurate protein structural class predictor using segmentation-based features, BMC Genom. 15 (2014) S2.
[27] T.G. Liu, X.Q. Zheng, J. Wang, Prediction of protein structural class for low-similarity sequence using support vector machine and PSI-BLAST profile, Biochimie 92 (2010) 1330–1334.
[28] T.G. Liu, X.B. Geng, X.Q. Zheng, R.S. Li, J. Wang, Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles, Amino Acids 42 (2012) 2243–2249.
[29] K. Chen, L.A. Kurgan, J.S. Ruan, Prediction of protein structural class using novel evolutionary collocation based sequence representation, J. Comput. Chem. 29 (2008) 1596–1604.
[30] S.L. Zhang, F. Ye, X.G. Yuan, Using principal component analysis and support vector machine to predict protein structural class for low similarity sequences via PSSM, J. Biomol. Struct. Dyn. 29 (2012) 634–642.
[31] Y.Y. Liang, S.Y. Liu, S.L. Zhang, Using correlation analysis and nonnegative matrix factorization to predict protein structural classes via position-specific scoring matrix, MATCH Commun. Math. Comput. Chem. 75 (2016) 743–758.
[32] X.Y. Xia, M. Ge, Z.X. Wang, X.M. Pan, Accurate prediction of protein structural class, PLoS One 7 (2012) e37653.
[33] S.Y. Ding, Y. Li, Z.X. Shi, S.J. Yan, A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile, Biochimie 97 (2014) 60–65.
[34] S.Y. Ding, S.J. Yan, S.H. Qi, Y. Li, Y.H. Yao, A protein structural classes prediction method based on PSI-BLAST profile, J. Theor. Biol. 353 (2014) 19–23.
[35] J.R. Wang, C. Wang, J.J. Cao, X.Q. Liu, Y.H. Yao, Q. Dai, Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features, Gene 554 (2015) 241–248.
[36] L.Q. Li, X. Cui, S.J. Yu, Y. Zhang, Z. Luo, H. Yang, Y. Zhou, X.Q. Zheng, PSSP-RFE: accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations, PLoS One 9 (2014) e92863.
[37] S.L. Zhang, Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC, Chemometr. Intell. Lab. Syst. 142 (2015) 28–35.
[38] P. Tao, T. Liu, X. Li, X.M. Chen, Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination, Amino Acids 47 (2015) 461–468.
[39] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition, J. Theor. Biol. 360 (2014) 109–116.
[40] Y.D. Cai, G.P. Zhou, Prediction of protein structural classes by neural network, Biochimie 82 (2000) 783–785.
[41] H.B. Shen, J. Yang, X.J. Liu, K.C. Chou, Using supervised fuzzy clustering to predict protein structural classes, Biochem. Biophys. Res. Commun. 334 (2005) 577–581.
[42] Z.X. Wang, Z. Yuan, How good is prediction of protein structural class by the component-coupled method, Proteins 38 (2000) 165–175.
[43] Y.F. Cao, S. Liu, L.D. Zhang, J. Qin, J. Wang, K.X. Tang, Prediction of protein structural class with rough sets, BMC Bioinf. 7 (2006) 20.
[44] Y.Y. Liang, S.Y. Liu, S.L. Zhang, Prediction of protein structural class based on different autocorrelation descriptors of position-specific scoring matrix, MATCH Commun. Math. Comput. Chem. 73 (2015) 765–784.
[45] J.W. Shen, J. Zhang, X.M. Luo, W.L. Zhu, K.X. Chen, Y.X. Li, H.L. Jiang, Predicting protein–protein interactions based only on sequences information, Proc. Nati. Acad. Sci. U. S. A. 104 (2007) 4337–4341.
[46] S.B. Zhang, Q.R. Tang, Protein-protein interaction inference based on semantic similarity of gene ontology terms, J. Theor. Biol. 401 (2016) 30–37.
[47] Z.S. Wei, K. Han, J.Y. Yang, H.B. Shen, D.J. Yu, Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests, Neurocomputing 193 (2016) 201–212.

[48] M. Bhasin, G.P. Raghava, GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors, Nucleic Acids Res. 32 (2004) 383–389.

[49] Z.C. Li, X. Zhou, Z. Dai, X.Y. Zhou, Classification of G proteins and prediction of GPCRs-G proteins coupling specificity using continuous wavelet transform and information theory, Amino Acids 43 (2012) 793–804.

[50] M.N. Davies, A. Secker, A.A. Freitas, M. Mendao, J. Timmis, D.R. Flower, On the hierarchical classification of G protein-coupled receptors, Bioinformatics 23 (2007) 3113–3118.

[51] J. Chen, H.M. Xu, P.A. He, Q. Dai, Y.H. Yao, A multiple information fusion method for predicting subcellular locations of two different types of bacterial protein simultaneously, Biosystems 139 (2016) 37–45.

[52] Q. Dai, S. Ma, Y.B. Hai, Y.H. Yao, X.Q. Liu, A segmentation based model for subcellular location prediction of apoptosis protein, Chemometr. Intell. Lab. Syst. 158 (2016) 146–154.

[53] G.Q. Jian, Y.S. Zhang, P.P. Qian, Prediction of subcellular localization for apoptosis protein: approached with a novel representation and support vector machine, MATCH Commun. Math. Comput. Chem. 67 (2012) 867–878.

[54] S.L. Zhang, Y.Y. Liang, Z.G. Bai, A novel reduced triplet composition based method to predict apoptosis protein subcellular localization, MATCH Commun. Math. Comput. Chem. 73 (2015) 559–571.

[55] B. Yu, S. Li, C. Chen, J.M. Xu, W.Y. Qiu, X. Wu, R.X. Chen, Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino composition, Chemometr. Intell. Lab. Syst. 167 (2017) 102–112.

[56] L.A. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, Pattern Recogn. 39 (2006) 2323–2343.

[57] K. Ahmad, M. Waris, M. Hayat, Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition, J. Membr. Biol. 249 (2016) 293–304.

[58] J. Zhang, P.P. Sun, X.W. Zhao, Z.Q. Ma, PECM: prediction of extracellular matrix proteins using the concept of Chou's pseudo amino acid composition, J. Theor. Biol. 363 (2014) 412–418.

[59] P.P. Zhu, W.C. Li, Z.J. Zhong, E.Z. Deng, H. Ding, W. Chen, H. Lin, Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition, Mol. BioSyst. 11 (2015) 558–563.

[60] S.P. Shi, J.D. Qiu, X.Y. Sun, J.H. Huang, S.Y. Huang, S.B. Suo, R.P. Liang, L. Zhang, Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction, BBA-Mol. Cell Res. 1813 (2011) 424–430.

[61] F. Ali, M. Hayat, Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition, J. Theor. Biol. 384 (2015) 78–83.

[62] G.L. Fan, Q.Z. Li, Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 334 (2013) 45–51.

[63] P. Liò, Wavelets in bioinformatics and computational biology: state of art and perspectives, Bioinformatics 19 (2003) 2–9.

[64] B. Yu, Y. Zhang, A simple method for predicting transmembrane proteins based on wavelet transform, Int. J. Biol. Sci. 9 (2013) 22–33.

[65] M.A. Rezaei, P. Abdolmaleki, Z. Karami, E.B. Asadabadi, M.A. Sherafat, H. Abrishami-Moghaddam, M. Fadaie, M. Forouzanfar, Prediction of membrane protein types 7by means of wavelet analysis and cascaded neural networks, J. Theor. Biol. 254 (2008) 817–820.

[66] R.P. Liang, S.Y. Huang, S.P. Shi, X.Y. Sun, S.B. Suo, J.D. Qiu, A novel algorithm novel algorithm combining support vector machine with the discrete wavelet transform or the prediction of protein subcellular localization, Comput. Biol. Med. 42 (2012) 180–187.

[67] S.G. Chang, B. Yu, M. Vetterli, Adaptive wavelet thresholding for image denoising and compression, IEEE Trans. Image Process. 9 (2000) 1532–1546.

[68] S.G. Chang, B. Yu, M. Vetterli, Spatially adaptive wavelet thresholding with context modeling for image denoising, IEEE Trans. Image Process. 9 (2000) 1522–1531.

[69] B. Yu, Y. Zhang, The analysis of colon cancer gene expression profiles and the extraction of informative genes, J. Comput. Theor. Nanosci. 10 (2013) 1097–1103.

[70] A. Kandaswamy, C.S. Kumar, R.P. Ramanathan, S. Jayaraman, N. Malmurugan, Neural classification of lung sounds using wavelet coefficients, Comput. Biol. Med. 34 (2004) 523–537.

[71] F. Luisier, T. Blu, M. Unser, A new SURE approach to image denoising: interscale orthonormal wavelet thresholding, IEEE. Trans. Image Process. 16 (2007) 593–606.

[72] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[73] L.C. Zhang, X.Q. Zhao, L. Kong, Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 355 (2014) 105–110.

[74] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27.

[75] J.P. Derrick, D.B. Wigley, The third IgG-binding domain from streptococcal protein G: an analysis by X-ray crystallography of the structure alone and in a complex with Fab, J. Mol. Biol. 243 (1994) 906–918.