Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition

Wenying Qiu , Shan Li , Xiaowen Cui , Zhaomin Yu , Minghui Wang , Junwei Du , Yanjun Peng , Bin Yu

PII: S0022-5193(18)30194-2
DOI: 10.1016/j.jtbi.2018.04.026
Reference: YJTBI 9441

To appear in: *Journal of Theoretical Biology*

Received date: 8 February 2018
Revised date: 10 April 2018
Accepted date: 16 April 2018

Please cite this article as: Wenying Qiu , Shan Li , Xiaowen Cui , Zhaomin Yu , Minghui Wang , Junwei Du , Yanjun Peng , Bin Yu , Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition, *Journal of Theoretical Biology* (2018), doi: 10.1016/j.jtbi.2018.04.026

Highlights

- A new method (PseAAC-PsePSSM-WD) to prediction protein submitochondrial locations.
- The protein sequence features are extracted by fusing the PseAAC and PsePSSM methods.
- 2-D wavelet denoising can effectively remove the redundant information in the protein sequences.
- We investigate the effect of the five different classifiers on the results.
- The proposed method increases the prediction performance over several methods.

Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition

Wenying Qiu [a, b, 1], Shan Li [a, b, 1], Xiaowen Cui [a, b], Zhaomin Yu [a, b], Minghui Wang [a, b], Junwei Du [c], Yanjun Peng [d*], Bin Yu [a, b*]

[a] College of MathemYatics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

[b] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

[c] College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China

[d] Shandong Province Key Laboratory of Wisdom Mining Information Technology, Shandong University of Science and Technology, Qingdao 266590, China

Abstract: Mitochondrion are important organelles of most eukaryotes and play an important role in participating in various life activities of cells. However, some functions of mitochondria can only be achieved in specific submitochondrial location, the study of submitochondrial locations will help to further understand the biological function of protein, which is a hotspot in proteomics research. In this paper, we propose a new method for protein submitochondrial locations prediction. Firstly, the features of protein sequence are extracted by combining Chou's pseudo-amino acid composition (PseAAC) and pseudo-position specific scoring matrix (PsePSSM). Then the extracted feature information is denoised by two-dimensional (2-D) wavelet denoising. Finally, the optimal feature vectors are input to the SVM classifier to predict the protein submitochondrial locations. We obtained the ideal prediction results by jackknife test and compared with other prediction methods. The results indicate that the proposed method is significantly better than the existing research results, which can provide a new method for other subcellular organ localization prediction. The source code and all datasets are available at https://github.com/QUST-BSBRC/PseAAC-PsePSSM-WD/ for academic use.

Keywords: Submitochondrial locations; Pseudo-amino acid composition; Pseudo-position specific

* Corresponding authors.
E-mail addresses: pengyanjuncn@163.com (Y. Peng), yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

scoring matrix; Two-dimensional wavelet denoising; Support vector machine; Machine learning

1. Introduction

Mitochondrion exist in most organelles of eukaryotes, which are divided into three regions: the inner membrane, outer membrane and matrix. They play important physiological functions during the process of life. Proteins are the major performer of cell life activities. Many research findings revealed that proteins express their functions within a specific submitochondrial structural domain, while abnormally functioning mitochondria lead to energy metabolism disorders, thus leading to a series of interacting damaging processes that ultimately lead to human disease. Among them, many diseases in human are directly related to mitochondria, such as Parkinson's disease (Burbulla et al., 2017), diabetes (Qi et al., 2017) and Alzheimer's disease (Mei and Wang, 2010). However, the localization of protein submitochondrial structure helps to design drug-related diseases caused by mitochondrial defects. Therefore, prediction of protein submitochondrial locations plays an important role in understanding the properties and functions of proteins, and also has long-term significance for biopharmaceutical and cell therapy.

Protein submitochondrial locations is an important research area of proteomics. With the rapid increase of protein sequence, protein submitochondrial locations can be directly obtained by experimental means, but the experimental results are often variable. Moreover, the experimental results tend to have variability. Also these experiments take a long time and costly, which can no longer meet the needs of the current study. Therefore, using computational methods for protein submitochondrial locations prediction becomes a good auxiliary means. When predicting the locations of protein submitochondrial through the combination of bioinformatics and computational methods, first, feature extraction is performed on the protein sequences in the datasets. Then, a predictive classification model is designed. The final assessment model gets the prediction result.

In recent years, the most widely used feature extraction methods in the prediction of protein subcellular locations are amino acid composition (Cedano et al., 1997), pseudo-amino acid composition (Cheng and Xiao, 2017a, 2017b; Yu et al., 2017; Zhang and Duan, 2018), position specific scoring matrix (Xiang et al., 2017; Zhai et al., 2017), pseudo-position specific scoring

matrix (Mei and Wang, 2010; Shen and Chou, 2007), Dipeptide composition (Bhasin and Raghava, 2004; Lin and Li, 2007) and gene ontology information (Cheng and Xiao, 2017c, 2017d; Chou and Cai, 2003; Du et al., 2012; Xiao et al., 2017) and so on. Nishikawa and Nakashima (Nakashima and Nishikawa, 1994) firstly proposed the amino acid composition (AAC) which is the characteristic expression method of protein sequence. It uses the simple 20-dimensional eigenvector to calculate the percentages of the 20 original amino acids contained in the protein sequence. Amino acid composition was initially used for the prediction of intracellular and extracellular proteins and subsequently extended to the field of protein structure and function prediction such as protein structure prediction (Saidijam et al., 2017) and protein subcellular localization prediction (Höglund et al., 2006). However, this method can't reflect the location and order information of the residues, and the order of amino acids in the primary structure of the protein is lost, thereby affecting the final prediction result. Huang and Li (Huang and Li, 2004) proposed that dipeptide composition can be used to characterize protein sequences. The frequency of two adjacent amino acid residues is used to describe the protein sequence and mapped into a 400-dimensional eigenvector. Although the amino acid sequence and coupling information are taken into account in extracting protein features, the dramatically increased number of dimensions, data calculated amount and noise can affect the advantages of the dipeptide composition. In order to remedy this defect, Chou (Chou, 2005) proposed the concept of pseudo amino acid composition (PseAAC), which is used in the prediction of protein subcellular locations. The PseAAC feature extraction method considers amino acid composition and similar protein sequences have similar structural types. It combines the sequence effect of the residues in the sequence with the physicochemical properties of the amino acids, reconstructs the PseAAC model by extending the information and combines it with the autocorrelation coefficient to construct the eigenvector. The eigenvectors of PseAAC usually exceed 20 dimensions. The first 20 features are the percentage of the original 20 amino acids in the protein sequence. The remaining features are used to reflect the order information of the protein sequence. At present, the method has been successfully applied in protein subcellular localization prediction (Cheng and Xiao, 2018; Cheng et al., 2017; Shi et al., 2008), protein structure prediction (Mei and Wang, 2010; Yu et al., 2017), protein-protein interaction prediction (Chou et al., 2016; Zhai et al., 2017), membrane protein recognition (Shen

et al., 2006), enzyme family classification (Cai and Chou, 2005; Lin et al., 2016) and protein function prediction (Bhasin and Raghava, 2004; Chou, 2011) and so on. Due to the evolution of protein sequences, some amino acid residue positions are conserved, and some residue positions are prone to mutation or amino acid substitutions, which will lead to changes in the function of proteins. Therefore, evolutionary information is introduced into the prediction of protein function. In 1999, Jones (Jones, 1999) proposed a position-specific iterative search algorithm and constructed a position-specific scoring matrix (PSSM) based on the algorithm, mainly filtering non-conserved sites based on PSSM and transforming matrix into a 20-dimensional vector. In recent years, with the rapid development of PseAAC feature extraction methods in various fields (Cheng and Xiao, 2017a, 2017b, 2017c; Chou and Shen, 2007b; Xiao et al., 2017) of protein function prediction, PseAAC has penetrated into PSSM, and the pseudo-position specific scoring matrix (PsePSSM) feature extraction method (Shen and Chou, 2007a) has emerged. The PsePSSM feature extraction method is used to generate a $20+20\times\xi$ dimension feature vector for each protein sequence. Then, the protein sequences of different lengths in the datasets are extracted through feature extraction into a vector with a uniform dimension. It promotes the further development of protein subcellular locations prediction research. In the field of bioinformatics, different biological databases may use different terms. In order to make the query consistency in different biological databases, Gene Ontology Consortium proposed Gene Ontology (GO) to construct database (Ashbumer et al., 2000). Among them, the database is divided into three parts: the biological process, the molecular function and the cellular component (Ashbumer et al., 2000). GO unit is term, what's more, each entry has a unique number tag and term name. This method is currently widely used in protein subcellular locations (Cheng and Xiao, 2017a, 2017b; Xiao et al., 2017) and protein-protein interactions (Deng et al., 2004) and other fields.

At present, based on classification algorithm of machine learning, it has made great progress in solving the prediction of protein subcellular location. The commonly used prediction algorithms include support vector machine (Uddin et al., 2018; Bhasin and Raghava, 2004; Xiang et al., 2017; Yu et al., 2017), K-nearest neighbor (Chon and Shen, 2006), Markov chain algorithm (Caragea et al., 2010), linear discriminant function (Chou and Elrod, 1998), neural network (Cai and Chou, 2000), genetic algorithm (Nanni and Lumini, 2008), Bayesian network (King et al., 2012) and

ensemble learning (Jia et al. 2015). Du et al. (Du and Li, 2006) firstly proposed the SubMito method to predict protein submitochondrial locations based on the sequence features of proteins and established a SVM localization prediction model. The overall prediction accuracy was 85.2%, the inner membrane prediction accuracy was 85.5%, the matrix prediction accuracy was 94.5%, but the prediction accuracy of the outer membrane was only 51.2%. They also released the first benchmarking dataset M317 in the paper, which is currently widely used to predict protein submitochondrial locations. Nanni and Lumini (Nanni and Lumini, 2008) constructed the GP-Loc method, using genetic algorithm and PseAAC to extract the feature of protein sequence, and then input it into the SVM classifier for location prediction. The overall prediction accuracy obtained was 4% higher than using the same dataset, but the accuracy of predicting inner membrane was 2.3% lower than that of the SubMito method. Zeng et al. (Zeng et al., 2009) established the PseAAC model based on the autocovariance algorithm, which combined the eight physicochemical properties of amino acids, and then imported the feature vectors extracted from the feature into the SVM classifier. The total accuracy of this method reached 89.7%, and the prediction accuracy of inner membrane was 7.6% higher than that of GP-Loc method. Du and Yu (Du and Yu, 2013) proposed the SubMito-PSPCP method through positional-specific physicochemical properties to predict protein submitochondrial locations and improved the overall prediction accuracy. Du et al. (Du and Jiao, 2017) also proposed a feature extraction method that combines functional domain enrichment score and positional-specific physicochemical properties (PSPCP) into SVM classifier. The overall prediction accuracy by jackknife test is higher, which is 93%. Lin et al. (Lin et al., 2013) used over-represented tetrapeptides selected in the protein sequence, getting optimal tetrapeptide combination by using binomial distribution, which has entered into the SVM classifier to predict protein submitochondrial locations. And a rigorous and low sequence homology benchmark dataset M495 and TetraMito predictor were constructed. In order to further improve the accuracy of prediction, Shi et al. (Shi et al., 2011) obtained the total prediction accuracy of protein submitochondrial location by combining SVM classifier and based on the discrete wavelet transform feature extraction method. The total accuracy reached 93.38%, especially mitochondrial outer membrane prediction accuracy was 82.93%. Zakeri et al. (Zakeri et al., 2011) fused 10 representative models of protein samples such as AAC, Dipeptide composition,

PseAAC, gene ontology information and the discrete models based on paired sequence similarity. They developed a Mito-Loc method based on an ordered weighted average with an overall prediction accuracy of 95%. Fan et al. (Fan and Li, 2012) predicted protein submitochondrial locations by combining AAC, dipeptide composition, gene ontology, evolutionary information and pseudo-average chemical shift, the overall prediction accuracy was 97.79%.

To develop a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications (Chen et al., 2016; Feng et al., 2017; Jia et al., 2015; Liu and Long, 2016; Liu and Xu, 2017; Liu et al., 2017; Liu and Yang, 2017; Xu and Li, 2017), one should observe the Chou's 5-step rule (Chou, 2011) i.e., making the following five steps very clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor; (ii) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) how to properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) how to establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one by one.

2. Materials and Methods

2.1. Datasets

That the high similarity between sequences may have an impact on the authenticity of the prediction results is considered. In order to make a reasonable comparison with the existing predictions of protein submitochondrial locations, three widely used and low-similarity protein submitochondrial datasets are selected for this study: M317, M495 and M983. Each dataset is divided into three regions: inner membrane, matrix, outer membrane. Fig. 1 shows the schematic structure of the protein submitochondrial.

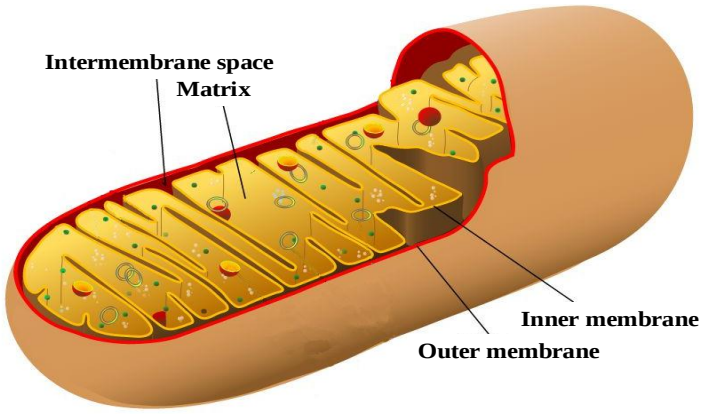

Fig. 1. Schematic illustration to show the locations of submitochondrial proteins: (1) inner membrane, (2) matrix, (3) outer membrane. Reproduced from website (Mariana, 2011).

The datasets M317 and M983 were constructed by Du and Li et al. (Du and Li, 2006; Du and Yu, 2013) with sequence similarity of no more than 40%. The dataset M495 was constructed by Lin et al. (Lin et al., 2013) with sequence similarity of no more than 25%. In order to test the performance of the model, we use the datasets M317 and M983 as the training datasets to select the parameters of the model, and the dataset M495 as the independent test dataset to test the model. The data composition of the three protein submitochondrial datasets are shown in Table 1.

Table 1

The detailed information of three protein submitochondrial datasets.

Datasets	Number of proteins			
	Inner membrane	Matrix	Outer membrane	Total
M317	131	145	41	317
M495	254	132	109	495
M983	661	177	145	983

2.2. Feature extraction

2.2.1. Pseudo-amino acid composition

Studies have shown that the structure and function of protein are related to the amino acid sequences that constitute them (Cedano et al., 1997; Cheng and Xiao, 2017a, 2017b). At present, researchers have proposed many feature extraction methods and developed corresponding online server and software (Liu et al., 2015; Liu et al., 2017; Shen and Chou, 2008) to promote the rapid development of protein subcellular locations prediction. Chou's pseudo-amino acid composition (PseAAC) (Chou, 2005) takes into account the order information of proteins and the

physicochemical properties of amino acids, which has been widely used in protein-protein interaction prediction (Chou et al., 2016; Zhai et al., 2017) and subcellular locations (Cheng and Xiao, 2018; Cheng and Zhao, 2017a, 2017b; Shi and Zhang, 2008; Zhu et al., 2015) and so on. They also developed the corresponding online server (Shen and Chou, 2008) to obtain different forms of pseudo-amino acid composition at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/. PseAAC uses $20+\lambda$ dimensional vector to represent the amino acid sequence of a protein. Among them, the first 20 dimensions are the frequency of occurrence in the sequence of the conventional 20 kinds of amino acids, and the latter λ dimension is sequence-related factors that reflect different levels of amino acid sequence information. Feature extraction process by PseAAC is shown as Eqs. (1), (2).

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

$$p_u=\begin{cases}\dfrac{f_\mu}{\sum_{\mu=1}^{20} f_\mu+\omega\sum_{k=1}^{\lambda}\tau_k}, 1\le\mu\le 20\\ \dfrac{\omega\tau_{\mu-20}}{\sum_{\mu=1}^{20} f_\mu+\omega\sum_{k=1}^{\lambda}\tau_k}, 21\le\mu\le 20+\lambda\end{cases} \tag{2}$$

where P is the feature vector, ω is the weight factor (Chou, 2001b), the value is 0.05. And f_μ is the frequency at which the μ-th amino acid in the protein appears. τ_k is the correlation factor of the k-th sequence, which can be calculated from the sequence correlation function in Eqs. (3) and (4).

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

$$\tau_k=\frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k}(k<L) \tag{4}$$

where L indicates the length of the protein sequence (the number of amino acids constituting the sequence), $H_1(R_i), H_2(R_i)$ indicate the hydrophobicity values and hydrophilicity values of the amino acid residues, respectively.

For the selection of parameter λ of PseAAC, $\lambda_{\max}$ is shorter than the length of the shortest sequence in the protein submitochondrial datasets. In this paper, the length of the shortest sequence in dataset M317 is 54, the length of the shortest sequence in dataset M983 is 102, and the length of the shortest sequence is 49 in dataset M495. So λ ranges from 0 to 49.

2.2.2. Pseudo-position specific scoring matrix

The pseudo-position specific scoring matrix (PsePSSM) feature extraction method was proposed by Chou and Shen (Chou and Shen, 2007a). PsePSSM not only considers the position information of the residues in the sequence, but also takes into account the biological significance of replacing residues in the sequence. It has been widely used in protein structure prediction (Chou and Shen, 2007a; Mei and Wang, 2010). Using the PsePSSM method to extract the features of a protein sequence, first, download the BLAST package locally at ftp://ftp.ncbi.nlm.nih.gov/blast/release/LATEST. Then call the appropriate executable file by the command line. Among them, PSI-BLAST, a multiple sequence homology alignment tool, is considered as the most widely used protein sequence similarity search algorithm, supports multiple iterations and allows to be set the threshold e to filter out protein sequences with lower similarity.

In this paper, we use the PSI-BLAST (Altschul et al., 1997) program to compare the imported submitochondrial protein sequences in FASTA format with the non-redundant protein database nr. The non-redundant protein database provided by NCBI, which contains about 85,107,862 protein sequences available at ftp://ftp.ncbi.nih.gov/blast/db/nr. In this paper, the parameter is set to 3 iterations, the value of e is equal to 0.001, and the other parameters adopt the default settings to obtain the PSSM of each protein sequence, as shown in Eq. (5).

$$P_{PSSM} = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,20} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ M_{i,1} & M_{i,2} & \cdots & M_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ M_{L,1} & M_{L,2} & \cdots & M_{L,20} \end{bmatrix} \tag{5}$$

where P_{PSSM} is a target protein matrix of $L \times 20$, L is the total number of residues of the target protein sequences, where each row represents the log-likelihood score of amino acid substitutions occurring at the corresponding positions in the query sequence. $M_{i,j}$ represents the score of the amino acid residue in the i-th position of the protein sequence being changed to amino acid type j-th during the evolution process. The PSSM matrix of protein sequence ranges from -9 to 11. In order to avoid the influence on the statistical analysis results due to the different dimension of the index, this paper first normalizes the PSSM matrix. The indicators in the matrix will be converted

to between 0 and 1 by Eq. (6).

$$f(x)=\frac{1}{1+e^{-x}} \tag{6}$$

Considering the sequence information of protein sequences, introducing Chou's PseAAC thought into PSSM, the P_{PSSM} of a single protein is converted into a feature vector $P_{PsePSSM}$ to represent the protein sequence. The lengths L of the protein sequences are different, and each protein sequence generates a matrix $L\times 20$. The PSSM matrices of different protein sequences need to be transformed into vectors with uniform dimensions so as to obtain PsePSSM, as shown in Eqs. (7) and (8).

$$\bar{M}_j=\frac{1}{L}\sum_{i=1}^{L}M_{i,j}\quad (j=1,2,\cdots,20) \tag{7}$$

$$G_j^{\xi}=\frac{1}{L-\xi}\sum_{i=1}^{L-\xi}\left[M_{i,j}-M_{(i+\xi),j}\right]^2\quad (j=1,2,\cdots,20,\xi<L) \tag{8}$$

In summary, PsePSSM method is used for protein sequence feature extraction, to obtain the Eq. (9).

$$P_{PsePSSM}=\left[\bar{M}_1\bar{M}_2\cdots\bar{M}_{20}G_1^{\xi}G_2^{\xi}\cdots G_{20}^{\xi}\right]^T\ (\xi=0,1,2,\cdots,L-1) \tag{9}$$

where $\bar{M}_j$ represents the average score of a protein sequence in which the amino acid residue at the i-th position in the evolution is mutated to j type of the amino acid. G_j^{ξ} is the correlation factor, which reflects the correlation between PSSM scores among amino acid residues separated by $i-1$ residues. And the value of ξ varies with different datasets. When $\xi=0$, $P_{PsePSSM}$ is transformed into a 20-dimensional eigenvector, which is given by Eq. (9). And a protein sequence can be expressed as an $20+20\times\xi$ dimensional vector that records the composition of a single protein without loss of sequence order information.

2.3. Two-dimensional wavelet denoising

Wavelet analysis is a new direction of rapid development in the field of mathematics in recent years. Its maximum advantage is that it can simultaneously analyze the signal in both time domain and frequency domain and has good ability of local information extraction, which have been made many achievements in the field of bioinformatics (Liò, 2003; Yu and Zhang, 2013a, 2013b; Yu et al., 2017). Wavelet denoising is one of the most important applications of wavelet analysis. It can

significantly improve the robustness of the prediction model when performing feature selection, which can reduce the amount of data learning, make the information more concentrate and understandable, and reduce the time of classification.

Wavelet denoising process is divided into three steps: wavelet decomposition, threshold quantization for decomposed high-frequency coefficients and signal reconstruction. Noise reduction depends on the selection of the wavelet function, decomposition scale, and threshold function. Among them, the most important thing is how to choose the threshold function and how to quantify the threshold (Chang et al., 2000a, 2000b). In the process of transforming protein sequences into numerical feature vectors of equal dimension by feature extraction, some redundant features and noise features are often included, leading to a decrease in prediction performance. Therefore, the feature selection using wavelet packet and wavelet function can eliminate redundant features, and apply the extracted feature information to various classifiers, which can effectively remove the redundant information and noise information. And it can extract characteristic signals of each type of protein itself which is of great importance for ensuring high prediction accuracy (Kandaswamy et al., 2004).

A two-dimensional (2-D) modelwith noise can be expressed as (Chang et al., 2000a):

$$s(i,j) = f(i,j) + \sigma e(i,j) \qquad i,j = 1,2,\cdots,m-1 \tag{10}$$

where (i,j) is the size of the signal, $f(i,j)$ is the true signal, $e(i,j)$ is the noise, $s(i,j)$ is the noise-containing signal and σ is the noise intensity. The purpose of wavelet noise reduction is to suppress $e(i,j)$, restore $f(i,j)$. Its specific decomposition process, as follows (Luisier et al., 2007):

First, the 2-D signal is decomposed by wavelet, using 2-D wavelet transform to decompose into four scales spatial frequency band. It obtains the orthogonal bases of protein sequences on different scales, is $f = f^1 \oplus d^H \oplus d^V \oplus d^D$. Among them, f^1 is the signal on the low frequency scale spatial frequency band, d^H is the signal on the horizontal scale spatial frequency band, d^V is the signal on the vertical scale spatial frequency band, and d^D is the signal on the diagonal scale space frequency band. Through the multi-resolution analysis of protein sequences, the low-frequency signal f^1 can be further decomposed into four directions: mitochondrial signal f^2 in the low-frequency scale space, signal $d^{1,H}$ in the spatial scale of the horizontal scale,

signal $d^{1,V}$ in the spatial scale of the vertical scale, signal $d^{1,D}$ in the spatial frequency band on the diagonal scale, and the decomposition process is still orthogonal decomposition $f^1 = f^2 \oplus d^{1,H} \oplus d^{1,V} \oplus d^{1,D}$. Since the signal energy of protein submitochondrial is finite, the number of decompositions of two-dimensional wavelet is also limited. The n-th wavelet decomposition process yields $f^{n-1} = f^n \oplus d^{n-1,H} \oplus d^{n-1,V} \oplus d^{n-1,D}$. Then, the decomposed high-frequency coefficients are threshold and threshold quantization are performed on the high-frequency coefficients of horizontal, vertical and diagonal directions from 1 to N by selecting an appropriate threshold. Finally, the 2-D signal is reconstructed. The reconstruction of 2-D signal is carried out according to the low-frequency coefficients of wavelet decomposition and the high-frequency coefficients after threshold quantization, so as to eliminate the interference noise of the effective signal.

The protein submitochondrial data used in this paper is a global feature. Firstly, fuse PseAAC and PsePSSM feature extraction methods to extract the features of each protein sequences, so that each sequence generates $(20+\lambda)+(20+20\times\xi)$ dimensional feature vectors. The $317\times((20+\lambda)+(20+20\times\xi))$ matrix and the $983\times((20+\lambda)+(20+20\times\xi))$ matrix is respectively generated for the datasets M317 and M983. Then 1-D wavelet denoising and 2-D wavelet denoising are respectively performed on the extracted protein feature vectors, which both use db6 wavelet function and the scale level is 6. Then the high-frequency coefficients after decomposition are threshold quantization. The thresholds are selected by wavelet function to obtain the default threshold. Finally, the signal reconstruction is performed based on the low-frequency coefficients of wavelet decomposition and the high-frequency coefficients after the threshold quantization.

In order to more intuitively compare the effects of 1-D wavelet denoising and 2-D wavelet denoising after feature extraction, this paper chooses the Q92406 protein in the inner membrane protein sequence as an example from the submitochondrial dataset M317 to get the result as shown in Fig. 2. As can be seen from Fig. 2, the Q92406 protein after 1-D wavelet denoising and 2-D wavelet denoising, using 2-D wavelet denoising not only can effectively remove redundant information in the protein sequences, but also extract characteristics signal of the protein itself which is more obvious than 1-D wavelet denoising effect. The decomposition of 1-D wavelet denoising band is not detailed enough. 2-D wavelet denoising can reduce the noise of the entire

dataset from low-frequency, horizontal, vertical and diagonal scale space. Decomposition is meticulous. Through many comparative analysis, we found that 2-D wavelet denoising can effectively improve the characteristics of each type of submitochondrial protein sequences. Therefore, we use 2-D wavelet denoising method for protein submitochondrial datasets.

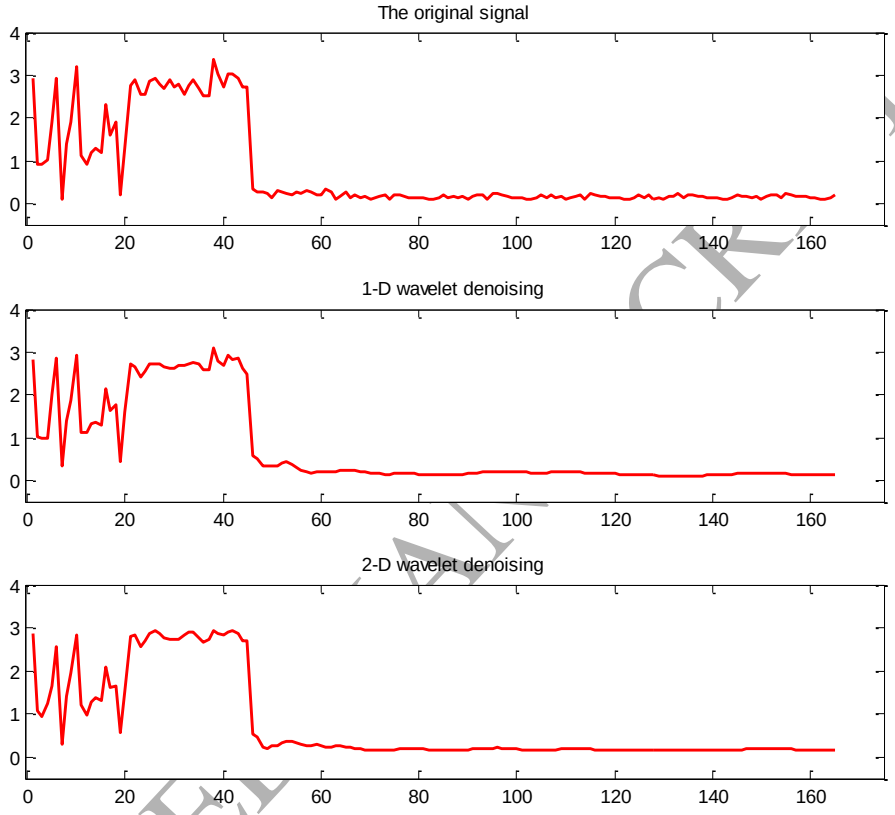

Fig. 2. Wavelet denoising using db6 wavelet function and decomposition scale of 6 on dataset M317 for >Q92406 protein. The y-axis represents the intensity of the signal and the x-axis represents the position of the sequence residues.

2.4. Support vector machine

SVM proposed by Vapnik et al. (Vapnik, 1995), is a machine learning method based on statistical learning theory. It has better generalization ability and higher ability of the classification accuracy in solving the problems of small sample, nonlinear and high dimensional bioinformatics data. SVM has been widely used in DNA function prediction (Kabir and Yu, 2017; Li et al., 2015), protein submitochondrial locations prediction (Du and Jiao, 2017; Du and Yu, 2013; Lin et al., 2013), protein subcellular locations prediction (Uddin, et al., 2018; Xiang et al., 2017; Yu et al., 2017), protein-protein interaction (Chou et al., 2016; Zhai et al., 2017) and protein structure

prediction (Mei and Wang, 2010; Saidijam et al., 2017; Yu et al., 2017) and other protein functions (Ding et al., 2013). Its basic principle (Vapnik, 1995) is to map the input sample set into high-dimensional space and construct the optimal hyperplane. By learning algorithm, it automatically finds out the support vectors that have good classification ability for the classification, making the distance between the hyperplane and two samples the largest, so as to achieve better classification results.

For a sample set $(x_i, y_i), i = 1,...n, x \in R^d, y \in \{+1, -1\}$ is satisfied as follow:

$$y_i\left[\left(\omega^t x_i\right) + b\right] - 1 \geq 0, i = 1,...,n \tag{11}$$

The optimal hyperplane with the largest margin can be found to separate y. For the problem of high-dimensional space classification, the SVM model should be able to correctly classify the two-class of samples and ensure the maximum classification interval. By using the appropriate kernel function $k(x_i, x_j)$, the inner product operation in high-dimensional space can be transformed into the function operation in low-dimensional space, the corresponding optimal classification of the decision-making function is also converted into:

$$f\left(x\right) = \text{sgn}\left\{\left(\omega, x\right) + b\right\} = \text{sgn}\left[\sum_{x_i \in sv} \alpha_i y_i k\left(x_i, x_j\right) + b\right] \tag{12}$$

The construction of kernel functions $k(x_i, x_j)$ is the key to SVM model. Different kernel functions directly affect the result of protein submitochondrial locations prediction. Commonly used kernel functions include linear kernel function, polynomial kernel function, radial basis function (RBF) and sigmoid kernel function and so on.

The protein submitochondrial locations prediction is a multi-class classification problem, and the number of different types of protein sequences in each dataset is unevenly distributed. Currently, SVM deals with multi-class classification problem. Frequently-used classification algorithms (Ding and Dubchak, 2001; Galar et al., 2011) are one-versus-one (OVO), one-versus-rest (OVR), direct acyclic graph (DAG) [83,84]. In this paper, we use OVR strategy in SVM to solve the multi-class classification problem. That is, to select any two types of dataset to establish SVM classifier, k class of samples need to build $k(k-1)/2$ SVM classifier. When categorizing an unknown sample, the category with the largest number of votes is the category

into which unknown samples are entered.

This study uses Chang and Lin (Chang and Lin, 2011) to develop the LIBSVM package, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.5. Performance evaluation and model building

In statistical theory, three validation tests are often used to test the validity of the model, which are self-consistency test, cross-validation test and jackknife test. Among them, jackknife test is considered the most objective evaluation method and widely used by the majority of researchers (Du and Jiao, 2017; Du and Yu, 2013; Lin et al., 2013; Shi et al., 2011) In this paper, we use jackknife test to evaluate the performance of the prediction model. The prediction models are evaluated by four evaluation metrics, which are sensitivity (Sn), specificity (Sp), overall accuracy (ACC) and Matthew's correlation coefficient (MCC). Those metrics are defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{FP + TN} \tag{14}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{16}$$

where TP represents correctly predicted sequence number of j-class submitochondrial protein locations and TN indicates the number of correctly predicted non-j-class submitochondrial protein sequences. FN represents the incorrectly predicted number of j-class submitochondrial protein locations. FP is the number of non-j-class submitochondrial protein locations but predicted to be the j-class of sequences, and N is the total number of sequences in the dataset.

But the above four metrics in Eqs. (13)-(16) are not easily understood by most biologists, particularly the one for Matthew's correlation coefficient (MCC). In order to evaluate the prediction model more intuitively, this paper uses Chou's symbols in studying signal peptides (Chou, 2001a). The aforementioned four metrics can be easily converted into a set of following formulations (Chen et al, 2013; Xu et al., 2013):

$$Sn = 1 - \frac{N_{-}^{+}}{N^{+}} \tag{17}$$

$$Sp = 1 - \frac{N_+^-}{N^-} \tag{18}$$

$$ACC = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{19}$$

$$MCC = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \tag{20}$$

where N^+ represents the number of positive samples, N^- represents the number of negative samples, N_-^+ represents the number of false negative samples, and N_+^- represents the number of false positive samples.

When $N_-^+ = 0$, it means that none of the true j-class protein submitochondrial locations are incorrectly predicted as non-j-class positions, we have the sensitivity $Sn = 1$. When $N_-^+ = N^+$, it means that all of the true j-class protein submitochondrial locations are incorrectly predicted as non-j-class positions, we have the sensitivity $Sn = 0$. However, when $N_+^- = 0$, it means that none of the non-j-class protein submitochondrial locations are incorrectly predicted as j-class positions, we have the specificity $Sp = 1$. When $N_+^- = N^-$, it means that all of the non-j-class protein submitochondrial locations are incorrectly predicted as j-class positions, we have the specificity $Sp = 0$. From the above discussion, it is much more intuitive and easier to understand when using the Eqs. (17)-(20) to examine a predictor for its sensitivity, specificity, overall accuracy, and Matthew's correlation coefficient. Particularly, its advantages have been analyzed and endorsed by a series of studies published very recently (Chen et al., 2015; Chen et al., 2016; Chen et al., 2017; Ehsan et al., 2018; Feng et al., 2018; Jia et al., 2015; Jia et al., 2016; Liu et al., 2018; Qiu et al., 2017a, 2017b) .

However, the set of metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology (Cheng and Xiao, 2017a, 2017b, 2017c, 2018; Cheng et al., 2017, Xiao et al., 2017) and system medicine (Cheng and Zhao, 2017a, 2017b) and biomedicine (Qiu and Sun, 2016), a completely different set of metrics as defined by Chou (Chou, 2013).

In addition, the ROC curve is a comprehensive index to evaluate the performance of the

model. The mapping of true positive rate and false positive rate when combining different critical points reflects the relationship between sensitivity and specificity under different critical values. The area under the ROC curve (AUC) is a reliable measure of predictive performance, with values ranging from 0 to 1. In general, the larger AUC value is, the better model prediction performance is.

For convenience, the protein submitochondrial locations prediction method proposed in this paper is called PseAAC-PsePSSM-WD. The general framework is shown in Fig. 3. We have implemented it in MATLABR2014a in windows Server 2012R2 running on a PC with system configuration Intel (R) Core (TM) i5-3210M CPU @ 2.50GHz 2.50 GHz 12.0GB of RAM.

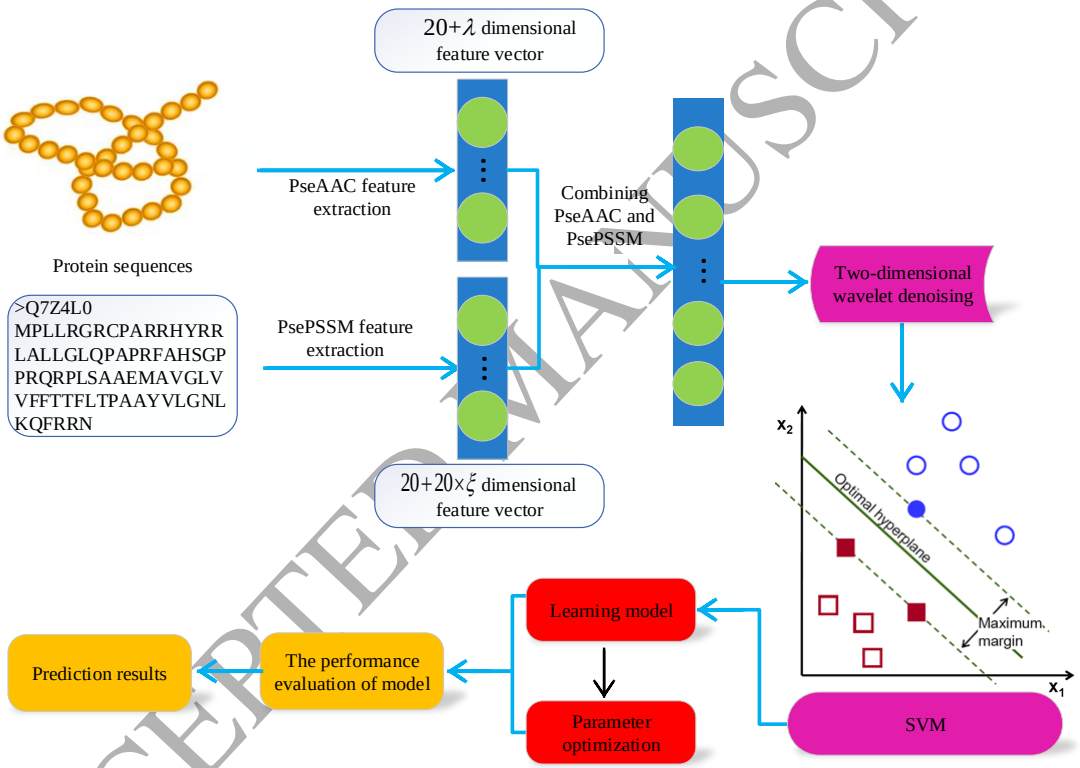

Fig. 3. Flow chart protein submitochondrial locations prediction based on PseAAC-PsePSSM-WD method.

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

1) Input protein sequences of two submitochondrial datasets M317 and M983, respectively, and the corresponding class labels;

2) Use PseAAC online service system developed by Chou and Shen to carry out feature extraction on the protein sequences and to obtain the $20+\lambda$ dimension feature vectors, where

$0 \le \lambda \le 49$;

3) PSI-BLAST tool is used to process each protein sequence to generate the corresponding PSSM, and the PsePSSM method is used to extract the features;

4) Wavelet denoising method is performed on the feature vectors obtained by fusion of PseAAC and PsePSSM and the redundant information in the sequence is removed to obtain the optimal feature vectors;

5) Input the optimal feature vectors after noise reduction into SVM classifier, and use jackknife test to predict the protein submitochondrial locations;

6) According to the obtained prediction accuracy, determine the optimal parameters of the model, including the value of λ and ξ, wavelet function selection, different decomposition scale and kernel function in SVM;

7) According to 6) the optimal model parameters, the Sn, Sp, ACC, MCC and AUC values are calculated and the predictive performance of the model is evaluated;

8) Using the model constructed in 1) -7), the independent test set M495 is used to test the prediction model.

3. Results and discussion

3.1. The choice of model parameters

3.1.1. Selection optimal parameters λ of PseAAC method

According to Chou's PseAAC concept (Chou, 2005), the choice of the weight factor ω and the number of amino acid residue intervals λ is an essential parameter for feature extraction. In this paper, we use the PseAAC method in the feature extraction process, the main consideration is to select the number of amino acid residue intervals λ. If the value of λ is too small to fully extract the features of submitochondrial protein sequences, the eigenvectors contain relatively little sequence information. If the value of λ is set too high, it will lead to more redundant information and reduce the accuracy of the prediction. If the value of λ exceeds the length of any one protein sequence in the dataset, the feature extraction of the protein will be meaningless. The shortest protein sequence length for the three standard submitochondrial datasets is 49, so we elect λ as the maximum value of 49. In order to find the optimal value of λ in the model, we set λ values from 0 to 49 in the protein submitochondrial datasets M317 and M983 respectively.

The SVM classifier is used to classify with RBF kernel function. The results are validated by jackknife test. The overall prediction accuracy of the two datasets at different values of λ and the average overall prediction accuracy of the two datasets are shown in Table 2.

Table 2

The prediction results of submitochondrial locations by selecting different λ.

Datasets	λ									
	0	5	10	15	20	25	30	35	40	45
M317	82.02	84.23	84.23	83.28	84.86	**85.80**	85.17	84.54	83.91	83.91
M983	71.62	74.36	74.06	73.75	74.16	**75.28**	74.67	73.86	74.26	74.26
Average ACC	76.82	79.30	79.15	78.52	79.51	**80.54**	79.92	79.20	79.09	79.09

As can be seen from Table 2, using the PseAAC method to predict protein submitochondrial locations, we only change the λ value when other conditions remain unchanged, which has a great impact on the prediction results. As the value of λ constantly changes, the accuracy of prediction also varies. Among them, the prediction accuracy of dataset M317 is between 82.02% and 85.80%, and when $\lambda = 25$, the highest prediction accuracy reaches 85.80%. The prediction accuracy of the dataset M983 is between 71.62% and 75.28%, and the highest prediction accuracy is 75.28% when $\lambda = 25$. The specific prediction results of dataset M317 and M983 in the inner membrane, matrix and outer membrane when using the PseAAC feature extraction method are shown in Supplementary materials Table S1, Table S2. When both datasets reach the highest average overall prediction accuracy parameter $\lambda = 25$. In order to more accurately evaluate the prediction results of PseAAC feature extraction methods under different λ, this paper considers the average prediction accuracy of two datasets. When $\lambda = 25$, M317 and M983 datasets achieve the highest average overall prediction accuracy of 80.54%.

In order to more intuitively find the optimal λ value, Fig. 4 shows the change of the overall prediction accuracy when different λ values are selected for datasets M317and M983. As can be seen from Fig. 4, as the value of λ changes, the prediction accuracy of the two datasets is also changing, and the datasets M317 and M983 achieve the highest average overall prediction accuracy when $\lambda = 25$. In summary, the optimal parameter λ value of 25, the use of PseAAC method for feature extraction of protein sequences, each of the protein sequence generates $20 + \lambda = 45$ dimensional feature vectors.

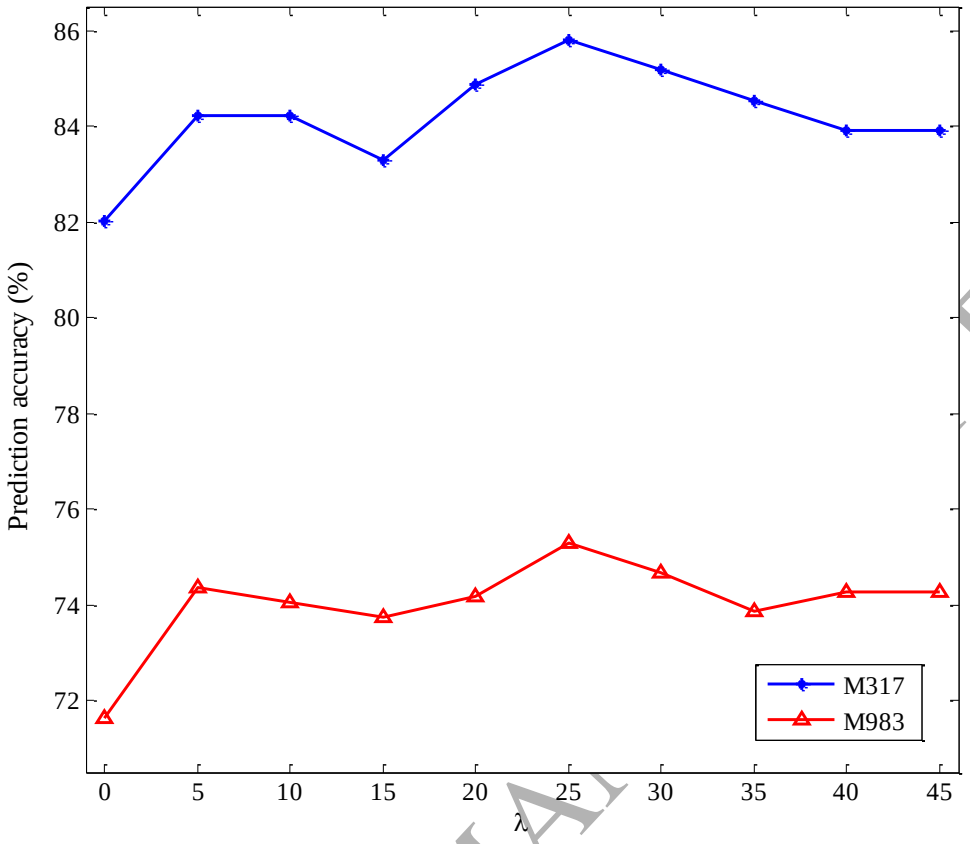

Fig. 4. Effect of selecting different values of λ on the overall prediction results of submitochondrial locations for datasets M317 and M983.

3.1.2. Selection optimal parameters ξ of PsePSSM method

In recent years, many researchers have used the PsePSSM method to predict protein subcellular locations (Shen and Chou, 2007; Mei and Wang, 2010). It not only considers the sequence information of protein sequences, but also takes into account the evolutionary information of proteins. The results show that it is an effective method of discretization of protein sequences. The value of parameter ξ in evolutionary information directly affects the prediction accuracy. In this paper, we use the PsePSSM method to extract protein sequence features of submitochondrial. In order to find the optimal ξ value in the model, we set the ξ value from 0 to 10 in the protein submitochondrial datasets M317 and M983, and use the RBF kernel function in the SVM classifier to predict. The results are validated by jackknife test. In order to make the choice of parameters in the model consistent, the average overall prediction accuracy of two datasets is calculated for further analysis. Table 3 shows the overall prediction accuracy and the

average overall prediction accuracy of the two datasets.

Table 3

The prediction results of submitochondrial locations by selecting different ξ.

Datasets	ξ										
	0	1	2	3	4	5	6	7	8	9	10
M317	84.86	86.44	86.12	87.07	87.07	**88.01**	87.07	87.07	87.38	86.75	86.75
M983	82.10	82.91	83.32	83.42	83.21	83.42	83.42	83.52	83.42	83.83	83.62
Average ACC	83.48	84.68	84.72	85.25	85.14	**85.72**	85.25	85.30	85.40	85.29	85.19

As can be seen from Table 3, using PsePSSM method to predict protein submitochondrial locations when other conditions remain unchanged, the prediction accuracy changes with ξ value changing. In dataset M317, the prediction accuracy is between 84.86% and 88.01%, and the highest prediction accuracy is 88.01% when $\xi = 5$. In the dataset M983, the prediction accuracy rate is between 82.10% and 83.83%, and the highest prediction accuracy rate is 83.83% when $\xi = 9$. The specific prediction results of datasets M317 and M983 in the inner membrane, matrix and outer membrane using the PsePSSM feature extraction method are shown in Supplementary materials Table S3, Table S4. Since the two datasets have different ξ values when the highest prediction accuracy is reached, the average overall prediction accuracy of the two datasets is taken into account and the ξ value corresponding to the highest average overall accuracy is chosen as the optimal parameter.

When $\xi = 5$, the highest average overall prediction accuracy of datasets M317 and M983 is 85.72%. Fig. 5 shows the change of the overall prediction accuracy when different values of ξ are selected for the two datasets. Therefore, with the optimal parameter ξ value of 5 and the use of the PsePSSM method, each protein sequence can be extracted $20 + 20 \times \xi = 120$ dimensional feature vectors.

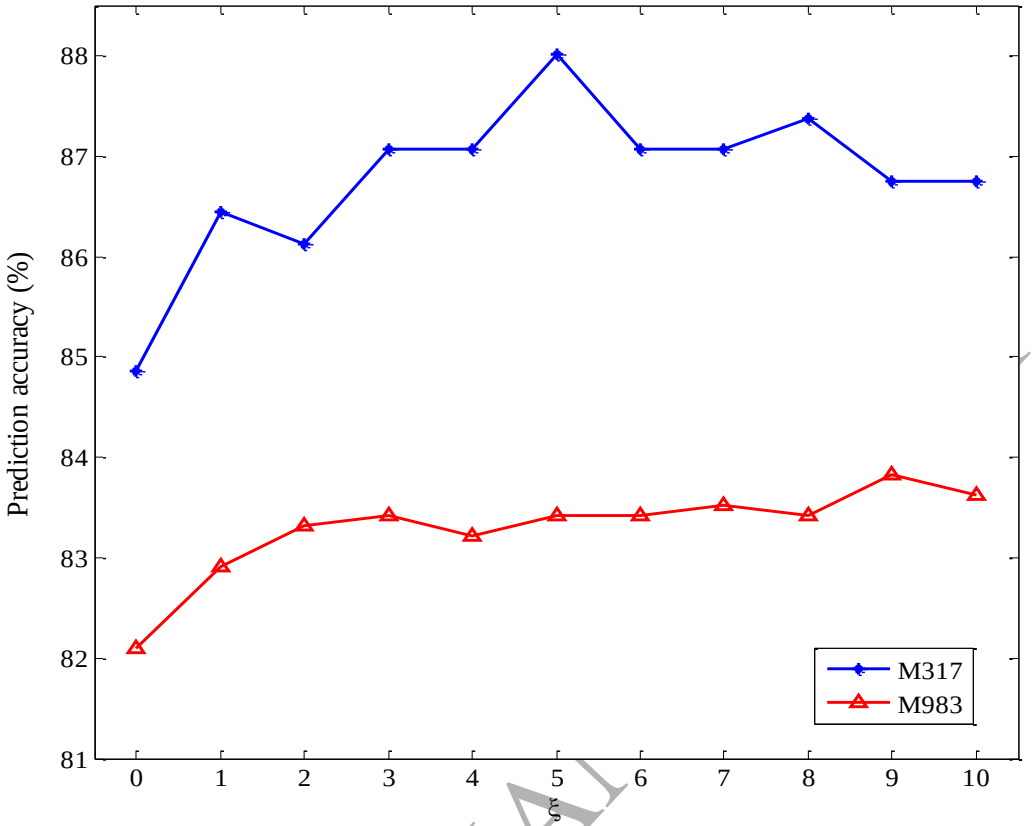

Fig. 5. Effect of selecting different values of ξ on the overall prediction results of submitochondrial locations for datasets M317 and M983.

3.2. Selection of wavelet functions and optimal decomposition scale

In protein submitochondrial locations prediction, the use of wavelet analysis for the decomposition of the obtained protein amino acid signal can effectively remove redundant information. In order to achieve the ideal prediction accuracy, we combine PseAAC and PsePSSM to extract the features of protein sequences. Each of the two datasets generated $(20+\lambda)+(20+20\times\xi)=45+120=165$ dimensional feature vector respectively, and then using wavelet denoising method to further extract the information. When using 2-D wavelet denoising to carry out feature extraction on datasets, the commonly used threshold model has a default threshold model and the threshold model obtained by the Birge-Massart strategy (Chang et al., 2000a, 2000b). Among them, the default threshold model can be obtained by selecting the wavelet function or wavelet packet to determine the default threshold. In the process of noise reduction, the wavelet coefficients are processed by threshold processing. The soft threshold is to make the continuity of the signal as a whole in a smooth way and to produce a smoother effect after noise

reduction. Hard thresholding preserves only some spikes in the signal, losing some of the information of the original signal itself. Therefore, we select soft threshold wavelet denoising method. For submitochondrial datasets M317 and M983, after many experiments, it is found that using the wavelet function to obtain the default threshold under different decomposition scales and different wavelet functions can obtain the highest prediction accuracy on the two datasets simultaneously (The specific results can be found in Supplementary materials Table S5, Table S6 and Fig. S1). In the following discussion, we use wavelet functions to obtain default thresholds for 2-D wavelet denoising.

Different wavelet functions have different families of wavelets, and different signal processing capabilities. At the same time, different decomposition scale analysis of protein sequences will have an impact on the prediction accuracy of the model. Decomposing a short sequence on a large scale will introduce redundant information, while decomposing a long sequence on a small scale will omit many details (Kandaswamy et al., 2004). Therefore, the choice of wavelet function and decomposition scale becomes an important step in the optimal performance of signal processing. In order to extract characteristic information of submitochondrial protein sequence more effectively, we select different wavelet functions and different decomposition scales for submitochondrial datasets M317 and M983, and use default thresholds obtained by wavelet function. RBF kernel function with SVM classifier is used to classify and the results are tested by jackknife test. In this paper, we select 11 representative wavelet functions db2, db6, db8, sym2, sym5, sym8, coif2, coif4, bior2.6, bior3.7 and bior6.8, select the decomposition scale from 3 to 6. The overall predicted accuracy of protein submitochondrial are obtained on datasets M317 and M983, as shown in Table 4 and Table 5.

Table 4

Predicted results submitochondrial locations on dataset M317 under different wavelet functions and decomposition scales.

Different wavelet functions	Jackknife test (%)			
	Scales			
	3	4	5	6
db2	99.68	99.37	99.37	99.37
db6	**100**	99.05	99.37	**100**
db8	99.68	99.37	**99.68**	99.68
sym2	99.68	99.37	99.37	99.37

sym5	99.37	99.05	99.37	99.37
sym8	**100**	99.37	99.37	99.68
coif2	**100**	99.37	**99.68**	99.68
coif4	**100**	99.37	**99.68**	99.37
bior2.6	**100**	99.37	99.37	99.05
bior3.7	99.05	**99.68**	**99.68**	99.68
bior6.8	99.68	99.68	99.05	98.42

As can be seen from Table 4, for the protein submitochondrial dataset M317, the overall accuracy of the model is between 98.42% and 100% under different wavelet functions and different decomposition scales. When the decomposition scale is 3, the highest overall prediction accuracy of protein submitochondrial locations is 100%, and the corresponding wavelet functions are db6, sym8, coif2, coif4 and bior2.6. When the decomposition scale is 4, the highest overall prediction accuracy is 99.68% and the corresponding wavelet functions are bior3.7 and bior6.8. When the decomposition scale is 5, the accuracy of the db8, coif2, coif4 and bior3.7 wavelet functions reaches the highest overall prediction accuracy of 99.68%. When the decomposition scale is 6, the db6 wavelet function is selected to obtain the highest overall prediction accuracy of 100%.

Table 5

Predicted results submitochondrial locations on dataset M983 under different wavelet functions and decomposition scales.

Different wavelet functions	Jackknife test (%)			
	Scales			
	3	4	5	6
db2	95.93	97.76	98.78	98.17
db6	96.74	98.98	**99.49**	**99.29**
db8	96.74	98.07	98.88	98.78
sym2	95.93	97.76	98.78	98.17
sym5	96.13	97.86	97.97	98.07
sym8	96.13	98.58	99.08	98.98
coif2	96.13	98.58	98.78	98.78
coif4	96.24	97.76	98.78	98.98
bior2.6	96.24	98.17	98.58	98.68
bior3.7	**98.68**	**99.29**	99.19	99.19
bior6.8	97.15	98.58	99.19	98.58

It can be seen from Table 5 that for the dataset M983, the accuracy of protein submitochondrial location prediction is between 95.93% and 99.69%. When the decomposition scale is 3, bior3.7 wavelet function is selected to obtain the highest overall prediction accuracy of

98.68%. When the decomposition scale is 4, the highest overall prediction accuracy of 99.29% is obtained using bior3.7 wavelet. When the decomposition scale is 5, db6 wavelet function is selected to obtain a higher overall prediction accuracy of 99.49%. When the decomposition scale is 6, the highest overall prediction accuracy of db6 wavelet is 99.69%, which is 1.22% higher than the prediction accuracy of sym5 wavelet function.

It can be seen from Table 4 and Table 5 that when the wavelet function is selected as db6 and the decomposition scale is 6, the overall accuracy of the two submitochondrial datasets M317 and M983 are 100% and 99.29%, respectively. In order to obtain the optimal performance of the model in wavelet denoising, we need to consider the wavelet functions and decomposition scales corresponding to the optimal prediction results of both two datasets under different wavelet scales and different decomposition scales. In order to more intuitively analyze the best wavelet function and decomposition scale in the two datasets, a bar graph of the overall prediction accuracy of protein submitochondrial locations prediction under different decomposition scales and different wavelet functions is drawn, as shown in Fig. 6 and Fig. 7.

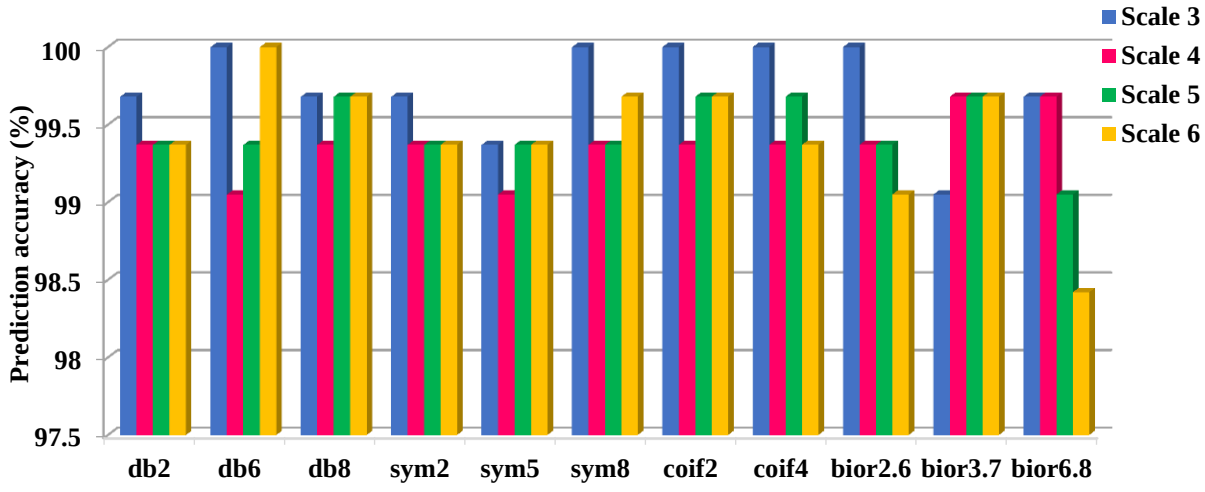

Fig. 6. Prediction performance of dataset M317 under different wavelet functions and different scales.

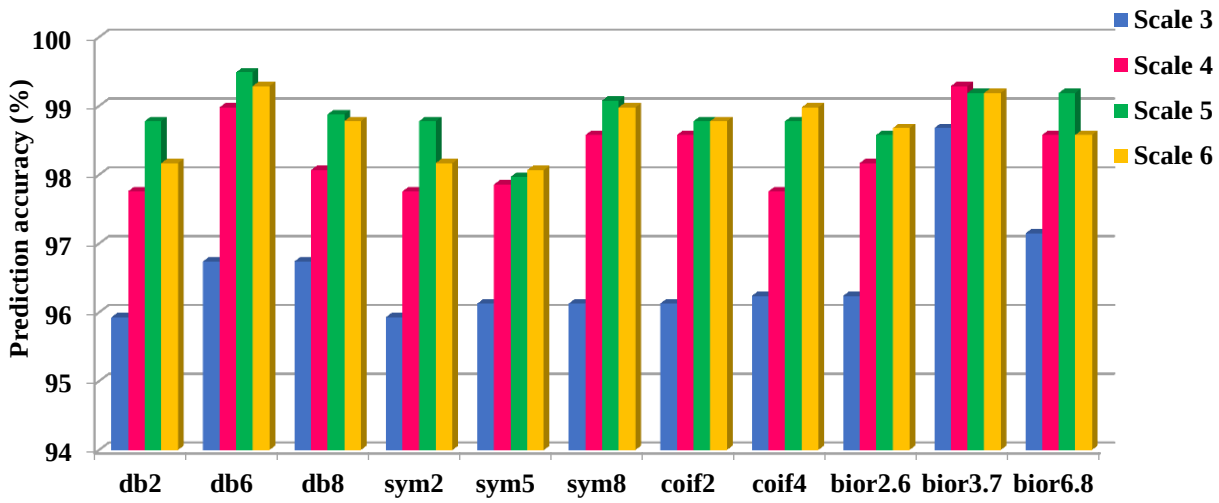

Fig. 7. Prediction performance of dataset M983 under different wavelet functions and different scales.

As can be seen from Fig. 6 and Fig. 7, for the datasets M317 and M983, the highest overall prediction accuracy is obtained when the decomposition scale is 6 and the db6 wavelet function is selected, which is obviously better than the results of other wavelet functions. This is because the dbN wavelet function has good regularity. The smooth error introduced by it as a sparse basis is not easy to be noticed, making the signal reconstruction process more smoothly, the location ability in the frequency domain is stronger, and the band division effect is better. Therefore, we choose db6 wavelet function as the optimal wavelet function, the optimal decomposition scale is 6, extract the information from the protein sequences in the dataset. In this way, it makes extraction of feature vector with a better way to express the original sequence information and improve the performance of protein submitochondrial locations prediction.

3.3. Effect of feature extraction algorithm on results

The selection of feature vector is particularly important for a predictive model. How to extract valid information from protein sequences becomes the key to the success or failure of the algorithm. PseAAC algorithm is based on amino acid residues for feature extraction algorithm, PsePSSM algorithm is based on sequence homology for feature extraction algorithm. The fusion of PseAAC algorithm and PsePSSM algorithm gets more information about the protein sequences, but also brings redundant variables. The 2-D wavelet denoising is used to remove the redundant information in the protein sequences and the characteristic signal of each type of protein is extracted. PseAAC algorithm is used for feature extraction. When selecting the best parameters

$\lambda = 25$, 45-D feature vectors are generated. PsePSSM algorithm is used for feature extraction. When selecting the best parameters $\xi = 5$, 120-dimensional feature vectors are generated. The two algorithms of PseAAC and PsePSSM are combined to generate a 165-dimensional feature vectors. For wavelet denoising, we choose db6 wavelet function, the decomposition scale is 6. PseAAC, PseAAC-PsePSSM, PseAAC-PsePSSM and PseAAC-PsePSSM-WD, four kinds of feature extraction methods are classified by RBF kernel function using SVM algorithm. The predicted results are got by jackknife test for protein submitochondrial locations in datasets M317 and M983 which are shown in Table 6 and Table 7.

In order to more intuitively compare the prediction results of the two datasets under the four feature extraction methods, we use the ROC curve to compare the prediction performance of the model under different feature extraction algorithms. ROC curves are generally used to evaluate the binary classifier, but protein submitochondrial locations is a multi-class prediction problem. In this paper, a one-versus-rest (OVR) strategy is used to convert the multi-classification problem into two-classification problems. Then, the true positive rate and the false positive rate of all the two classes are averaged as the final output of the method. Fig. 8 and Fig. 9 show the ROC curves of the datasets M317 and M983 respectively under the four different feature extraction methods.

Table 6

Comparison of different feature extraction algorithms for prediction results of dataset M317.

Algorithm	Jackknife test (%)			
	Inner membrane	Matrix	Outer membrane	ACC
PseAAC	87.02	97.24	41.46	85.80
PsePSSM	90.08	97.24	48.78	88.01
PseAAC-PsePSSM	92.89	67.80	64.83	84.23
PseAAC-PsePSSM-WD	100	100	100	**100**

Table 7

Comparison of different feature extraction algorithms for prediction results of dataset M983.

Algorithm	Jackknife test (%)			
	Inner membrane	Matrix	Outer membrane	ACC
PseAAC	94.55	36.16	35.17	75.28
PsePSSM	93.04	68.36	57.93	83.42
PseAAC-PsePSSM	90.08	99.31	48.78	88.96
PseAAC-PsePSSM-WD	99.55	98.87	98.62	**99.29**

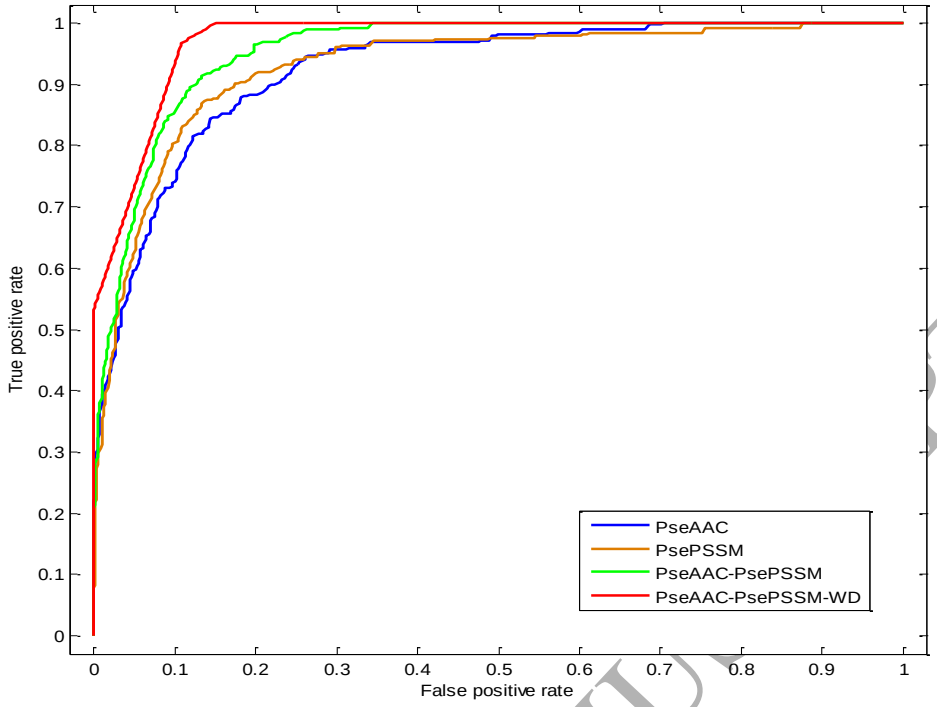

Fig. 8. This graph shows the ROC curve of the protein submitochondrial dataset M317.

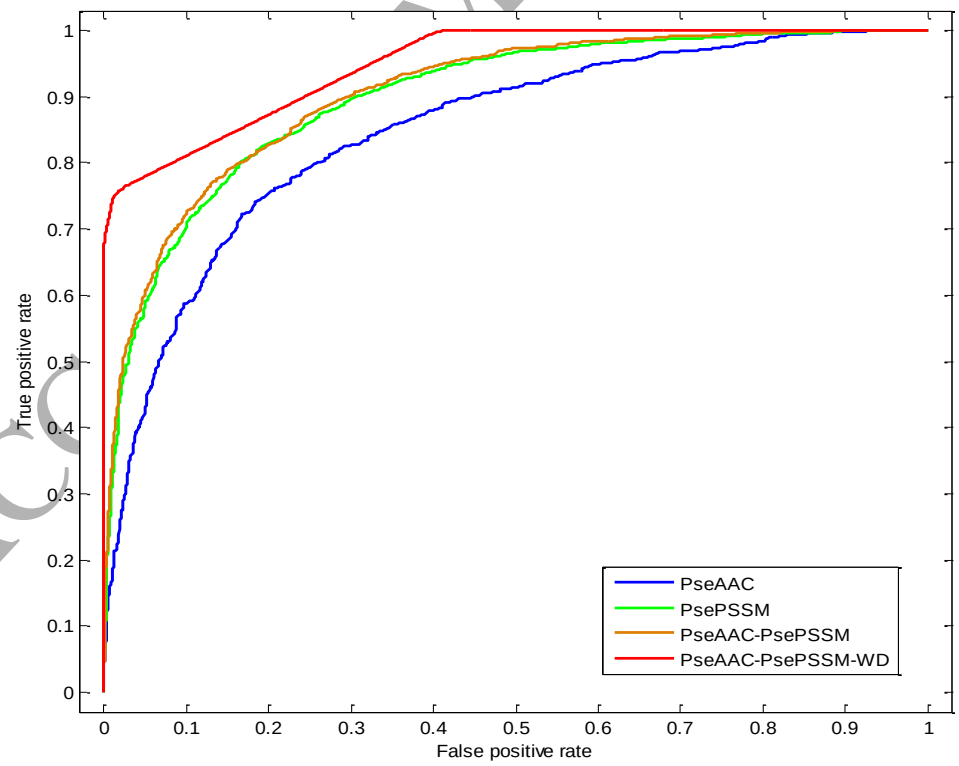

Fig. 9. This graph shows the ROC curve of the protein submitochondrial dataset M983.

As can be seen from Table 6, there are significant differences in the prediction results of protein submitochondrial localization using different feature extraction algorithms. In dataset M317, the overall prediction accuracy obtained by the PseAAC and PsePSSM feature extraction algorithms was 85.80% and 88.01%, respectively. Among them, the overall prediction accuracy obtained by using PsePSSM algorithm is 2.21% higher than that obtained by using PseAAC algorithm, probably because the PsePSSM algorithm combines the original information such as sequence and evolution to improve the prediction accuracy. After the fusion of PseAAC and PsePSSM, the overall prediction accuracy is 84.23%, which is 1.57% lower than PseAAC and 3.78% lower than PsePSSM. This may be because the two algorithms only extract more information from the protein sequence, which cannot eliminate the redundant information and reduce the prediction accuracy. The overall prediction accuracy is 100% using PseAAC-PsePSSM-WD feature extraction algorithm, which is much higher than the other three feature extraction algorithms, which is 14.20% higher than the PseAAC algorithm with the lowest accuracy rate and 11.99% higher than PsePSSM algorithm, 15.77% higher than PseAAC-PsePSSM algorithm. The predicted results of the submitochondrial structure of the three proteins have been significantly improved, with the prediction accuracy of 100% respectively. Among them, the sensitivity of this method to the prediction of the location of the inter membrane is 9.92% higher than that of the PsePSSM algorithm, and the sensitivity to the prediction of the localization of the matrix is increased by 32.20% compared with the PseAAC-PsePSSM algorithm. The sensitivity to the prediction of the outer membrane increased by 58.54% than that using PseAAC algorithm. As can be seen from Fig. 8, the area AUC of the ROC curve using the PseAAC-PsePSSM-WD feature extraction algorithm for the dataset M317 is 0.9724, which is significantly higher than that of the other three methods. The AUC values of PseAAC, PsePSSM, PseAAC-PsePSSM are 0.9241, 0.9290 and 0.9548, respectively.

It can be seen from Table 7 that in the dataset M983, the PseAAC algorithm is used to get the overall prediction accuracy of 75.28%, which is 8.14% lower than that of the PsePSSM algorithm. The overall prediction accuracy of PseAAC-PsePSSM algorithm is 88.96%, which is 5.66% and 0.41% higher than that of PseAAC algorithm and PsePSSM algorithm respectively. Compared with the PseAAC, PsePSSM and PseAAC-PsePSSM algorithms, the accuracy of the 2-D wavelet

denoising of the feature vectors fused by the two algorithms is 99.29%, which are respectively 24.01%, 15.87% and 10.33% higher than the PseAAC-PsePSSM algorithm. Among them, the sensitivity of our prediction method, for the prediction of the inner membrane, compared with the use of PseAAC-PsePSSM algorithm increased by 9.47%. For the prediction of the matrix location, the sensitivity obtained by using PseAAC algorithm increased by 62.71%. The outer membrane locations prediction, the sensitivity obtained by using PseAAC algorithm increased by 63.45%. It can be seen from Fig. 9 that for the M983 dataset, the coverage area of the ROC curve obtained by using the PseAAC-PsePSSM-WD feature extraction algorithm is the largest, that is, the maximum AUC value is 0.9485. It is significantly higher than the other three feature extraction algorithms. The AUC values of PseAAC, PsePSSM and PseAAC-PsePSSM are 0.8483, 0.8992 and 0.9057, respectively. By comparing the effect of different feature extraction algorithms on the results and combining the robustness of ROC curves to the four different feature extraction algorithms, PseAAC-PsePSSM-WD is the optimal feature extraction algorithm in this paper.

3.4. Effect of kernel function on results

The performance of SVM using different kernel functions is quite different. The selection of kernel function is the key to using SVM classifier. This paper uses PseAAC-PsePSSM feature extraction, select $\lambda = 25$, $\xi = 5$. 2-D wavelet denoising using db6 wavelet function, decomposition scale of 6. The different kernel functions are chosen to obtain the predictions of protein submitochondrial location in datasets M317 and M983, as shown in Table 8 and Table 9. Where 0 represents a linear kernel function, 1 represents a cubic polynomial kernel function, 2 represents a RBF kernel function, and 3 represents a sigmoid kernel function. Fig. 10 shows the overall prediction accuracy of two datasets under different kernel functions.

Table 8

Prediction results of submitochondrial locations of different kernel functions in dataset M317.

Locations	Jackknife test (%)			
	Different kernel functions			
	0	1	2	3
Inner membrane	100	87.02	100	100
Matrix	99.31	100	100	100
Outer membrane	97.56	70.73	100	100
ACC	99.37	90.85	100	100

As can be seen from Table 8 that for dataset M317, the prediction effect using RBF kernel function and sigmoid kernel function is obviously better than the other two kernel functions. And the overall prediction accuracy obtained is 100%. The sensitivity to the predicted location of inner membrane, matrix and outer membrane are 100%, respectively. The overall prediction accuracy obtained by linear kernel function and cubic polynomial kernel function are 99.37% and 90.85%, respectively, which are 0.63% and 9.15% lower than those obtained by using RBF kernel function and sigmoid kernel function respectively. Among them, the linear kernel function of the inner membrane prediction is better. Sensitivity reaches 100%. Sensitivity is 100% by using the cubic polynomial kernel function to predict matrix.

Table 9

Prediction results of submitochondrial locations of different kernel functions in dataset M983.

Locations	Jackknife test (%)			
	Different kernel functions			
	0	1	2	3
Inner membrane	99.24	99.70	99.55	93.19
Matrix	98.87	74.01	98.87	88.14
Outer membrane	98.62	80.00	99.62	63.45
ACC	99.08	92.17	99.29	87.89

It can be seen from Table 9 that for the dataset M983, the predicted overall accuracy of submitochondrial location using the RBF kernel function is 99.29%, which is 11.4% higher than the overall prediction accuracy obtained using the sigmoid kernel function. Among them, the prediction sensitivity of the RBF kernel function to the location of inner membrane, matrix and outer membrane are 99.55%, 98.87% and 99.62%, respectively, far higher than those predicted by other kernel functions. In addition, the overall prediction accuracy obtained by linear kernel function and cubic polynomial kernel function are 99.08% and 92.17% respectively, which are 0.21% and 7.12% lower than the overall prediction accuracy obtained by using RBF kernel function respectively. As can be seen more intuitive in Fig. 10, when using the SVM classifier to classify datasets M317 and M983, the prediction using RBF kernel function is significantly better than the other kernel function. Considering the impact of using different kernel functions on datasets M317 and M983, RBF kernel function is selected as the kernel function of SVM classifier.

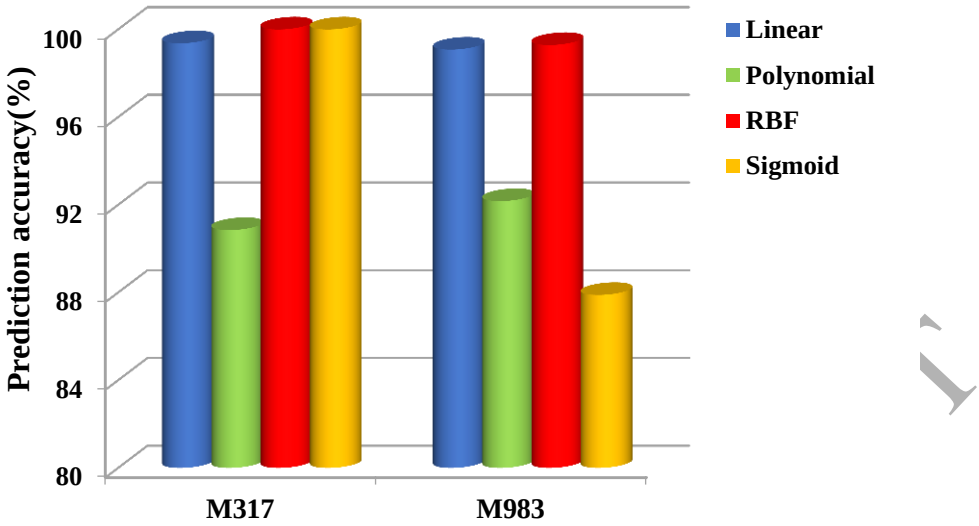

Fig. 10. The overall prediction accuracy of different kernel functions in datasets M317 and M983.

3.5. Selecting classification algorithm

Because of the large number of protein sequences and hidden information which difficult to reveal, the performance of the prediction algorithm has high requirements. In recent years, pattern recognition methods such as statistics and machine learning have been widely used in prediction algorithms. In this paper, we focus on five classification algorithms: support vector machine (SVM), k-nearest neighbor (KNN), naïve Bayes, decision tree (DT) and linear discriminant analysis (LDA). The SVM algorithm uses the RBF kernel function, the KNN algorithm uses the Euclidean distance, and the number of nearest neighbor points is three. Naïve Bayes, DT, and LDA all use default parameters. For the datasets M317 and M983, the features extracted from the PseAAC and PsePSSM for protein sequences are merged respectively, and then the feature vectors are further processed using 2-D wavelet denoising. The db6 wavelet function is used to decompose the data to 6. Under the jackknife test, the main prediction results of five different classification algorithms are obtained, as shown in Table 10 and Table 11. Specific results of the five kinds of classification algorithms for predicting the submitochondrial locations of datasets M317 and M983 are given in Supplementary materials Table S7, Table S8.

Table 10

Prediction results of submitochondrial locations of the M317 in different classification algorithms.

Classifiers	Jackknife test (%)			
	Sn (%)	Sp (%)	MCC	ACC (%)
SVM	100	100	1	100
KNN	98.93	99.61	0.9889	99.37

LDA	98.96	99.69	0.9864	99.37
DT	96.36	98.6	0.9529	97.48
Naïve Bayes	95.37	97.56	0.9101	94.95

Table 11

Prediction results of submitochondrial locations of the M983 in different classification algorithms.

Classifiers	Jackknife test (%)			
	Sn (%)	Sp (%)	MCC	ACC (%)
SVM	99.01	99.59	0.9852	99.29
KNN	98.54	99.53	0.9778	98.98
DT	94.99	97.77	0.9318	96.74
LDA	95.42	97.78	0.922	96.13
Naïve Bayes	78.71	88.51	0.6043	78.13

As can be seen from Table 10 that for dataset M317, using SVM algorithm for protein submitochondrial locations prediction, the overall prediction accuracy rate of this model is 100% with sensitivity and specificity of 100% and MCC of 1. Through these evaluation index values, it can be concluded that the SVM classifier predicts significantly better than the other four classifiers. The predicted accuracy of Naïve Bayes algorithm is 94.95%, which is 5.05% lower than that of SVM algorithm. The accuracy of KNN algorithm and LDA are 99.37%, which is 0.63% lower than the prediction accuracy of SVM algorithm.

It can be seen from Table 11 that for dataset M983, the overall prediction accuracy is 99.29%, the sensitivity and specificity are 99.01% and 99.59% respectively, and the MCC is 0.9852. In contrast, the prediction results using SVM algorithm is 0.31%-21.16% higher than the other four algorithms. Among them, the overall prediction accuracy using naïve Bayes algorithm is 78.13%, which is 21.16% lower than using SVM algorithm. Fig. 11 shows the overall prediction accuracy of the datasets M317 and M983 under different classification algorithms. As can be clearly seen from Fig. 11, the overall prediction accuracy of the two datasets is significantly higher than the other four prediction algorithms when using SVM as the prediction algorithm. Because SVM has good generalization performance, it can effectively avoid over-fitting and other advantages. Therefore, this paper selects SVM algorithm to predict protein submitochondrial locations.

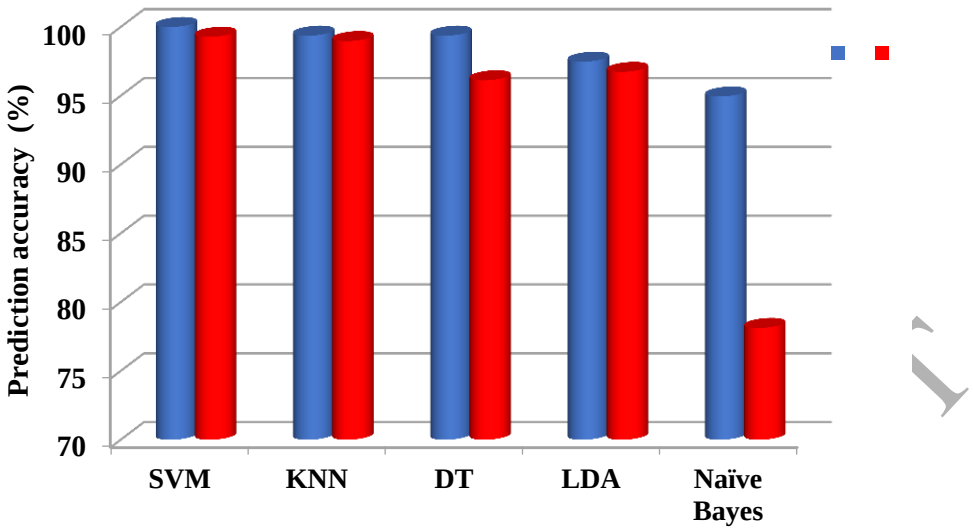

Fig. 11. The overall prediction accuracy of submitochondrial locations of the five classifications in datasets M317 and M983.

3.6. Performance of prediction model

In this paper, the fusion of PseAAC and PsePSSM methods for protein sequence feature extraction, the optimal parameters λ and ξ are 25 and 5, respectively, and then select the optimal feature vectors based on 2-D wavelet denoising, db6 wavelet function, the decomposition scale of 6. The RBF kernel function in SVM classifier is used to predict the submitochondrial datasets M317 and M983. The main results obtained by the jackknife test are shown in Table 12 and Table 13.

Table 12

Predicted performance of the protein submitochondrial dataset M317.

Dataset	Locations	Sn (%)	Sp (%)	MCC	ACC (%)
M317	Inner membrane	100	100	1	100
	Matrix	100	100	1	
	Outer membrane	100	100	1	

Table 13

Predicted performance of the protein submitochondrial dataset M983.

Dataset	Locations	Sn (%)	Sp (%)	MCC	ACC (%)
M983	Inner membrane	99.55	99.38	0.99	99.29
	Matrix	98.87	99.50	0.98	
	Outer membrane	98.62	99.88	0.99	

It can be seen from Table 12 that for the protein submitochondrial dataset M317, the overall prediction accuracy of this method is 100%, and the sensitivity, specificity and MCC for each type

of protein are respectively 100% and 100% % and 1. It can be seen from Table 13 that for the dataset M983, the overall prediction accuracy reaches 99.29%. The MCC of inter membrane and outer membrane are both 0.99, and the sensitivity to inter membrane, matrix and outer membrane are 99.55%, 98.87% and 98.62% respectively. The experimental results show that the PseAAC-PsePSSM-WD model can accurately predict the location of inter membrane, matrix and outer membrane in both datasets. In order to further verify the predictive ability of this model, we use the independent test set M495 to evaluate the PseAAC-PsePSSM-WD prediction model based on the jackknife test. The results obtained are shown in Table 14. In order to more intuitively see the prediction effect of the model, Fig. 12 shows the ROC curve of the dataset M495.

Table 14

Predicted performance of the protein submitochondrial dataset M495.

Dataset	Locations	Sn (%)	Sp (%)	MCC	ACC (%)
M495	Inner membrane	100	99.33	0.98	98.99
	Matrix	97.73	100	0.99	
	Outer membrane	98.17	99.74	0.98	

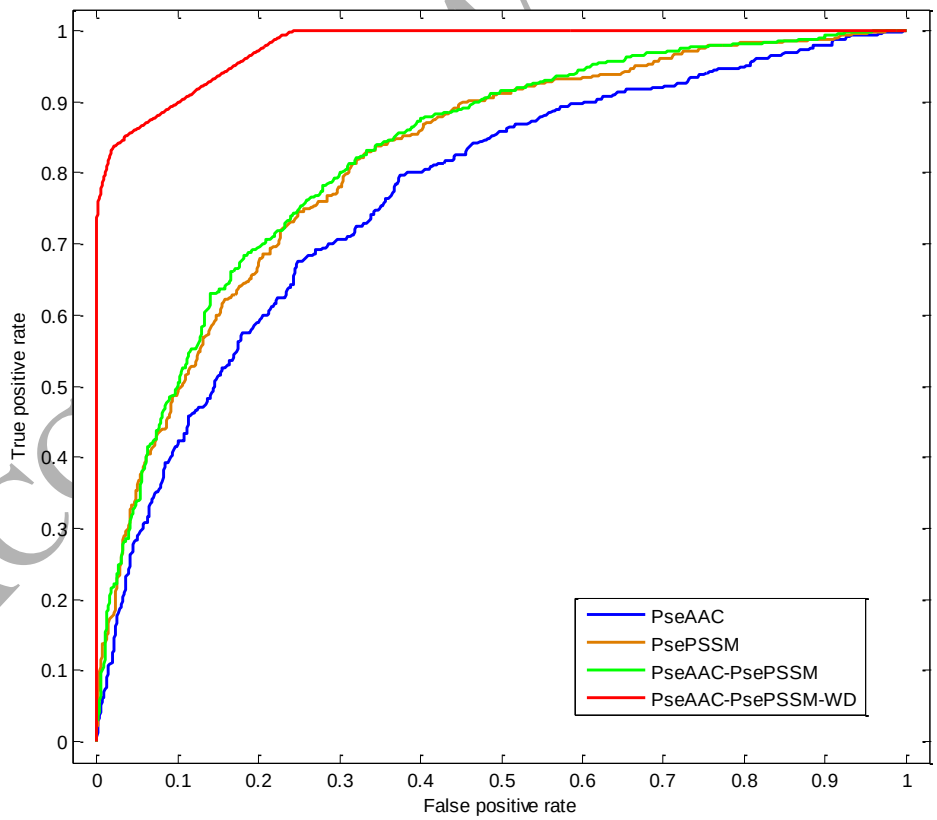

Fig. 12. This graph shows the ROC curve of the protein submitochondrial dataset M495.

As can be seen from Table 14, the overall prediction accuracy of the PseAAC-PsePSSM-WD model for the protein submitochondrial dataset M495 is 98.99%. Among them, the sensitivity of the inter membrane reached 100%, 2.27% higher than the matrix. The specificity of the matrix reached 100%. As can be seen from Fig. 12, the curve coverage area of the PseAAC-PsePSSM-WD model is obviously higher than that of the PseAAC, PsePSSM and PseAAC-PsePSSM feature extraction models, with an AUC of 0.9786. The AUC of the other three feature extraction methods are 0.7702, 0.8202 and 0.8276, respectively. The above results show that the proposed PseAAC-PsePSSM-WD model is an effective tool for protein submitochondrial locations prediction.

3.7. Comparison with other methods

A variety of prediction methods have been proposed for protein submitochondrial locations prediction research. In order to evaluate the prediction performance of the PseAAC-PsePSSM-WD model presented in this paper more objectively, under the dataset M317 and M983, based on the jacknife test method, the prediction results in this paper are compared with those in other papers. The PseAAC-PsePSSM-WD prediction model is further evaluated using an independent test set M495. Table 15 details the predicted results of the proposed method and other prediction methods on the datasets M317, M983 and M495.

Table 15

Comparison of different methods predicted results on three protein submitochondrial datasets.

Datasets	Methods	Sensitivity for each class (%)			ACC(%)
		Inner membrane	Matrix	Outer membrane	
M317	SubMito	85.5	94.5	51.2	85.2
	GP-Loc	83.2	97.2	78.1	89.0
	Predict_SubMito	91.8	96.4	66.1	89.7
	MitoLoc	97.7	99.0	68.3	94.7
	SubIdent	91.6	97.3	82.9	93.1
	Fan and Li	94.7	99.3	80.5	94.9
	SubMito-PSPCP	98.6	93.9	70.7	93.1
	Du et al.	93.9	98.6	78.1	94.0
	PseAAC-PsePSSM-WD	100.0	100.0	100	**100**
M983	SubMito-PSPCP	95.6	77.9	74.0	89.0
	Ahmad et al	87.1	98.6	99.6	95.1
	PseAAC-PsePSSM-WD	99.6	98.9	98.6	**99.3**
M495	Lin et al	98.8	86.4	78.9	91.1

PseAAC-PsePSSM-WD	100.0	97.7	98.2	**99.0**

It can be seen from Table 15 that the overall prediction accuracy of the PseAAC-PsePSSM-WD method for protein submitochondrial dataset M317 is 100%, which is 5.1% -14.8% higher than the accuracy of other prediction methods, 14.8% higher than the SubMito prediction method proposed by Du et al. (Du and Li, 2006) and 6.9% higher than the overall prediction accuracy of the SubIdent algorithm proposed by Shi et al. (Shi et al., 2011), 11% higher than the GP-Loc algorithm proposed by Nanni et al. (Nanni and Lumini, 2008) Among them, the sensitivity of method of this article on the inter membrane, matrix and outer membrane are 100%, that significantly higher than other methods in these proteins on the sensitivity. For outer membrane prediction, the SubIdent algorithm has a sensitivity of 51.2%, which is 48.8% lower than the PseAAC-PsePSSM-WD method. The results show that the overall prediction accuracy of the PseAAC-PsePSSM-WD method is obviously higher, and a better prediction result is obtained on the M317 dataset.The prediction accuracy of the dataset M983 is 99.3%, which is 10.3% higher than the SubMito-PSPCP prediction algorithm proposed by Du et al. (Du and Jiao, 2017). Among them, the sensitivity of inter membrane, matrix and outer membrane obtained by this method is significantly higher than those predicted by SubMito-PSPCP method, which were 4%, 21% and 24.6% higher respectively. Compared with the method proposed by Ahmad et al. (Ahmad et al., 2016), the overall prediction accuracy rate increased by 4.2%, but the sensitivity to the outer membrane was even lower by 1%, probably due to the poor permeability of the mitochondrial outer membrane. For the independent test set M495, the overall prediction accuracy of the PseAAC-PsePSSM-WD method is 99.0%, which is 7.9% higher than the algorithm proposed by Lin et al. (Lin et al., 2013), in which the sensitivity of the inter membrane reaches 100% and the sensitivity of the matrix and outer membrane 97.7% and 98.2% respectively, which are 11.3% and 19.3% higher than those proposed by Lin et al. Respectively. In conclusion, the PseAAC-PsePSSM-WD method achieves the highest prediction accuracy on all three submitochondrial locations datasets by comparison. These results fully demonstrate that the prediction model we constructed can significantly improve the prediction accuracy of protein submitochondrial locations, and the prediction result is satisfactory.

4. Conclusion

With the advent of the big data era, protein sequence data in biological databases have been rapidly increasing. It has been far from satisfying the needs of life science research to obtain protein sequence information through experimental methods. Therefore, it is more and more important to study the prediction of protein structure and function based on machine learning. Research on protein submitochondrial locations will help researchers design drug-related diseases caused by mitochondrial defects, and protein submitochondrial locations prediction will still be a challenging task. In this paper, a new protein submitochondrial locations prediction method is proposed. Firstly, PseAAC and PsePSSM are fused to extract the features of the protein sequences, and then 2-D wavelet denoising is performed on the fused feature vectors. Finally, the optimal noise-reduced feature vectors are input to the SVM classifier to predict the position of protein submitochondrial. In this paper, a new protein mitochondria localization prediction method is proposed. PseAAC feature extraction contains the frequency of various amino acids in a protein sequences, the sequence of protein sequences and various physicochemical properties. PsePSSM feature extraction can get the evolutionary information of the protein sequences, but also retains the sequence information of the amino acid residues in the protein sequences. 2-D wavelet denoising can effectively remove the redundant information in the protein sequences and extract the characteristic signals of each type of protein, which is of great significance for ensuring a high prediction accuracy. SVM classification algorithm can handle high-dimensional data, avoid over-fitting and effectively remove non-support vector. With the most rigorous jackknife test, the overall prediction accuracy of the datasets M317, M983 and M495 reach the highest prediction accuracy at the moment and is compared with other prediction methods. The experimental results show that the proposed method not only can effectively improve the prediction accuracy of protein submitochondrial locations, but also can accurately distinguish the mitochondrial inter membrane, matrix and outer membrane three regions of the proteins. We expect this method to be useful not only for the prediction of protein subcellular structures but also as a powerful tool in related fields such as bioinformatics, proteomics and molecular biology. We will make great effect in the future to provide a publicly accessible web server for our approach.

As pointed out in (Chou and Shen, 2009) and demonstrated in a series of recent publications (Chen et al., 2017; Cheng and Xiao, 2008, 2017a, 2017b, 2017c; Cheng et al., 2017; Cheng and

Zhao, 2017a; Chou and Shen, 2007b; Liu et al., 2015; Liu et al., 2016; Qiu et al., 2017; Xiao et al. 2017), user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have increased impacts on medical science (Chou, 2015), driving medicinal chemistry into an unprecedented revolution (Chou, 2017), we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper.

Acknowledgements

The authors sincerely thank the two anonymous reviewers for their many valuable comments that have improved this manuscript. This work was supported by the National Natural Science Foundation of China (Nos. 51372125 and 51572136), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), and the Key Laboratory Open Foundation of Shandong Province.

References

Ahmad, K., Waris, M., Hayat, M., 2016. Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition. J. Membr. Biol. 249, 1-12.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Miller, W., Lipman, D.J., 1997. Gappe BLAST and PSI-BLAST: a new general database search program. Nucleic Acids Res. 25, 3389-3402.

Ashbumer, M., Catherine, A.B., Judith, A.B., David, B., Heather, B., Michael, J.C., Allan, P.D., Kara, D., Selina, S.D., Janan, T.E., Midori, A.H., David, P.H., Laurie, I.T., Andrew, K., Suzanna, L., John, C.M., Joel, E.R., Martin, R., Gerald, M.R., Gavin, S., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25-29.

Bhasin, M., Raghava, G.P., 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. 32, 414-419.

Burbulla, L.F., Song, P., Mazzulli, J.R. E., Zampese, E., Wong, Y.C., Jeon, S., Santos, D.P., Blanz, J., Obermaier, C.D., Strojny, C., Savas, J.N., Kiskinis, E.., Zhuang, X., Krüger, R., Surmeier, D.J., Krainc, D., 2017. Dopamine oxidation mediates mitochondrial and lysosomal dysfunction in Parkinson's disease. Science 357, 1255-1261.

Cai, Y.D., Chou, K.C., 2000. Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Biochem. Biophys. Res. Commun. 4, 172-173.

Cai, Y.D., Chou, K.C., 2005. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J. Proteome Res. 4, 967-971.

Caragea, C., Caragea, D., Silvescu, A., Honavar, V., 2010. Semi-supervised prediction of protein

subcellular localization using abstraction augmented Markov models. BMC Bioinform. 11, 1-13.

Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E., 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266, 594-600.

Chang, S.G., Yu, B., Vetterli, M., 2000a. Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. Image Process. 9, 1532-1546.

Chang, S.G., Yu, B., Vetterli, M., 2000b. Spatially adaptive wavelet thresholding with context modeling for image denoising, IEEE Trans. Image Process. 9, 1522-1531.

Chang,C.C. Lin, C.J., 2011. LIBSVM: a library for support machines. ACM Trans. Intell. Syst. Technol. 2, 1-27.

Chen, W., Feng, P.M., Lin, H., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68.

Chen, W., Feng, P., Ding, H., Lin, H., 2015. iRNA-Methyl: identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. 490, 26-33.

Chen, W., Feng, P., Ding, H., 2016. Using deformation energy to analyze nucleosome positioning in genomes. Genomics. 107, 69-75.

Chen, W., Tang, H., Ye, J., Lin, H., 2016. iRNA-PseU: identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids. 5, e332.

Chen, W., Feng, P., Yang, H., Ding, H., 2017. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 8, 4208-4217.

Cheng, X., Xiao, X., 2017a. pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC. Mol. Biosyst. 13, 1722-1727.

Cheng, X., Xiao, X., 2017b. pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene 628, 315-321.

Cheng, X., Xiao, X., 2017c. pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics. doi:10.1016/j.ygeno.2017.10.002.

Cheng, X., Xiao, X., 2017d. pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics 11, 1-9.

Cheng, X., Xiao, X., 2018. pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics. 110, 50-58.

Cheng, X., Zhao, S.G., 2017a. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics 33, 341-346.

Cheng, X., Zhao, S.G., 2017b. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget. 8, 58494-58503.

Cheng, X., Zhao, S.G., Lin, W.Z., 2017. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics 33, 3524-3531.

Chou, K.C., 2001a. Using subsite coupling to predict signal peptides, Protein Eng. 14, 75-79.

Chou, K.C., 2001b. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Bioinform. 43, 246-255.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme

subfamily classes. Bioinformatics 21, 10-19.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236-247.

Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092-1100.

Chou, K.C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218-234.

Chou, K.C., 2017. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr. Top. Med. Chem. 17, 2337-2358.

Chou, K.C., Cai, Y.D., 2003. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem. Biophys. Res. Commun. 311, 743-747.

Chou, K.C., Elrod, D.W., 1998. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem. Biophs. Res. Commun. 252, 63-68.

Chou, K.C., Shen, H.B., 2006. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888-1897.

Chou, K.C., Shen, H.B., 2007a. MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Comm. 360, 339-345.

Chou, K.C., Shen, H.B., 2007b. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1-16.

Chou, K.C., Shen, H.B., 2009. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1, 63-92.

Chou, K.C., Jia, J., Liu, Z., Xiao, X., Liu, B. 2016. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn. 34, 1946-1961.

Deng, M., Tu, Z., Sun, F., Chen, T., 2004. Mapping gene ontology to proteins based on protein-protein interaction data. Bioinformatics 20, 895-902.

Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349-358.

Ding, H., Guo, S.H., Deng, E.Z., Yuan, L.F., Guo, F.B., Huang, J., Rao, N.N., Chen, W., Lin, H., 2013. Prediction of golgi-resident protein types by using feature selection technique. Chemom. Intell. Lab. Syst. 124, 9-13.

Du, P.F., Jiao, Y.S., 2017. Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. J. Theor. Biol. 416, 81-87.

Du, P.F., Li, Y.D., 2006. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinform. 7, 1-8.

Du, P., Tian, Y., Yan, Y., 2012. Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. J. Theor. Biol. 313, 61-67.

Du, P.F., Yu, Y., 2013. SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. Biomed. Res. Int. 3, 263829.

Ehsan, A., Mahmood, K., Khan, Y.D., Khan, S.A., 2018. A novel modeling in mathematical biology for classification of signal peptides. Sci. Rep. 8, 1039.

Fan, G.L., Li, Q.Z., 2012. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids 43, 545-555.

Feng, P., Ding, H., Yang, H., Chen, W., 2017. iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol. Ther. - Nucleic Acids. 7, 155-163.

Feng, P., Yang, H., Ding, H., Lin, H., 2018. iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics. doi:10.1016/j.ygeno.2018.01.005.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recognit. 44, 1761-1776.

Höglund, A., Dönnes, P., Blum, T., Adolph, H.W., Kohlbacher, O., 2006. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22, 1158-1165.

Huang, Y., Li, Y., 2004. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20, 21-28.

Jia, J., Liu, Z., Xiao, X., 2015. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor. Biol. 377, 47-56.

Jia, J., Liu, Z., Xiao, X., Liu, B., 2016. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal. Biochem. 497, 48-56.

Jones, D.T., 1999, Protein prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.

Kabir, M., Yu, D.J., 2017. Predicting DNase I hypersensitive sites via un-biased pseudo trinucleotide composition. Chemom. Intell. Lab. Syst. 167, 78-84.

Kandaswamy, A., Kumar, C.S., Ramanathan, R.P., Jayaraman, S., Malmurugan, N., 2004. Neural classification of lung sounds using wavelet coefficients, Comput. Biol. Medi. 34, 523-537.

King, B.R., Vural, S., Pandey, S., Barteau, A., Guda, C., 2012. ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. BMC Res. Notes 5, 1-7.

Lin, H., Chen, W., Yuan, L.F., Li, Z.Q., Hui, D., 2013. Using over-represented tetrapeptides to predict protein submitochondria locations. Acta Biotheor. 61, 259-268.

Lin, H., Li, Q.Z., 2007. Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components. J. Comput. Chem. 28, 1463-1466.

Lin, H., Wu, Y., Tang, H., Chen, W., 2016. Predicting Human enzyme family classes by using pseudo amino acid composition. Curr. Proteomics. 13, 99-104.

Li, W.C., Deng, E.Z., Ding, H., Chen, W., Lin, H., 2015. iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition, Chemom. Intell. Lab. Syst. 141, 100-106.

Liò, P., 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19, 2-9.

Liu, B., Fang, L., Liu, F., Wang, X., 2015. Identification of real microRNA precursors with a

pseudo structure status composition approach. PLoS ONE 10, e0121501.

Liu, B., Fang, L., Long, R., Lan, X., 2016. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32, 362-369.

Liu, B., Liu, F., Wang, X.L., Chen, J.J., Fang, L.Y., Chou, K.C., 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 43, 65-71.

Liu, B., Long, R., 2016. iDHS-EL: identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics 32, 2411-2418.

Liu, B., Wang, S., Long, R., Chou, K.C., 2017. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics 33, 35-41.

Liu, B., Wu, H., Chou, K.C., 2017. Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and Protein Sequences. Nat. Sci. 9, 67-91.

Liu, B., Yang, F., 2017. 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Mol. Ther. Nucleic Acids. 7, 267-277.

Liu, B., Yang, F., Huang, D.S., 2018. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 34, 33-40.

Liu, L.M., Xu, Y., 2017. iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med Chem. 13, 552-559.

Luisier, F., Blu, T., Unser, M., 2007. A new sure approach to image denoising: interscale orthonormal wavelet thresholding. IEEE Trans. Image Process.16, 593-606.

Mariana, R.V.L.H., 2011. Animal mitochondrion diagram, https://commons.wikimedia.org/wiki/File:Animal_mitochondrion_diagram_zh_ml.svg.

Mei, S., Wang, F., 2010. Amino acid classification based spectrum kernel fusion for protein subnuclear localization. BMC Bioinform. 11, 1-8.

Nakashima, H., Nishikawa, K., 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. 238, 54-61.

Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria location. Amino Acids. 34, 653-660.

Qi, H., Casalena, G., Shi, S., Yu, L., Ebefors, K., Sun, Y., Zhang, W., D'Agati, V., Schlondorff, D., Haraldsson, B., Böttinger, E., Daehn, I., 2017. Glomerular endothelial mitochondrial dysfunction is essential and characteristic of diabetic kidney disease susceptibility. Diabetes 66 763-778.

Qiu, W.R., Jiang, S.Y., Sun, B.Q., 2017a. iRNA-2methyl: identify RNA 2′-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier. Med. Chem. 13, 734-743.

Qiu, W.R., Sun, B.Q., 2016. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32, 3116-3123.

Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., 2017. iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics. Doi:10.1016/j.ygeno.2017.10.008.

Saidijam, M., Azizpour, S., Patching, S.G., 2017. Amino acid composition analysis of human secondary transport proteins and implications for reliable membrane topology prediction. J. Biomol. Struct. Dyn. 35, 929-949.

Shen, H.B., Chou, K.C., 2007. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng. Des. Sel. 20, 561-567.

Shen, H.B., Chou, K.C., 2008. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386-388.

Shen, H.B., Yang, J., Chou, K.C., 2006. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J. Theor. Biol. 240, 9-13.

Shi, S.P., Qiu, J.D., Sun, X.Y., Huang, J.H., Huang, S.Y., Suo, S.B., Liang, R.P., Zhang, L., 2011. 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, 424-430.

Uddin, M.R., Sharma, A., Farid, D.M., Rahman, M.M., Dehzangi, R., Shatabda, S., 2018. EvoStruct-Sub: an accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features. J. Theor. Biol. 443, 138-146.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Xiang, Q.L., Liao, B., Li, X.H., Xu, H.M., Chen, J., Shi, Z.X., Dai, Q., Yao, Y.H., 2017, Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. Artif. Intell. Med. 78, 41-46.

Xiao, X., Cheng, X., Su, S., Nao, Q., 2017. pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Nat. Sci. 9, 331-349.

Xu, Y., Ding, J., Wu, L.Y., 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, e55844.

Xu, Y., Li, C., 2017. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med Chem. 13, 544-551.

Yu, B., Zhang, Y., 2013a. The analysis of colon cancer gene expression profiles and the extraction of informative genes. J. Comput. Theor. Nanosci. 10, 1097-1103.

Yu, B., Zhang, Y., 2013b. A simple method for predicting transmembrane proteins based on wavelet transform. Int. J. Biol. Sci. 9, 22-33.

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

Yu, B., Lou, L.F., Li, L., Zhang,Y., Qiu, W.Y., Wu, X., Wang, M.H., Tian, B.G., 2017. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J. Mol. Graph. Model. 76, 260-273.

Yu, B., Li, S., Qiu, W.Y., Chen, C., Chen, R.X., Wang, L., Wang, M.H., Zhang, Y., 2017. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget 8, 107640-107665.

Zakeri, P., Moshiri, B., Sadeghi, M., 2011. Prediction of protein submitochondria locations based on data fusion of various features of sequences. J. Theor. Biol. 269, 208-216.

Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang L., Yu, L, L.Z., Li, M.L.,2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259, 366-372.

Zhai, J.X., Cao, T.J., An, J.Y., Bian, Y.T., 2017. Highly accurate prediction of protein self-interactions by incorporating the average block and PSSM information into the general PseAAC. J. Theor. Biol. 432, 80-86.

Zhang, T.L., Ding, Y.S., Chou, K.C., 2008. Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J. Theor. Biol. 250, 186-193.

Zhang, S.L., Duan, X., 2018. Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J. Theor. Biol. 437, 239-250.

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