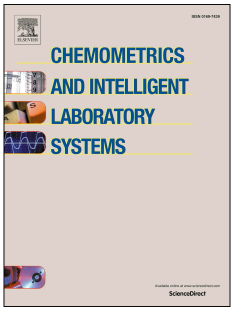

Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition

Bin Yu, Shan Li, Cheng Chen, Jiameng Xu, Wenying Qiu, Xue Wu, Ruixin Chen

PII: S0169-7439(16)30420-8

DOI: 10.1016/j.chemolab.2017.05.009

Reference: CHEMOM 3440

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 27 October 2016

Revised Date: 5 May 2017

Accepted Date: 7 May 2017

Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition

Bin Yu [a, b*], Shan Li[a, b], Cheng Chen[a, b], Jiameng Xu[a, b], Wenying Qiu[a, b], Xue Wu[a, b], Ruixin Chen[a, b]

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

[b] Bioinformatics Laboratory, College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

Abstract: Information on the subcellular localization of Gram-negative bacterial proteins is of great significance to study the pathogenesis, drug design and discovery of certain diseases. Protein subcellular localization is an important part of proteomics, while providing new opportunities and challenges for chemometrics. Since the prediction of protein subcellular localization can help to understand their function and the role played by their metabolic processes, a number of protein subcellular localization prediction methods have been developed in recent years. In this paper, we propose a novel method by combining wavelet denoising with support vector machine to predict the subcellular localization of proteins for the first time. Firstly, the features of the protein sequence are extracted by Chou's pseudo amino acid composition (PseAAC), then the feature information of the extracted is denoised by two-dimensional (2-D) wavelet. Finally, the optimal feature vectors are input to the SVM classifier to predict subcellular location of the Gram-negative bacterial proteins. Quite promising predictions are obtained using the jackknife test and compared with other predictive methods. The results indicate that the method proposed in this paper can remarkably improve the prediction accuracy of protein subcellular localization, and it can be used to predict the other attributes of proteins.

Keywords: Gram-negative bacteria; Subcellular localization; Pseudo amino acid composition; Two-dimensional wavelet denoising; Support vector machine

1. Introduction

Protein is involved in various forms of life activities. The living organism's growth, development, reproduction and other life activities are inseparable from the role of the protein. Protein also maintains a highly ordered operation of the protection of the cell system [1]. Bacteria are a special kind of protein that

*Corresponding author. Tel./fax: +86 532 88958923.
E-mail address: yubin@qust.edu.cn (B. Yu).

is divided into two major groups according to the cell wall: Gram-negative bacteria and Gram-positive bacteria [2]. By the use of Gram staining, those bacteria which were stained purple were called Gram-positive bacteria; those which were dyed red or pink were called Gram-negative bacteria. Protein subcellular localization information of Gram-negative bacteria plays an important role in the research of pathogenesis of certain diseases and new drug discovery [3].

At present, life sciences research has entered into the post-genome era, with the depth of the study, the accumulated protein sequence data in database of protein increase exponentially. Protein subcellular location information is of great significance for the understanding of the biological function of the protein and also a hot topic in chemical biology research. Because the newly synthesized protein in the living organism is only transported to the correct subcellular location, can it plays its biological function. Subcellular location information can help to understand the role of the protein in the metabolic process for proteins with known basic functions [4]. Although the subcellular location of proteins can be determined by such experimental methods [5] as cell fractionation, electron microscopy, and fluorescence microscopy, these traditional experimental methods are not only time-consuming, costly and the repeatability is relatively poor [6]. With the rapid growth of biological data such as nucleic acid and protein sequences, it is far from enough to determine the subcellular localization of protein by experimental methods. Therefore, in order to speed up the annotation process of protein structure and function, it is an arduous and challenging task to study how to use computational intelligence to predict protein subcellular localization for researchers.

During the past two decades, with the rapid development of computational intelligence, the prediction of protein subcellular localization has become a hot research field of bioinformatics. In general, protein subcellular localization prediction is essentially a multi-class classification problem in pattern recognition. The study of the problem generally include the following three steps: (1) establish an objective representative benchmark dataset to obtain the feature information of each kind of subcellular localization and extract the feature parameters or characteristics; (2) introduce or develop a powerful prediction algorithm; (3) predict protein by prediction algorithm and evaluate prediction results.

Up to now, with the continuous improvement of prediction accuracy, a lot of research results have been achieved [7]. As early as 1991, Nakai and Kanehisa [8] establish a Gram-negative bacterial proteins subcellular localization prediction system by the use of protein N-terminal sequence information. In 1999, PSORT prediction software was developed also based on the N-terminal signal sequence feature information [9]. Subsequently, based on the N-terminal sequence information and using the neural network approach, Emanuelsson et al. [10] designed the TargetP system to predict protein subcellular localization. Although the N-terminal sorting signal can characterize the protein to some extent, its effectiveness depends on the completeness of the protein sequences. Once the N-terminal signal of the protein sequence is incomplete or the signal is lost, the prediction effect of the method becomes unsatisfactory. Nakashima et al. [11] proposed a prediction method based on the amino acid composition and the frequency of residue pairs. By using the 20 amino acid composition of proteins, they can distinguish intracellular and extracellular proteins, and found the relationship between subcellular localization of protein and its amino acid composition. It is because of their research that many researchers put forward more protein subcellular localization prediction method using amino acid composition principle [4,12,13]. For example, Reinhardt et al. [12] used amino acid components as feature information to predict protein subcellular locations in prokaryotes and eukaryotes, and constructed the first artificial neural network prediction system. Hua et al. [13] constructed their respective classifiers to predict protein subcellular locations using amino acid compositions and all obtained good results. The method of amino acid component extraction is convenient, but it does not make full use of the amino acid residues sequence and various physical and chemical properties, so it can not describe the protein in a comprehensive way and the related improvement methods begin to appear. One of the most representative is the pseudo-amino acid composition (PseAAC) proposed by Chou et al. [14]. Chou et al. [14,15] defined the PseAAC using the physicochemical properties of amino acids, such as hydrophilicity, hydrophobicity, etc., in combination with the amino acid composition. The predicted subcellular localization accuracy has been significantly improved on the basis of the original. They developed an online server [16] that calculates the composition of pseudo-amino acids at http://www.csbio.sjtu.edu.cn/bioinf/PseAA/. In this server, researchers can choose different parameters to get access to a variety of different forms of PseAAC. Recently, the researchers have made extensive use of the PseAAC methods in predicting protein structure [17,18], functional prediction [19,20], protein-protein interaction [21,22] and protein subcellular localization prediction [23-25].

It is hard to make a breakthrough by solely using one certain characteristic. In recent years, researchers are more inclined to fuse multiple features to characterize the protein sequence, with a view to synthesize the advantages of each sequence coding method to obtain more protein sequences feature information. Gardy et al. [26] proposed a coding method PSORT-B containing amino acid composition, N-terminal sorting signal and motifs to predict the subcellular localization of Gram-negative proteins. Chou and Cai [27] predicted protein subcellular localization by fusing PseAAC and functional domain information as feature vectors. Subsequently, they combined GO annotation to predict protein subcellular localization [28], and achieved good prediction effect. Zhang et al. [29] proposed a new approach named as encoding based on grouped weight (EBGW) for protein sequence based on the idea of coarse-grained description and grouping. Chou and Shen [30] proposed a coding method combined by GO annotation, functional domain and evolutionary information to predict eukaryotic protein sequences containing single-location or multiple-location subcellular fraction sites. Yu et al. [25] used amino acid substitution matrix and auto covariance transformation to extract the sequence features of proteins, constructed the feature vectors and proposed a new pseudo-amino acid model to predict the subcellular localization of apoptosis proteins. Liu et al. [31] proposed a method for predicting the subcellular localization of apoptosis proteins based on tri-gram encoding of PSSM matrix, which incorporates evolution information of proteins. Li et al. [32] proposed a novel PseAAC model to predict the subcellular localization of three bacterial proteins by fusing features from PSSM matrix, GO information, and PROFEAT. Chen et al. [33] proposed a multiple information fusion method for predict subcellular locations of two different types of bacterial protein by combining physicochemical properties, auto covariance transformation of the PSSM matrix, and GO information, and obtained a better prediction effect. Wang et al. [34] carried out feature extraction on protein sequences by fusing PSSM and Chou’s PseAA composition. Linear discriminant analysis (LDA) was used to reduce the dimension. The K-NN (K-Nearest Neighbor) classifier was employed to identify Gram-negative bacterial proteins subcellular location and the prediction accuracy was 93.57%. Wang et al. [35] extended this method to kernel linear discriminant analysis (KLDA) dimensionality reduction. Using K-NN classifier, the highest recognition rate of Gram-negative bacterial protein subcellular location was 98.77%.

Not only protein sequence information, another important factor for protein subcellular location prediction is classification algorithm. Successful classification algorithm should be able to efficiently and correctly separate proteins with different subcellular locations. In recent years, pattern recognition methods such as statistics and machine learning have been widely used in prediction algorithms, such as K-nearest neighbor (K-NN) [34-36], neural network [10,12], hidden Markov model (HMM) [14,37], Bayesian network [38,39], and support vector machine (SVM) [13,24,31-33] and so on. Among them, SVM has the advantages of fast computation speed, strong extraction ability of implicit information in training set and excellent generalization performance, which makes SVM a preferred classifier for many researchers. It is widely used to predict membrane protein types [40,41], protein structural class [18,42], protein-protein interaction [43,44], protein subcellular localization [31-33,45], G protein-coupled receptors [46,47], protein post-translational modification nitrosylation sites [48,49] and other protein function research.

This paper presents a new method for predicting protein subcellular localization--support vector machine prediction model based on wavelet denoising (WD-SVM). Firstly, Chou's PseAAC is used for feature extraction on the sequence of Gram-negative bacterial proteins. Then the feature vectors of the extracted proteins are denoised by two-dimensional (2-D) wavelet to make the features of each kind of proteins more prominent. Finally, the optimal feature vectors after wavelet denoising are input to support vector machine classifier for prediction. By jackknife test, different effects on the results are compared due to choosing different λ values, wavelet functions, wavelet decomposition of different scales and kernel functions. Through the comparative analysis, the optimal parameters of the prediction model are determined. The overall prediction accuracies are 99.08% and 98.6% on A653 and C1114 datasets, respectively. The experimental results show that the proposed WD-SVM model can remarkably improve the prediction effect of protein subcellular localization.

2. Material and Methods

2.1. Dataset

Two different benchmark datasets of Gram-negative bacteria are chosen in this study, which are taken from [2]. They can be obtained by downloading the online Supporting Information A and C from the website at http://www.csbio.sjtu.edu.cn/bioinf/Gneg/. The first dataset contains 653 Gram-negative bacterial protein sequences, of which 152 are cytoplasm proteins, 76 extracellular proteins, 12 fimbrium proteins, 6 flagellum proteins, 186 inner membrane proteins, 6 nucleoid proteins, 103 outer membrane proteins, and 112 periplasm proteins. The dataset is recorded as A653. The second dataset contains 1114 Gram-negative bacterial proteins, of which 140 are cytoplasm proteins, 74 extracellular proteins, 687 inner membrane proteins, 97 outer membrane proteins, and 116 periplasm proteins. The dataset is recorded as C1114. A schematic drawing to show the eight types of Gram-negative bacterial proteins is shown in Fig. 1. In this paper, the first dataset A653 as a training dataset, used to select the parameters of the prediction model, the second dataset C1114 as an independent testing dataset, used to test the applicability of the model.

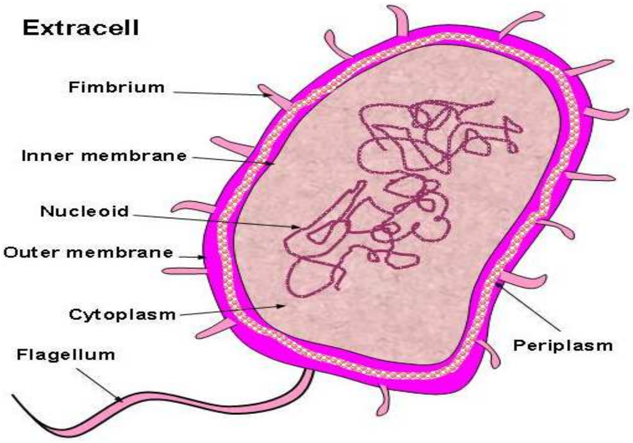

Fig. 1. Schematic illustraction to show the eight subcellular locations of Gram-negative bacterial proteins: (1) cytoplasm, (2) extracell, (3) fimbrium, (4) flagellum, (5) inner membrane, (6) nucleoid, (7) outer membrane, (8) periplasm. Reproduced from Chou and Shen [2].

2.2. PseAAC

To avoid losing sequence-order information of proteins, Chou [14] proposed pseudo-amino acid composition (PseAAC) approach to solve the problem of extraction of feature vectors in protein sequences. According to Chou's PseAAC discrete model, the protein sequence could be represented by $(20+\lambda)-D$ (dimensional) vector.

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

where the $(20+\lambda)$ components are given by

$$p_\mu=\begin{cases}\dfrac{f_\mu}{\sum\limits_{\mu=1}^{20}f_\mu+\omega\sum\limits_{k=1}^{\lambda}\tau_k}, & 1\le\mu\le20\\[2ex] \dfrac{\omega\tau_{\mu-20}}{\sum\limits_{\mu=1}^{20}f_\mu+\omega\sum\limits_{k=1}^{\lambda}\tau_k}, & 20+1\le\mu\le20+\lambda\end{cases} \tag{2}$$

where ω is the weight factor, which was set at 0.05 in this article[14]. τ_k is the k-tier sequence correlation factor, which reflects the sequence-order information. f_μ are the occurrence frequencies of μ amino acids in proteins. As can be seen from the above formula, the first 20 dimension of the feature vector P is the amino acid composition and the posterior λ dimension is the sequence correlation factor reflecting the different levels of amino acid sequence information. Sequence correlation factor can be calculated from the hydrophobicity index, hydrophilicity index, and side chain mass of the amino acids residue. Because the length of the shortest protein sequence in the benchmark dataset is 50, the maximum value allowed for λ in Eqs. (1) and (2) is 49.

In this study, we used PseAAC web-server developed by Shen and Chou [16] to carry out feature extraction on protein sequences. On this web-server, by selecting different parameters λ, the optimal λ can be determined from the accuracy of the prediction results.

2.3. Two-dimensional wavelet denoising

In recent years, wavelet analysis has been applied to a variety of biological information [51]. Especially for sequence analysis and protein structure studies [52-54]. It has the characteristics of multiresolution analysis, but also has relatively analysis methods with the changeable time-frequency window, which has very good localization properties in the time and frequency intra-areas. Wavelet analysis is very suitable for the detection of biological signal singularity, and open up a new way for the extraction and transformation of biological information data. Wavelet denoising (WD) is one of the most important applications of wavelet analysis. For a large class of signals, their energy is highly concentrated in a small amount of wavelet coefficients or wavelet packet coefficients, while due to its randomness of the noise, the energy is more evenly distributed in the wavelet transform domain. So threshold method is used to suppress the noise to the maximum while preserving the signal in the transform domain [55,56]. In order to eliminate the noise genes effectively and select the more accurate classification informative genes, we used the multi-scale wavelet threshold method to denoise the data in the gene expression profiles, and got the satisfying tumor classification accuracy [57]. The decomposition of protein amino acids signal by wavelet analysis can effectively remove the redundant information and extract the characteristic signals of each kind of proteins, which is very important to ensure the high accuracy of prediction [58].

A two-dimensional (2-D) model with noise [55,59] can be expressed as:

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

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

In this paper, we use the PseAAC algorithm to extract the features of each protein sequence in the dataset, so that each sequence generates $20+\lambda$ dimension features, and finally get a matrix $653\times(20+\lambda)$, treat it as a 2-D signals, which denoise the entire 2-D signals. The thresholds can be selected using a uniform global threshold, or can be divided into three directions (horizontal, vertical and diagonal direction), so that all the direction of the noise can be separated, so that the protein specificity of each data set is more significant. 2-D wavelet denoising flow chart is shown in Fig. 2.

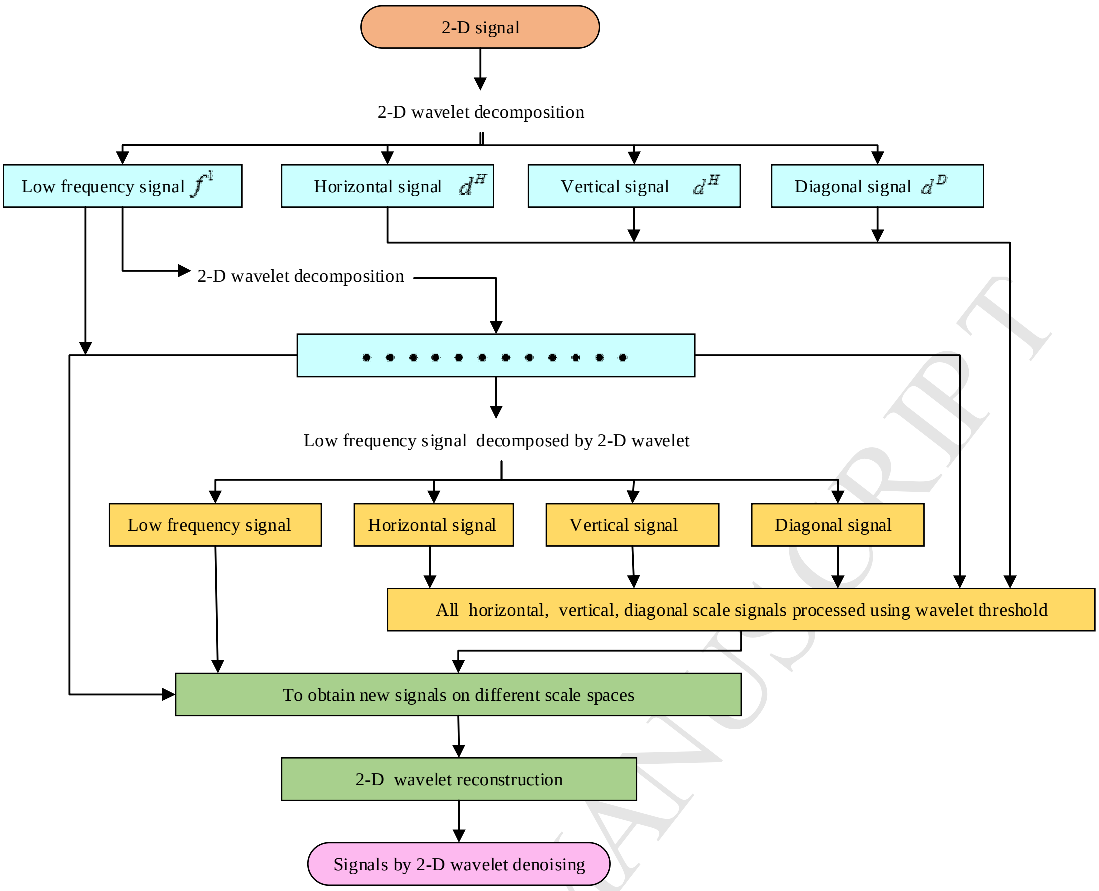
2-D signal
2-D wavelet decomposition
Low frequency signal f^1
Horizontal signal d^H
Vertical signal d^H
Diagonal signal d^D
2-D wavelet decomposition
Low frequency signal decomposed by 2-D wavelet
Low frequency signal
Horizontal signal
Vertical signal
Diagonal signal
All horizontal, vertical, diagonal scale signals processed using wavelet threshold
To obtain new signals on different scale spaces
2-D wavelet reconstruction
Signals by 2-D wavelet denoising

Fig. 2. Flow chart of two-dimensional wavelet denoising.

Two-dimensional wavelet denoising steps are as follows:

(1) Wavelet decomposition of 2-D signal. Using 2-D wavelet transform, make 2-D data decompose to the four scale spatial frequency band, that is, low-frequency scale spatial frequency band signals f^1, horizontal scale spatial frequency band signals d^H, the signals on the vertical scale spatial frequency band d^V, the signals on the diagonal scale spatial frequency band d^D. It is orthogonal decomposition process on different scales, that is, $f = f^1 \oplus d^H \oplus d^V \oplus d^D$ .The low-frequency signals f^1 can also continue to decompose into four directions: the signals of Gram-negative bacterial proteins in the low-frequency scale space f^2, the signals in the spatial scale of the horizontal scale $d^{1,H}$, the signals in the spatial scale of the vertical scale $d^{1,V}$, the signals in the spatial frequency band of the diagonal scale $d^{1,D}$. The decomposition process is still orthogonal decomposition $f^1 = f^2 \oplus d^{1,H} \oplus d^{1,V} \oplus d^{1,D}$. The signals energy of Gram-negative bacterial is limited, so the number of decomposition of 2-D wavelet is also limited, the n process of wavelet decomposition $f^{n-1} = f^n \oplus d^{n-1,H} \oplus d^{n-1,V} \oplus d^{n-1,D}$.

(2) Threshold quantization of the decomposed high-frequency coefficients. Threshold quantization is performed by selecting the appropriate thresholds for the high-frequency coefficients from 1 to the N horizontal, vertical and diagonal directions of each layer.

(3) Reconstruct the 2-D signals. The 2-D signals are reconstructed based on the low-frequency coefficients of the wavelet decomposition and the high-frequency coefficients after the threshold quantization. The reconstructed signals is the noise reduction signals.

In these three steps, the most critical is how to select the threshold and threshold quantization. Since the selection of thresholds directly affects the quality of noise reduction, a variety of theoretical and empirical models are proposed, but none of them is general. 2-D signals commonly used threshold models: default threshold determination model, threshold model determined by Birge-Massart strategy. For the default threshold, we can choose the wavelet function or the wavelet packet function to determine. In the process of numerical experiment, this study aims at improving the overall accuracy of the selection of appropriate methods. In the process of noise reduction, the wavelet coefficients are subjected to threshold processing operation, and the selection of soft thresholding and hard thresholding has certain influence on the results of wavelet denoising [60]. Soft thresholding is a smoother way to produce a smoother effect after noise reduction, while hard thresholding processing preserves features such as spikes in the signal, we choose soft thresholding wavelet denoising method.

In this study, we use the PseAAC to carry out feature extraction on protein sequences, and then extract the feature vectors of the proteins by 1-D and 2-D wavelet denoising respectively. The specific process is as follows: Firstly, the wavelet decomposition function of two denoising methods both use bior2.4 wavelet, the scale level is 3. Then the high-frequency coefficients are processed by threshold, and the thresholds are selected by wavelet function to get the default threshold. Finally, the signals are reconstructed based on the low-frequency coefficients of the wavelet decomposition and the high-frequency coefficients after the threshold quantization. In order to more intuitively compare the influence of 1-D and 2-D wavelet denoising on the dataset, we picked the protein No. P41121 in the cytoplasm proteins sequence as an example, and the results are shown in Fig. 3. From Fig. 3, it can be seen that the P41121 protein can be effectively removed redundant information in protein sequence by using 2-D wavelet denoising after 1-D and 2-D wavelet denoising, so as to extract the characteristic signal of the protein itself. And from the one-dimensional and 2-D wavelet denoising principle, we can see that

1-D wavelet denoising decomposition of the frequency band is not detailed enough. 2-D wavelet denoising can deal with the whole dataset. The data is denoised from the low-frequency, horizontal, vertical and diagonal scale space. We found that 2-D wavelet denoising can effectively improve the feature of each Gram-negative bacterial protein sequences, so we adopt 2-D wavelet denoising method for Gram-negative bacteria dataset.

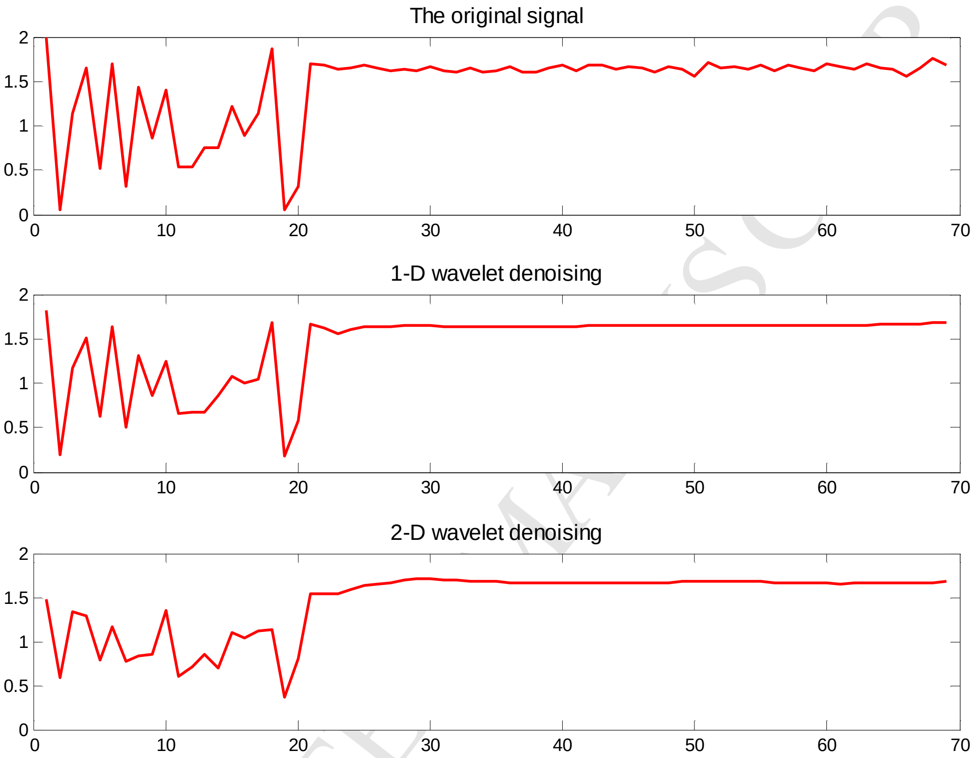

Fig. 3. Wavelet denoising using bior2.4 wavelet function and decomposition scale is $j=3$. The y axis indicates the intensity of signal; the x axis indicates the residue position along the sequence.

2.4. Support vector machine

Support vector machine (SVM) is a machine learning method based on statistical learning theory proposed by Vapnik et al. [61]. SVM has been widely used in many fields of bioinformatics for solving bioinformatics data with high dimensionality, small sample size and non-linearity, and it has excellent learning performance under the principle of structural risk minimization. The problem of subcellular localization prediction and functional classification have made ideal prediction results [31,43].

For a two-class classification problem, suppose a training set with n samples

$$G=\{(x_i,x_j)\,|\,i=1,2,\cdots,n\} \qquad x_i\in R^d, y_i\in\{+1,-1\} \tag{4}$$

where x_i is d dimension feature vectors of sample i, y_i is the class labels of sample i. In order to complete the classification of the two classes of samples, SVM maps the samples of the input vector into a high-dimensional feature space using some nonlinear kernel function, and constructing an optimal separating hyperplane in this space to make the two samples linearly separable.

The kernel function $K(x_i, x_j)$ is used to replace the dot product in the optimal classification hyperplane, the decision function is given by:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i y_i K\left(x_i, x_j\right) + b\right\} \tag{5}$$

where α_i denotes Lagrange multiplier, b is classification threshold value, $K(x_i, x_j)$ is the kernel function.

The selection of kernel function is very important in the process of training the support vector. It can effectively overcome the "curse of dimensionality". The appropriate kernel function can improve the prediction accuracy of the classification model. Generally, four kinds of kernel functions, i.e. linear kernel function, polynomial kernel function, radial basis function (RBF), and sigmoid kernel function, can be available to perform prediction.

SVM was originally designed for two-class classification, and protein subcellular localization prediction is a multi-class classification problem. Currently, there are three main strategies for multi-classification: one-versus-one (OVO), one-versus-rest (OVR) [62] and direct acyclic graph SVM (DAGSVM) [63]. In this paper, we use the OVO strategy. For a N - classification problem, we need to train $N \times (N-1)/2$ two-class SVM classifiers, one classifier is used to distinguish which class it belongs to respectively. Votes for each class are counted and the class with the most votes as the predict category. This study used LIBSVM software developed by Chang and Lin [64], which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.5. Performance evaluation and model building

In statistical prediction, several methods, such as jackknife test, self-consistency test, independent test, and k-fold cross-validation, are often used to evaluate the prediction performance [18,53]. In our study, jackknife test and self-consistency test method are used to examine the performance of the prediction model.

Sensitivity (S_{in}), specificity (S_{ip}) and Matthew's correlation coefficient (MCC) and overall accuracy (OA) are used to evaluate the results of the prediction system. These measures can be defined as follows:

$$S_{in} = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$S_{ip} = \frac{TP_i}{TP_i + FP} \tag{7}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i)\times(TP_i + FN_i)\times(TN_i + FP_i)\times(TN_i + FN_i)}} \tag{8}$$

$$OA = \frac{\sum_{i=1}^{n} TP_i}{N} \tag{9}$$

where TP_i denotes the number of true positives in i th subcellular location, FN_i denotes the number of false negatives in i th subcellular location, FP_i denotes the number of false positives and TN_i denotes the number of true negatives in i th subcellular location. n is the class number, N is the total number of sequences in the dataset.

For convenience, the method for predicting protein subcellular localization proposed in this paper is called WD-SVM, and the general procedure is shown in Fig. 4. We have implemented it in MATLAB R2014a in windows 10 running on a PC with system configuration Intel (R) Core (TM) i7-4510U CPU @ 2.00 GHz 2.60 GHz with 8.00GB of RAM.

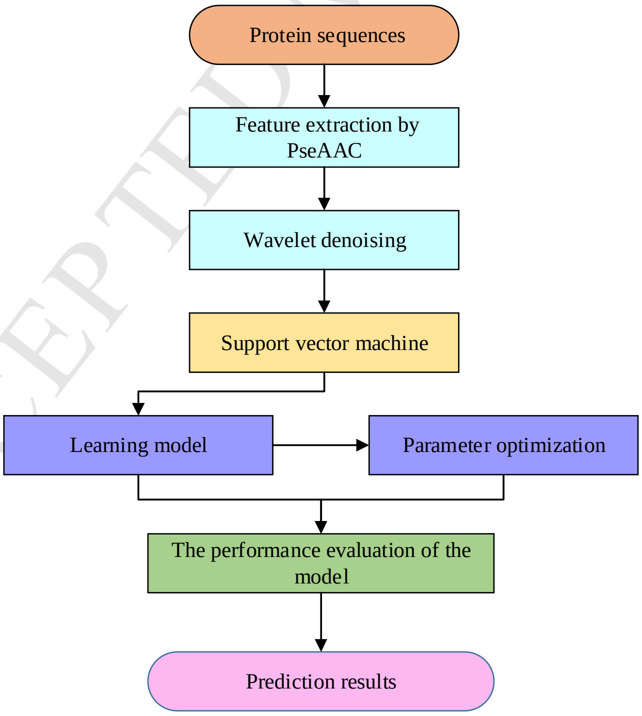

Fig. 4. Flow chart of protein subcellular localization prediction based on WD-SVM method.

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

1) Input the protein sequences of Gram-negative bacteria, and the class label corresponding to 8 kinds of proteins;

2) Using PseAAC online service system developed by Shen et al. [16] to carry out feature extraction on the protein sequences and the protein sequences were transformed into numerical sequences;

3) Wavelet denoising method is used to denoise the numerical sequences extracted in 2), and the redundant information in the sequence is eliminated;

4) The optimal feature vectors of noise reduction are input to the SVM classifier to predict the subcellular localization of the proteins;

5) According to the accuracy of prediction, choose the optimal parameters of the model, including the λ values of parameters, wavelet function selection, different decomposition scales and kernel functions in SVM;

6) According to the optimal parameters of the model, S_{in}, S_{ip}, MCC and OA are calculated, and the prediction performance of the model is evaluated.

3. Results and discussion

3.1. Selection of optimal parameter λ

In current study, we used PseAAC to carry out feature extraction on protein sequences of Gram-negative bacteria. In the process of feature extraction, the selection of values λ played an important role in the results. If the value λ is set too small, the sequence information contained in the feature vectors will be very small, and the features of the Gram-negative bacterial protein sequences can not be extracted completely. If the value λ is set too large, it will bring more redundant information, thus affecting the prediction results. And when the value λ exceeds the length of a protein sequence in the dataset, it will not make sense to use PseAAC to feature extraction of protein sequences. Since the length of the shortest protein in the benchmark dataset is 50, the maximum value of λ for this paper is 49. To find the optimal value λ in the model, set the values λ from 0 to 49 in turn. For the different values of λ, the WD-SVM model is constructed using db1 wavelet function in the wavelet denoising, the decomposition scale level is 3, the SVM is used to classify and using the linear kernel function, the results are tested by jackknife method. The overall prediction accuracy of each class of the A653 dataset is shown in Table 1. The effect of the different values of λ on the results are shown in Fig. 5.

Table 1 Prediction results of subcellular localization of the A653 dataset by selecting different values of λ .

λ / Locations	Jackknife test (%)										
	0	5	10	15	20	25	30	35	40	45	49
Cytoplasm	96.71	97.37	98.68	100	100	100	100	100	100	100	**100**
Extracell	78.95	82.89	86.84	76.31	44.73	44.74	55.26	63.16	73.68	73.68	**84.21**
Fimbrium	100	83.33	100	100	100	66.67	66.67	66.67	100	66.67	**100**
Flagellum	100	100	100	100	100	100	100	100	100	100	**100**
inner	94.62	98.92	100	98.92	93.55	98.92	98.92	94.62	94.62	94.62	**98.92**
Nucleoic	66.67	66.67	100	100	100	100	100	100	100	100	**100**
outer	75.73	81.55	81.55	83.49	82.52	82.52	74.76	90.29	90.29	91.26	**91.26**
Periplasm	83.93	83.03	78.57	86.61	86.61	83.03	**97.32**	90.18	90.18	90.18	90.18
OA	88.36	90.66	91.58	92.04	86.68	86.98	89.43	90.35	92.19	91.73	**94.79**

As shown from Table 1, when other conditions of WD-SVM model are certain, by changing the value of λ, the results have a great impact. With the constant change of λ value, the accuracy of the model for the prediction of each class of proteins in Gram-negative bacteria is also constantly changing. For each class of proteins, when $\lambda = 49$, the prediction accuracy of cytoplasm proteins, fimbrium proteins, flagellum and nucleoid proteins reach 100%. The overall accuracy of prediction of extracell proteins, inner membrane and outer membrane proteins are 84.21%, 98.92% and 91.26% respectively, which all reach the highest value predicted by the model. We also found that for the periplasm proteins, the overall prediction accuracy is 97.32%, 7.14% higher than when $\lambda = 49$. This indicates that the feature information extracted from the protein sequences is suitable for most proteins to classify with the increase of the parameter value, while a little protein may affect the classification effect due to the excessive information extracted. Considering that, when $\lambda = 49$, the prediction model for the seven types of Gram-negative bacterial proteins reach the predicted maximum value. The overall average prediction accuracy rate reach 94.79%. As can be seen from Fig. 5, when $\lambda = 49$, the prediction accuracy of the model is the highest, so the optimal parameter value $\lambda = 49$ of the WD-SVM model is selected.

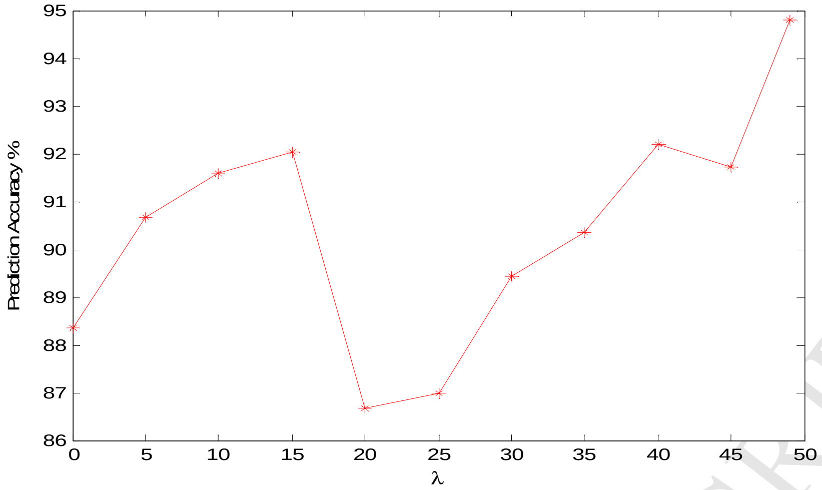

Fig. 5. Effect of selecting different values of λ on the prediction results of subcellular localization.

3.2. Selection of wavelet function and optimal decomposition scale

After the feature extraction of the protein sequences, the wavelet denoising method was used to extract information from the dataset. When dealing with the data, it can be found that choosing different wavelet basis functions will have certain influence on the prediction model. Different wavelet functions are different for different signal processing. If the characteristics of the wavelet function can better match message structure of the signal information, the better feature information can be extracted from the sequences [53]. When analyzing the protein sequences, different decomposition scales have different prediction results. Decomposing a longer sequence with too low a decomposition scale level would omit much detailed information, and decomposing a shorter sequence with too high a decomposition scale would inevitably introduce redundant information [65]. In order to extract the protein sequences feature information of Gram-negative bacteria more effectively, the influence of different wavelet functions and different decomposition scales on the prediction model was investigated.

When 2-D denoising method is used to carry out feature extraction on datasets, there are two different methods for threshold selection, the default threshold, and the threshold gained by Birge-Massart strategy. In the process of noise reduction, using the default threshold will result in a uniform global threshold, whereas thresholds obtained using the Birge-Massart strategy will generate three different thresholds in horizontal, vertical and diagonal. For the Gram-negative bacterial proteins dataset, we found that the default threshold obtained by wavelet packet function obtained the highest overall prediction precision under different wavelet functions and different decomposition scale levels. (The specific results see Supplementary materials Table S1). In the following discussion, we use the wavelet packet to obtain the default thresholds for 2-D noise reduction. When PseAAC was used to extract feature information, $\lambda = 49$ is chosen and classify the samples using linear kernel function of SVM. The results was validated by jackknife test, and the prediction results of subcellular localization of Gram-negative bacterial proteins under different wavelet functions and different decomposition scales were obtained. As is shown in Table 2 (The specific results see Supplementary materials Table S2-S5).

Table 2 Prediction results of subcellular localization in the A653 dataset by different wavelet functions and different decomposition scales.

Scales / Functions	Jackknife test (%)			
	3	**4**	**5**	**6**
db1	**94.79**	79.79	92.80	92.80
db5	92.96	95.87	97.40	97.70
db9	92.34	91.27	97.09	98.01
sym6	88.67	91.24	98.47	98.47
sym8	84.53	87.90	94.64	98.77
coif1	90.96	92.80	97.09	98.32
coif4	86.22	95.87	92.80	97.24
bior2.2	90.35	**97.40**	96.78	97.24
bior2.4	92.34	94.03	**98.77**	98.77
bior2.6	85.76	91.12	96.48	**99.08**

It can be seen from Table 2 that different wavelet functions and different decomposition scales can affect the accuracy of the model. For subcellular localization of Gram-negative bacterial proteins, Table 2 shows that when the decomposition scale is 3, the highest prediction accuracy is 94.79%, and db1 is the most suitable wavelet function. When the decomposition scale level is 4, the highest overall average prediction accuracy is 97.40%. At this time, bior2.2 wavelet function is selected. When the decomposition scale is 5, the highest average prediction accuracy of the model is 98.77%, and bior2.4 wavelet function is chosen. When the decomposition scale is 6, the highest average prediction accuracy of the model is 99.08%, and bior2.6 is the optimal wavelet function. So the wavelet function suitable for the prediction model changes with the different decomposition scale levels. As we all know, wavelet function has many excellent properties, such as compactly support, orthogonality, symmetry, smoothness and high order of vanishing moments. In practice, wavelet function selection has some conflicting constraints, and none of these wavelet functions share simultaneously all of these properties. As is shown in Table 2, when we choose bior2.6 wavelet function and the decomposition scale is 6, the prediction accuracy is the highest. This is because the bior2.6 wavelet is a symmetric biorthogonal spline wavelet function with compact support. This function can select the suitable approximate coefficient and detail coefficient by using Mallat algorithm [66]. Moreover, bior2.6 wavelet function has good locality characteristics in both time and frequency domain, which can also effectively remove high frequency noise and reduce dimensionality, eliminate redundant information of protein sequences, reduce leakage and aliasing of feature information, to make the feature vectors represent the original sequence information well and improve the prediction performance of protein subcellular localization.

In order to more intuitively analyze the bior2.6 wavelet in the choice of decomposition scale 3, 4, 5, 6, the best classification effect of the eight types of Gram-negative bacterial proteins, Fig. 6 shows the overall prediction accuracy of the subcellular localization of eight classes of proteins in Gram-negative bacteria (The specific results see Supplementary materials Table S6). "1-8" in the abscissa indicate eight category of Gram-negative bacteria: cytoplasm, extracell, fimbrium, flagellum, inner membrane, nucleoid, outer membrane, periplasm proteins. It can be seen from Fig. 6 that when the decomposition scale is 6, the prediction accuracy of nucleoid proteins is obviously lower than that when the decomposition scale is 4. Since there are only 6 protein sequences in nucleoid proteins in A653 dataset. When the decomposition scale is 6 of this protein, the subcellular localization prediction accuracy is 5/6 = 83.33%. When the protein was decomposition scale is 4, the subcellular localization prediction accuracy is 6/6 = 100%, only one protein sequence is not predicted. While the other seven kinds of protein in the decomposition scale of 6, the prediction accuracy are ideal compared with that in other decomposition scales. When using bior2.6 wavelet function to denoise the protein sequences, the select of 6 is the optimal decomposition scale.

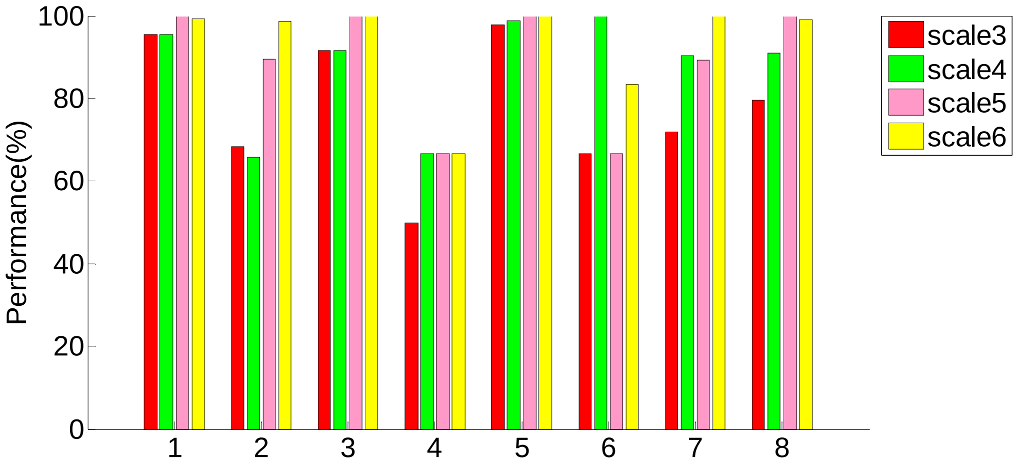

Fig. 6. The performance of different decomposition scales by using the bior2.6 wavelet function.

3.3. Effect of kernel function on results

When SVM is used to classify, the selection of kernel function will have important influence on the prediction results. In this paper, using PseAAC for feature information extraction, choosing $\lambda = 49$ and wavelet denoising using bior2.6 wavelet function, the decomposition scale of 6, select the different kernel functions to get prediction results of subcellular localization of Gram-negative bacterial proteins in the A653 dataset, as is shown in Table 3 . In Table 3, 0 denotes a linear kernel function, 1 denotes a cubic polynomial kernel function, 2 denotes an RBF kernel function, and 3 denotes a sigmoid kernel function. It can be seen from Table 3 that, when the linear kernel function is used, the prediction accuracy of the cytoplasm proteins is 0.66% lower than that of the RBF kernel function, however, for other protein subcellular localization of Gram-negative bacteria, the prediction accuracy is much better than other kernel functions. The overall average prediction accuracy of WD-SVM model is 99.08%, which is 17.76% higher than that of polynomial kernel function, 5.36% higher than RBF kernel function and 33.38% higher than sigmoid kernel function. By comparison, the linear kernel function is chosen as the kernel function of SVM algorithm.

Table 3 Prediction results of subcellular localization of the A653 dataset under different kernel functions.

Functions / Locations	Jackknife test (%)			
	0	**1**	**2**	**3**
cytoplasm	99.34	86.18	100	52.63
extracell	98.68	47.33	97.37	84.21
fimbrium	100	33.33	66.67	0
flagellum	66.67	0	0	0
inner membrane	100	93.55	100	100
nucleoid	83.33	0	0	0
outer membrane	100	100	85.44	60.19
periplasm	99.10	74.11	92.86	33.03
OA	99.08	81.32	93.72	65.70

3.4. Selecting classification algorithms

For a query protein, how to quickly and effectively identify which category it belongs to? Many different machine learning algorithms have been developed at present. In this study, we investigate the effect of six classification algorithms on the prediction of protein subcellular localization. Through many experiments, SVM classification algorithm uses the linear kernel function and principal component analysis (PCA) [67] extracts 25D most important features. Kernel principal component analysis (KPCA)

[68] extracts 22D most important features. The k-nearest neighbor (KNN) algorithm [35,36] choose $K=3$, and the distance is correlation distance. In the feature extraction of the protein sequences, choosing $\lambda=49$ and 2-D wavelet denoising using bior2.6 wavelet function, the decomposition scale of 6. Using jackknife test method, the prediction results obtained by different algorithms, as shown in Table 4.

Table 4 Prediction results of subcellular localization of the A653 dataset under different classification algorithms.

Algorithms / Locations	Jackknife test (%)					
	SVM	PCA-SVM	KPCA-SVM	KNN	WD-KNN	WD-SVM
cytoplasm	76.32	79.61	81.58	79.61	93.42	99.34
extracell	40.79	28.95	36.84	36.84	98.68	98.68
fimbrium	25	16.67	8.33	0	91.67	**100**
flagellum	16.67	16.67	33.33	16.67	83.33	66.67
inner	86.56	83.33	80.65	81.72	98.92	**100**
nucleoid	16.67	16.67	16.67	0	83.33	83.33
Outer	52.43	49.51	50.49	54.37	86.41	**100**
Periplasm	39.28	50.89	53.57	40.18	100	99.10
OA	62.94	62.79	64.01	61.72	95.41	**99.08**

As is shown in Table 4, the overall prediction accuracy of the WD-SVM model with wavelet denoising is 99.08%. The prediction accuracy of the SVM classification algorithm is 62.94%, and the prediction accuracy of the KNN algorithm is 61.72%, which are 36.14% and 37.36% lower than WD-SVM model, respectively. We can see that WD-SVM is much better than SVM and KNN in predicting the accuracy of each class proteins. This shows that using wavelet denoising method to feature information processing of protein sequence data, more conducive to extract each class of protein sequence information, and improve the prediction accuracy of the model. Three different classification models, PCA-SVM, KPCA-SVM and WD-SVM, were only different in different way of dealing with protein sequence information from PseAAC. It can be seen from Table 4 that the prediction accuracy of each class of protein positions is higher than that of PCA-SVM, KPCA-SVM model and WD-SVM model by using wavelet denoising method to extract further information. The overall accuracy of WD-SVM model is 36.29% higher than that of PCA-SVM and 35.07% higher than KPCA-SVM. The feature information are also processed by wavelet denoising, but the accuracy obtained by using different classification algorithm is not the same. From Table 4, it can be seen that WD-KNN algorithm can predict the flagellum proteins of Gram-negative bacteria to be 83.33%, which is 16.66% higher than WD-SVM classification model. The overall prediction accuracy of WD-SVM is 3.67% higher than that of WD-KNN. This indicates that the WD-SVM model constructed by combining the SVM classification algorithm with 2-D wavelet denoising technique can effectively predict protein subcellular localization.

3.5. Performance of prediction model

In this study, the protein sequences are extracted features by PseAAC, and the subcellular localization of proteins is predicted by SVM based on wavelet denoising. According to the above analysis, we choose $\lambda = 49$ using PseAAC to extract feature information and used the bior2.6 wavelet function, the decomposition scale of 6 to decompose the feature data by 2-D wavelet packet denoising, and selected the linear kernel function as the kernel function of SVM. In this paper, self-consistency and jackknife test were used to test Gram-negative bacteria dataset A653, the main results obtained are shown in Table 5.

Table 5 Prediction performance of different test method for protein subcellular localization on the A653 dataset.

Test / Locations	Self-consistency test			Jackknife test		
	S_{in} (%)	S_{ip} (%)	MCC	S_{in} (%)	S_{ip} (%)	MCC
cytoplasm	100	100	1.00	99.34	100	0.9957
extracell	100	100	1.00	98.68	98.68	0.9851
fimbrium	100	100	1.00	100	85.71	0.9244
flagellum	100	100	1.00	66.67	100	0.8152
inner membrane	100	100	1.00	100	99.46	0.9962
nucleoid	100	100	1.00	83.33	100	0.9121
outer membrane	100	100	1.00	100	98.09	0.9886
periplasm	100	100	1.00	99.1	100	0.9946
OA		100			99.08	

As can be seen from Table 5, using the self-consistency method to test the dataset, the overall prediction accuracy rate is 100%, and the dataset for each class of protein sensitivity, specificity and MCC are 100%, 100% and 1, respectively. The experimental results show that the predictive model proposed in this paper can grasp the autocorrelation of subcellular localization of each Gram-negative bacteria. Using jackknife cross-validation to test the model, as can be seen from Table 5, fimbrium, inner membrane and outer membrane proteins' sensitivity all reach 100%. Cytoplasm, flagellum, nucleoid and periplasm proteins' specificity reach 100%. And Matthew's correlation coefficient (MCC) of cytoplasm, extracell, inner membrane, outer membrane and periplasm proteins reach 0.9957, 0.9851, 0.9962, 0.9886 and 0.9946, respectively. The jackknife method was used to test the model, and the overall prediction accuracy of the model is 99.08%. These results suggest that the proposed WD-SVM model is an effective tool for protein subcellular localization prediction.

3.6. Comparison with other methods

In order to evaluate the effectiveness of the proposed WD-SVM model more objectively, the prediction accuracy of the model is compared with the prediction methods with the same dataset based on jackknife test. Table 6 provides a detailed comparison of the results of this method and other prediction methods on the A653 dataset. The WD-SVM$_1$ model uses PseAAC to carry out feature extraction, $\lambda = 49$, using bior2.4 wavelet function, decomposition scale of 5 to denoise data, SVM uses linear kernel function. WD-SVM model uses PseAAC feature extraction, $\lambda = 49$, using bior2.6 wavelet function, decomposition of the scale 6 to reduce noise of the data, SVM selection linear kernel function.

Table 6 Prediction results with different methods on the A653 dataset using jackknife test.

Methods	Jackknife test (%)				
Locations	l_1 LPP-KNN[50]	LDA-KNN[34]	KLDA-KNN[35]	WD-SVM$_1$	WD-SVM
cytoplasm	$\frac{151}{152}=99.34$	$\frac{148}{152}=97.37$	$\frac{152}{152}=100$	$\frac{152}{152}=100$	$\frac{151}{152}=99.34$
extracell	$\frac{74}{76}=97.37$	$\frac{73}{76}=96.05$	$\frac{73}{76}=96.05$	$\frac{75}{76}=98.68$	$\frac{75}{76}=98.68$
fimbrium	$\frac{10}{12}=83.33$	$\frac{10}{12}=83.33$	$\frac{12}{12}=100$	$\frac{10}{12}=83.33$	$\frac{12}{12}=100$
flagellum	$\frac{5}{6}=83.33$	$\frac{5}{6}=83.33$	$\frac{6}{6}=100$	$\frac{6}{6}=100$	$\frac{4}{6}=66.67$
inner membrane	$\frac{179}{186}=96.24$	$\frac{171}{186}=91.94$	$\frac{186}{186}=100$	$\frac{186}{186}=100$	$\frac{186}{186}=100$
nucleoid	$\frac{6}{6}=100$	$\frac{6}{6}=100$	$\frac{6}{6}=100$	$\frac{5}{6}=83.33$	$\frac{5}{6}=83.33$
outer membrane	$\frac{99}{103}=96.12$	$\frac{95}{103}=92.23$	$\frac{100}{103}=97.09$	$\frac{100}{103}=97.09$	$\frac{103}{103}=100$
periplasm	$\frac{109}{112}=97.32$	$\frac{103}{112}=91.96$	$\frac{110}{112}=98.21$	$\frac{111}{112}=99.11$	$\frac{111}{112}=99.11$
OA	$\frac{633}{653}=96.94$	$\frac{611}{653}=93.57$	$\frac{645}{653}=98.77$	$\frac{645}{653}=98.77$	$\frac{647}{653}=99.08$

As shown from Table 6, the overall prediction accuracy of WD-SVM$_1$ model is 98.77%, which are 1.83% and 5.2% higher than those of the l_1 LPP-KNN and LDA-KNN models, respectively. It is the same with the prediction accuracy of KLDA-KNN model. The overall prediction accuracy of WD-SVM model is 99.08%, which are 2.14%, 5.51% and 0.31% higher than those of the l_1 LPP-KNN, LDA-KNN and KLDA-KNN models, respectively. The WD-SVM$_1$ model proposed by this paper predicted subcellular localization of cytoplasm, extracell, flagellum, inner membrane and periplasm proteins, and the prediction accuracy of each category reach 100%, 98.68%, 100% , 100% and 99.11%, respectively, and reached the maximum of several different prediction methods. We also found that WD-SVM model predicted the subcellular localization of extracell, fimbrium, inner membrane, outer membrane and periplasm proteins, and the prediction accuracy of each class reach 98.68% , 100%, 100%, 100% and 99.11%, respectively, also reached the highest accuracy among the different prediction methods. As shown in Table 6, WD-SVM shows poor prediction results for flagellum and nucleoid proteins, which is lower than other methods, but WD-SVM$_1$ model could predict flagellum by 100%. Considering that only the six protein sequences are present in the dataset, the number of protein sequences is too small, which affects the construction of the training model.

In order to further validate the actual predictive ability of WD-SVM model, we used the independent testing dataset C1114 and compared with other reported prediction methods. Table 7 shows the predictive results of the subcellular localization on the C1114 dataset (The specific results of the dataset C1114 are shown in Table S7 of Supplementary materials). PSORT-B [26] was constructed by fusing the amino acid composition, N-terminal sorting signal, motif and other characteristics of the coding method, combined with SVM algorithm to establish Gram-negative bacteria proteins subcellular localization prediction model. Gneg-Ploc [2] was based on KNN using GO and PseAAC to establish protein subcellular localization prediction model. For WD-SVM model, we chose $\lambda = 28$ using PseAAC to extract features on the protein sequences (since the length of the shortest protein in the test dataset is 29, the maximum value of λ for this paper is 28), and used bior2.6 wavelet function, the decomposition scale is 6. The 2-D wavelet packet denoising is applied to obtain feature information, and the linear kernel function is used as the kernel function of SVM. As shown from Table 7, the WD-SVM model achieves the highest overall prediction accuracy of 98.6% on the C1114 dataset, which are 19.2% and 3.2% higher than those of the PSORT-B and Gneg-Ploc model, respectively. In addition, the predictive accuracy of cytoplasm, extracell, inner membrane and outer membrane proteins are higher than that of other models by the use of WD-SVM. For example, the prediction accuracy of the cytoplasm proteins, WD-SVM model is 100%, which is 16.4% and 6.4% higher than those of the PSORT-B and Gneg-Ploc prediction models, respectively. It indicates that the model of this paper has excellent properties for the prediction subcellular localization of Gram-negative bacteria proteins. In summary, our proposed method has achieved satisfactory prediction results.

Table 7 Performance comparison of the independent testing dataset by the jackknife test on the C1114 dataset.

Methods / Locations	Jackknife test (%)		
	PSORT-B[26]	Gneg-Ploc[2]	WD-SVM
cytoplasm	117/140=83.6	131/140=93.6	140/140=100
extracell	23/74=31.1	67/74=90.5	68/74=91.9
inner membrane	570/687=83.0	670/687=97.5	686/687=99.9
outer membrane	80/97=82.5	85/97=87.6	94/97=96.9
periplasm	94/116=81.0	110/116=94.8	110/116=94.8
OA	884/1114=79.4	1063/1114=95.4	1098/1114=98.6

4. Conclusion

With the advent of the big data age, massive protein sequences are exponentially growing into the database. Therefore, traditional biological experimental methods are unable to meet the need of life science research, so it is more and more important to study protein subcellular localization prediction method based on computational intelligence. In this paper, a novel protein subcellular localization prediction method is proposed. Using Chou's PseAAC to carry out feature extraction on protein sequences of Gram-negative bacteria, and subcellular localization of proteins is predicted by SVM algorithm based on 2-D wavelet denoising. PseAAC feature extraction can avoid the loss of protein sequence-order information, and get detailed protein sequence information. 2-D wavelet denoising can effectively remove redundant information in protein sequences, and extract the characteristic signals of each class of protein, which is very important for guaranteeing high prediction accuracy. SVM algorithm can easily deal with high-dimensional data, to avoid over-fitting and effectively eliminate non-support vector. By jackknife test, the overall prediction accuracies of the two benchmark datasets reach 99.08% and 98.6%, respectively, and compared with the method proposed by Wang et al. [34,35], Zheng et al. [50], Gardy et al. [26] and Chou et al. [2]. The results show that the proposed method not only can effectively improve the prediction accuracy of protein subcellular localization, but also is expected to be used for the prediction of other attributes of proteins.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [18,69], we shall make efforts in our future work to provide a web-server for the method presented in this paper. The source code for implementing in this study is available from the author upon request.

Acknowledgements

The authors sincerely thank the anonymous referees for their many valuable comments that have improved this manuscript. This work was supported by the National Natural Science Foundation of China (Nos. 51372125, 21571112), The Natural Science Foundation of Shandong Province of China (No. ZR2014FL021), A Project of Shandong Province Higher Educational Science and Technology Program (No. J12L156), The Applied Basic Research Project of Qingdao of China (No. 15-9-1-83-JCH) and College Students' Innovative Entrepreneurial Training Program (No. 201606002).

References

[1] F. Eisenhaber, P. Bork, Wanted: subcellular localization of proteins based on sequence, Trends cell biol. 8 (4) (1998) 169-170.

[2] K.C. Chou, H.B. Shen, Large-scale predictions of gram-negative bacterial protein subcellular locations, J. Proteome Res. 5 (12) (2006) 3420-3428.

[3] J.L. Gardy, F.S. Brinkman, Methods for predicting bacterial protein subcellular localization, Nat. Rev. Microbiol. 4 (10) (2006) 741-751.

[4] A. Garg, M. Bhasin, G.P.S. Raghava, Support vector machine-based method for subcellular location of human proteins using amino acid compositions, their order and similarity search, J. Biol. Chem. 280 (15) (2005) 14427-14432.

[5] Z. P. Feng, An overview on predicting the subcellular location of a protein, In Silico Biol. 2 (3) (2002) 291-303.

[6] K.C. Chou, Y.D. Cai, Using function domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (48) (2002) 45765-45769.

[7] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (1) (2007) 1-16.

[8] K. Nakai, M. Kanehisa, Expert system for predicting protein localization sites in Gram- negative bacteria, Proteins: Struct. Funct. Bioinform. 11 (2) (1991) 95-110.

[9] K. Nakai, P. Horton, PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem. Sci. 24 (1) (1999) 34-36.

[10] O. Emanuelsson, H. Nielsen, S. Brunk, G. Von Heijne. Predicting subcelluar localization of proteins based on their N-terminal amino acids sequences, J. Mol. Biol. 300 (4) (2000) 1005-1016.

[11] H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, J. Mol. Biol. 238 (1) (1994) 54-61.

[12] A. Reinhardt, T. Hubbard, Using neural networks for prediction of the subcellular location of ptoteins, Nucleic Acids Res. 26 (9) (1998) 2230-2236.

[13] S.J Hua, Z.R. Sun, Support vector machine approach for protein subcellular localization prediction, Bioinformatics 17 (8) (2001) 721-728.

[14] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Struct. Funct. Bioinform. 43 (3) (2001) 246-255.

[15] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2005) 10-19.

[16] H.B. Shen, K.C. Chou, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2) (2008) 386-388.

[17] L. Nanni, S. Brahnam, A. Lumini, Prediction of protein structure classes by in corporating different protein descriptors into general Chou's pseudo amino acid composition, J. Theor. Biol. 360 (2014) 109-116.

[18] S.L. Zhang, Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC, Chemometrics and Intelligent Laboratory Systems, 142 (15) (2015) 28-35.

[19] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol., 273 (1) (2011) 236-247.

[20] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou’s amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (3) (2007) 546-551.

[21] L. Liu, Y.D. Cai, W.C. Lu, K.Y. Feng, C.R. Peng, B Niu, Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection, Biochem. Biophys. Res. Commun. 380 (2) (2009) 318-322.

[22] Y.N. Zhang, X.Y. Pan, Y. Huang, H.B. Shen, Adaptive compressive learning for prediction of protein-protein interactions from primary sequence, J. Theor. Biol. 283 (1) (2011) 44-52.

[23] P.P. Zhu, W.C. Li, Z.J. Zhong, E.Z Deng, H. Ding, W. Chen, H. Lin, Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition, Mol. BioSyst. 11 (2) (2015) 558-563.

[24] S.B. Wan, M. W. Mak, S.Y. Kung, GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition, J. Theor. Biol. 323 (2013) 40-48.

[25] X.Q. Yu, X.Q. Zheng, T.G. Liu, Y.C. Dou, J. Wang, Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation, Amino Acids 42 (5) (2012) 1619-1625.

[26] J.L. Gardy, C. Spencer, K. Wang, M. Ester, G.E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria, Nucleic Acids Res. 31 (13) (2003) 3613-3617.

[27] K.C Chou, Y.D. Cai, Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition, J. Cell. Biochem. 91 (6) (2004) 1197-1203.

[28] K.C. Chou, Y.D. Cai, A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology, Biochem. Biophys. Res. Commun. 311 (3) (2003) 743-747.

[29] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, FEBS Lett. 580 (26) (2006) 6169-6174.

[30] K.C. Chou, H.B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0, PloS ONE 5 (4) (2010) e9931.

[31] T.G. Liu, P.Y. Tao, X.W. Li, Y.F. Qin, C.H. Wang, Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based in PSSM and recursive feature elimination, J. Theor. Biol. 366 (2015) 8-12.

[32] L.Q. Li, S.J. Yu, W.D. Xiao, H. Yang, Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach, Biochimie 104 (2014) 100-107.

[33] J. Chen, H.M. Xu, P.A. He, Y.H. Yao, A multiple information fusion method for prediction subcellular location of two different types of bacterial protein simultaneneously, BioSystems 139 (2016) 37-45.

[34] T. Wang, J. Yang, Predicting subcellular localization of gram-negative bacterial proteins by linear dimensionality reduction method, Protein Pept. Lett. 17 (1) (2010) 32-37.

[35] T. Wang, J. Yang, Using the nonlinear dimensinality reduction method for the prediction of subcellular localization of gram-negative bacterial proteins, Mol. Divers. 13 (2009) 475-481.

[36] Y. Huang, Y.D. Li, Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 20 (1) (2004) 21-28.

[37] H. Saini, G. Raicar, A. Dehzangi, S. Lal, A. Sharma, Subcellular localization for Gram positive and Gram negative bacterial proteins using linear interpolation smoothing model, J. Theor. Biol. 386 (2015) 25-33.

[38] M.S. Scott, D.Y. Thomas, M.T. Hallett, Predicting subcellular localization via protein motif co-occurrence, Genome Res. 14 (2004) 1957-1966.

[39] B.R. King, S Vural, S. Pandey, A. Barteau, C. Guda, ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes, BMC Res. Notes 5 (2012) 351.

[40] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophysical J. 84 (5) (2003) 3257-3263.

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

[42] L.C. Zhang, L. Kong, X.D, Han, J.F. Lv, Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure, J. Theor. Biol. 400 (2016) 1-10.

[43] J.W. Shen, J. Zhang, X.M. Luo, W.L. Zhu, K.X. Chen, Y.X. Li, H.L. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Nati. Acad. Sci. USA 104 (11) (2007) 4337-4341.

[44] Y.Z. Guo, L.Z. Yu, Z.N. Wen, M.L. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucl. Acids Res. 36 (9) (2008) 3025-3030.

[45] Q. Dai, S. Ma, Y.B. Hai, Y.H Yao, X.Q. Liu, A segmentation based model for subcellular location prediction of apoptosis protein, Chemometrics and Intelligent Laboratory Systems 158 (2016) 146-154.

[46] Z.C. Li, X. Zhou, Z. Dai, X.Y. Zou, Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm, BMC Bioinforma. 11 (2010) 325.

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

[48] Z. Chen, Y. Zhou, Z.D. Zhang, J.G. Song, Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features, Briefings in Bioinformatics 16 (4) (2015) 640-657.

[49] Z.X. Liu, J. Cao, Q. Ma, X.J. Cao, J. Ren, Y. Xue, GPS-YNo2: computational prediction of tyrosine nitration sites in proteins, Mol. BioSyst. 7 (4) (2011) 1197-1204.

[50] Z.L. Zheng, J. Yang, Subcellular localization of gram-negative bacterial proteins using sparse learning, Protein J. 29 (3) (2010) 195-203.

[51] P. Liò, Wavelets in bioinformatics and computational biology: state of art and perspectives, Bioinformatics 19 (1) (2003) 2-9.

[52] B. Yu, Y. Zhang, A simple method for predicting transmembrane proteins based on wavelet transform, Int. J. Biol. Sci. 9 (1) (2013) 22-33.

[53] R.P. Liang, S.Y. Huang, S.P. Shi, X.Y. Sun, S.B. Suo, J.D. Qiu, A novel algorithm novel algorithm combining support vector machine with the discrete wavelet transform or the prediction of protein subcellular localization, Comput. Biol. Med. 42 (2) (2012) 180-187.

[54] J.D. Qiu, S.H. Luo, J.H. Huang, R.P. Liang, Using support vector machines for prediction of protein structural classes based on discrete wavelet transform, J. Comput. Chem. 30 (8) (2009) 1344-1350.

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

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

[57] B. Yu, Y. Zhang, The analysis of colon cancer gene expression profiles and the extraction of informative genes, J. Comput. Theor. Nanosci. 10 (5) (2013) 1097-1103.

[58] A. Kandaswamy, C.S. Kumar, R.P. Ramanathan, S. Jayaraman, N. Malmurugan, Neural classification of lung sounds using wavelet coefficients, Comput. Biol. Med. 34 (6) (2004) 523-537.

[59] F. Luisier, T. Blu, M. Unser, A new SURE approach to image denoising: interscale orthonormal wavelet thresholding, IEEE Trans. Image Process. 16 (3) (2007) 593-606.

[60] D.L. Donoho, De-Noising by soft-thresholding, IEEE Trans. Information Theory 41 (3) (1995) 613-627.

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

[62] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes, Pattern Recognition 44 (8) (2011) 1761-1776.

[63] J.C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, Adv. Neural Inf. Process. Syst. 12 (3) (2000) 547-553.

[63] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1-27.

[65] Z.N. Wen, K.L. Wang, M.L. Li, F.S. Nie, Y. Yang, Analyzing functional similarity of protein sequences with discrete wavelet transform, Comput. Biol. Chem. 29 (3) (2005) 220-228.

[66] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674-693.

[67] Q.S. Du, Z.Q. Jiang, W.Z. He, D.P. Li, K.C. Chou, Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction, J. Biomol. Struct. Dyn. 23 (6) (2006) 635-640.

[68] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299-1319.

[69] K.C. Chou, H.B. Shen, Review: recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2) (2009) 63-92.

Highlights

- A new method (WD-SVM) to prediction protein subcellular localization.
- 2-D wavelet denoising can effectively remove redundant information in protein sequences.
- The optimal parameters of the prediction model are discussed from several aspects.
- We investigate the effect of the six prediction algorithms on the results.
- The proposed method increases the prediction performance over several methods.