Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach

Baoguang Tian , Xue Wu , Cheng Chen , Wenying Qiu , Qin Ma , Bin Yu

PII: S0022-5193(18)30564-2
DOI: https://doi.org/10.1016/j.jtbi.2018.11.011
Reference: YJTBI 9713

To appear in: *Journal of Theoretical Biology*

Received date: 13 September 2018
Revised date: 8 November 2018
Accepted date: 15 November 2018

Please cite this article as: Baoguang Tian , Xue Wu , Cheng Chen , Wenying Qiu , Qin Ma , Bin Yu , Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach, *Journal of Theoretical Biology* (2018), doi: https://doi.org/10.1016/j.jtbi.2018.11.011

Highlights

- A new method (PPIs-WDSVM) for prediction protein-protein interactions.
- The protein sequence features are extracted by fusing the PseAAC, AC and EBGW methods.
- 2-D wavelet denoising can effectively remove the redundant information in the protein sequences.
- We compared the effect of the five different classifiers on the results.
- The proposed method increases the prediction performance over several methods.

Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach

Baoguang Tian[a,b], Xue Wu[a,b], Cheng Chen[a,b], Wenying Qiu[a,b], Qin Ma[c], Bin Yu[a,b,d,e,*]

[a]College of Mathematics 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]Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD 57007, USA

[d]School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

[e]Department of Biochemistry & Molecular Biology, Medical Genetics, and Oncology, University of Calgary, Calgary T2N 4N1, Canada

[*]Corresponding author: *E-mail address*: yubin@qust.edu.cn (B. Yu).

Abstract: Research on protein-protein interactions (PPIs) not only helps to reveal the nature of life activities but also plays a driving role in understanding the mechanisms of disease activity and the development of effective drugs. The rapid development of machine learning provides new opportunities and challenges for understanding the mechanism of PPIs. It plays an important role in the field of proteomics research. In recent years, an increasing number of computational methods for predicting PPIs have been developed. This paper proposes a new method for predicting PPIs based on multi-information fusion. First, the pseudo-amino acid composition (PseAAC), auto-covariance (AC) and encoding based on grouped weight (EBGW) methods are used to extract the features of protein sequences, and the extracted three groups of feature vectors were fused. Secondly, the fused feature vectors are denoised by two-dimensional (2-D) wavelet denoising. Finally, the denoised feature vectors are input to the support vector machine (SVM) classifier to predict the PPIs. The ACC of PPIs of Helicobacter pylori (H. pylori) and Saccharomyces cerevisiae (S. cerevisiae) datasets were 95.97% and 95.55% by 5-fold cross-validation test and compared with other prediction methods. The experimental results show that the proposed multi-information fusion prediction method can effectively improve the prediction performance of PPIs. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/PPIs-WDSVM/.

Keywords: Protein-protein interactions; Pseudo-amino acid composition; Multi-information fusion; Two-dimensional wavelet denoising; Support vector machine; Machine learning

1. Introduction

PPIs play a central role in life activities. Numerous life processes such as DNA replication, transcription, protein translation, modification and localization, and information transmission are closely related to PPIs (Hu et al., 2011; Zhang et al., 2012; Yin et al., 2014; Zhang et al., 2016; Murakami et al., 2017). The study of PPIs not only reveals the rules of cell activities such as growth and development, metabolism, signal transduction, differentiation, and apoptosis at the molecular level but also provides an important theoretical basis for exploring disease mechanism, disease treatment, disease prevention, drug target discovery, and drug design. The technology of PPIs has been mature, and considerable achievements have been made, and advances in proteomics have been promoted. However, these methods are constrained by the current experimental science, which is time-consuming and laborious, and brings a large number of "false positive" and "false negative" data while detecting a large number of PPIs (Han et al., 2005; You et al., 2010; You et al., 2015; Huang et al., 2016). Therefore, research on PPIs has turned to seek machine learning method predictions. To improve the prediction accuracy of PPIs, it is of great significance to develop reliable machine learning methods (Shoemaker and Panchenko, 2007; Wei et al., 2017).

In order to solve the defects of biological experiment technology, assist the biological experiment, and analyze the interaction data obtained by the experiment, using the machine learning (Szilágyi et al., 2005; Hu et al., 2012; Ren et al., 2011; Wang et al., 2014; Folador et al., 2014; Zhu et al., 2015; Zhang and Wei, 2015; Liu et al., 2015) method to study PPIs has become a hot spot for researchers. Its primary job is to extract feature information of proteins. The characteristics of proteins mainly include amino acid sequence information (Chou and Cai, 2006; Hamp and Rost, 2015; Zhang et al., 2011), structural information (Binny et al., 2013; Zhang et al.,2016; Mier et al., 2017; Yu et al., 2017c), evolutionary information (Zhang and Tang, 2016) and so on. Protein sequence information is the most basic and easy to obtain. Its composition determines the structural information of the protein, and the structural information determines the function of the protein. At present, many studies are based on the sequence information of proteins. For example, Guo et al. (2008) first proposed the use of AC coding of hydrogen acid sequences to predict PPIs. First, they get the 7 kinds of physiochemical properties of 20 kinds of amino acids, including hydrophobicity, hydrophilicity, volumes of side chains of amino acids, polarity, polarizability, solvent-accessible surface area (SASA) and net charge index (NCI) of side chains of amino acids, and then these vectorized data are encoded according to the AC method. Finally, the prediction accuracy of the *Yeast* dataset obtained by the SVM based on the radial basis function (RBF) as the classifier is 88.09%. Wang et al. (2017) proposed a DNN-LCTD method based on protein sequence. The local conjoint triad descriptor (LCTD) is used as the feature

extraction method. LCTD combines the advantages of local description and conjoint triad. It is capable of interpreting the interaction between residues in consecutive and discontinuous regions of the amino acid sequence. Deep neural networks (DNNs) not only learn the right features from the data by themselves but also learn and discover hierarchical representations of the data. The *S. cerevisiae* dataset was processed by the DNN-LCTD method with a prediction accuracy of 93.12%. Wang et al. (2018) extracted features from protein sequences using continuous and discrete wavelet transforms and combined with weighted sparse representation-based classifiers (WSRC) to predict the *Yeast*, *Human*, and *H. pylori* datasets. These methods search for the rules followed by PPIs by studying the sequence characteristics of proteins and building corresponding machine learning models to determine whether two proteins interact.

However, the single type of feature information cannot adequately reflect the function of the protein, which affects the prediction accuracy of PPIs. In recent years, researchers have been more inclined to fusion multiple feature information to describe protein sequences, in order to integrate the advantages of each sequence coding method, obtain more protein sequence feature information, and improve the prediction accuracy of PPIs (Lei et al., 2012; Li et al., 2012; Zhu et al., 2013a). For example, Du et al. (2017) proposed a method for deep neural networks for PPIs prediction (Deep PPI). They use amino acid composition (ACC), dipeptide composition (DC), composition, transition and distribution, Quasi-Sequence-Order Descriptors and amphiphilic pseudo amino acid composition (APAAC) as feature extraction method, predicting with a deep neural network (DNN) as a classifier, the prediction accuracy of the *S. cerevisiae*, *H. pylori* and *H. sapien*s datasets were 94.43%, 86.23% and 98.14%, respectively. Wei et al. (2017) artificially generated two PPIs datasets from the PPIs database and proposed two novel feature extraction methods, one based on the physicochemical properties of proteins, and the other based on the secondary structural information. LibD3C developed by Lin et al. (2014) is an integrated classifier. Based on the physicochemical properties of proteins, the prediction accuracy of the *Negatome*, *RandomParis*, and *RecommendationPairs* datasets are 83.20%, 99.30%, and 56.70%, respectively. The prediction accuracy of the *Negatome*, *RandomParis* and *RecombinePairs* datasets obtained by the secondary structure information are 81.00%, 96.70%, and 57.50%, respectively. Göktepe et al. (2018) applied a new technique called weighted skip-sequential conjoint triads, which consists of Bi-Gram representations, pseudo-amino acid composition and different forms of Conjoint triad as feature extraction method, the extracted three sets of feature vectors were fused, and principal component analysis (PCA) was used for dimensionality reduction, and SVM was used as the prediction classifier to obtain the prediction accuracy of HPRD (Human protein reference database) which was 93.45%.

Not only feature extraction methods, but classification algorithms also affect the prediction

accuracy of PPIs. In recent years, the pattern recognition methods such as statistics and machine learning (Zhu et al., 2013b; Luo et al., 2015; Angermueller et al., 2016; An et al., 2016; Gao et al., 2016; Wang et al., 2017; Wen et al., 2017) have been widely used in the study of PPIs prediction. The commonly used methods include K-neighbor method, neural network, the Bayesian network, random forest, support vector machine and ensemble learning. For example, Yousef et al. (2013) used twelve physicochemical properties of amino acids to convert sequences of protein pairs into different eigenvectors, using principal component analysis (PCA) to denoise eigenvectors, and using the learning vector quantization (LVQ) as the classifier, the prediction accuracy of *S. cerevisiae*, *H. pylori* and independent test sets were 93.88%, 90.03%, and 89.72%, respectively. Huang et al. (2015) used the discrete cosine transform (DCT) on the substitution matrix representation (SMR) to extract the feature data of PPIs. They use weighted sparse representation-based classifier (WSRC) as the classifier, which is validated effectively on the datasets of *Yeast*, *Human*, and *H. pylori*. Jia et al. (2015a) developed a new predictive classifier called iPPI-Emsl, which extracted the feature of the data by the PseAAC method, and transformed the numerical sequences using discrete wavelet transform (DWT). Using the random forest (RF) as the ensemble classifier, the prediction accuracy of the *H. pylori* dataset using the 10-fold cross-validation test is 90.75%. Ding et al. (2016) used the multivariate mutual information (MMI) method to extract the features of the data and also used the RF as the classifier to obtain the prediction accuracy of the *H. pylori* and *S. cerevisiae* datasets which are 87.59% and 95.01%, respectively. Li et al. (2017) used the improved weber local descriptor (IWLD) combined with the position-specific scoring matrix (PSSM) to extract the features of the *H. pylori* and *Yeast* datasets. Using the discriminative vector machine classifier (DVM) as the classifier, the prediction accuracy of the two datasets are 96.52% and 91.80%, respectively. Wang et al. (2017) proposed a computational method based on the probabilistic classification vector machine (PCVM) model and the zernike moment (ZM) descriptor to predict PPIs called PCVMZM, the prediction accuracy of the *Yeast* and *H. pylori* datasets were 94.48% and 91.25%, respectively. In recent years, many prediction methods for studying PPIs have been developed, but their prediction effects are not prominent enough. Therefore, the development of new prediction methods poses new challenges for studying PPIs.

In this paper, we propose a new prediction method for predicting PPIs called PPIs-WDSVM based on multi-information fusion. Firstly, the PseAAC, ACC and EBGW methods were used to extract the features of the protein sequences in the *H. pylori* and *S. cerevisiae* datasets, and the extracted feature vectors were fused. Secondly, the 2-D wavelet denoising was performed on the fused protein feature vectors to eliminate redundant information. Finally, the best feature vector after noise reduction is input to the SVM classifier for prediction. By comparing the different

parameter values of feature extraction methods, wavelet functions, decomposition scales and the effects of classifiers on the results and after comparison analysis, determine the optimal parameters of the prediction model. Using SVM as a classifier, the average ACC of the *H. pylori* and *S. cerevisiae* datasets are 95.97% and 95.55% and compared with the other four classifiers. Finally, the prediction results of *H. pylori* and *S. cerevisiae* datasets are compared with other methods. The experimental results show that the proposed PPIs-WDSVM prediction model based on multi-information fusion is better than the existing research results and can significantly improve the predictive effect of PPIs.

To develop a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications (Liu et al., 2017a; 2017b; 2017c; Liu et al., 2018; Cheng et al., 2017a, 2017b, 2017c; 2017d; 2017e; 2017f; 2017g; Cheng et al., 2018a; 2018b; Feng et al., 2017; Feng et al., 2018; Su et al., 2018; Chen et al., 2018a, 2018b; Yang et al., 2018; Chou et al., 2018), one should observe the Chou's (2011) 5-step rule 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 robust 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

To make a reasonable comparison with the existing PPIs prediction results, this paper uses a wide range of *H. pylori* and *S. cerevisiae* PPIs datasets to evaluate the pros and cons of the model.

The *H. pylori* dataset was constructed by Martin et al. (2005) and was used by many researchers to test the validity of the computational method. This dataset contains 2,916 protein pairs, of which 1,458 protein pairs are in both positive and negative samples. The *S. cerevisiae* dataset is from the DIP (www.dip.doe-mbi.ucla.edu) core database, version DIP_20070219 (Xenarios et al., 2002). In order to reduce the fragment and sequence similarity, the samples with fewer than 50 residues in the positive sample are eliminated, and the protein sequence with sequence similarity greater than or equal to 40% is filtered, thereby reducing the redundant information existing in the dataset, so 5594 PPIs pairs were obtained as positive sample data, and 5594 protein sequences with different subcellular location information were selected as negative samples, and 11188 protein pairs were finally obtained, and both positive and negative samples had 5594 protein pairs.

2.2. Pseudo-amino acid composition

With the advent of the post-genomic era, the focus of bioinformatics research has shifted from genome sequencing to the annotation of the functions of sequenced genomes, especially the study of PPIs. The most crucial question is how to convert protein sequence information into a numerical vector and retain the characteristic information of most sequences. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a comprehensive review (Chou, 2015). Chou (2005) proposed the PseAAC feature extraction method in 2005. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics (Huang et al., 2011; Dehzangi et al., 2015; Behbahani et al., 2016; Meher et al., 2017; Chou, 2017), and three powerful open access soft-wares, called 'PseAAC-Builder', 'propy', and 'PseAAC-General', were established: the former two are for generating various modes of Chou's special PseAAC (Chou, 2009); while the 3rd one for those of Chou's general PseAAC (Chou, 2011), including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs. (9) and (10) of (Chou, 2011)), "Gene Ontology" mode (see Eqs. (11) and (12) of (Chou, 2011)), and "Sequential Evolution" or "PSSM" mode (see Eqs. (13) and (14) of (Chou, 2011)). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) (Chen et al., 2014) was developed for generating various feature vectors for DNA/RNA sequences (Wei et al., 2015; Liu et al., 2017c) that have proved very useful as well. Recently a very powerful web-server called 'Pse-in-One' (Liu et al., 2015) and its updated version 'Pse-in-One2.0' (Liu et al., 2017d) have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies. We choose the PseAAC method as one of the feature extraction methods. The method comprehensively considers the sequence information of protein sequences and the physicochemical properties of amino acids, and is widely used for PPIs (Chou et al., 2016; Zhao et al., 2012; Zhai et al., 2017; Jia et al., 2018), subcellular localization (Chou and Shen, 2007; Liao et al., 2011; Mandal et al., 2015; Cheng et al., 2017d; Yu et al., 2017a, 2017b), submitochondrial localization (Du and Yu, 2013; Du and Jiao, 2017; Ahmad et al., 2016; Qiu et al., 2018), protein structural prediction (Yu et al., 2017c; Liang et al., 2014; Zhang, 2015) and other areas (Xu et al., 2013; Liu et al., 2014; Xu et al., 2015; Khan et al., 2017). They developed the corresponding online server (Shen and Chou, 2008; Liu et al., 2015; Liu et al., 2017d). The specific steps of the PseAAC method are shown in the following equations:

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

Each of these components is given as follows:

$$P_u = \begin{cases} \dfrac{f_u}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k}, & 1 \le u \le 20 \\ \dfrac{\omega_{\tau_{u-20}}}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k}, & 20+1 \le u \le 20+\lambda \end{cases} \quad (2)$$

where P represents the feature vector, ω is the weighting factor, the value is 0.05 (Chou, 2001a). τ_k is the correlation factor of the k-th sequence, which reflects the sequence information between the amino acids in the sequence. f_u $(u = 1, 2, \ldots, 20)$ is expressed as the frequency of occurrence of the u-th amino acid in the protein in the sequence. It can be seen from the above formula that the first 20 dimensions of the feature vector P are amino acid compositions, and the latter λ dimension is a sequence correlation factor reflecting different levels of amino acid sequence information. The sequence-related factor can be calculated from the hydrophobic index, the hydrophilic index, and the side chain molecular weight of the residue. The value of λ must be less than the length of the shortest protein sequence in the training set. In the case of $\lambda = 0$, PseAAC degenerates into the traditional AAC.

In this paper, PseAAC was selected as one of the methods for protein sequence feature extraction. Since the shortest sequence of *H. pylori* and *S. cerevisiae* datasets has a length of 12, λ ranges from 0 to 11. The protein sequence is predicted by selecting different parameters λ, and the optimal parameter λ is determined by the ACC, and the extracted feature vector is the $2 \times (20 + \lambda)$ dimension.

2.3. Auto covariance

The auto-covariance (AC) coding method proposed by Guo et al. (2008) was used as one of the feature extraction methods for protein interaction. This method fully considers the long-range interaction among amino acids in the sequence and can effectively improve the prediction accuracy of PPIs. The seven physicochemical properties of the amino acids were first extracted, including hydrophobicity, hydrophilicity, volumes of side chains of amino acids, polarity, polarizability, SASA and NCI of side chains of amino acids. Subsequently, the AC is used to convert protein sequences of different lengths into data mats of the same length. Its calculation formula is as follows:

$$AC_{lag,j} = \frac{1}{n - lag} \sum_{i=1}^{n=lag} (X_{i,j} - \frac{1}{n} \sum_{i=1}^{n} X_{i,j}) \times (X_{(i+lag),j} - \frac{1}{n} \sum_{i=1}^{n} X_{i,j}) \quad (3)$$

where lag represents the separation distance of amino acid disability at two different positions in the sequence, i is the amino acid disability position, j represents an amino acid descriptor,

and n is the sequence length.

Combining the physicochemical properties of amino acids with protein sequences is beneficial to improve the predictive performance of PPIs model. In this paper, the AC method is used to extract the features of protein sequences. The optimal lag is determined from the accuracy of the prediction result by selecting different parameters lag. The difference in lag value affects the dimension of the extracted feature vector, and the dimension is $2\times(7\times lag)$. Although this paper is a prediction study of PPIs, the application of amino acid physicochemical properties can be easily extended to the field of bioinformatics (Lapinsh et al., 2002; Guo et al., 2006; Liu et al., 2010; Wu et al., 2010; Yang et al., 2010; Liu et al., 2012; Gao et al., 2016), such as protein subcellular localization, protein structure prediction.

2.4. Encoding based on grouped weigh

Grouping weight coding (EBGW) was proposed by Zhang et al. (2005). Based on the physicochemical properties of amino acids, they reduced the protein sequences to binary feature sequences using the idea of "coarsening" in physics and introduced normal weight function to realize the grouping weight coding of protein sequences. This method is used as one of the feature extraction methods in this paper.

Proteins are mainly composed of 20 amino acids and are generally treated as a string of 20 letters. Considering the physicochemical characteristics of amino acid hydrophobicity and charge, it can be divided into four categories:

Neutral non-polar amino acid: $C1=\{G,A,V,L,I,M,P,F,W\}$,

Neutral polar amino acid: $C2=\{Q,N,S,T,Y,C\}$,

Acidic amino acid: $C3=\{D,E\}$,

Basic amino acid: $C4=\{H,K,R\}$.

There are three ways to divide the above four categories: $\{C1,C2\}$ and $\{C3,C4\}$, $\{C1,C3\}$ and $\{C2,C4\}$, $\{C1,C4\}$ and $\{C2,C3\}$.

Let $A=a_1a_2\cdots a_n$ be a protein sequence, and use the above classification method to convert the protein sequence A into three binary 0-1 sequences by the three mappings $\phi_i(A)$ $(i=1,2,3)$ defined below:

$$\phi_i(A)=\phi_i(a_1)\phi(a_2)\cdots\phi_i(a_n) \tag{4}$$

$$\phi_1(a_j)=\begin{cases}1\ if\ a_j\in\{C1,C2\}\\0\ if\ a_j\in\{C3,C4\}\end{cases} \tag{5}$$

$$\phi_2(a_j)=\begin{cases}1\ if\ a_j\in\{C1,C3\}\\0\ if\ a_j\in\{C2,C4\}\end{cases} \tag{6}$$

$$\phi_3(a_j)=\begin{cases}1 \ if \ a_j \in \{C1, C4\} \\ 0 \ if \ a_j \in \{C2, C3\}\end{cases} \tag{7}$$

$\phi_i(A)$ $(i=1,2,3)$ is the three sequence features of the protein sequence A. Given a sequence feature of length n, the number of occurrences of "1" in the sequence is weight, and the frequency is a regular weight. Taking L, the sequence features are divided into L sub-sequences, and the normal weight of each sub-sequence is calculated to obtain an L -dimensional vector H_i $(i=1,2,3)$, and the transition process from $\phi_i(A)$ $(i=1,2,3)$ to H_i $(i=1,2,3)$ is the encoding based on the grouped weight of the sequence features. Thus, the three sequence features are obtained as three vectors, which are combined to form an $3L$ -dimensional vector, denoted a $W=[w_1, w_2, \cdots, w_{3L}]$, and called W as encoding based on the grouped weight of protein sequence A. The group weight coding method based on the "coarse granulation" idea reflects the distribution of amino acids in certain types of protein sequences and can well characterize the essential characteristics of protein sequences, so it is widely used in proteomics research (Zhang et al., 2006; Nanni and Lumini, 2009; Shi et al., 2012; Zou, 2014).

2.5. Two-dimensional wavelet denoising

In recent years, wavelet analysis has been widely used in proteomics research (Liò, 2003; Qiu et al., 2010; Li et al., 2012; Jia et al., 2015b, 2015c; Yu and Zhang, 2013a, 2013b; Yu et al., 2017a, 2017b, 2017c; Qiu et al., 2018). Wavelet denoising (WD) is one of the important applications of wavelet analysis. The so-called threshold denoising is to decompose the signal, then perform threshold processing on the decomposed coefficients, and finally reconstruct the denoised signal. In the wavelet denoising process, the wavelet function, the decomposition scale, and the choice of the threshold estimation method are the key factors affecting the final noise reduction effect (Chang et al., 2000a, 2000b).

A 2-D model with noise can be expressed as:

$$s(i,j)=f(i,j)+\sigma e(i,j) \ \ i,j=1,2,\cdots,m-1 \tag{8}$$

where (i,j) is the size of the signal, $f(i,j)$ is the real signal, $e(i,j)$ is the noise, $s(i,j)$ is the noise-containing signal and σ is the noise, m is the signal length. The purpose of wavelet denoising is to suppress $e(i,j)$ and recover $f(i,j)$.

The algorithm for 2-D wavelet denoising is as follows:

(1) Decompose the data into signals on the low-frequency scale, the horizontal scale, the vertical scale, and the diagonal scale spatial frequency band under the 2-D wavelet transform.

(2) Threshold processing of the decomposed high-frequency coefficients. The appropriate threshold is selected to perform threshold quantization processing on the high-frequency coefficients in the horizontal, vertical, and diagonal directions of each layer.

(3) Reconstruct the 2-D signal. The 2-D signal is reconstructed according to the low-frequency coefficients of the wavelet decomposition and the high-frequency coefficients processed by the threshold quantization.

Commonly used threshold functions are mainly hard threshold functions and soft threshold functions. In general, the hard threshold method can well preserve local features such as signal edges. The soft threshold processing is relatively smooth, but it will cause distortion such as edge blur. Therefore, this paper chooses the hard threshold function for wavelet denoising.

To visually compare the difference between the 1-D wavelet and the 2-D wavelet to reduce the feature vectors after fusion, this paper selects the PPIs data pairs with the sequence names >RA7SMKP1015 and >RA7VEC8D014 from the *H. pylori* dataset. The original sequence, the 1-D wavelet denoising, and the 2-D wavelet denoising sequence are drawn after the feature fusion method. The default threshold method determined by the wavelet function is used, the wavelet function is db8, and the decomposition scale is 6. The result is shown in Fig.1. It can be seen from Fig. 1 that the use of 2-D wavelet noise reduction can not only effectively remove redundant information in the protein sequence, but also make the characteristic signal of the extracted protein sequence more obvious. 2-D wavelet denoising can denoise the entire dataset from low-frequency scale space, horizontal scale space, vertical scale space, and diagonal scale space, which is more detailed than the 1-D wavelet noise reduction decomposition frequency band. Through repeated comparative analysis, we found that the 2-D wavelet denoising method can effectively improve the prediction accuracy of PPIs. Therefore, we use the 2-D wavelet method to denoise PPIs data.

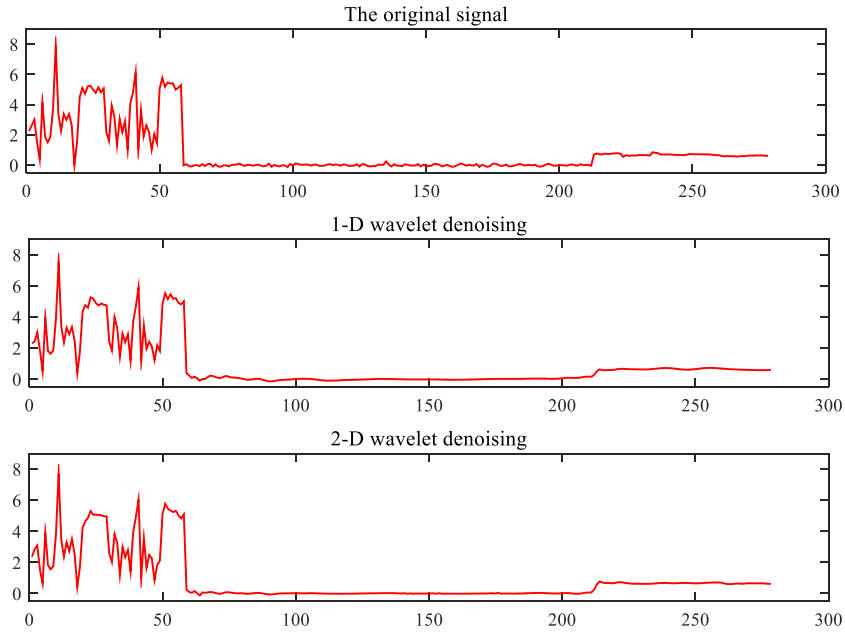

Fig. 1. Wavelet denoising using db8 wavelet function and decomposition scale of 6 on *H. pylori* dataset for >RA7SMKP1015 and >RA7VEC8D014 protein pair. The y-axis represents the intensity of the signal, and the x-axis represents the position of the sequence residues.

2.6. Support vector machine

SVM is a machine learning method proposed by Vapnik (1995) based on statistical learning theory. Its central idea is to establish an optimal decision hyperplane, which maximizes the distance between the two types of samples on the two sides of the plane and provides a good generalization ability for classification problems. At present, the SVM method has been widely used in PPIs (Jia et al., 2015b; Guo et al., 2008; Donaldson et al., 2003; Yang et al., 2010; Chatterjee et al., 2011; Li et al., 2014; Zhai et al., 2017; Daberdaku and Ferrari, 2018), protein subcellular localization (Yu et al., 2017a, 2017b; Yu et al., 2018; Xiang et al., 2017; Tung et al., 2017; Uddin., 2018;), protein submitochondrial localization (Du and Li, 2006; Lin et al., 2013; Du and Yu, 2013; Li et al., 2014; Du and Jiao, 2017; Qiu et al., 2018) and other proteomics fields.

For a two-class sample classification problem, assume a training set for n sample:

$$G = \{(x_i, y_i) | i = 1, 2, \cdots, n\} \quad x_i \in R^d, y_i \in \{+1, -1\} \tag{9}$$

where x_i is the d -dimensional feature vector of the i sample, then R^d is the d -dimensional feature vector space of the x_i , and y_i is the category tag of the sample i . To accomplish the classification purposes of the two types of samples, the SVM first maps the samples of the input space to a high-dimensional feature space through a nonlinear kernel function relationship and constructs the optimal classification hyperplane in this space to make the two types of samples linearly separable.

The kernel function is used to replace the dot product in the optimal classification plane. The optimal classification function is:

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{n} a_i y_i K(x_i, x_j) + b\right\} \tag{10}$$

where a_i represents the Lagrange multiplier, b is the classification threshold, and $K(x_i, \mathrm{x}_j)$ is the kernel function. The selection of the kernel function is very important in the process of training the support vector. It can effectively overcome the "dimension disaster" problem, and the appropriate kernel function can improve the prediction accuracy of the classification model. Commonly used kernel functions mainly include linear kernel function, polynomial kernel function, radial basis kernel function, and sigmoid kernel function. This study used Chang and Lin (2011) to develop the LIBSVM software, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.7. Model building and performance evaluation

The flow chart of the PPIs prediction model based on the multi-information fusion proposed in this paper is shown in Fig.2. Our experimental environment is Intel (R) Core (TM) i5-4258U CPU @2.40GHz 4.00GB of RAM and uses MATLAB2014a programming.

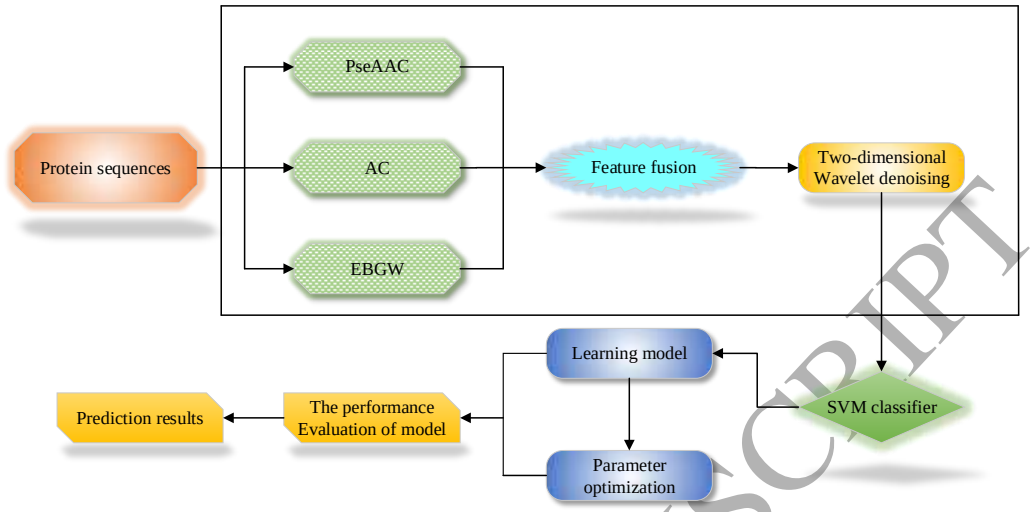

Fig. 2. Flowchart of PPIs prediction method.

The steps of the PPIs prediction method based on the PPIs-WDSVM method are described as:

1) Input protein sequences of the *H. pylori* and *S. cerevisiae* datasets, respectively, and the class labels for the positive and negative samples.

2) The PseAAC, AC, and EBGW methods are used to extract the feature of the protein sequence. The three sets of feature vectors are respectively brought into the SVM classifier for prediction to determine the optimal parameters of the three methods. The parameters include the λ value in the PseAAC method, the lag value in the AC method, and the L value in the EBGW method.

3) Combine the three sets of numerical feature vectors extracted by the best parameters in 2).

4) The 2-D wavelet denoising method is used to denoise the feature vectors in 3), and the SVM classifier is used to predict the denoised data to determine the optimal wavelet function and decomposition scale.

5) The optimal wavelet function and decomposition scale obtained in 4) was used to conduct 2-D wavelet denoising on the fused feature vectors. The denoised feature vectors were input into the SVM classifier and combined with 5-fold cross-validation test to determine the optimal kernel function according to the predicted results.

6) Based on the optimal parameters of the model obtained in 2), 4), and 5), we calculated the average Se, Pre, MCC, ACC, and AUC values of the *H.pylor*i and *S. cerevisiae* datasets and evaluated the predicted performance of the model.

In the statistical learning theory, the jackknife test, the re-substitution test, the independent

test, and the k-fold cross-validation test are usually used to verify the validity of the model. This paper uses 5-fold cross-validation to test the performance of the predictive model.

In order to evaluate the predictive model more intuitively, this paper combines the evaluation index formula used in the latest research results, using the symbols introduced by Chou (2001b) in the study of signal peptides, and deduces the conventional formula into a set of intuitive formulas to evaluate protein interaction prediction models (Chen et al., 2013; Yan et al., 2013), as given in Eq. (14) of (Chen et al., 2013) or in Eq. (19) of (Yan et al., 2013). Ever since then, the set of Chou's intuitive metrics have been concurred and applauded by a series of recent publications (Chou, 2001b; Chen et al., 2013; Feng et al., 2013). The formula is as follows:

$$Se = 1 - \frac{N_-^+}{N^+} \tag{11}$$

$$Sp = 1 - \frac{N_+^-}{N^-} \tag{12}$$

$$ACC = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{13}$$

$$MCC = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \tag{14}$$

from that we get:

$$Pre = \frac{N^+ - N_-^+}{N^+ - N_-^+ + N_+^-} \tag{15}$$

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 no positive samples are wrongly predicted as negative samples, then $Se = 1$; When $N_-^+ = N^+$, it means that all positive samples are predicted as negative samples, then $Se = 0$ and $Pre = 1$; When $N_+^- = 0$, it means that no negative samples are wrongly predicted as positive samples , then $Sp = 1$ and $Pre = 0$; When $N_+^- = N^-$, it means that all the negative samples are predicted as positive samples, then $Sp = 0$. When the metric formula (11-15) is used to detect the sensitivity (Se), specificity (Sp), overall accuracy (ACC) and Matthews correlation coefficient (MCC) of the predictor, the prediction results are more intuitive and easier to understand. However, these formulas are only effective for single-label systems, and a completely different set of metrics is required for multi-label systems that frequently occur in fields such as system biology (Xiao et al., 2011; Wang and Li, 2012; Chou

et al., 2012; Chou, 2013; Lin et al., 2013; Zou, 2014).

Both the ROC (receiver operating characteristic) and PRC (precision-recall curve) curves are used to evaluate the classification performance of machine learning algorithms for a given dataset, each containing a fixed number of positive and negative samples. The ROC curve and the AUC (area under the curve) value are often used to evaluate the merits of a binary classifier. The ROC curve is a comprehensive indicator of continuous variables of sensitivity and specificity, and each point on the curve reflects the susceptibility to the same signal stimulus. The area under the ROC curve (AUC) is often used to measure the predictive performance of the classifier. The closer its value is to 1, the better the classification effect. The PRC curve shows the relationship between precision and recall, and the closer the area under the curve AUPR (area under precision-recall) is to 1, the better the classification effect.

3. Results and discussion

3.1. Parameter selection of feature extraction method

3.1.1. Selection optimal parameters λ of PseAAC method

In this paper, we use the PseAAC method to extract the features of the protein sequences in the *H. pylori* and *S. cerevisiae* datasets. In the process of feature extraction, the selection of the parameter λ value in the PseAAC method affects the prediction result, and also affects the dimension of the feature vector after extraction. In this method, the value of λ must be less than the minimum length of the protein sequence. The shortest length of the protein sequence in the *H. pylori* and *S. cerevisiae* datasets is 12, so the maximum value of this paper is 11 and the range of λ is $0 \le \lambda \le 11$. To find the optimal parameter in the model, we set the λ value from 0 to 11 in turn. In the case of setting different parameter values, we select the SVM classifier based on the RBF kernel function for prediction and use the 5-fold cross-validation method to test the results. The evaluation indicators of the two datasets under different parameters are shown in Table 1. The specific prediction results are shown in the Supplementary materials Table S1.

Table 1

The prediction results of two datasets by selecting different λ in PseAAC method.

Datasets	5-fold cross-validation						
		1	3	5	7	9	11
H. pylori	ACC (%)	76.95	77.13	77.30	76.89	77.78	76.99
	MCC	0.5393	0.5433	0.5461	0.5381	0.5559	0.5399
S. cerevisiae	ACC (%)	82.02	81.94	82.62	83.66	83.55	84.28
	MCC	0.6404	0.6388	0.6525	0.6732	0.6711	0.6856

Table 1 lists the model evaluation indicators obtained by the *H. pylori* and *S. cerevisiae* datasets at different values of λ. As the value of λ is different, the prediction results of the two

datasets obtained by using the SVM classifier are also different. It can be seen from Table 1 that when $\lambda=9$, the ACC of the *H. pylori* dataset reaches the highest of 77.78%, which is 0.48%-0.89% higher than the ACC obtained by other parameters. The MCC reaches the highest of 0.5559, which was 0.98%-1.78% higher than the MCC obtained by other parameters. When $\lambda=11$, the ACC of the *S. cerevisiae* dataset reaches the highest of 84.28%, which is 0.62%-2.34% higher than the ACC obtained by other parameters. The MCC reaches the highest of 0.6856, which was 1.24%-4.68% higher than the MCC obtained by other parameters. To compare the prediction efficiency of the PseAAC method on two datasets, it is necessary to select the same parameters for feature extraction of data. In this paper, we comprehensively consider the average ACC and average MCC of two datasets under different parameters. When $\lambda=9$, the average ACC of the two datasets reaches the highest of 80.67%, the average MCC reached the highest of 0.6135. Therefore, the optimal parameter λ value of PseAAC feature extraction method is 9.

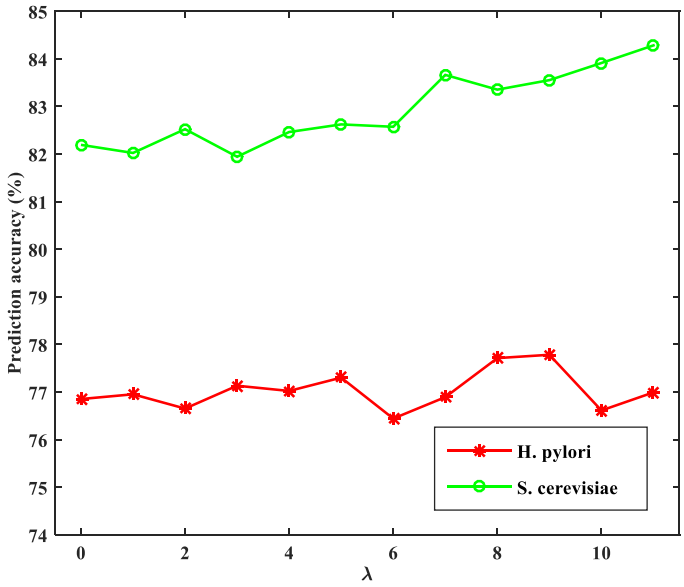

Fig. 3. Effect of selecting different values of λ on the ACC results of *H. pylori* and *S. cerevisiae* datasets in the PseAAC method.

Fig. 3 is a line graph showing the change rate of ACC obtained by classifying *H. pylori* and *S. cerevisiae* datasets with different λ, which can display the optimal λ more intuitively. It can be seen from Fig. 3, as the value of λ changes, the ACC of the two datasets also changes. When $\lambda=9$, the ACC of the *H. pylori* dataset reaches the highest. When $\lambda=11$, the ACC of the *S. cerevisiae* dataset reaches the highest. In this paper, we use the average ACC to determine the uniform optimal parameters. Therefore, the optimal parameter λ in the PseAAC method should be chosen to be 9, and the extracted feature vector dimension is $2\times(20+\lambda)=58$. It means that the feature vector dimensions of the *H. pylori* and *S. cerevisiae* datasets obtained by the PseAAC method are 2916×58 and 11188×58, respectively.

3.1.2. Selection optimal parameters lag *of the AC method*

The AC method is one of the methods for feature extraction of protein sequences in the *H. pylori* and *S. cerevisiae* datasets. In the process of feature extraction, the selection of the parameter lag value in the AC method affects the prediction result, and also affects the dimension of the extracted feature vector. For example, when $lag = 3$, the extracted feature vector dimension is 42; when $lag = 10$, the extracted feature vector dimension is 140. For the AC method, the value of lag must be less than the shortest length of the protein sequence and greater than 0. The shortest length of the protein sequence in the *H. pylori* and *S. cerevisiae* datasets is 12, so the maximum lag is 11, and the range is $0 < lag < 12$. To find the optimal parameter in the model, we set the lag value from 1 to 11. In the case of setting different values of lag, the SVM classifier selects the RBF kernel function and uses the 5-fold cross-validation method to test the result. The evaluation parameters of the datasets under different parameters are shown in Table 2. The specific prediction results are shown in Supplementary materials Table S2.

Table 2

The prediction results of two datasets by selecting different lag in AC method.

Datasets	5-fold cross-validation						
		1	3	5	7	9	11
H. pylori	ACC (%)	66.60	73.42	76.37	76.58	77.19	77.95
	MCC	0.3332	0.4694	0.5277	0.5319	0.5441	0.5593
S. cerevisiae	ACC (%)	63.03	73.40	77.34	80.09	81.55	83.05
	MCC	0.2617	0.4680	0.5469	0.6023	0.6319	0.6621

It can be seen from Table 2 that with the different values of lag, the results of the prediction of the two data sets are different. When $lag = 11$, the ACC of *H. pylori* dataset reaches the highest value of 77.95%, which is 0.76% to 11.35% higher than that of other parameters, and the MCC value also reaches the highest value of 0.5593, which is 1.52% to 22.61% higher than that of other parameters. Similarly, when $lag = 11$, the ACC of the *S. cerevisiae* dataset reaches the highest value of 83.05%, which is 1.50%-20.02% higher than the ACC obtained by other parameters, and the MCC also reaches the highest value of 0.6621, which is 3.02%-40.04% higher than the MCC obtained by other parameters.

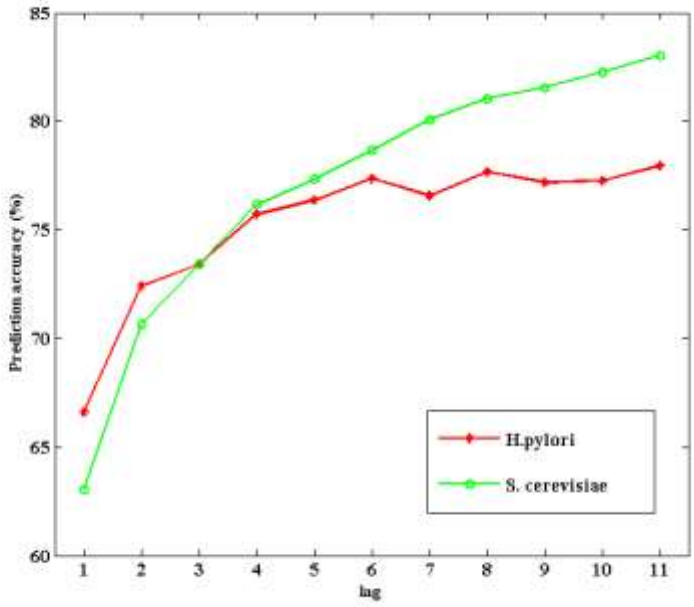

Fig.4 Effect of selecting different values of lag on the ACC results of *H. pylori* and *S. cerevisiae* datasets in the AC method.

Fig. 4 is a line graph showing the change rate of ACC obtained by classifying *H. pylori* and *S. cerevisiae* datasets with different lag, which can display the optimal lag more intuitively. It can be seen from Fig. 4 that with the change of lag, the ACC of the two datasets is also changing. When $lag=11$, the ACC of the two datasets is the highest. In summary, the optimal parameter lag in the AC feature extraction method is 11, then the extracted feature vector dimension is $2\times(7\times lag)=154$, which means that the feature vector dimensions of *H. pylori* and *S. cerevisiae* datasets are 2916×154 and 11188×154 after the AC method feature extraction.

3.1.3. Selection optimal parameters L of EBGW method

The EBGW method is one of the methods for feature extraction of protein sequences in *H. pylori* and *S. cerevisiae* datasets. In this method, the L must be less than the shortest length of the protein sequence and greater than 0, and the range is $0<L<12$. In the case of setting different parameter values, we use the SVM based on the RBF kernel function as the classifier and use the 5-fold cross-validation method to test the results. The evaluation indicators of the two datasets under different parameters are shown in Table 3. The specific prediction results are shown in Supplementary materials Table S3.

Table 3

The prediction results of two datasets by selecting different L in EBGW method.

Datasets	5-fold cross-validation						
		1	3	5	7	9	11
H. pylori	ACC (%)	67.15	70.51	70.30	71.06	71.98	72.15

	MCC	0.3455	0.4111	0.4076	0.4222	0.4404	0.4439
S. cerevisiae	ACC (%)	67.56	72.94	73.83	73.66	75.32	76.11
	MCC	0.3526	0.4606	0.4778	0.4745	0.5080	0.5230

It can be seen from Table 3 that with the different values of L, the results of predicting the two datasets using the EBGW method are also different. When $L=11$, the ACC of the *H. pylori* dataset reaches the highest value of 72.15%, which is 0.17%-5.00% higher than the ACC obtained by other parameters, the MCC also reached a maximum of 0.4439, which is 0.35%-9.84% higher than the MCC obtained by other parameters. Similarly, when $L=11$, the ACC of the *S. cerevisiae* dataset reaches the highest value of 76.11%, which is 0.79%-8.55% higher than the ACC obtained by other parameters, and the MCC also reaches the highest value of 0.5230, which is 1.50%-17.04% higher than the MCC obtained by other parameters.

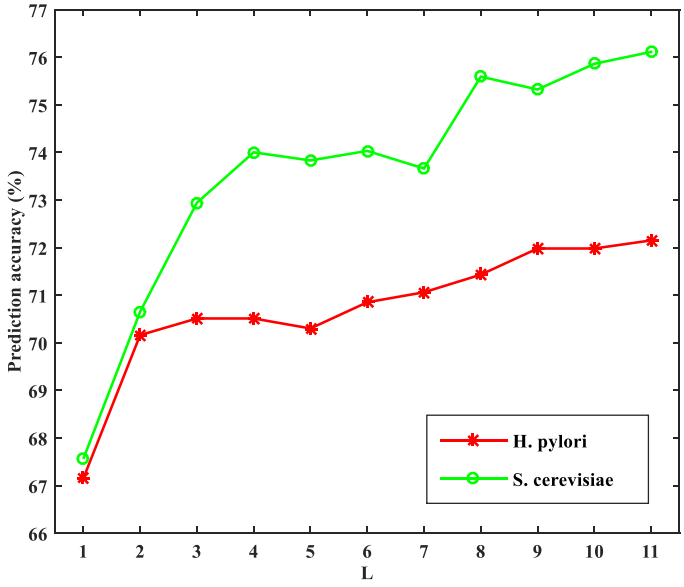

Fig. 5. Effect of selecting different values of L on the ACC results of *H. pylori* and *S. cerevisiae* datasets in EBGW method.

Fig. 5 is a line graph showing the change rate of ACC obtained by classifying *H. pylori* and *S. cerevisiae* datasets with different L values, which can display the optimal L value more intuitively. It can be seen from Fig. 5 that with the change of the L, the ACC of the two datasets obtained by the EBGW method also changes. When $L=11$, the ACC of the two datasets reaches the highest. In summary, the optimal parameter L in the EBGW feature extraction method is selected as 11, and the extracted feature vector dimension is $(3\times L)\times 2=66$. It means that the feature vector dimensions of the *H. pylori* and *S. cerevisiae* datasets obtained by feature extraction using the EBGW method are 2916×66 and 11188×66, respectively.

3.2. Selection of wavelet functions and optimal decomposition scale

In this paper, we use the hard threshold method under 2-D wavelet denoising to denoise the

fused feature vector and eliminate redundant information. The results of the 2-D wavelet denoising method are related to the wavelet function and the decomposition scale. Once improperly selected, the local feature information in the protein sequence will be weakened, resulting in the loss of detailed information about the data. Generally, the wavelet function is selected to be considered comprehensively from the support length, vanishing moment, symmetry, regularity, and similarity. Wavelet functions have their characteristics when processing signals and no wavelet function can achieve optimal denoising effect for all types of signals.

In this paper, we consider the influence of different wavelet functions and different decomposition scales on the prediction results under the default threshold method determined by the wavelet function. We use the 2D wavelet denoising method to decompose the fused feature vector. The wavelet function in this method selects Daubechies (dbN) wavelet, Symlet (symN) wavelet, Coiflet (coifN) wavelet and Biorthogonal (biorNr.Nd), the decomposition scale is selected from 1 to 6, and the hard threshold processing method is adopted. In the feature extraction stage, when we use PseAAC, AC and EBGW methods to extract features from two datasets, the optimal parameters are 9, 11 and 11, respectively, and the obtained feature vector dimensions are 58, 154 and 66 respectively. We use the first method described above to fuse the feature vectors obtained by the three methods. This method connects the feature vectors obtained by the PseAAC, AC and EBGW methods into a vector whose dimension is equal to the sum of the vector dimensions of three methods. Finally, we get the dimension of the fused vector is 278, which means that the feature vector size after extraction of the *H. pylori* dataset is 2916×278, and the feature vector size after extraction of the *S. cerevisiae* dataset is 11188×278. We perform 2-D wavelet denoising on the fused feature vector. In the case of selecting different wavelet functions and decomposition scales, the SVM classifier based on RBF kernel function is used to predict the datasets after denoising, and the prediction results are tested by 5-fold cross-validation. Under the default threshold method determined by the wavelet function, the partial prediction results of the two datasets obtained by selecting different wavelet functions and different decomposition scales are shown in Table 4 and Table 5. The specific prediction results are shown in Supplementary materials Tables S4 and S5.

Table 4

Predicted results PPIs on *H. pylori* dataset under different wavelet functions and decomposition scales.

Wavelet functions	5-fold cross-validation (%)					
	Scales					
	1	2	3	4	5	6
db1	87.83	90.95	92.97	93.90	91.98	91.84
db5	88.89	90.88	93.96	93.18	93.72	93.62
db8	88.24	91.70	94.48	94.31	95.61	**96.33**

sym2	87.45	92.18	92.49	93.93	93.31	93.90
sym4	87.72	90.84	93.52	93.86	93.18	93.79
sym7	89.13	92.52	93.76	94.92	95.30	94.58
coif1	87.76	90.74	93.72	93.76	94.17	94.34
coif2	87.79	91.26	91.77	92.90	93.83	93.18
coif4	88.24	91.39	93.21	93.59	94.62	94.79
bior3.1	86.53	89.52	91.39	92.18	91.87	91.70
bior3.3	88.34	92.22	92.80	92.63	92.87	92.66
bior5.5	88.10	91.32	92.90	93.18	93.35	93.48

Table 5

Predicted results PPIs on *S. cerevisiae* dataset under different wavelet functions and decomposition scales.

Wavelet functions	5-fold cross-validation (%)					
	Scales					
	1	2	3	4	5	6
db1	91.93	93.00	94.20	94.49	94.50	94.06
db5	91.37	93.22	94.22	94.99	94.90	94.61
db8	91.91	93.22	94.86	95.08	95.06	95.60
sym2	90.49	93.22	93.62	94.36	94.98	95.02
sym4	91.46	93.66	94.23	94.90	94.89	95.24
sym7	91.32	93.63	94.26	94.71	95.30	95.24
coif1	90.85	93.07	93.85	94.32	95.16	95.37
coif2	91.11	93.39	93.91	95.18	95.48	95.33
coif4	91.50	93.60	94.57	95.34	95.65	95.71
bior3.1	90.09	90.60	90.30	90.39	90.34	90.43
bior3.3	91.19	92.27	91.61	91.57	91.97	91.82
bior5.5	90.85	92.79	94.70	95.39	95.58	**96.03**

Tables 4 and 5 show the ACC of the *H.pylori* and *S. cerevisiae* datasets under partial wavelet functions and the decomposition scale of 1 to 6, respectively. It can be seen that different wavelet functions and different decomposition scales result in different prediction results. Table 4 shows the results of PPIs prediction of the *H. pylori* dataset. It can be seen that the ACC is between 86.53% and 96.33%. When the wavelet function selects db8, and the decomposition scale is 6, the highest prediction rate of protein-protein interaction is 96.33%, which is 9.80% higher than that of the bior3.1 wavelet and the decomposition scale is 1. Table 5 shows the results of PPIs prediction of the *S. cerevisiae* dataset. It can be seen that the ACC is between 89.94% and 96.03%. When the wavelet function selects bior5.5, and the decomposition scale is 6, the highest ACC of PPIs is 96.03%, which is 0.43% higher than the ACC when the db8 wavelet is selected, and the decomposition scale is 6. The prediction results obtained under two different parameters are not much different. Considering comprehensively, to facilitate the superiority of the comparison algorithm on the two datasets, we should choose a unified wavelet function and decomposition scale. Therefore, the optimal wavelet function we selected in this paper is db8, and the

decomposition scale is 6. To more intuitively analyze the optimal wavelet function and decomposition scale selection in the two datasets, the predictions of the two datasets under different wavelet functions and 4 to 6 decomposition scales are drawn under the threshold method obtained by the default strategy. The histogram of the ACC is shown in Figs. 6 and 7.

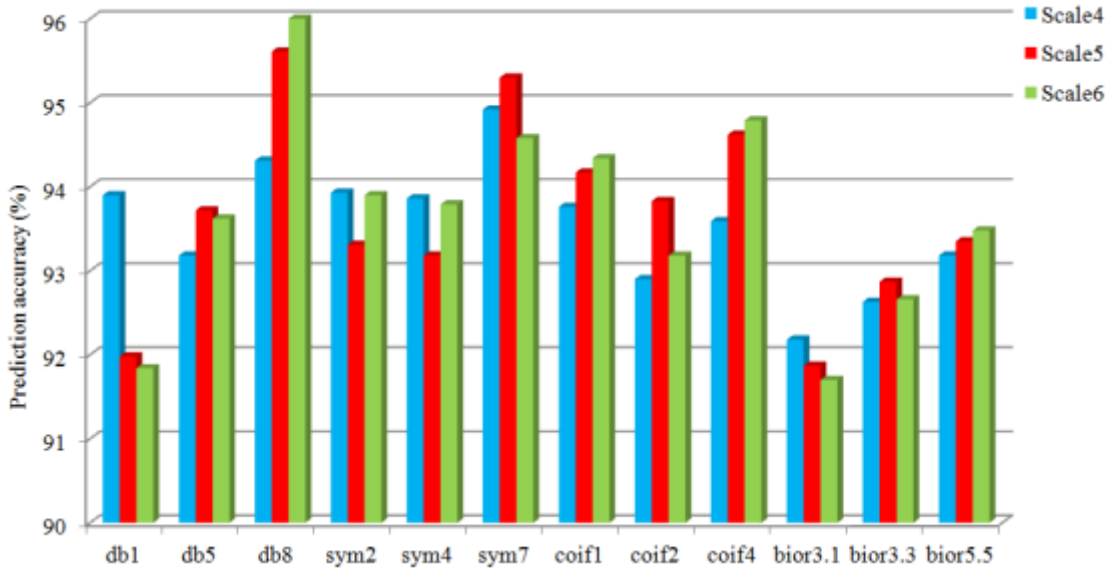

Fig. 6. Prediction performance of *H. pylori* dataset under different wavelet functions and different scales.

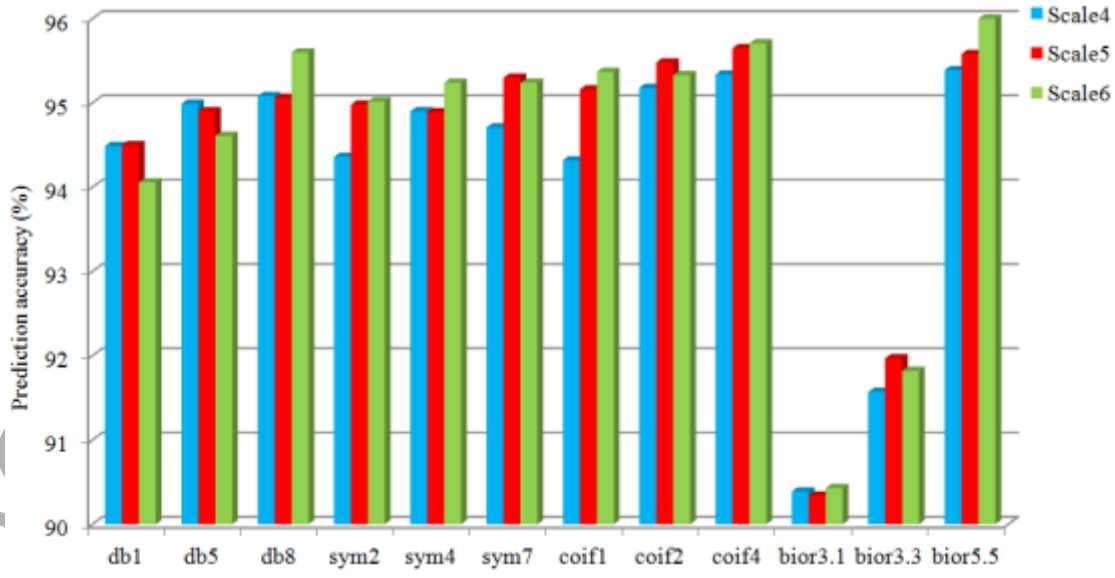

Fig. 7. Prediction performance of *S. cerevisiae* dataset under different wavelet functions and different scales.

It can be seen from Fig. 6 that for the *H. pylori* dataset when the wavelet function is db8, and the decomposition scale is 6, the ACC is the highest. In Fig. 7, for the *S. cerevisiae* dataset, when the wavelet function is bior5.5, and the decomposition scale is 6, the ACC is the highest. Because

the dbN wavelet has good regularity and good band division effect, the signal weight process is relatively smooth, and the ACC of the *S. cerevisiae* dataset is relatively high when the db8 wavelet function is selected. Considering comprehensively, the wavelet function db8 and the number of decomposition layers are finally selected to be 6 for the two datasets. At this time, the ACC of the PPIs of the *H. pylori* and *S. cerevisiae* datasets are 96.33% and 95.60%, respectively.

3.3. Effect of kernel function on results

For the feature vectors after 2-D wavelet denoising, we use SVM method for classification prediction, and the choice of kernel function in the SVM classifier also affects the prediction results. In the feature extraction method, the optimal parameter λ selected by the PseAAC method is 9, the optimal parameter lag selected by the AC method is 11, and the optimal parameter L selected by the EBGW method is 11, and the total dimension of the feature vectors after the fusion of the three methods is 278. We use the eigenvectors after the fusion to perform two-dimensional wavelet denoising, using a hard threshold function, a default threshold method, a wavelet function db8, and a decomposition scale of 6. We select the linear kernel function, the cubic polynomial kernel function, the RBF function and the sigmoid kernel function in the SVM classifier to predict after 2-D wavelet denoising data and use the 5-fold cross-validation to test. The PPIs prediction results of the two datasets are shown in Table 6, and the specific results are shown in Supplementary materials Table S6.

Table 6

Prediction results of PPIs different kernel functions in *H. pylori* and *S. cerevisiae*.

Datasets	5-fold cross-validation				
	Kernel functions	Linear	Polynomial	RBF	Sigmoid
H. pylori	ACC (%)	84.64	84.26	**96.33**	78.33
	MCC	0.6928	0.7163	**0.9269**	0.5666
S. cerevisiae	ACC (%)	84.48	93.40	**95.60**	71.13
	MCC	0.6897	0.8704	**0.9121**	0.4226

Table 6 shows the results of PPIs predictions for the two datasets by different kernel functions in the SVM classifier. It can be seen from Table 6 that for the *H. pylor*i dataset, the AAC obtained when the select RBF function is the highest, which is 96.33%. It is 11.69%, 12.07%, and 18.00% higher than the linear kernel function, the cubic polynomial, and the sigmoid kernel function, respectively. The MCC is also the highest value of 0.9269, which is 23.41%, 21.06%, and 36.03% higher than the linear kernel function, the cubic polynomial kernel function and the Sigmoid kernel function, respectively. For the *S. cerevisiae* dataset, the highest ACC is 95.60% when selecting the RBF kernel function, which is 11.12%, 2.20%, and 24.47% higher than the linear kernel function, the cubic polynomial kernel function and the sigmoid kernel function. The MCC is also the highest value of 0.9121, which is 22.24%, 4.17%, and 48.95% higher than the

linear kernel function, the cubic polynomial kernel function, and the sigmoid kernel function, respectively. The prediction results of the two datasets show the superiority of the RBF kernel function. The kernel function of RBF can map a sample to a higher-dimensional space. Compared with the polynomial kernel function, the kernel function of RBF needs to determine relatively few parameters, and the number of kernel parameters directly affects the complexity of the function, thus reducing the difficulty of numerical calculation.

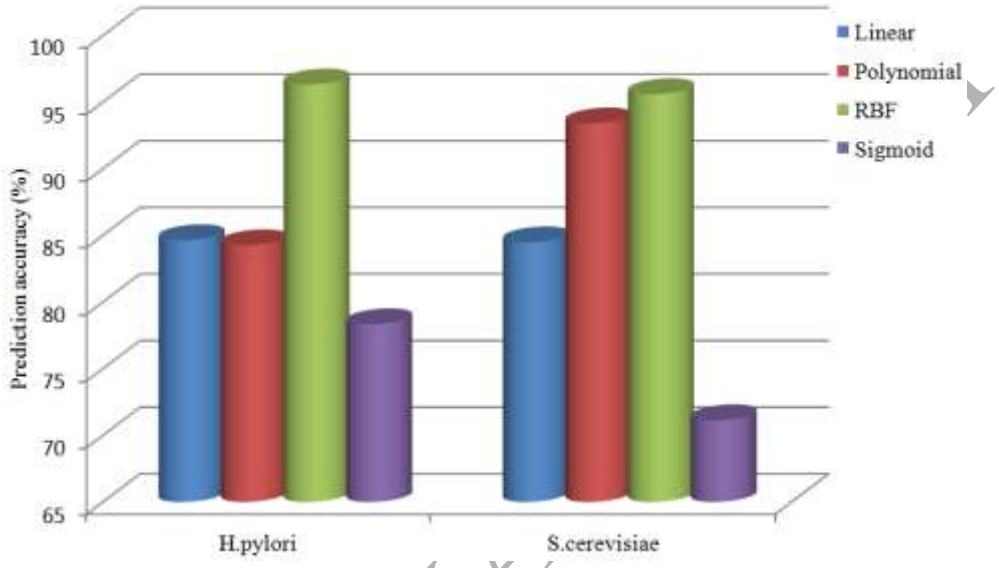

Fig. 8. The ACC of different kernel functions in *H. pylori* and *S. cerevisiae* datasets.

It can be seen from Fig. 8 that the *H. pylori* and *S. cerevisiae* datasets obtained with the RBF kernel function have the highest prediction accuracy rates of 96.33% and 95.60%, respectively. The prediction accuracy obtained by the sigmoid kernel function is relatively low. By comparison, RBF is chosen as the kernel function of the SVM classifier.

3.4. Effect of feature extraction methods on results

In this paper, we use PseAAC, AC, and EBGW methods to extract the features of *H. pylori* and *S. cerevisiae* datasets, and then perform 2-D wavelet denoising on the three sets of feature vectors. The wavelet function is selected as db8, and the decomposition scale is 6. The SVM classifier based on RBF kernel function is used to predict the data after 2-Dl wavelet denoising, and the prediction result is tested by the 5-fold cross-validation method. The predicted PPIs prediction results are shown in Tables 7 and 8.

Table 7

Comparison of different feature extraction methods for the prediction results of *H. pylori* dataset.

Methods	5-fold cross-validation			
	ACC (%)	Se (%)	Pre (%)	MCC
PseAAC	92.76	92.18	93.27	0.8553
AC	89.95	88.75	90.93	0.7993
EBGW	74.55	75.45	74.12	0.4912

PseAAC+AC+EBGW	**96.33**	**96.16**	**96.52**	**0.9269**

Table 8

Comparison of different feature extraction methods for the prediction results of *S. cerevisiae* dataset.

Methods	5-fold cross-validation			
	ACC (%)	Se (%)	Pre (%)	MCC
PseAAC	92.57	92.89	92.31	0.8515
AC	88.20	86.88	89.24	0.7643
EBGW	77.33	79.75	76.07	0.5473
PseAAC+AC+EBGW	**95.60**	**95.66**	**95.56**	**0.9121**

Table 7 shows the PPIs prediction results of the *H. pylori* dataset under four feature extraction methods. It can be seen from Table 7 that the ACC, Se, Pre, and MCC of the feature vectors obtained by the fusion method are all highest at 96.33%, 96.16%, 96.52% and 0.9269, respectively. Its ACC is 3.57%, 6.38%, and 21.78% higher than that obtained by the PseAAC method, the AC method, and the EBGW method, respectively. The MCC values of the fusion method were 7.16%, 12.76%, and 43.57% higher than those obtained by the PseAAC, AC, and EBGW methods, respectively. It can be seen from the prediction results obtained from the three single methods that the ACC obtained by the PseAAC method is 2.81% and 18.21% higher than the AC and EBGW methods, respectively. This indicates that the method can fully consider the protein sequence information compared with the other two methods and provides an important contribution to the high ACC obtained by the fusion method.

Table 8 shows the prediction results of the *S. cerevisiae* dataset under four feature extraction methods. It can be seen from Table 8 that the highest values of ACC, Se, Pre and MCC obtained by the feature vectors obtained by the fusion method are 95.60%, 95.66%, 95.56% and 0.9121, respectively. Its ACC is 3.03%, 7.40%, and 18.27% higher than that obtained by PseAAC, AC, and EBGW methods, respectively. The MCC of the fusion method was 6.06%, 14.78%, and 36.48% higher than the PseAAC, AC, and EBGW methods, respectively. It can be seen from the prediction results obtained from the three single methods that the ACC obtained by the PseAAC method are 4.37% and 15.24% higher than the AC and EBGW methods, respectively. This indicates that the method can fully consider the protein sequence information compared with the other two methods and provides an essential contribution to the high ACC obtained by the fusion method.

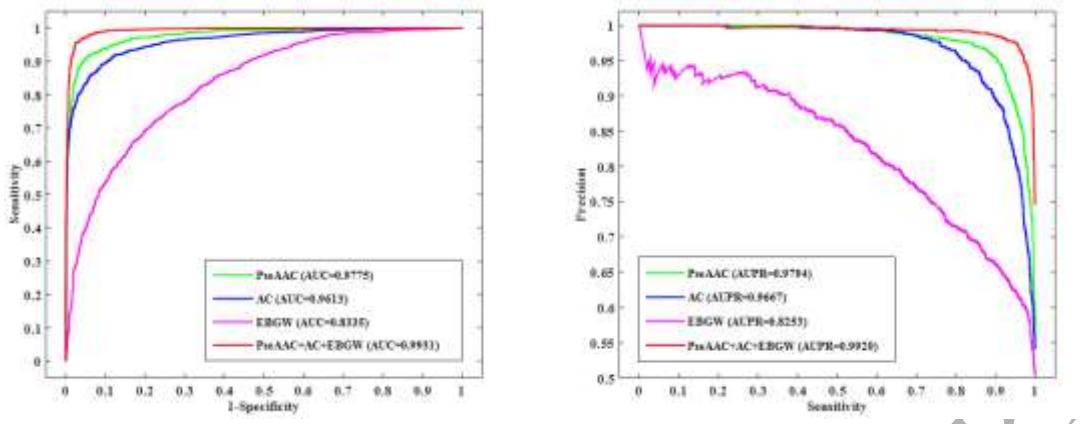

Fig. 9. ROC and PRC curves of *H. pylori* dataset under four feature extraction methods.

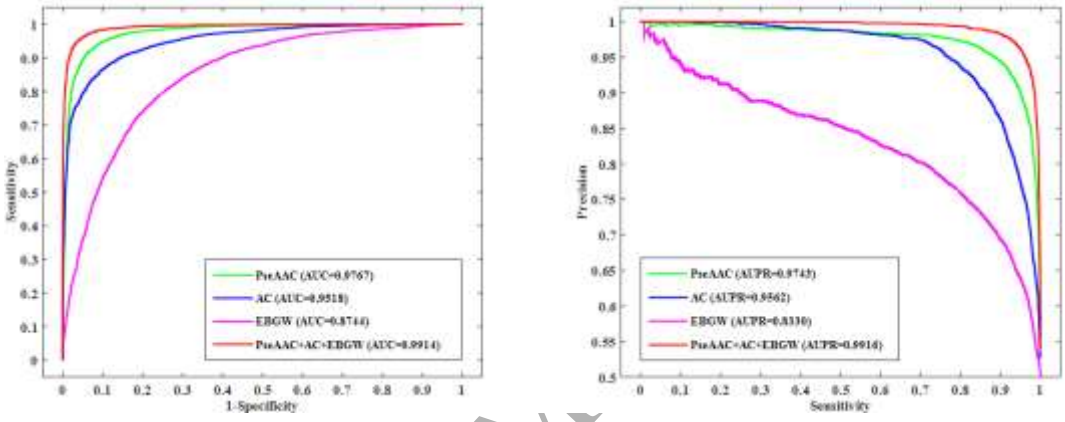

Fig. 10. ROC and PRC curves of *S. cerevisiae* dataset under four feature extraction methods.

Fig. 9 shows the ROC and PRC curves of the *H. pylori* dataset under four feature extraction methods. It can be seen from Fig. 9 that the area under the ROC curve obtained by the fusion of PseAAC, AC and EBGW has the largest AUC value of 0.9931, which are 1.56%, 3.18%, and 15.96% higher than that of the PseAAC, AC and EBGW methods, respectively. The area under the PRC curve obtained by the fusion method also has a maximum AUPR value of 0.9920, which are 1.26%, 2.53%, and 16.67% higher than the PseAAC, AC, and EBGW methods, respectively.

Fig. 10 shows the ROC and PRC curves of the *S. cerevisiae* dataset under four feature extraction methods. It can be seen from Fig. 10 that the area under the ROC curve obtained by the fusion of PseAAC, AC and EBGW has the largest AUC value of 0.9914, which are 1.47%, 3.96%, and 11.70% higher than that of the PseAAC, AC and EBGW methods, respectively. The area under the PRC curve obtained by the fusion method also has a maximum AUPR value of 0.9916, which are 1.73%, 3.54%, and 15.86% higher than the PseAAC, AC, and EBGW methods, respectively.

In summary, we use the fusion method to obtain the highest ACC, which indicates that it can fully consider the protein sequence information and improve the performance of the PPIs prediction model.

3.5. Selecting the classification algorithm

At present, machine learning methods have been successfully applied to the study of PPIs prediction. In this paper, we investigate the effects of five classification algorithms on PPIs predictions. The five classification algorithms are support vector machine (SVM), naïve Bayes (NB), logistic regression (LR), decision Tree (DT), and Adaboost algorithm. We choose the SVM classifier based on RBF kernel function and other classifiers to choose default parameters. The five algorithms used a 5-fold cross-validation to predict PPIs for the *H. pylori* and *S. cerevisiae* datasets, resulting in predictions as shown in Tables 9 and 10.

Table 9

Prediction results of the *H. pylori* dataset under five classification algorithms.

Classifiers	5-fold cross-validation			
	ACC (%)	Se (%)	Pre (%)	MCC
NB	80.32	94.58	73.61	0.6327
LR	85.73	87.11	84.83	0.7153
DT	91.74	93.35	90.46	0.8354
Adaboost	93.59	93.90	93.34	0.8719
SVM	**96.33**	**96.16**	**96.52**	**0.9269**

Table 10

Prediction results of the *S. cerevisiae* dataset under five classification algorithms.

Classifiers	5-fold cross-validation			
	ACC (%)	Se (%)	Pre (%)	MCC
NB	76.7	61.94	87.89	0.5589
LR	84.79	84.93	84.69	0.6958
DT	90.03	90.15	89.95	0.8007
Adaboost	89.32	89.11	89.49	0.7865
SVM	**95.60**	**95.66**	**95.56**	**0.9121**

It can be seen from Table 9 that the ACC of the *H. pylori* dataset obtained by using SVM as a classifier is 96.33%, which is 2.74%-16.01% higher than the other four classifiers. Among them, it is 16.01% higher than the NB classifier, 10.60% higher than the LR classifier, 4.59% higher than the DT classifier, and 2.74% higher than the AdaBoost classifier. The MCC obtained by the SVM classifier is 0.9269, which is 5.50%-29.42% higher than the other four classifiers. It can be seen from Table 10 that the ACC of the *S. cerevisiae* dataset obtained by using SVM as a classifier is 95.60%, which is 5.57%-18.90% higher than the other four classifiers. Among them, it is 18.90% higher than NB classifier, 10.81% higher than LR classifier, 5.57% higher than DT classifier, and 6.28% higher than AdaBoost classifier. The MCC obtained by the SVM classifier is 0.9121, which is 11.14%-35.32% higher than the other four classifiers.

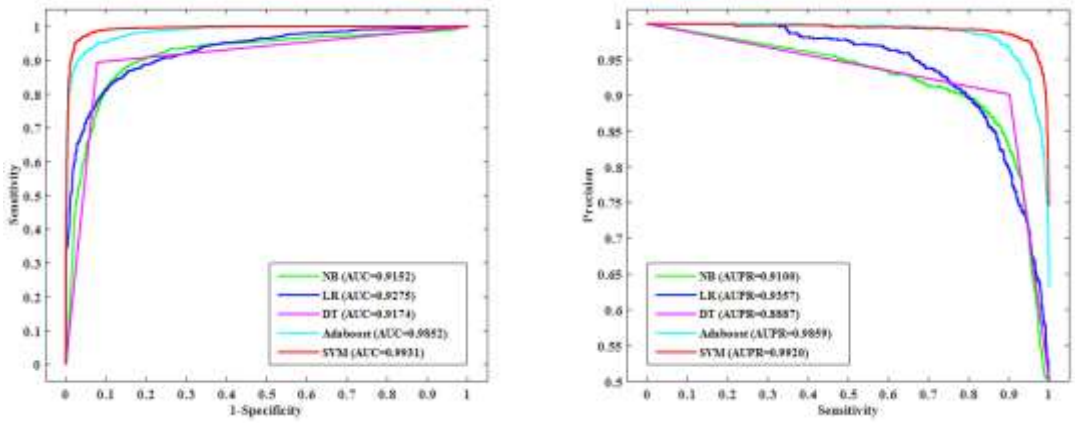

Fig. 11. ROC and PRC curves of *H. pylori* dataset under five classifiers.

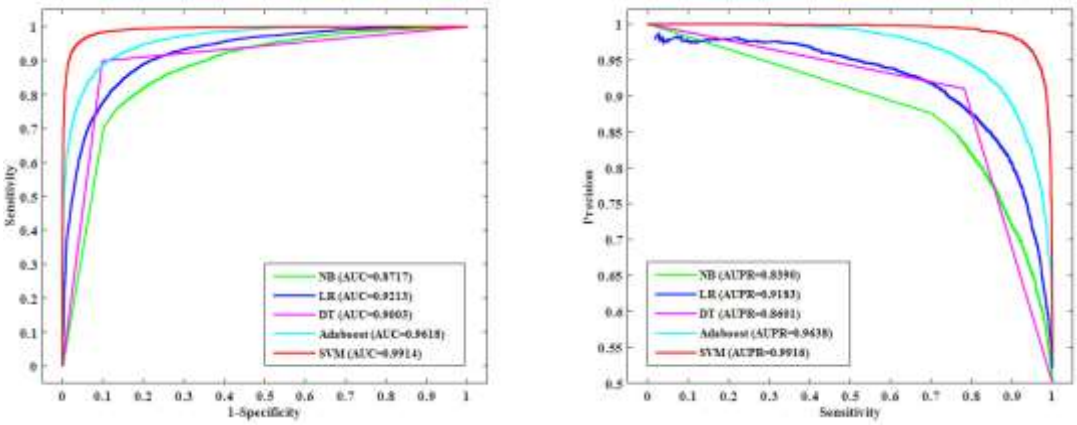

Fig. 12. ROC and PRC curves of *S. cerevisiae* dataset under five classifiers.

Fig. 11 shows the ROC and PRC curves of the *H. pylori* dataset under five classifier algorithms. It can be seen that the area under the ROC curve obtained by the SVM classifier has the largest AUC value of 0.9931, which are 7.79%, 6.56%, 7.57%, and 0.79% higher than the NB, LR, DT and AdaBoost classifiers, respectively. The area under the PRC curve obtained by the SVM classifier is the largest AUPR value of 0.9920, which are 8.20%, 5.63%, 10.33%, and 0.61% higher than the NB, LR, DT, and AdaBoost classifiers, respectively.

Fig. 12 shows the ROC and PRC curves for the *S. cerevisiae* dataset under five classifier algorithms. It can be seen that the area under the ROC curve obtained by the SVM classifier has the largest AUC value of 0.9914, which are 11.97%, 7.01%, 9.11%, and 2.96% higher than the NB, LR, DT and AdaBoost classifiers, respectively. The area under the PRC curve obtained by the SVM classifier is the largest AUPR value of 0.9916, which are 15.26%, 7.33%, 13.15%, and 2.78% higher than the NB, LR, DT and AdaBoost classifiers, respectively. The AUC and AUPR values obtained by the SVM classifier of the two datasets are close to 1, which indicates the unique superiority of the SVM classifier. It has good generalization performance and can effectively avoid over-fitting. Considering this, we choose the SVM algorithm as a classifier for predicting PPIs.

3.6. Performance of prediction model

In this paper, we first use PseAAC, AC, and EBGW to extract the features of *H. pylori* and *S. cerevisiae* datasets and fuse the three sets of feature vectors extracted by the three methods. Secondly, we choose the wavelet function as db8 and the decomposition scale to 6 to perform 2-D wavelet denoising on the fused feature vector. Finally, we use the SVM classifier based on the RBF kernel function to classify the data after 2-D wavelet denoising five times and use the 5-fold cross-validation method to test the results. The PPIs prediction results of the two datasets obtained are shown in Table 11.

Table 11

Prediction performance of *H. pylori* and *S. cerevisiae* datasets.

Datasets	5-fold cross-validation					
		ACC (%)	Se (%)	Pre (%)	MCC	AUC
H. pylori	1	96.33	96.16	96.52	0.9269	0.9931
	2	95.88	95.47	96.27	0.9177	0.9920
	3	95.75	95.54	95.94	0.9150	0.9922
	4	95.82	95.47	96.13	0.9163	0.9906
	5	96.06	96.02	96.09	0.9211	0.9926
	Average	95.97 ± 0.23	95.73 ± 0.33	96.19 ± 0.22	0.9194 ± 0.0048	0.9921 ± 0.0009
S. cerevisiae	1	95.60	95.66	95.56	0.9121	0.9914
	2	95.66	96.05	95.32	0.9133	0.9915
	3	95.37	95.57	95.19	0.9074	0.9912
	4	95.51	95.76	95.29	0.9103	0.9910
	5	95.59	96.00	95.23	0.9119	0.9915
	Average	95.55 ± 0.11	95.81 ± 0.21	95.32 ± 0.14	0.9100 ± 0.0023	0.9900 ± 0.0002

Table 11 shows the results of five classification predictions for the two datasets. It can be seen from Table 11 that the average ACC, Se, Pre, MCC and AUC of the *H. pylori* dataset obtained by the 5-fold cross-validation method are 95.97%, 95.73%, 96.19%, 0.9194 and 0.9921, respectively. The average ACC of the *S. cerevisia*e dataset is 95.55%, and the average Se, Pre, MCC, and AUC are 95.81%, 95.32%, 0.9100, and 0.9900, respectively. The results show that the PPIs-WDSVM prediction model based on multi-information fusion is an effective method to study PPIs.

3.7. Comparison with other methods

At present, many scholars use the *H. pylori* and *S. cerevisiae* datasets for PPIs prediction studies. In this paper, we use the 5-fold cross-validation method to compare the proposed PPIs prediction model based on multi-information fusion with other prediction models using the same datasets. First, we use PseAAC, AC, and EBGW as the feature extraction methods, and combine the feature vectors obtained by the three methods. Secondly, we choose the db8 wavelet function

and the decomposition scale is 6 and perform 2-D wavelet denoising on the fused feature vector. Finally, we use the SVM classifier based on the RBF kernel function to predict the wavelet denoised data. The prediction results of the *H. pylori* and *S. cerevisiae* datasets are compared with other methods as shown in Tables 12 and13, respectively.

Table 12

Comparison of different methods predicted results on *H. pylori* dataset.

Methods	ACC (%)	Pre (%)	Se (%)	MCC
Phylogenetic bootstrap (Bock and Gough, 2003)	75.80	80.20	69.80	N/A
HKNN (Nanni and Lumini, 2006)	84.00	84.00	86.00	N/A
Signature products (Martin et al., 2005)	83.40	85.70	79.90	N/A
Boosting (Shi et al., 2010)	79.52	81.69	80.37	0.7064
WSRC (Huang et al., 2016)	84.28	80.45	90.54	0.7325
PCVM (Wang et al., 2017)	91.25	90.06	92.05	0.8404
DVM (Li et al., 2017)	91.80	92.15	91.47	0.8360
PPIs-WDSVM	**95.97**	**96.19**	**95.73**	**0.9194**

It can be seen from Table 12 that the ACC of the PPIs-WDSVM prediction model proposed in the *H. pylori* dataset is 95.97%, which is 4.17%-20.17% higher than other methods. Among them, it is 4.17% higher than the DVM prediction model proposed by Li et al. (2017) and 20.17% higher than the Phylogenetic bootstrap model proposed by Bock and Gough. (2003). Similarly, our proposed prediction model has the highest Pre, Se, and MCC, which are 96.19%, 95.73%, and 0.9194, respectively. Among them, MCC is 21.30% higher than the Boosting model proposed by Shi et al. (2010) and 7.90% higher than the PCVM model proposed by Wang et al. (2017).

Table 13

Comparison of different methods predicted results on *S. cerevisiae* dataset.

Methods	ACC (%)	Pre (%)	Se (%)	MCC
ACC (Guo et al., 2008)	89.33	88.87	89.93	N/A
AC (Guo et al., 2008)	87.36	87.82	87.30	N/A
Cod1 (Yang et al., 2010)	75.08	74.75	75.81	N/A
Cod2 (Yang et al., 2010)	80.04	82.17	76.77	N/A
Cod3 (Yang et al., 2010)	80.41	81.66	78.14	N/A
Cod4 (Yang et al., 2010)	86.15	90.24	81.03	N/A
SVM+LD (Zhou et al., 2011)	88.56	89.50	87.37	0.7715
WSRC (Huang et al., 2016)	92.50	95.87	88.82	0.8609
RF+PR+LPQ (Wong et al., 2015)	93.92	**96.45**	91.10	0.8856
PCVM (Wang et al., 2017)	94.48	93.92	95.13	0.8958
PPIs-WDSVM	**95.55**	95.32	**95.81**	**0.9100**

It can be seen from Table 13 that the ACC of the RF+PR+LPQ model proposed by Wong et al.(2015) is 96.45%, which is 1.13% higher than the PPIs-WDSVM model proposed in this paper.

However, our proposed PPIs-WDSVM prediction model has the highest ACC, Se, and MCC on the S. cerevisiae dataset, which are 95.55%, 95.81% and 0.9100, respectively. Among them, ACC is 1.07%-20.47% higher than other prediction models. It is 6.22% higher than the ACC model proposed by Guo et al. (2008), 3.05% higher than the WSRC model proposed by Huang et al. (2016), and 1.63% higher than the RF+PR+LPQ model proposed by Wong et al. (2015). The MCC is 13.85% higher than the SVM+LD model proposed by Zhou et al. (2011), which is 4.91% higher than the WSRC model proposed by Huang et al. (2016) and 2.44% higher than the RF+PR+LPQ model proposed by Wong et al. (2015). The results show that the PPIs-WDSVM model based on multi-information fusion has achieved satisfactory prediction results on both datasets.

4. Conclusion

The study of PPIs not only provides clues for annotating the biological functions of unknown proteins but also provides the necessary information for understanding the mechanisms of life activities. The PPIs data presents big data characteristics. It takes time and energy to identify PPIs through the traditional high-throughput technology. Moreover, due to the limitations of experimental techniques and experimental conditions, there are a large number of proteins that are difficult to be studied experimentally in model organisms. Therefore, it is necessary to vigorously develop a PPIs prediction method based on machine learning to expand the understanding of protein function. This paper proposes a new protein-protein interaction prediction method based on multi-information fusion. Firstly, we use PseAAC, AC, and EBGW to extract the feature data of the protein sequence and fuse the extracted feature vectors to integrate the advantages of each feature extraction method and avoid losing important protein sequence feature information. Secondly, we perform 2-D wavelet denoising on the fused data, and the wavelet denoising method decomposes the obtained protein amino acid sequence, which can effectively remove redundant information and extract important characteristic signals in the protein sequence. It is of great significance for ensuring high prediction accuracy. Finally, we input the data after 2-D wavelet denoising to the SVM classifier for prediction. SVM can simplify the usual classification and regression problems, which not only can grasp the key samples but also the method is simple and has good robustness. The PPIs-WDSVM prediction model proposed in this paper has an average prediction accuracy of 95.97% and 95.55% for the *H. pylori* and *S. cerevisiae* datasets, respectively, and is compared with other prediction methods. The experimental results show that the PPIs-WDSVM prediction model based on multi-information fusion can effectively improve the prediction performance of PPIs. We hope that this method can be used not only for the prediction of PPIs but also as a powerful tool for related fields such as bioinformatics and proteomics.

As Chou and Shen (2009) pointed out, user-friendly and publicly accessible web-servers

represent the future direction of reporting important computational analysis and discovery (Shen and Chou, 2008; Liu et al., 2016; Liu et al., 2017a; Cheng and Xiao, 2017; Cheng et al, 2018a, 2018b, 2018c; Chou et al., 2018; Chen et al., 2018a; Jia et al., 2018). In fact, they significantly enhance the impact of computational biology on medical science (Chou, 2015), pushing medicine to an unprecedented revolution (Chou, 2017). In our future work, we will work hard to build a web-server for the prediction approach proposed in this study.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (Nos. 61863010 and 11771188), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), and the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159). This work was supported by National Science Foundation/EPSCoR Award No. IIA-1355423, the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University (SDSU). This work was also supported by Hatch Project: SD00H558-15/project accession No. 1008151 from the USDA National Institute of Food and Agriculture. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (grant number ACI-1548562).

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, 293-304.

An, J., Meng, F., You, Z., Chen, X., Yan, G., Hu, J., 2016. Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. Protein Sci. 25, 1825-1833.

Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O., 2016. Deep learning for computational biology. Mol. Syst. Biol. 12, 878.

Behbahani, M., Mohabatkar, H., Nosrati, M., 2016. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. J. Theor. Biol. 411, 1-5.

Binny, P.S., Saha, S., Anishetty, R., Anishetty, S., 2013. A matrix based algorithm for protein-protein interaction prediction using domain-domain associations. J. Theor. Biol. 326, 36-42.

Bock, J.R., Gough, D.A., 2003. Whole-proteome interaction mining. Bioinformatics 19, 125-135.

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

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.

Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., Plewczynski, D., 2011. PPI_SVM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. Cell. Mol. Biol. Lett. 16, 264-278.

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

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

Cheng, X., Xiao, X., Chou, K.C., 2017e. 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., Xiao, X., Chou, K.C., 2017f. pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics 7543, 30102-30107.

Cheng, X., Xiao, X., Chou, K.C., 2017g. pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics 34, 1448-1456.

Cheng, X., Xiao, X., 2017. 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., Lin, W.Z., Xiao, X., Chou, K.C., 2018a. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinformatics doi: 10.1093/bioinformatics/bty628.

Cheng, X., Xiao, X., Chou, K.C., 2018b. pLoc_bal-mGneg: predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC. J. Theor. Biol. 458, 92-102.

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

Chen, W., Ding, H., Zhou, X., Lin, H., Chou, K.C., 2018a. iRNA(m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition. Anal. Biochem. 15, 561-562, 59-65.

Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., Chou, K.C., 2018b. iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites. Mol. Ther. Nucleic Acids 11, 468-474.

Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013. iRSpot-PseDNC: identify recombination spots

with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68.

Chen, W., Lei, T.Y., Jin, D.C., Lin, H., Chou, K.C., 2014. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 456, 53-60.

Chou, K.C., 2001a. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 44, 246-255.

Chou, K.C., 2001b. Prediction of signal peptides using scaled window. Peptides 22, 1973-1979.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21,10-19.

Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteom. 6, 262-274.

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., 2006. Predicting protein-protein interactions from sequences in a hybridization space. J. Proteome Res. 5, 316-322.

Chou, K.C., Cheng, X., Xiao, X., 2018. pLoc_bal-mHum: Predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. Genomics doi: 10.1016/j.ygeno.2018.08.007.

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.

Chou, K.C., Shen, H.B., 2007. 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., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629-641.

Daberdaku, S., Ferrari, C., 2018. Exploring the potential of 3D Zernike descriptors and SVM for protein-protein interface prediction. BMC Bioinf. 19, 35.

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol. 364, 284-294.

Ding, Y., Tang, J., Guo, F., 2016. Predicting protein-protein interactions via multivariate mutual information of protein sequences. BMC Bioinf. 17, 398.

Donaldson, I., Martin, J., Bruijn, B.D., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T., Hogue, C.W., 2003. Prebind and textomy-mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinf. 4, 11.

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 Bioinf. 7, 1-8.

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.

Du, X., Sun, S., Hu, C., Yu, Y., Yan, Y., Zhang, Y. 2017. DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. J. Chem. Inf. Model 57, 1499-1510.

Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., Chou, K.C., 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.M., Chen, W., Lin, H., Chou, K.C., 2013. iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal. Biochem. 442, 118-125.

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

Folador, E.L., Hassan, S.S., Lemke, N., Barh, D., Silva, A., Ferreira, R.S., Azevedo, V., 2014. An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. Integr. Biol. 6, 1080-1087.

Gao, Z.G., Wang, L., Xia, S.X., You, Z.H., Yan, X., Zhou, Y., 2016. Ens-PPI: a novel ensemble classifier for predicting the interactions of proteins using auto covariance transformation from PSSM. Biomed. Res. Int. 2016, 4563524.

Göktepe, Y.E., Kodaz, H., 2018. Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing 303, 68-74.

Guo, Y., Li, M., Lu, M., Wen, Z., Huang, Z., 2006. Predicting G-protein coupled receptors G-protein coupling specificity based on autocross-covariance transform. Proteins 65, 55-60.

Guo, Y., Yu, L., Wen, Z., Li, M., 2008. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic. Acids Res. 36, 3025-3030.

Hamp, T., Rost, B., 2015. Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics 31, 1945-1950.

Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., Vidal, M., 2005. Effect of sampling on topology predictions of protein-protein interaction networks. Nat. Biotechnol. 23, 839-944.

Huang, L., Liao, L., Wu, C.H., 2016. Protein-protein interaction prediction based on multiple kernels and partial network with linear programming. BMC. Syst. Biol. 10, 45.

Huang, T., Chen, L., Cai, Y.D., Chou, K.C., 2011. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS One 6, e25297.

Huang, Y.A., You, Z.H., Chen, X., Yan, G.Y., 2016. Improved protein-protein interactions prediction via weighted sparse representation model combining continuous wavelet descriptor and PseAA composition. BMC Syst. Biol. 10, 485-494.

Huang, Y.A., You, Z.H., Gao, X., Wong, L., Wang, L., 2015. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. Biomed. Res. Int. 2915, 902198.

Hu, L., Huang, T., Shi, X., Lu, W.C., Cai, Y.D., Chou, K.C., 2011. Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS One 6, e14556.

Hu, L.L., Feng, K.Y,, Cai, Y.D., Chou, K.C., 2012. Using protein-protein interaction network information to predict the subcellular locations of proteins in budding yeast. Protein Pept. Lett. 19, 644-651.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2015a. 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., Chou, K.C., 2015b. 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.

Jia, J., Xiao, X., Liu, B., 2015c. Prediction of protein-protein interactions with physicochemical descriptors and wavelet transform via random forests. J. Lab. Autom. 21, 368-377.

Jia, J., Li, X., Qiu, W., Xiao, X., Chou, K.C., 2018. iPPI-PseAAC (CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC. J. Theor. Biol. 60, 195-203.

Khan, M., Hayat, M., Khan, S.A., Iqbal, N., 2017. Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. J. Theor. Biol. 415, 13-19.

Lapinsh, M., Gutcaits, A., Prusis, P., Post, C., Lundstedt, T., Wikberg, J.E.S., 2002. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci. 11, 795-805.

Lei, Y.K., You, Z.H., Ji, Z., Zhu, L., Huang, D.S., 2012. Assessing and predicting protein interactions by combining manifold embedding with multiple information integration. BMC Bioinformatics 13, S3.

Liang, K., Zhang, L., Lv, J., 2014. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 344, 12-18.

Liao, B., Jiang, J.B., Zeng, Q.G., Zhu, W., 2011. Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition. Protein. Pept. Lett 18, 1086-1092.

Li, B.Q., Huang, T., Liu, L., Cai, Y.D., Chou, K.C., 2012. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS One, 7, e33393.

Li, L., Yu, S., Xiao, W., Li, Y., Hu, W., Huang, L., Zheng, X., Zhou, S., Yang, H., 2014. Protein submitochondrial localization from integrated sequence representation and SVM-based backward feature extraction. Mol. Biosyst. 11, 170-177.

Li, L., Zhang, P., Zheng, T., Zhang, H., Jiang, Z., Huang, D., 2014. Integrating semantic information into multiple kernels for protein-protein interaction extraction from biomedical literatures. PLoS One 9, e91898.

Lin, C., Chen, W., Qiu, C., Wu, Y., Krishnan, S., Zou, Q., 2014. LibD3C: ensemble classifiers with

a clustering and dynamic selection strategy. Neurocomputing 123, 424-435.

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, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2013. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. 9, 634-644.

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

Liu, B., Fang, L., Long, R., Lan, X., Chou, K.C., 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., Li, K., Huang, D.S., Chou, K.C., 2018. iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics doi: 10.1093/bioinformatics/bty458.

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., 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, W65-71.

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

Liu, B., Yang, F., Chou, K.C., 2017b. 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., Chou, K.C., 2017c. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 34, 33-40.

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

Liu, B., Weng, F., Huang, D.S., Chou, K.C. 2018. iRO-3wPseKNC: identify DNA replication origins by three-window-based PseKNC. Bioinformatics 34,3086-3093.

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., Chou, K.C., 2014. iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9, e106691.

Liu, G.H., Shen, H.B., Yu, D.J., 2015. Prediction of protein-protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures. J. Membr. Biol. 249, 141-153.

Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., 2012. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42, 2243-2249.

Liu, T., Zheng, X., Wang, C., Wang, J., 2010. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept. Lett. 17, 1263-1269.

Li, Z., Zhou, X., Dai, Z., Zou, X., 2012. Classification of G proteins and prediction of GPCRs-g proteins coupling specificity using continuous wavelet transform and information theory. Amino Acids 43, 793-804.

Li, Z.W., You, Z.H., Chen, X., Li, L.P., Huang, D.S., Yan, G.Y., Nie, R., Huang, Y.A., 2017. Accurate prediction of protein-protein interactions by integrating potential evolutionary

information embedded in PSSM profile and discriminative vector machine classifier. Oncotarget 8, 23638-23649.

Luo, X., You, Z., Zhou, M., Li, S., Leung, H., Xia, Y., Zhu, Q., 2015. A highly efficient approach to protein interactome mapping based on collaborative filtering framework. Sci. Rep. 5, 7702.

Mandal, M., Mukhopadhyay, A., Maulik, U., 2015. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Comput. 53, 331-344.

Martin, S., Roe, D., Faulon, J.L., 2005. Predicting protein-protein interactions using signature products. Bioinformatics 15, 441-446.

Meher, P.K., Sahu, T.K., Saini, V., Rao, A.R., 2017. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci. Rep. 7, 42362.

Mier, P., Alanis-Lobato, G., Andrade-Navarro, M.A., 2017. Protein-protein interactions can be predicted using coiled coil co-evolution patterns. J. Theor. Biol. 412, 198-203.

Murakami, Y., Tripathi, L. P., Prathipati, P., Mizuguchi, K., 2017. Network analysis and in silico prediction of protein-protein interactions with applications in drug discovery. Curr. Opin. Struct. Biol. 44, 134-142.

Nanni, L., Lumini, A., 2006. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Oxford University Press 22, 1207-1210.

Nanni, L., Lumini, A., 2009. An ensemble of reduced alphabets with protein encoding based on grouped weight for predicting DNA-binding proteins. Amino Acids 36, 167-175.

Qiu, J.D., Luo, S.H., Huang, J.H., Sun, X.Y., Liang, R.P., 2010. Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine. Amino Acids 38, 1201-1208.

Qiu, W., Li, S., Cui, X., Yu, Z., Wang, M., Du, J., Peng, Y., Yu, B., 2018. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. J. Theor. Biol. 450, 86-103.

Ren, L.H., Shen, Y.Z., Ding, Y.S., Chou, K.C., 2011. Bio-entity network for analysis of protein-protein interaction networks. Asian J. Control 13, 726-737.

Shen, H.B., Chou, K.C., 2008. HIVcleave: a web-server for predicting human immunodeficiency virus protease cleavage sites in proteins. Anal. Biochem. 375, 388-390.

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.

Shi, M.G., Xia, J.F., Li, X.L., Huang, D.S., 2010. Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids 38, 891-899.

Shi, S.P., Qiu, J. D., Sun, X.Y., Suo, S.B., Huang, S.Y., Liang, R.P., 2012. A method to distinguish between lysine acetylation and lysine methylation from protein sequences. J. Theor. Biol. 310, 223-230.

Shoemaker, B.A., Panchenko. A.R., 2007. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS. Comput. Biol. 3, e43.

Su, Z.D., Huang, Y., Zhang, Z.Y., Zhao, Y.W., Wang, D., Chen, W., Chou, K.C., Lin, H., 2018.

iLoc-IncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. doi: 10.1093/bioinformatics/bty508.

Szilágyi, A., Grimm, V., Arakaki, A.K., Skolnick, J., 2005. Prediction of physical protein-protein interactions. Phys. Biol. 2, S1-16.

Tung, C.H., Chen, C.W., Sun, H.H., Chu, Y.W., 2017. Predicting human protein subcellular ocalization by heterogeneous and comprehensive approaches. PLoS. One 12, e0178832.

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.

Wang, D.D., Wang, R., Yan, H., 2014. Fast prediction of protein-protein interaction sites based on extreme learning machines. Neurocomputing 128, 258-266.

Wang, J., Zhang, L., Jia, L., Ren, Y., Yu, G., 2017. Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. Int. J. Mol. Sci. 18, 2373.

Wang, T., Li, L., Huang, Y.A., Zhang, H., Ma, Y., Zhou, X., 2018. Prediction of protein-protein interactions from amino acid sequences based on continuous and discrete wavelet transform features. Molecules 23, 823.

Wang, X., Li, G.Z., 2012. A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins. PLoS One 7, e36317.

Wang, Y., You, Z., Li, X., Chen, X., Jiang, T., Zhang, J., 2017. PCVMZM: using the probabilistic classification vector machines model combined with a zernike moments descriptor to predict protein-protein interactions from protein sequences. Int. J. Mol. Sci. 18, 1029.

Wei, C., Hao, L., Chou, K.C., 2015. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol. Biosyst. 11, 2620-2634.

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., Guo, F., 2017. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. 83, 67-74.

Wen, Y.T., Lei, H.J., You, Z.H., Lei, B.Y., Chen, X., Li, L.P., 2017. Prediction of protein-protein interactions by label propagation with protein evolutionary and chemical information derived from heterogeneous network. J. Theor. Biol. 430, 9-20.

Wong, L., You, Z.H., Ming, Z., Li, J., Chen, X., Huang, Y.A., 2015. Detection of interactions between proteins through rotation forest and local phase quantization descriptors. Int. J. Mol. Sci. 17, 21.

Wu, J., Li, M.L., Yu, L.Z., Wang, C., 2010. An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition. Protein J. 29, 62-67.

Xenarios, I., Łukasz Salwínski, Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D., 2002. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions Nucleic. Acids. Res. 30, 303-305.

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., Wu, Z.C., Chou, K.C., 2011. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol.

284, 42-51.

Xu, R., Zhou, J., Liu, B., He, Y., Zou, Q., Wang, X., Chou, K.C., 2015. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. J. Biomol. Struct. Dyn. 33, 1720-1730.

Xu, Y., Ding, J., Wu, L.Y., Chou, K.C., 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.

Yang, H., Qiu, W.R., Liu, G., Guo, F.B., Chen, W., Chou, K.C., Lin, H., 2018. iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiaeby incorporating hexamer composition into general PseKNC. Int. J. Biol. Sci. 14, 883-891.

Yang, L., Li, Y., Xiao, R., Zeng, Y., Xiao, J., Tan, F., Li, M., 2010. Using auto covariance method for functional discrimination of membrane proteins based on evolution information. Amino Acids 38, 1497-1503.

Yang, Z., Lin, H., Li, Y., 2010. BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. J. Biomed. Inform. 43, 88-96.

Yan, X., Shao, X.J., Wu, L.Y., Deng, N.Y., Chou, K.C., 2013. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1, e171.

Yin, Z., Deng, T., Peterson, L.E., Yu, R., Lin, J., Hamilton, D.J., Reardon, P.R., Sherman. V., Winnier. G., Zhan. M., Lyon. C.J., Wong. S.T., Hsueh. W.A., 2014. Transcriptome analysis of human adipocytes implicates the NOD-like receptor pathway in obesity-induced adipose inflammation. Mol. Cell. Endocrinol 394, 80-87.

Yousef, A., Moghadam, C.N., 2013. A novel method based on new adaptive LVQ neural network for predicting protein-protein interactions from protein sequences. J. Theor. Biol. 336, 231-239.

You, Z.H., Lei, Y.K., Gui, J., Huang, D.S., Zhou, X., 2010. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics 26, 2744-2751.

You, Z.H., Li, J., Gao, X., He, Z., Zhu, L., Lei, Y.K., Ji, Z.W., 2015. Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. Biomed Res. Int. 2015, 1-9.

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

Yu, B., Lou, L., Li, S., Zhang, Y., Qiu, W., Wu, X., Wang, M., Tian, B., 2017c. 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., Wang, M., Du, J., Zhang, Y., Chen, X., 2018. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. BMC Genomics, 19, 478.

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.

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, P., Tao, L., Zeng, X., Qin, C., Chen, S., Zhu, F., Li, Z., Jiang, Y., Chen, W., Chen, Y.Z., 2016. A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. Brief. Bioinform 18, 1057-1070.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., Honig, B., 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490, 556-560.

Zhang, S., 2015. Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC. Chemom. Intell. Lab. Syst. 142, 28-35.

Zhang, S.B., Tang, Q.R., 2016. Protein-protein interaction inference based on semantic similarity of gene ontology terms. J. Theor. Biol. 401, 30-37.

Zhang, S.W., Wei, Z.G., 2015. Some remarks on prediction of protein-protein interaction with machine learning. Med. Chem. 11, 254-264.

Zhang, X., Xiong, J., Song, J., Chang, S., 2016. Prediction of human protein-protein interaction by a domain-based approach. J. Theor. Biol. 396, 144-153.

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

Zhang, Z.H., Wang, Z.H., Wang, Y.X., 2005. A new encoding scheme to improve the performance of protein structural class prediction. Lect. Notes Comput. Sc. 1164-1173.

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

Zhao, X.W., Ma, Z.Q., Yin, M.H., 2012. Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition. Protein Pept. Lett. 19, 492-500.

Zhou, Y. Z., Gao, Y., Zheng, Y.Y., 2011. Prediction of protein-protein interactions using local description of amino acid sequence. Adv. Comput. Sci. Educ. Appl. 202, 254-262.

Zhu, L., You, Z.H., Huang, D.S., 2013a. Increasing the reliability of protein-protein interaction networks via non-convex semantic embedding. Neurocomputing 121, 99-107.

Zhu, L., You, Z.H., Huang, D.S., Wang, B., 2013b. t-LSE: a novel robust geometric approach for modeling protein-protein interaction networks. PLoS One 8, e58368.

Zhu, Z., Zhang, Y., Ji, Z., He, S., Yang, X., 2015. High-throughput DNA sequence data compression. Brief. Bioinform. 16, 1-15.

Zou, H.L., 2014. A multi-label classifier for prediction membrane protein functional types in animal. J. Membr. Biol. 247, 1141-1148.