Computational Prediction of Protein-Protein Interactions in Plants Using Only Sequence Information

Jie. Pan
Xijing University
Zhu Hong. You (zhuhongyou@ms.xjb.ac.cn)
Xijing University
Li Ping. Li
Xijing University
Chang-Qing. Yu
Xijing University
Xin-Ke. Zhan
Xijing University

Research Article

Keywords: Protein-protein interactions (PPIs), position specific scoring matrix (PSSM), inverse fast Fourier transform (IFFT)

Posted Date: April 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-411601/v1

Computational prediction of protein-protein interactions in plants using only sequence information

Jie. Pan[1], Zhu Hong. You[1]*, Li Ping. Li[1], Chang-Qing. Yu[1], Xin-Ke. Zhan[1]

1 College of Information Engineering, Xijing University, Xi'an 710123, China;
* Correspondence: zhuhongyou@ms.xjb.ac.cn (Zhu-Hong. You.)

Abstract
Protein-protein interactions (PPIs) in plants plays a significant role in plant biology and functional organization of cells. Although, a large amount of plant PPIs data have been generated by high-throughput techniques, but due to the complexity of plant cell, the PPIs pairs currently obtained by experimental methods cover only a small fraction of the complete plant PPIs network. In addition, the experimental approaches for identifying PPIs in plants are laborious, time-consuming, and costly. Hence, it is highly desirable to develop more efficient approaches to detect PPIs in plants. In this study, we present a novel computational model combining weighted sparse representation-based classifier (WSRC) with a novel inverse fast Fourier transform (IFFT) representation scheme which was adopted in position specific scoring matrix (PSSM) to extract features from plant protein sequence. When performed the proposed method on the plants PPIs dataset of *Mazie*, *Rice* and *Arabidopsis thaliana* (*Arabidopsis*), we achieved excellent results with high accuracies of 89.12%, 84.72% and 71.74%, respectively. To further assess the prediction performance of the proposed approach, we compared it with the state-of-art support vector machine (SVM) classifier. To the best of our knowledge, we are the first to employ protein sequences information to predict PPIs in plants. Experimental results demonstrate that the proposed method has a great potential to become a powerful tool for exploring the plant cell function.

1. Introduction

In plants, the prediction of protein-protein interactions (PPIs) provides important information for understanding the molecular mechanisms underlying biological processes. Recently, a large number of high-throughput experimental approaches have been developed to identified PPIs, such as affinity-purification coupled to mass spectrometry (AP-MS) [1] and yeast two-hybrid (Y2H) [2-5] screens methods. Although we have accumulated a large amount of plant PPIs data [6-8], these experimental approaches also some inevitable drawbacks, which are not only costly, but also laborious and time-consuming. Moreover, these traditional biochemical experiments always suffer from high false positive rates and high false negative rates. And due to the complexity of plant growth and development systems, large-scale prediction experimental methods could not be adopted in plant domain, and now only a small fraction of the whole plant PPIs network can be detected. Therefore, it is very significance to develop the efficient computational approaches to identify PPIs in plants [9, 10].

In recent years, much effort has been made to develop PPIs identification methods based on different data types, including literature mining knowledge [11], gene fusion [12] and protein structure information [13, 14]. A large amount of PPIs dataset has been built, such as TAIR [15], PRIN [16], and MINT [17]. There are also some approaches that combine data and information from different sources [18-20] to predict PPIs. However, without prior knowledge of corresponding proteins, these methods cannot be implemented.

Recently, the PPIs prediction methods, which extract information directly from amino acid sequences have received much attention [21-24]. Many researchers have worked to provide sequences-based methods to detect novel PPIs, and experimental results indicated that PPIs in plants can be accurately identified using only sequence information [25-28]. For example, Sun *et al.* [29] presented a method that using a type of deep-learning algorithm called stacked autoencoder (SAE) to use sequence-based approaches for predicting PPIs in human datasets. This model obtained the best results on 10-fold cross-validation which was based on protein sequence autocovariance coding. One of the excellent works that utilizing the protein sequence information to

predict PPIs is presented by Shen *et al.* [30]. This method is based on a SVM model that combine with a conjoint triad feature and a kernel function for describing amino acids. Specifically, according to the volumes and dipoles, the 20 amino acid sequence will be clustered into seven classes. Then the conjoint triad method will abstract the features of protein pairs. Wang *et al.* [31] proposed a novel computational method for detecting PPIs adopting sequence information, and combining Zernike moments descriptor with stacked autoencoder. First, they employed Zernike moment feature representation on a position specific scoring matrix. Secondly, a stacked autoencoder was used for noise reduction. Finally, a powerful model, the probabilistic classification vector machines model (PCVM) was used to handle the classification problem. You *et al.* [32] also developed a novel computational approach called PCA-EELM to predict PPIs. The main improvement of this study is that they adopted the PCA method to construct the most discriminative new feature set. In addition, many methods based on amino acid sequences have been developed in the literature [33, 34]. While these studies have achieved some progress, there is still room for improvement in terms of the efficiency and accuracy of the models.

In the present work, we provided a novel computational method to detect the PPIs in plants from protein sequence information, which employing a novel position specific scoring matrix (PSSM) and combining the weighted sparse representation-based classifier (WSRC) with inverse fast Fourier transform (IFFT). This feature representation approach combined with the WSRC has remarkably performed in the prediction of the PPIs in plants. Furthermore, the main idea of our proposed model includes three steps. First, the plant protein sequence could be represented as a position specific scoring matrix so that we can obtain the biological evolutionary information between different types of amino acids. Second, utilizing the inverse Fast Fourier transform (IFFT) method to extracted a 400-dimensional vector from each plant proteins PSSM matrix. As a result, each protein pairs will be described as an 800-dimensional feature vector. Thirdly, a powerful classifier, weighted sparse representation-based classifier, is employed to perform PPIs predictions on three plants PPIs datasets, including Maize, Rice and Arabidopsis thaliana (Arabidopsis). We also compared the proposed model with the state-of-the-art support vector machine (SVM) classifier to further evaluate the prediction performance. The experiments results demonstrated that our approach performs significantly well in distinguishing interacting and noninteracting plants protein pairs. These experimental results further shows that the proposed approach is promising and reliable for the prediction of protein-protein interactions in plants. The source codes and datasets explored in this work are available at: https://github.com/jie-pan111/protein_sequence.

2. Results

2.1. Evaluation Criteria.

To demonstrate the prediction performance of the proposed approach, four evaluation criteria was used in this work, including accuracy (Acc.), sensitivity (Sen.), precision (Prec.), and Matthews correlation coefficient (MCC) [35-37]. Their corresponding calculating formulas are defined as follows:

$$Acc. = \frac{TN + TP}{TP + FP + FN + TN} \tag{1}$$

$$Sen. = \frac{TP}{TP + FN} \tag{2}$$

$$PR. = \frac{TP}{TP + FP} \tag{3}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP \times FP) \times (TN + FN) \times (TN + FP) \times (TP \times FN)}} \tag{4}$$

where true positive (TP) denotes the number of plants protein-protein pairs classified as interacting correctly while true negative (TN) stands for the number of non-interacting PPIs pairs predicted correctly; false positive (FP) denotes the number of samples classified as interacting incorrectly, and false negative (FN) denotes the count of interacting plants PPIs pairs that predict to have no interaction. In addition, we also adopted the receiver operating characteristic (ROC) curves to assess the prediction performance of the proposed approach, and the area under the

Receiver Operating Characteristic curve (AUC) is calculated used for demonstrating the quality of prediction model.

2.2. Assessment of Prediction Ability.

In this article, we used 5-fold cross-validation to evaluate the predictive ability of our model in three plant data sets involving *Maize*, *Rice* and *Arabidopsis*. In this way, we can prevent overfitting and test the stability of the proposed method. More specifically, the whole data set is partitioned into five roughly equal parts, four of them were used to construct a training set and the rest one was adopted as a testing set. Thus, five models can be generated for the five sets of data. The cross validation has the advantages that it can minimize the impact of data dependency and improved the reliability of the results.

The five-fold cross validation results of the proposed approach on the three plants datasets are listed in Table 1-3. Form Table 1, we can observe that when applying the proposed method to the *Mazie* data set, we obtained best prediction results of average accuracy, precision, sensitivity, and MCC were 89.12%, 87.49%, 91.32%, and 80.59%, with corresponding standard deviations 0.59%, 1.38%, 0.64%, and 0.94%, respectively. When exploring the proposed method on the *Rice* dataset, we yield the good results of average accuracy, precision, sensitivity, MCC of 84.72%, 85.04%, 84.44% and 84.10%, respectively. The standard deviations of these criteria values are 0.73%, 0.85%, 0.65% and 1.00% respectively. When predicting PPIs of *Arabidopsis* dataset, the proposed approach obtained good results of average accuracy, precision, sensitivity, MCC of 71.74%, 69.33%, 77.02% and 58.97% and the standard deviations are 0.48%, 0.58%, 1.15% and 0.38%, respectively. Figure. 1-3 shows the ROC curves for the proposed approach on *Maize*, *Rice* and *Arabidopsis.* The average AUC values range from 79.19% to 93.76% (*Maize*: 93.76%, *Rice*: 88.75% and *Arabidopsis:*79.19%), suggesting that our method is fit well for our purposes to predict PPIs in plants from amino acid sequences.

These good results collectively demonstrate that using the information of protein sequence alone to predict PPIs in plants is sufficient enough, and that powerful prediction capability for predicting PPIs can be yielded by adopting weighted sparse representation-based classifier combined IFFT features. This strong prediction performance derives from the feature extraction method for plant protein sequences and the choice of machine learning classifier. The high accuracies and low standard deviations of this criterion values indicate that our proposed model is feasible and effective for predicting PPIs in plants.

Test set	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
1	89.16	87.81	90.84	80.66	93.64
2	88.64	85.94	91.84	79.83	93.64
3	88.56	87.19	90.89	79.70	93.24
4	89.20	86.84	92.17	80.71	94.05
5	90.04	89.65	90.85	82.06	94.21
Average	**89.12 ± 0.59**	**87.49 ± 1.38**	**91.32 ± 0.64**	**80.59 ± 0.94**	**93.76 ± 0.38**

Table 1. 5-fold cross-validation results achieved on the ***Maize*** dataset using the proposed method.

Test set	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
1	84.22	84.66	83.70	73.42	88.96
2	85.63	84.76	84.95	75.32	89.16
3	84.74	85.51	84.73	74.12	89.37
4	85.21	86.24	85.03	74.77	88.43
5	83.80	84.04	83.78	72.85	87.82
Average	**84.72 ± 0.73**	**85.04 ± 0.85**	**84.44 ± 0.65**	**84.10 ± 1.00**	**88.75 ± 0.62**

Table 2. 5-fold cross-validation results achieved on the ***Rice*** dataset using the proposed method.

Test set	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
1	71.31	69.00	76.58	58.88	77.95
2	71.55	68.59	77.65	59.03	80.56
3	71.15	69.94	76.72	58.61	78.71
4	71.07	69.22	75.56	58.72	78.74
5	72.28	69.89	78.59	59.59	80.02
Average	**71.74 ± 0.48**	**69.33 ± 0.58**	**77.02 ± 1.15**	**58.97 ± 0.38**	**79.19 ± 1.06**

Table 3. 5-fold cross-validation results achieved on the ***Arabidopsis*** dataset using the proposed method.

2.3. Comparison of the proposed model with different classifiers.

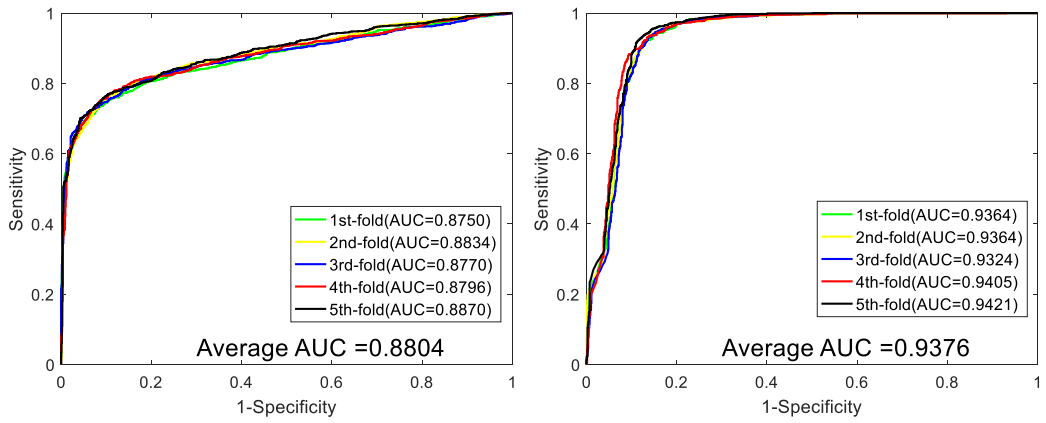
Although the WSRC model obtained better performance in predicting PPIs of plants, we also need to further verify the prediction ability of the proposed method. We compared the prediction accuracy of the WSRC model with that of the state-of-art SVM model via the same feature extraction approach based on the *Maize*, *Rice* and *Arabidopsis* datasets, respectively. We applied the same feature extraction approach on the *Maize*, *Rice* and *Arabidopsis* datasets and compared the prediction accuracy of the WSRC model with the state-of-the-art SVM. We employed the LIBSVM tool to run this classification, and 5-fold cross-validation was also adopted in these experiments. In order to obtain better performance of SVM classifier, we should optimize several parameters of SVM classifier. In this study, the penalty parameter C and the kernel parameter g of SVM model was optimized by the gird search method. In the experiments of *Maize* and *Rice* dataset, we set c=5, g=0.5 and c=6, g=0.5. when applying on *Arabidopsis* dataset, we set c=7, g=0.03.

As shown in Table 4, it is clearly seen that when applied the SVM model to predict PPIs of *Maize* dataset, we yield good results with average accuracy, precision, sensitivity, MCC and AUC of 81.77%, 83.10%, 79.78%, 70.16% and 88.04%, respectively. When identifying PPIs of Rice dataset, the SVM classifier yield good results with average accuracy, precision, sensitivity, MCC and AUC of 79.13%, 78.27%, 80.91%, 66.93% and 86.62%, respectively. When exploring the *Arabidopsis* dataset, the average accuracy, precision, sensitivity, MCC and AUC come to be 62.55%, 63.49%, 58.97%, 53.03% and 66.87%, respectively. For the three plant datasets, the classification results yield by the SVM-based models are lower than those by the proposed approach. In summary, it is obvious that the overall prediction results of WSRC model is better than that of SVM-based approach.

Dataset	Classifier	Acc. (%)	PR. (%)	Sen. (%)	MCC (%)	AUC (%)
Mazie	WSRC	**89.12 ± 0.59**	**87.49 ± 1.38**	**91.32 ± 0.64**	**80.59 ± 0.94**	**93.76 ± 0.38**
	SVM	81.77 ± 0.57	83.10 ± 1.30	79.78 ± 0.92	70.16 ± 0.71	88.04 ± 0.48
Rice	WSRC	**84.72 ± 0.73**	**85.04 ± 0.85**	**84.44 ± 0.65**	**84.10 ± 1.00**	**88.75 ± 0.62**
	SVM	79.13 ± 0.81	78.27 ± 2.13	80.91 ± 0.55	66.93 ± 0.94	86.62 ± 0.91
Arabidopsis	WSRC	**71.74 ± 0.48**	**69.33 ± 0.58**	**77.02 ± 1.15**	**58.97 ± 0.38**	**79.19 ± 1.06**
	SVM	62.55 ± 1.44	63.49 ± 1.68	58.97 ± 2.85	53.03 ± 0.81	66.87 ± 1.52

Table 4: The comparison of the WSRC method with the SVM-based method on three plant datasets.

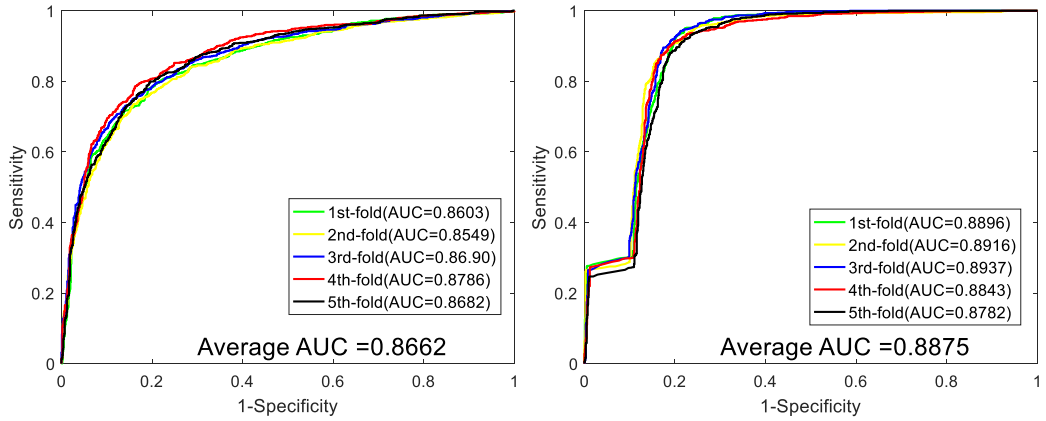
Meanwhile, the ROC curves of these experiments are also shown in Figs 1-3. From Figure 1, it is obvious to see that the AUC value of SVM model on the *Maize* dataset is 0.8804 and that of the WSRC is 0.8912. From Figure 2, we can see that the average AUC of SVM classifier is 0.8662 and that of WSRC method is 0.8875. From Figure 3 we can see that when predicting PPIs of *Arabidopsis* dataset, the SVM-based method can obtain good results with average AUC of 0.6687 and that of WSRC method is 0.7919. All these experiments indicates that the average AUC value of WSRC method is so large than that of the SVM-based method. From all of these experiments results, we can draw the following conclusion that the weighted sparse representation-based classifier is an effective and robust model for PPIs prediction in plants.



(a). ROC of SVM method (b). ROC of WSRC method

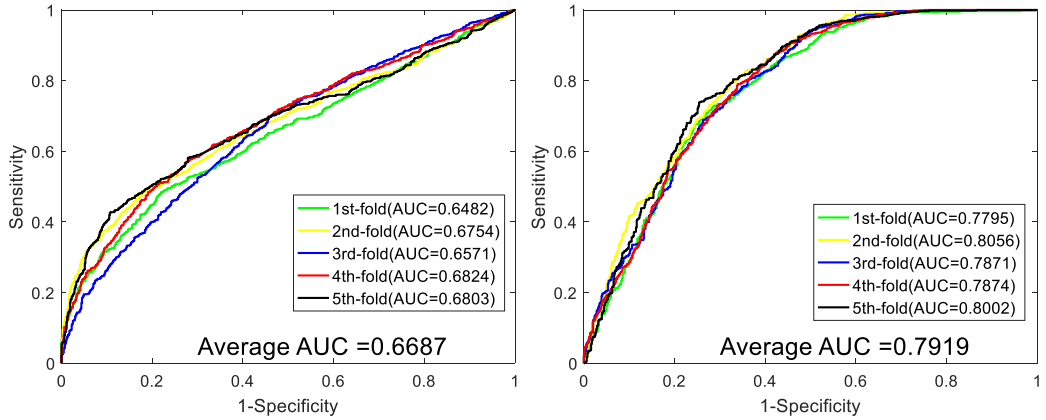
Figure 1. Comparison of the ROC curves obtained by WSRC and SVM-based method on ***Maize*** dataset (5-fold

cross validation). (a) shows the ROC curves performed by SVM method on ***Mazie*** PPIs dataset. (b) shows the ROC curves performed by WSRC method on ***Maize*** dataset.



(a). ROC of SVM method (b). ROC of WSRC method

Figure 2. Comparison of the ROC curves obtained by WSRC and SVM-based method on ***Rice*** dataset (5-fold cross validation). (a) shows the ROC curves performed by SVM method on ***Rice*** PPIs dataset. (b) shows the ROC curves performed by WSRC method on ***Rice*** dataset.



(a). ROC of SVM method (b). ROC of WSRC method

Figure 3. Comparison of the ROC curves obtained by WSRC and SVM-based method on ***Arabidopsis*** dataset (5-fold cross validation). (a) shows the ROC curves performed by SVM method on ***Arabidopsis*** PPIs dataset. (b) shows the ROC curves performed by WSRC method on ***Arabidopsis*** dataset.

3. Methods

3.1. Data collection and data set construction.

We verify the proposed model on three plants PPIs dataset. The first dataset is *Maize*. Maize is one of the most important food, feed and industrial crops in the world and also an excellent model for plant genetics. In order to better demonstrate the prediction performance of the proposed model and understand the molecular mechanisms underlying various traits of *Maize*, we select the maize as the third plant data set in this study. We collected the *Maize* dataset from agriGO [38] and *Protein-Protein Interaction Database for Maize* (PPIM) [39], which covers 2,762,560 interactions among 14,000 proteins. After the strict inclusion and exclusion screening, we select 6250 protein pairs from 6497 maize proteins. As a result, the whole *Maize* dataset is constructed by 12500 maize protein pairs.

Rice is one of the most important staple foods for more than half of the world's population. To validate the generality of the proposed method, we also performed our method on the *Rice* PPIs dataset. We collected the *Rice* dataset from the protein reference database agriGO [38] and PRIN [16]. In order to construct the negative dataset, we selected 4800 additional protein pairs which work in different subcellular and assumed that they will not interact with each other. As a result, the whole *Rice* data set is constructed by 9600 protein pairs from 3760 *Rice* proteins.

Arabidopsis thaliana (*Arabidopsis*) is a well-known model plant and we chose it as the third

dataset in this study, which we collected from public PPIs databases TAIR [15], IntAct [40] and BioGRID [41]. After removing redundant PPIs, we yield the remaining 4120 protein pairs to build the positive data set, which containing 6013 *Arabidopsis* proteins [42]. For constructing the negative data set, we randomly selected the same number of non-interacting protein pairs. On this foundation, the entire *Arabidopsis* dataset is constructed by 8240 protein pairs.

3.2. Position-Specific Scoring Matrix (PSSM).

Through the Position-Specific Scoring Matrix (PSSM) which is reported by Gribskov [43] and it achieved great success in protein binding site prediction, protein secondary structure prediction and prediction of disordered regions [44-46]. The structure of PSSM can be represented as a matrix of N rows and 20 columns. Each protein sequence can be transformed as follows:

$$M = \left\{ M_{\alpha,\beta}, \alpha = 1,\mathrm{K}\ , N, \quad \beta = 1,\mathrm{K}\ , 20 \right\} \tag{5}$$

where N denotes the length of a given plant protein sequence and column 20 represents the number of 20 amino acid. For each query sequence, the value $M_{\alpha,\beta}$, which could be described as β-th amino acid, will be set up by PSSM at the position of α. Thus, $M_{\alpha,\beta}$ can be calculated as:

$$M_{\alpha,\beta} = \sum_{k=1}^{20} p(\alpha,k) \times q(\beta,k)\ , \tag{6}$$

Thus, the value of Dayhoff's mutation matrix between the β-th and k-th amino acids can be described as $q(\beta,k)$, and the occurrence frequency score of the k-th amino acid in the position of α with the probe can be represented by $p(\alpha,k)$. Hence, a high value means a strongly conservative position; otherwise, it will imply a weakly conservative position.

In this study, we employed the *Position-Specific Iterated BLAST* (*PSI-BLAST*) tool [47] to generate the PSSM for each protein sequence. we assigned the e-value to 0.001 and selected 3 iterations in the process. In addition, all other parameters were set to default values to obtain highly and widely homologous sequences.

3.3. Inverse Fast Fourier transform.

In the fields of computational science and engineering, the Fast Fourier Transform (FFT) [48] is one of the most important algorithms. It is an indispensable algorithm in the field of Digital Signal Processing. However, FFT algorithm is not suitable in many practical applications when the data are not uniformly sampled. For this reason, we adopted the inverse fast Fourier Transform (IFFT) [49] method to obtain the transient response in time domain.

In the FFT, the irregularities of the twiddle factors can be solved by the Sine and Cosine transform of the signal. The Cosine and Sine transformations of the input signal are added together to obtain the FFT of the two-dimensional signal. As shown in Equation (7) and Equation (8), the required Sine matrix and Cosine matrix for FFT and IFFT can be defined as:

$$C(u+1, x+1) = \cos((pi/4)^{*}(u * x)) \tag{7}$$

$$S(u+1, x+1) = \sin((pi/4)^{*}(u * x)) \tag{8}$$

Hence, by the rules, the 2-D FFT can be yield by adding the Sine and Cosine transform as shown in Equation (9):

$$F(u,v) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \left\{ \begin{matrix} F(u,v)(\cos(2\pi(ux+vy)/n) \\ +j\sin(2\pi(ux+vy)/\mathrm{N})) \end{matrix} \right\} \tag{9}$$

for $k = 0,1,2\mathrm{K}\ , N-1$.

So, the IFFT for 2-D image can be described as follows:

$$f(x,y) = (1/MN) \sum_{u=0}^{M-1} \sum_{V=0}^{N-1} \left\{ \begin{matrix} F(u,v)(\cos(2\pi(ux+vy)/N) \\ -j\sin(2\pi(ux+vy)/N)) \end{matrix} \right\} \tag{10}$$

for $n = 0,1,2\mathrm{K}\ , N-1$.

In our study, each protein sequence in the three plant datasets, will be converted into a 400-dimensional vector by means of an inverse fast Fourier transform.

3.4. Weighted sparse representation-based classifier.

Recently, with the improvement of linear representation methods (LRBM) and compressed sensing (CS) theory, sparse representation-based classification (SRC) [50, 51] algorithm has been proven to widely applied in signal processing, pattern recognition and computer vision. The SRC assumes that there is a training sample matrix $X \in R^{d \times n}$, which denotes *n* training set and *d*-dimensional feature vectors, and it also assumes that there are sufficient training samples belonging to the kth class and set up $X_k = [l_{k1}, \mathrm{L}\ l_{kn_k}]$, where n_k denotes the sample number of *kth* class and l_i represents the label of *ith* sample. Thus, the sample matrix *X* could be defined as $X = [X_1 \mathrm{K}\ X_K]$. The SRC algorithm can described the test sample $y \in R^d$ with the linear combination of *k*th-class training samples as:

$$y = \alpha_{k,1} l_{k,1} + \alpha_{k,2} l_{k,2} + \mathrm{L} + \alpha_{k,nk} l_{k,nk} \tag{11}$$

when the whole training set representation are taking into account, it can be further symbolized as follows:

$$y = X_{\alpha 0} \tag{12}$$

where $\alpha_0 = [0, \mathrm{L}, 0, \alpha_{k.1}, \alpha_{k,2} \mathrm{L}\ \alpha_{k,nk}, 0, \mathrm{L}, 0]^T$. It is well known that the nonzero entries in α_0 are only relevant to the *kth* class, so if the samples size is too large, α_0 will become sparse.

For SRC algorithm, the key of it is to search the α vector that formula (12) can satisfy and can minimize the l_0-norm of itself. It can be represented as:

$$\hat{\alpha}_0 = \arg\min \|\alpha\|_0 \quad \text{subject to} \quad y = X_\alpha \tag{13}$$

Since problem (13) is a NP-hard problem and it is difficult to be solved accurately. According to the CS theory, if the α is sparse enough, we can solve the related convex l_1-minimization problem instead of dealing with the solution of l_0-minimization problem directly:

$$\hat{\alpha}_1 = \arg\min \|\alpha\|_1 \quad \text{subject to} \quad y = X_\alpha \tag{14}$$

To deal with the occlusion, the Eq (14) can be extended to the stable l_1-minimization problem:

$$\hat{\alpha}_1 = \arg\min \|\alpha\|_1 \quad \text{subject to} \quad \|y - X_\alpha\| \le \varepsilon \tag{15}$$

where $\varepsilon > 0$ represents the tolerance for reconstruction error. We can solve the Eq. (15) by using the standard linear programming methods.

After achieving the sparsest solution $\hat{\alpha}_1$, SRC can assign the test sample *y* to class *k* via the following rule:

$$\min_k r_k(y) = \left\| y - X \hat{\alpha}_1^k \right\|, \quad k = 1\mathrm{K}\ K \tag{16}$$

where $X \hat{\alpha}_1^k$ is the reconstruction which is built by training samples of class k and the class number of the whole samples can be defined as *K*. Then the SRC set a test sample as a sparse combination of training sample. Finally, we assigned it to the class which can minimizes the residual between itself and $X \hat{\alpha}_1^k$.

However, some studies [52-54] have reported that in some cases, locality structure of data is more important than sparsity. Moreover, the traditional SRC could not be guaranteed to be local.

To solve this problem, Lu *et al.* [55] developed a novel variant of SRC called weighted sparse representation-based classifier (WSRC). The main improvement of this method is that it combines the locality structure of data with sparse representation. Through mapping the training data into a higher-dimensional kernel include feature space, it can yield a better performance of classification. Gaussian kernel-based distance was used in WSRC to calculate the weights:

$$d_G = (S_1, S_2) = e^{-\|s_1 - s_2\|^2 / 2\sigma^2} \tag{17}$$

where $s_1, s_2 \in R^d$ denotes two samples; σ is the Gaussian kernel width. In this way, WSRC can preserve the locality structure of data and it can address the following questions:

$$\hat{\alpha}_1 = \arg\min \|W\alpha\|_1 \quad \text{subject to} \quad y = X\alpha \tag{18}$$

and specifically,

$$diag(W) = \left[d_G(y, x_1^1), \mathrm{K}, d_G(y, x_{n_k}^k) \right]^T \tag{19}$$

where n_k is the sample number of training set in class k and W represents a block diagonal matrix about locality adaptor. Dealing with occlusion, we would finally solve the following stable l_1-minimization problem:

$$\hat{\alpha}_1 = \arg\min \|W\alpha\|_1 \quad \text{subject to} \quad \|y - X\alpha\| \le \varepsilon \tag{20}$$

where $\varepsilon > 0$ denotes the tolerance value.
To summarize, the WSRC algorithm can be stated as follows:

Algorithm. Weighted Sparse Representation-based Classifier (WSRC)
1. Input: the matrix of training samples $X \in R^{d \times n}$ and a test sample $y \in R^d$.
2. Normalize the columns of X to have unit l_2-norm.
3. Calculate the Gaussian distances between y and each sample in X and employ them to adjust the training samples matrix X to X'.
4. Solving the stable l_1-minimization problem defined in Eq. (19).
5. Compute the residuals $r_k(y) = \left\| y - X \hat{\alpha}_1^k \right\| (k = 1, 2, \mathrm{K}, K)$.
6. Output: the prediction label of y as $identify(y) = \arg\min(r_k(y))$.

3.5. Support Vector Machine.

There are various methodologies for machine learning models to predict PPIs and support vector machine is one of the most popular classifiers. In 1995, SVM was first developed by Cortes and Vapnik *et al*. [56] and it is a generalized linear model usually used for classification and regression tasks. The ideal of SVM algorithm is to find the optimal hyperplane that maximally separates training data from the two classes. Hence, we can convert it to a convex quadratic programming problem. The formal definition of SVM can be expressed as:

$$\begin{gathered} Minimize \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} S_i \\ s.\ t: y_i(w^T x_i + b) \ge 1 - S_i, \quad \forall i \in \{1, \mathrm{K}, n\} \\ w, b, S_i \ge 0 \end{gathered} \tag{21}$$

where w represents the normal vector which defined the hyperplane and the classifier parameter can be defined as C; n denotes the number of vectors in the training dataset; x_i are the training vectors with m features; S_i are the slack variables; y_i is either 1 or -1 and it is the classification of each x_i; and b denotes the coefficient which determines the axis intercepts.

The maximization of the margin can be expressed as the first part of the objective function in equation (21). The $1/\|w\|^2$ represents the margin, which is determined by the distance between the nearest vectors and hyperplane. Thus, the maximization of $\|w\|^2$ is the minimization of $1/\|w\|^2$. The goal of objective function is to maximize the margin. Because as the margin increases, so too will the variability between the classes, which ensure a cleaner separation. However, if the margins are increased, the probability of misclassification will also be increased. The rate of misclassification is estimated by the slack variable S_i, which is set it to be 0 for well-classified vectors, between 0 and 1 for vectors located in the separation region, and above 1 for misclassified vectors. The training examples which defined the separating hyperplane are called support vectors. The parameter *C* denotes the weight vector: if the value of *C* is too high, it will lead to an increase of the penalties for misclassification and so that the area of margin will be reduced. The flow chart of our method is shown as Figure 4.

Figure 4. Flow chart of the proposed method

Conclusions

In this study, we present an effective and accurate computational method that utilize the information of amino acid sequence for predicting PPIs in plants. This method is based on a weighted sparse representation-based classifier combining with inverse fast Fourier transform and a position-specific-scoring-matrix. The main point of this approach is to employ the unique of WSRC method including better generalization, simply and considering the sparsity and continuity of plants protein sequence data. The whole prediction model is composed of the following steps. Firstly, all the plant protein sequences were converted as the PSSM so that the evolutionary information from each sequence can be obtained. Secondly, we employed the inverse fast Fourier transform to extract feature vector from PSSM. Finally, weighted sparse representation-based classifier would be used as machine learning classifier. The proposed approach performs significantly well on three plants PPIs datasets, including *Maize*, *Rice* and *Arabidopsis*. In order to prove the efficient and reliability efficient of the proposed model, we also compare it prediction performance with the state-of-the-art SVM model. All of these experiments results indicates that our method can improve the accuracy of the PPIs prediction in plants. In conclusion, the proposed method is a reliable, efficient and powerful prediction model for future proteomics research. To be the best of our knowledge, this is the first time to use computational methods to predict PPIs in plants.

Data availability

The source codes and datasets explored in this work are available at:

https://github.com/jie-pan111/protein_sequence.

References

1 Chen, Y. & Weckwerth, W. Mass spectrometry untangles plant membrane protein signaling networks. *Trends in plant science*. https://doi.org/10.1016/j.tplants.2020.03.013 (2020).

2 Matiolli, Cleverson Carlos, and Maeli Melotto. A comprehensive Arabidopsis yeast two-hybrid library for protein-protein interaction studies: a resource to the plant research community. *Molecular plant-microbe interactions* **31** 899-902. https://doi.org/10.1094/MPMI-02-18-0047-A (2018)

3 Janik, K. & Schlink, K. Unravelling the function of a bacterial effector from a non-cultivable plant pathogen using a yeast two-hybrid screen. *Journal of visualized experiments: JoVE*. https://dx.doi.org/10.3791%2F55150 (2017).

4 Singh, Raksha, et al. Rice mitogen-activated protein kinase interactome analysis using the yeast two-hybrid system. *Plant physiology* **160**.1 477-487. https://dx.doi.org/10.2307/23274708 (2012).

5 Velásquez-Zapata, V., Elmore, J. M., Banerjee, S., Dorman, K. S. & Wise, R. P. Y2H-SCORES: A statistical framework to infer protein-protein interactions from next-generation yeast-two-hybrid sequence data. *bioRxiv*. https://doi.org/10.1101/2020.09.08.288365 (2020).

6 Di Silvestre, D., Bergamaschi, A., Bellini, E. & Mauri, P. Large scale proteomic data and network-based systems biology approaches to explore the plant world. *Proteomes* **6**, 27. https://doi.org/10.3390/proteomes6020027 (2018).

7 Waese, J. et al. ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *The Plant Cell* **29**, 1806-1821. https://doi.org/10.1105/tpc.17.00073 (2017).

8 Ding, Z. & Kihara, D. Computational identification of protein-protein interactions in model plant proteomes. *Scientific reports* **9**, 1-13. https://doi.org/10.1038/s41598-019-45072-8 (2019).

9 Mahood, E. H., Kruse, L. H. & Moghe, G. D. Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences* **8**, e11376. https://doi.org/10.1002/aps3.11376 (2020).

10 Sahu, S. S., Weirick, T. & Kaundal, R. Predicting genome-scale Arabidopsis-Pseudomonas syringae interactome using domain and interolog-based approaches. *BMC bioinformatics*, **15**, 1-8 BioMed Central. https://doi.org/10.1186/1471-2105-15-S11-S13 (2014).

11 Hartmann, Julia, et al. The effective design of sampling campaigns for emerging chemical and microbial contaminants in drinking water and its resources based on literature mining. *Science of the Total Environment* **742** 140546. https://doi.org/10.1016/j.scitotenv.2020.140546 (2020).

12 An, D., Cao, H. X., Li, C., Humbeck, K. & Wang, W. Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes. *Genes* **9**, 43. https://doi.org/10.3390/genes9010043 (2018).

13 Chou, K.-C. & Shen, H.-B. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PloS one* **5**, e11335. https://doi.org/10.1371/journal.pone.0011335 (2010).

14 Nelson, C. J. & Millar, A. H. Protein turnover in plant biology. *Nature plants* **1**, 1-7. https://doi.org/10.1038/nplants.2015.17 (2015).

15 Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research* **40**, D1202-D1210. https://doi.org/10.1093/nar/gkr1090 (2012).

16 Gu, H., Zhu, P., Jiao, Y., Meng, Y. & Chen, M. PRIN: a predicted rice interactome network. *BMC bioinformatics* **12**, 1-13. https://doi.org/10.1186/1471-2105-12-161 (2011).

17 Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic acids research* **40**, D857-D861. https://doi.org/10.1093/nar/gkr930 (2012).

18 Huang, Y.-A., You, Z.-H., Chen, X. & Yan, G.-Y. Improved protein-protein interactions prediction via weighted sparse representation model combining continuous wavelet descriptor and PseAA composition. *BMC systems biology* **10**, 485-494. https://doi.org/10.1186/s12918-016-0360-6 (2016).

19 Wei, L. et al. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artificial Intelligence in Medicine* **83**, 67-74. https://doi.org/10.1016/j.artmed.2017.03.001 (2017).

20 Li, J. et al. A conserved NAD+ binding pocket that regulates protein-protein interactions during aging. *Science* **355**, 1312-1317. https://doi.org/10.1126/science.aad8242 (2017).

21 Hashemifar, S., Neyshabur, B., Khan, A. A. & Xu, J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* **34**, i802-i810. https://doi.org/10.1093/bioinformatics/bty573 (2018).

22 Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic acids research* **45**, W291-W299. https://doi.org/10.1093/nar/gkx366 (2017).
23 Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660-668. https://doi.org/10.1093/bioinformatics/btx624 (2018).
24 Wei, L., Zhou, C., Chen, H., Song, J. & Su, R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34**, 4007-4016. https://doi.org/10.1093/bioinformatics/bty451 (2018).
25 Pelay-Gimeno, M., Glas, A., Koch, O. & Grossmann, T. N. Structure-based design of inhibitors of protein–protein interactions: mimicking peptide binding epitopes. *Angewandte Chemie International Edition* **54**, 8896-8927. https://doi.org/10.1002/anie.201412070 (2015).
26 Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.-C. iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. *Molecules* **21**, 95. https://doi.org/10.3390/molecules21010095 (2016).
27 Wei, Z.-S., Han, K., Yang, J.-Y., Shen, H.-B. & Yu, D.-J. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing* **193**, 201-212. https://doi.org/10.1016/j.neucom.2016.02.022 (2016).
28 Charoenkwan, P. et al. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. *Scientific reports* **11**, 1-13. https://doi.org/10.1038/s41598-021-82513-9 (2021).
29 Sun, T., Zhou, B., Lai, L. & Pei, J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC bioinformatics* **18**, 1-8. https://doi.org/10.1186/s12859-017-1700-2 (2017).
30 Shen, J. et al. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences* **104**, 4337-4341. https://doi.org/10.1073/pnas.0607879104 (2007).
31 Wang, Y.-B. et al. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Molecular BioSystems* **13**, 1336-1344. https://doi.org/10.1039/C7MB00188F (2017).
32 You, Zhu-Hong, et al. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC bioinformatics* **14**, BioMed Central. https://doi.org/10.1186/1471-2105-14-S8-S10 (2013)
33 Skoblov, M. et al. Protein partners of KCTD proteins provide insights about their functional roles in cell differentiation and vertebrate development. *Bioessays* **35**, 586-596. https://doi.org/10.1002/bies.201300002 (2013).
34 Xia, J.-F., Zhao, X.-M. & Huang, D.-S. Predicting protein–protein interactions from protein sequences using meta predictor. *Amino acids* **39**, 1595-1599. https://doi.org/10.1007/s00726-010-0588-1 (2010).
35 Basith, S., Manavalan, B., Shin, T. H. & Lee, G. iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Computational and structural biotechnology journal* **16**, 412-420. https://doi.org/10.1016/j.csbj.2018.10.007 (2018).
36 Manavalan, B., Subramaniyam, S., Shin, T. H., Kim, M. O. & Lee, G. Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. *Journal of proteome research* **17**, 2715-2726. https://doi.org/10.1021/acs.jproteome.8b00148 (2018).
37 Wei, L., Luan, S., Nagai, L. A. E., Su, R. & Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **35**, 1326-1333. https://doi.org/10.1093/bioinformatics/bty824 (2019).
38 Tian, T. et al. agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids research* **45**, W122-W129. https://doi.org/10.1093/nar/gkx382 (2017).
39 Zhu, G. et al. PPIM: a protein-protein interaction database for maize. *Plant physiology* **170**, 618-626. https://doi.org/10.1109/COASE.2017.8256085 (2016).
40 Kerrien, S. et al. The IntAct molecular interaction database in 2012. *Nucleic acids research* **40**, D841-D846. https://doi.org/10.1093/nar/gkr1088 (2012).
41 Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic acids research* **47**, D529-D541. https://doi.org/10.1093/nar/gky1079 (2019).
42 Yang, S., Li, H., He, H., Zhou, Y. & Zhang, Z. Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods. *Briefings in bioinformatics* **20**, 274-287. https://doi.org/10.1093/bib/bbx123 (2019).
43 Gribskov, M., McLachlan, A. D. & Eisenberg, D. Profile analysis: detection of distantly related

proteins. *Proceedings of the National Academy of Sciences* **84**, 4355-4358. https://doi.org/10.1073/pnas.84.13.4355 (1987).
44 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **292**, 195-202. https://doi.org/10.1006/jmbi.1999.3091 (1999).
45 Chen, X.-w. & Jeong, J. C. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **25**, 585-591. https://doi.org/10.1093/bioinformatics/btp039 (2009).
46 Jones, D. T. & Ward, J. J. Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Bioinformatics* **53**, 573-578. https://doi.org/10.1002/prot.10528 (2003).
47 Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in biochemical sciences* **23**, 444-447. https://doi.org/10.1016/S0968-0004(98)01298-5 (1998).
48 Nussbaumer, Henri J. The fast Fourier transform. *Fast Fourier Transform and Convolution Algorithms*. Springer, Berlin, Heidelberg, 80-111. https://doi.org/10.1109/PROC.1967.5957 (1981).
49 T. Anitha and S. Ramachandran. Novel algorithms for 2-D FFT and its inverse for image compression. *in 2013 International Conference on Signal Processing, Image Processing & Pattern Recognition* 62-65: IEEE. https://doi.org/10.1109/ICSIPR.2013.6497959 (2013).
50 Liao, B. et al. Learning a weighted meta-sample based parameter free sparse representation classification for microarray data. *PLoS One* **9**, e104314. https://doi.org/10.1371/journal.pone.0104314 (2014).
51 Wright, J., Ganesh, A., Zhou, Z., Wagner, A. & Ma, Y. Demo. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, **31**, 2. https://doi.org/10.1109/TPAMI.2008.79 (2008).
52 Wang, J. et al. Locality-constrained linear coding for image classification. *in 2010 IEEE computer society conference on computer vision and pattern recognitio* 3360-3367: IEEE. https://doi.org/10.1109/CVPR.2010.5540018 (2010).
53 Sharma, A. & Paliwal, K. K. A deterministic approach to regularized linear discriminant analysis. *Neurocomputing* **151**, 207-214. https://doi.org/10.1016/j.neucom.2014.09.051 (2015).
54 Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science* **290**, 2323-2326. https://doi.org/10.1126/science.290.5500.2323 (2000).
55 Lu, C.-Y., Min, H., Gui, J., Zhu, L. & Lei, Y.-K. Face recognition via weighted sparse representation. *Journal of Visual Communication and Image Representation* **24**, 111-116. https://doi.org/10.1016/j.jvcir.2012.05.003 (2013).
56 Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273-297. https://doi.org/10.1007/BF00994018 (1995).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61722212 and Grant 61873212.

Author Contributions

J.P., Z-H.Y., and L-P. L. conceived the algorithm, carried out the analyses, prepared the data sets, carried out experiments, and wrote the manuscript. J.P., C-Q.Y., X-K.Z., designed, performed and analyzed experiments and wrote the manuscript. All authors read and approved the final manuscript.

Competing Interests: The author declares that there is no conflict of interest regarding the publication of this paper.

Additional Information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Figures

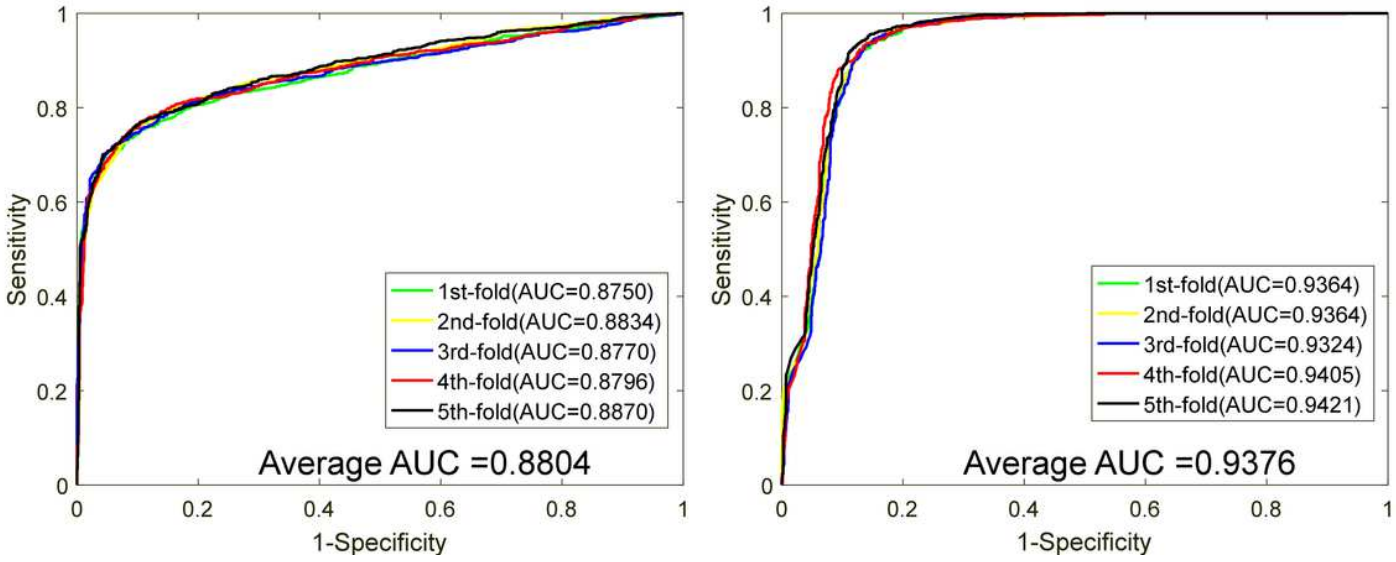


Figure 1

Comparison of the ROC curves obtained by WSRC and SVM-based method on Maize dataset (5-fold cross validation). (a) shows the ROC curves performed by SVM method on Mazie PPIs dataset. (b) shows the ROC curves performed by WSRC method on Maize dataset.

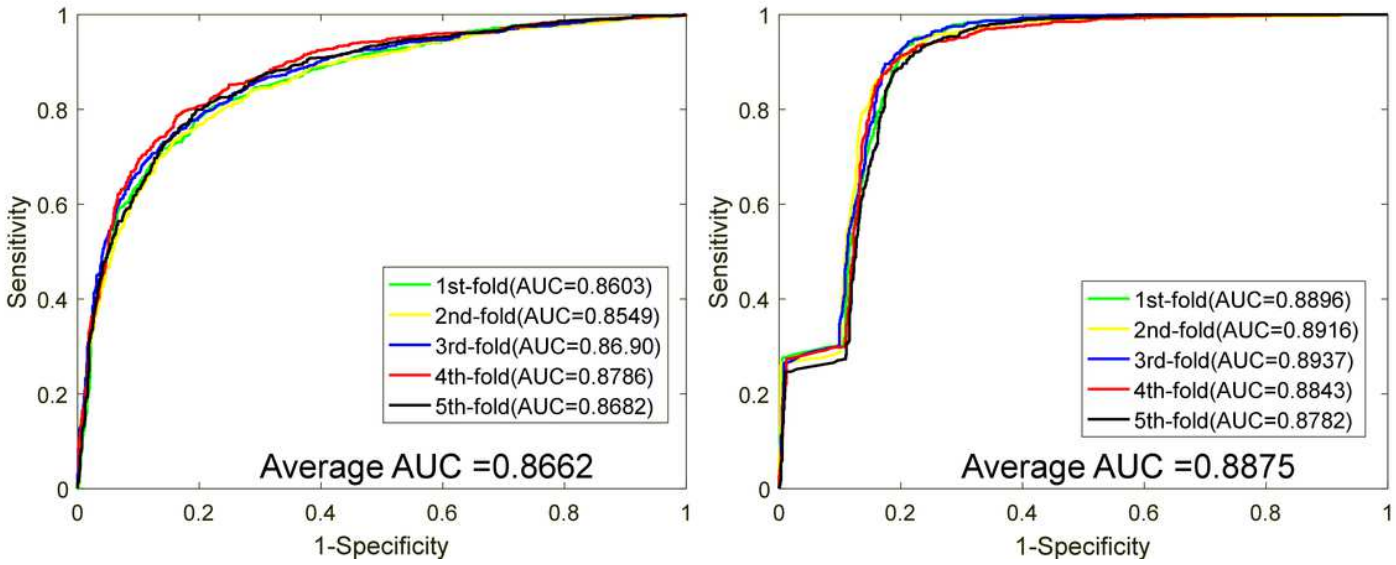


Figure 2

Comparison of the ROC curves obtained by WSRC and SVM-based method on Rice dataset (5-fold cross validation). (a) shows the ROC curves performed by SVM method on Rice PPIs dataset. (b) shows the ROC curves performed by WSRC method on Rice dataset.

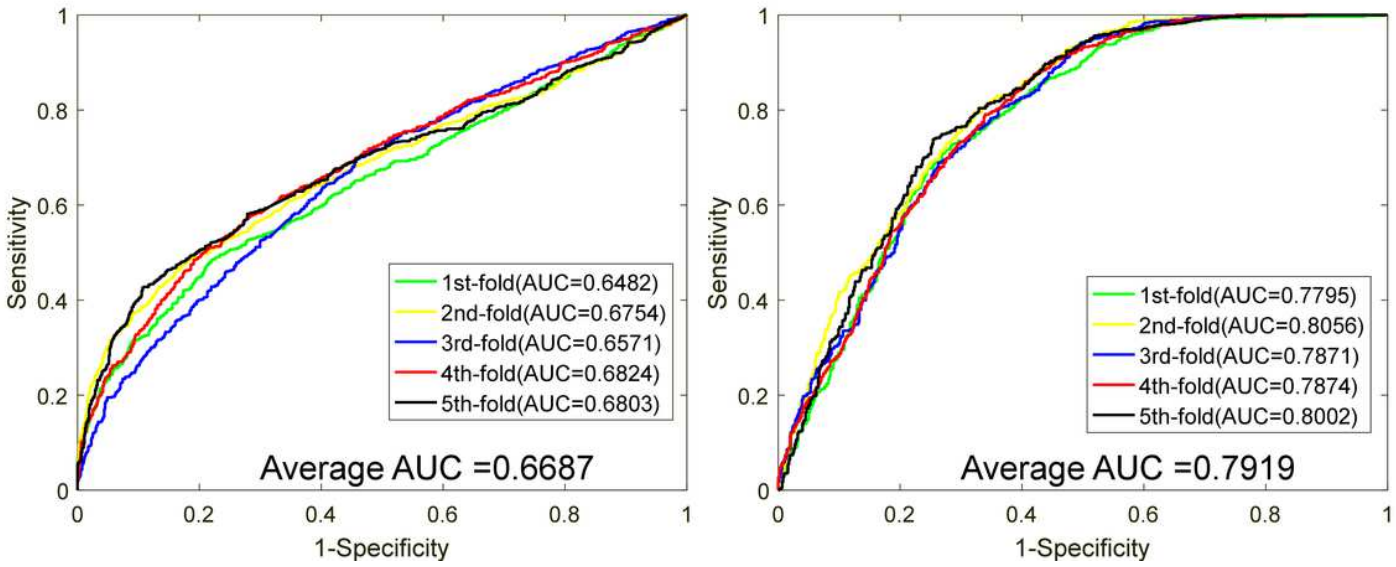


Figure 3

Comparison of the ROC curves obtained by WSRC and SVM-based method on Arabidopsis dataset (5-fold cross validation). (a) shows the ROC curves performed by SVM method on Arabidopsis PPIs dataset. (b) shows the ROC curves performed by WSRC method on Arabidopsis dataset.

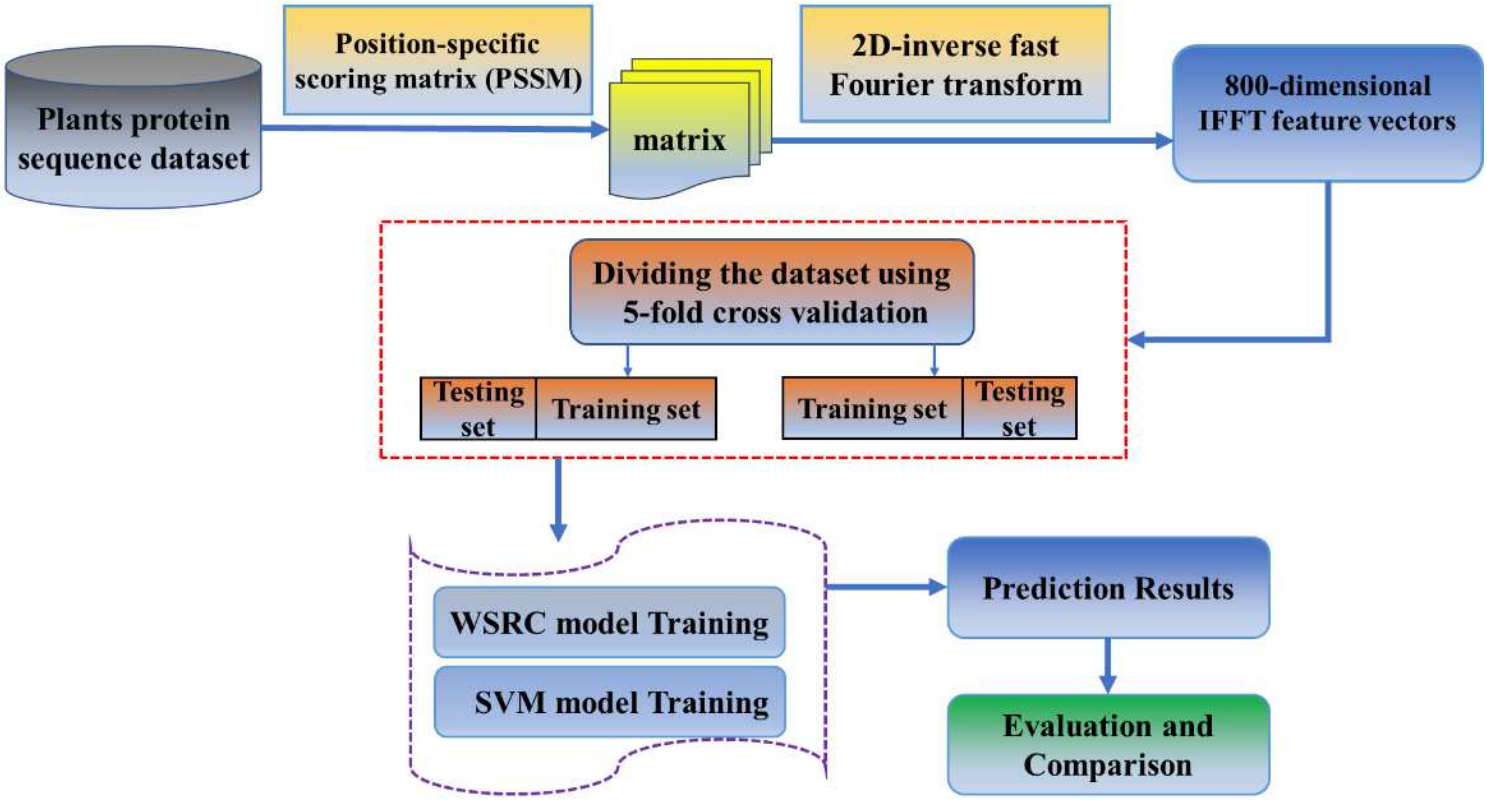


Figure 4

Flow chart of the proposed method