Hindawi
Scientific Programming
Volume 2021, Article ID 1607946, 11 pages
https://doi.org/10.1155/2021/1607946

Research Article

FWHT-RF: A Novel Computational Approach to Predict Plant Protein-Protein Interactions via an Ensemble Learning Method

Jie Pan, Li-Ping Li, Chang-Qing Yu, Zhu-Hong You, Zhong-Hao Ren, and Jing-Yu Tang

School of Information Engineering, Xijing University, Xi'an 710123, China

Correspondence should be addressed to Li-Ping Li; cs2bioinformatics@gmail.com

Received 26 May 2021; Accepted 14 July 2021; Published 22 July 2021

Academic Editor: Pengwei Wang

Protein-protein interactions (PPIs) in plants are crucial for understanding biological processes. Although high-throughput techniques produced valuable information to identify PPIs in plants, they are usually expensive, inefficient, and extremely time-consuming. Hence, there is an urgent need to develop novel computational methods to predict PPIs in plants. In this article, we proposed a novel approach to predict PPIs in plants only using the information of protein sequences. Specifically, plants' protein sequences are first converted as position-specific scoring matrix (PSSM); then, the fast Walsh–Hadamard transform (FWHT) algorithm is used to extract feature vectors from PSSM to obtain evolutionary information of plant proteins. Lastly, the rotation forest (RF) classifier is trained for prediction and produced a series of evaluation results. In this work, we named this approach FWHT-RF because FWHT and RF are used for feature extraction and classification, respectively. When applying FWHT-RF on three plants' PPI datasets *Maize*, *Rice*, and *Arabidopsis thaliana* (*Arabidopsis*), the average accuracies of FWHT-RF using 5-fold cross validation were achieved as high as 95.20%, 94.42%, and 83.85%, respectively. To further evaluate the predictive power of FWHT-RF, we compared it with the state-of-art support vector machine (SVM) and *K*-nearest neighbor (KNN) classifier in different aspects. The experimental results demonstrated that FWHT-RF can be a useful supplementary method to predict potential PPIs in plants.

1. Introduction

Protein-protein interactions (PPIs) in plants underlie many biological processes, including cellular organization, signal transduction [1], metabolic cycles [2], and plant defense [3]. Thus, detecting and characterizing the protein interactions are critically important for understanding the relevant molecular mechanisms inside the plant cells. With the scientific and technological advances, a multitude of experimental approaches had been developed to identify PPIs in plants, such as yeast two-hybrid (Y2H) [4], bimolecular fluorescence complementation (BiFC) [5], tandem affinity purification (TAP) [6], and some other high-throughput DNA sequencing technology for PPIs detection. Therefore, a huge and ever-increasing of experimental data about plants PPIs has been accumulated. However, these approaches have some inevitable shortcomings; they are particularly expensive, time-consuming, and always present problems with high false-negative rates. Besides, it is also difficult to apply large-scale experiments on plants due to the complexity of interactions in plant cells. As a result of these shortcomings, developing accurate computational methods to predict PPIs would be of great value to plant biologists.

In recent years, many computational methods and ensemble learning algorithms have been established to offer complementary and supporting information for previous experimental approaches [7–9]. These methods can be broadly classified into three categories: protein structure-based method, docking-based method, and sequence-based method. Generally, the first two methods usually need structural details. However, many proteins do not have information about the prior knowledge, such as 3D structural information and protein homology. In addition, with the rapid advance in high-throughput sequencing technology, more and more plant protein sequence data are

available, which lead to a great interest in sequence-based methods for PPIs prediction.

To date, many sequence-based approaches have been presented for predicting PPIs and many ensemble-learning algorithms have been proposed for classification [10–12]. For example, Yi et al. [13] proposed a method called RPI-SAN, which adopts the deep-learning stacked auto-encoder network to mine the features from RNA and protein sequences and then employs the rotation forest classifier to predict ncRNA binding proteins. Hashemifar et al. [14] developed a novel deep learning framework named DPPI. DPPI combined random projection and data augmentation with a deep, Siamese-like convolutional neural network to predict PPIs. Zhang et al. [15] presented the EnsDNN (Ensemble Deep Neural) method, which first employed the local descriptor, covariance descriptor, and multiscale continuous and discontinuous local descriptor together to explore the interactions between proteins. Then, it trained the deep neural networks (DNNs) based on different configurations of each descriptor. Finally, they adopted a two-hidden layers neural network to integrate these DNNs to predict potential PPIs. Wei et al. [16] combined the novel negative samples, features, and an ensemble classifier to predict PPIs. They report two types of novel feature extraction methods. One is the based on physicochemical properties of proteins, and the other is based on the secondary structure information. Sun et al. [17] applied a deep-learning algorithm, stacked autoencoder, to identify PPIs from the protein sequence. Kulmanov et al. [18] developed a method called DeepGO, which combined a deep ontology-aware classifier with amino acid sequence information to detect protein functions and interactions. Despite these advances, there is still room for improvement in the prediction performance of PPIs' model [19].

In this article, we present a novel sequence-based computational approach, namely, FWHT-RF, to predict potential protein-protein interactions in plants. More specifically, we first transformed the plants protein sequences as position-specific scoring matrix (PSSM). Then, in order to fully characterize the evolutionary information of protein pairs, we performed the fast Walsh–Hadamard transform (FWHT) on the PSSM to extract features' vectors. Although FWHT plays an essential role in image analysis and pattern recognition, but as we know, it is first time to be applied in plant biology for the purpose of PPIs' prediction. Lastly, a powerful classification model, rotation forest (RF), was used to train the models. The major contributions of FWHT-RF are as follows: (1) FWHT-RF did not depend on unique subspaces in the studied proteomic space or known PPIs' samples because it extracts features directly from PSSM of the plant protein sequence. (2) Since these characteristics are linked to the evolutionary past of plant proteins, they have more power to detect PPIs than many other approaches. (3) The basic features from PSSM for each plant proteins were extracted using a novel statistical selection feature mechanism and converted into a 400-dimensional feature vector. As a result, the feature vectors of these two proteins are integrated to create an 800-dimensional feature vector for each protein pair. (4) Finally, this work suggested to use the RF classifier for training these features, which can improve the accuracy of PPIs prediction. This model has been well investigated in three plants' datasets (*Maize*, *Rice,* and *Arabidopsis thaliana* (*Arabidopsis*)) and yields a high prediction accuracy of 95.20%, 94.42%, and 83.85%, respectively. To further evaluate the predictive performance of FWHT-RF, we compared FWHT-RF with the state-of-art support vector machine (SVM) and k-nearest neighbor (KNN) classifier. The experimental results indicated that FWHT-RF can be a complement tool to large-scale prediction of PPIs in plants.

2. Materials and Methods

2.1. Benchmark Datasets Collection. Although many experiments and databases have been developed to identify and store the PPIs data in plants [20, 21], however, false positive interactions are typical in these data. These false positive data may have a negative impact for the computational methods. Therefore, the construction of benchmark datasets to improve the accuracy of plant PPIs prediction is necessary. In this paper, we evaluate the FWHT-RF approach through three plants' benchmark datasets, including *Maize*, *Rice,* and *Arabidopsis thaliana* (*Arabidopsis*).

As we all know, maize is one of the most important cereal crops in the world and a model plant for genomic studies of PPIs. The *Maize* dataset was gathered from the *Protein-Protein Interaction Database for Maize* (*PPIM*) [22] and *agriGO* [23]. We obtained 14,800 nonredundant maize protein pairs which built the positive dataset. In order to construct the negative dataset, we selected 14,800 additional maize protein pairs of different subcellular localizations. Consequently, the whole *Maize* dataset consists of 29,600 protein pairs.

To further demonstrate the feasibility of the proposed method, two different types of plant PPIs' datasets were also adopted in this study. The first one is *Rice*, which was gathered from the PRIN [24] database, which consisted of 9600 nonredundant rice protein interaction pairs (4800 interacting pairs and 4800 noninteracting pairs). The second is the popular model plant *Arabidopsis*. We collected *Arabidopsis* PPIs from public PPI databases IntAct [25], BioGRID [26], and TAIR [27]. After the removal of redundant sequences, we obtained 28,110 interactions from 7437 *Arabidopsis* proteins. The negative protein pairs are generated by randomly pairing the proteins without evidence of interactions. In this way, the whole *Arabidopsis* dataset is constructed by 56,220 protein pairs.

2.2. Representation of Target Proteins. The position-specific scoring matrix (PSSM) [28] was firstly proposed for testing the distantly related proteins. In recent years, PSSM has been widely used for mining the evolutionary information of protein sequences [29]. PSSM is a $P \times 20$ matrix. The

number of amino acids in the proteins is represented by P, and the naive amino acids are represented by 20 columns. Suppose that $L = \{\varphi\_(i, j)\colon i = 1, \ldots, P,\ j = 1, \ldots, 20\}$, and the following is a summary of each matrix:

$$L = \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} & \cdots & \varphi_{1,j} & \cdots & \varphi_{1,20} \\ \varphi_{2,1} & \varphi_{2,2} & \cdots & \varphi_{2,j} & \cdots & \varphi_{2,20} \\ \vdots & \vdots & \vdots & \vdots & & \\ \varphi_{i,1} & \varphi_{i,2} & \cdots & \varphi_{i,j} & \cdots & \varphi_{i,20} \\ \vdots & \vdots & \vdots & \vdots & & \\ \varphi_{p,1} & \varphi_{p,2} & \cdots & \varphi_{p,j} & \cdots & \varphi_{p,20} \end{bmatrix}, \quad (1)$$

where $\varphi_{1,1}$ in the i row of PSSM indicates the probability of the ith residue being mutated into jth native amino acid.

In this study, we adopted the Position-Specific Iterated BLAST (PSI-BLAST) [30] tool to generate the PSSM for the purpose of extracting evolutionary information. To achieve broad and high homologous sequences, the expectation value (e value) was set to 0.001, the number of iterations was set to 3, and other parameters were maintained as the default values.

2.3. 2D Fast Walsh–Hadamard Transform. Walsh–Hadamard transform (WHT) [31] is employed in many applications such as image analysis and signal processing. It is recognized as a generalized type of Fourier transforms (FT) and has three popular orderings: (1) *Natural Ordering* (*Hadamard Ordering*), (2) *Dyadic Ordering* (*Paley Ordering*), and (3) *Sequency Ordering* (*Walsh Ordering*) [32–34]. In this study, we will focus on the WHT of *Natural Ordering*. The WHT matrix consists only by ±1. Since no multiplication operation is required in the computation, the computational complexity is greatly reduced. In the encrypted domain, this algorithm can avoid quantization error and thus WHT can ensure perfect reconstruction of the encrypted image. Therefore, WHT is better and more effective than transformations such as DFT [35] or DCT [36].

Suppose that $Q(i, j)$ represented the input image with $a \times b$ size, where a and b were used to describe the same and the power of 2. The two-dimensional fast Walsh–Hadamard transform (FWHT) [37] of *Natural Ordering* (*Hadamard Ordering*) can be defined as follows:

$$\psi(p,k) = \frac{1}{a}\sum_{i=0}^{a-1}\sum_{j=0}^{b-1} H_\eta(p,i)Q(i,j)H_\eta(j,k), \quad p = 0,1,2,3,\ldots,a-3,a-2,a-1 \text{ and } n = 0,1,2,3,\ldots,b-3,b-2,b-1, \quad (2)$$

where $\mu = \log_2 p$ and H_η denotes the Hadamard matrix. H_η can be generated by the core matrix:

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad (3)$$

and the Kronecker product recursion is as follows:

$$H_\eta = H_1 \otimes H_{\eta-1} = \begin{bmatrix} H_{\eta-1} & H_{\eta-1} \\ H_{\eta-1} & -H_{\eta-1} \end{bmatrix}, \quad (4)$$

where $\otimes$ is the Kronecker product operator [38]. The 2D FWHT [39] is a separable transformation which can be further divided into two 1D transforms. When applying the 2D FWHT on the input image, $Q(i, j)$ is equivalent to applying 1D FWHT on all columns of the input image initially and then using 1D FWHT on all rows of achieved results. For 2D FWHT, the computational complex is $a \log a$. In this study, $Q(i, j)$ is the input signal matrix, and here is the $P \times 20$ PSSM matrix. By this way, the plant protein sequence can be represented by FWHT feature descriptors.

2.4. Ensemble Rotation Forest Classifier. Rotation forest (RF) was introduced by Rodriguez et al. [40], which is an ensemble learning algorithm based on an independently trained decision tree. The main advantage of RF is that it can balance diversity and accuracy at the same time. RF first randomly divided the samples into different subsets. Then, principal component analysis (PCA) [41] was used to transform the attribute subsets to increase the difference between the subsets. At last, the transformed subsets will be fed into the decision trees. The results of RF can be achieved via a voting method by these trees. The specific steps of RF are as follows.

Suppose that $\{q_i, p_i\}$ contains T samples, of which $q_i = (q_{i1}, q_{i2}, q_{i3}, \ldots, q_{iL})$ be an L-dimensional feature vector. Let Z represents the training sample set containing T training samples and forming a matrix of $T \times L$. Let U represents the feature set and M denotes the label set. Assume that the number of decision trees is S; then, the decision trees can be denoted as $D_1, D_2, D_3, \ldots, D_S$. The rotation forest algorithm is implemented as follows:

(1) Choose the suitable parameter M, which can randomly split divide U into M disjointed subsets, and the number of features contained in the feature subset is L/M.

(2) Let $U_{i,j}$ represent the jth feature subset and be used to train the classifier D_i. The sample subset $Z'_{i,j}$ is

constructed by a nonempty subset, which is randomly picked out from a certain proportion.

(3) Apply PCA on $Z'_{i,j}$ to order the coefficients, which is stored in matrix $\lambda_{i,j}$.

(4) The coefficients achieved from the matrix $\lambda_{i,j}$ are used to construct a sparse rotation matrix φ_i, which can be defined as follows:

$$\varphi_i = \begin{bmatrix} a_{i,1}^{(1)}, \ldots, a_{i,1}^{(S_1)} & 0 & \cdots & 0 \\ 0 & a_{i,1}^{(1)}, \ldots, a_{i,1}^{(S_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i,1}^{(1)}, \ldots, a_{i,1}^{(S_M)} \end{bmatrix}. \tag{5}$$

During the prediction process, a test sample g is given, which is generated by the classifier D_i of $R_{i,j}(Z\varphi_i^a)$ which is introduced to indicate that g belongs to class p_i. Then, the class of confidence is calculated via the average combination, and the formula can be expressed as follows:

$$V_j(g) = \frac{1}{S}\sum_{i=1}^{S} R_{i,j}(Z\varphi_i^a). \tag{6}$$

Then, assign the category with the largest $V_j(g)$ value to g. The overview of FWHT-RF workflow is presented in Figure 1.

3. Results and Discussion

3.1. Validation Measures. In this work, we employed multiple evaluation indicators to access the effectiveness of FWHT-RF, including accuracy (Acc.), sensitivity (Sen.), precision (Prec.), Matthews correlation coefficient (MCC), and area under the receiver operating characteristic curve (AUC). Correspondingly, the first four formulas can be represented as follows:

$$\begin{aligned} \text{Acc.} &= \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}, \\ \text{Sen.} &= \frac{\text{TP}}{\text{FN} + \text{TP}}, \\ \text{Prec.} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{MCC} &= \frac{\text{TN} \times \text{TP} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TN} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FN})}}, \end{aligned} \tag{7}$$

where TP (true positive) represents the number of true plant PPIs that are correctly identified (positive samples), FP (false positive) refers to the number of noninteraction plant protein pairs (negative samples), and TN (true negative) denotes the number of correct classification of positive samples, while FN (false negative) refers to the number of incorrect classification of negative samples.

To provide a more comprehensive assessment of the FWHT-RF method, the receiver operating characteristic (ROC) curves, which are suitable for accessing the performance of the proposed method, were computed. The area under the ROC curves (AUC) was also calculated to test the predictive ability of FWHT-RF. AUC denotes the probability that a positive sample is ahead of a negative one. The AUC value closer to 1.0 indicates the better predictive performance of the FWHT-RF method [42].

3.2. Assessment of Prediction Ability. In this article, we adopt 5-fold cross-validation technique to comparatively access the prediction performance of FWHT-RF in three plant datasets involving *Maize*, *Rice*, and *Arabidopsis*. By this way, we can prevent overfitting and test the stability of the proposed method. More specifically, each plant PPIs' dataset is randomly split into five subsets, one of them is used as a testing set in turn and the other four subsets are adopted as training sets. 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 improve the reliability of the results.

The 5-fold cross validation results of the proposed approach on the three plants datasets are listed in Tables 1–3. From Tables 1–3, we can observe that when applying the proposed method to the *Maize* dataset, we obtained the best prediction results of average accuracy, precision, sensitivity, and MCC as 95.20%, 97.29%, 92.99%, and 90.85% with corresponding standard deviations 0.38%, 0.26%, 0.62%, and 0.69%, respectively. When performing FWHT-RF on the *Rice* dataset, we yielded good results of average accuracy, precision, sensitivity, and MCC of 94.42%, 94.63%, 94.17%, and 89.46%, respectively. The standard deviations of these criteria values are 0.56%, 0.84%, 0.72%, and 0.99%, respectively. When performing FWHT-RF on the *Arabidopsis* dataset, the proposed approach obtained good results of average accuracy, precision, sensitivity, and MCC of 83.85%, 89.29%, 76.95%, and 72.66% and the standard deviations are 0.35%, 0.62%, 1.16%, and 0.52%, respectively. Figures 2–4 show the ROC curves for the proposed approach on *Maize*, *Rice,* and *Arabidopsis*. The average AUC values range from 90.55% to 97.50% (*Maize*: 97.50%, *Rice*: 96.90%, and *Arabidopsis*: 90.55%), demonstrating that FWHT-RF is fitting well for predicting PPIs in plants from amino acid sequences.

These good results collectively indicated that it is sufficient to predict PPIs in plants only using protein sequence information and that powerful prediction capability can be generated by combining the RF classifier with FWHT features' descriptors. The high accuracies and low standard deviations of these criterion values indicate that FWHT-RF is feasible and effective for predicting potential PPIs in plants.

3.3. Comparison of RF with SVM and KNN Classifiers. There are various methodologies for machine learning models to identify PPIs, and most of them are based on traditional classifiers. To further access the predictive performance of FWHT-RF, we compared it by using the same feature extraction approach with the state-of-art SVM and KNN classifier in the same three plants' datasets. The main idea of the SVM algorithm is to find the optimal hyperplane that maximally

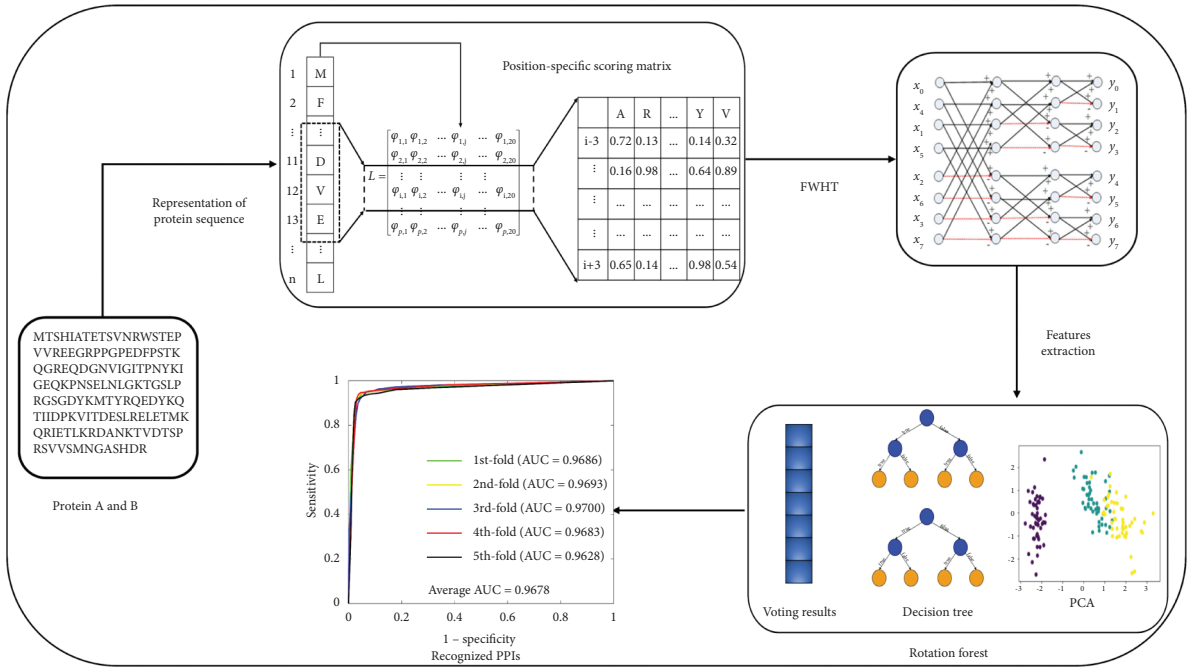

Figure 1: The workflow of FWHT-RF for predicting protein-protein interactions in plants.

Table 1: 5-fold cross-validation results obtained on the *Maize* dataset using FWHT-RF.

Testing set	Acc. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	94.68	92.06	97.24	89.91	97.06
2	95.06	92.97	96.93	90.60	97.40
3	95.56	93.49	97.65	91.50	97.95
4	95.59	93.60	97.27	91.55	97.62
5	95.12	92.82	97.37	90.70	97.45
Average	**95.20 ± 0.38**	**92.99 ± 0.62**	**97.29 ± 0.26**	**90.85 ± 0.69**	**97.50 ± 0.33**

Table 2: 5-fold cross-validation results obtained on the *Rice* dataset using FWHT-RF.

Testing set	Acc. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	94.90	94.19	95.75	90.31	96.42
2	93.70	92.98	94.44	88.19	96.89
3	94.17	94.45	93.65	89.01	96.80
4	94.27	94.32	94.12	89.20	96.94
5	95.05	94.91	95.21	90.59	97.46
Average	**94.42 ± 0.56**	**94.17 ± 0.72**	**94.63 ± 0.84**	**89.46 ± 0.99**	**96.90 ± 0.37**

Table 3: 5-fold cross-validation results obtained on the *Arabidopsis* dataset using FWHT-RF.

Testing set	Acc. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
1	84.00	77.27	89.37	72.88	90.73
2	83.48	75.96	89.71	72.13	90.21
3	84.04	77.77	88.88	72.97	90.91
4	83.48	75.54	90.01	72.09	90.00
5	84.25	78.22	88.46	73.24	90.88
Average	**83.85 ± 0.35**	**76.95 ± 1.16**	**89.29 ± 0.62**	**72.66 ± 0.52**	**90.55 ± 0.41**

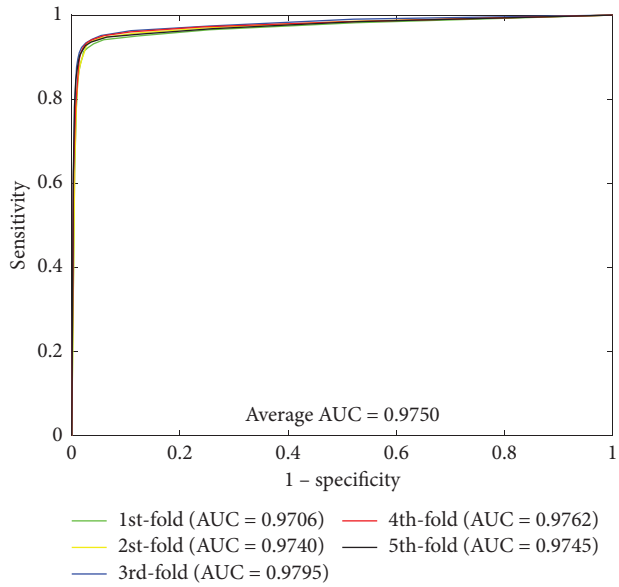

Figure 2: ROC curve for FWHT-RF on *Maize* PPIs dataset.

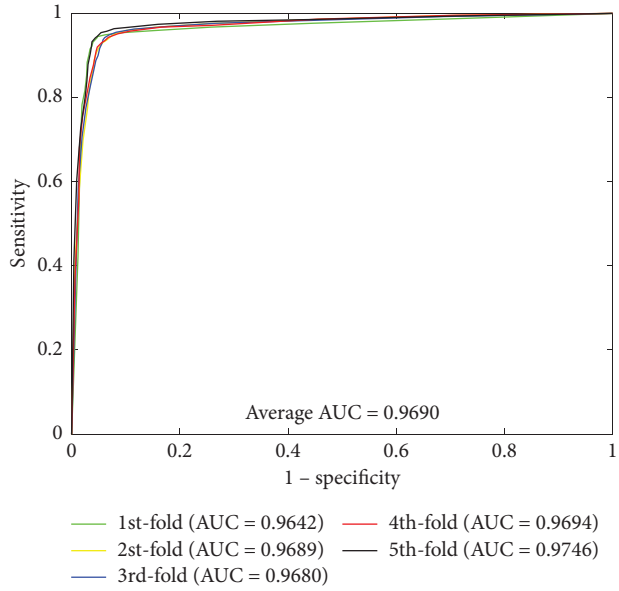

Figure 3: ROC curve for FWHT-RF on *Rice* PPIs dataset.

separates training data from the two classes, and it is effective for solving classification prediction problems. *K*-nearest neighbor is a supervised machine learning technique, and it can solve the classification task. The LIBSVM tool was selected in this paper to training the SVM model. At the same time, there are two parameters c and g that need to be optimized. In the experiment of the *Maize* and *Rice* dataset, we set $c = 5$, $g = 0.3$, $c = 7$, and $g = 0.4$, respectively. When applying the FWHT-RF on the *Arabidopsis* dataset, we set $c = 5$ and $g = 0.7$. The KNN model needs to choose the neighbor k and distance measuring function. In this paper, k is set to be 1 and the distance measuring function is selected as $L1$.

Figure 5 shows the experimental results of RF, SVM, and KNN models in three plants datasets *Maize*, *Rice,* and *Arabidopsis*. From Figures 5(a)–5(d), it can be concluded that the results of the RF classifier are significantly better than those of SVM and KNN classifiers. For example, the accuracy gaps between SVM and RF on the *Maize*, *Rice*, and *Arabidopsis* were 7.98%, 8.53%, and 3.26%, respectively. Similarly, the accuracy gaps between KNN and RF are 11.72%, 15.36%, and 10.40%, respectively. The ROC curves achieved by the SVM and KNN classifiers on the three plants datasets are shown in Figures 6–8. All the experimental results are listed in Table 4.

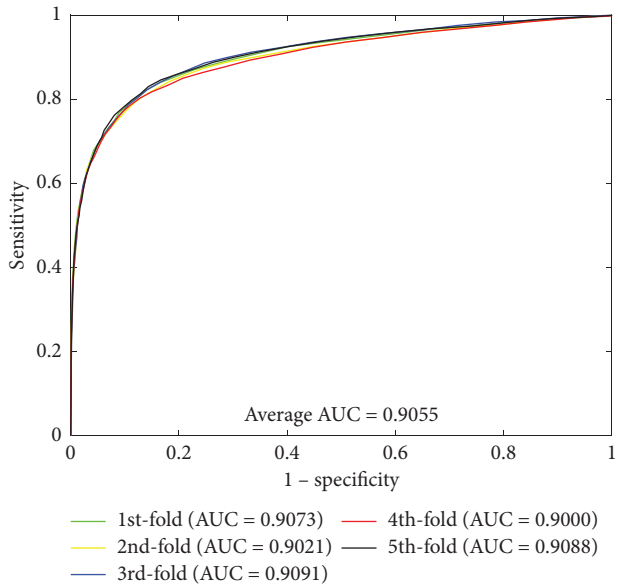

Figure 4: ROC curve for FWHT-RF on *Arabidopsis* PPIs dataset.

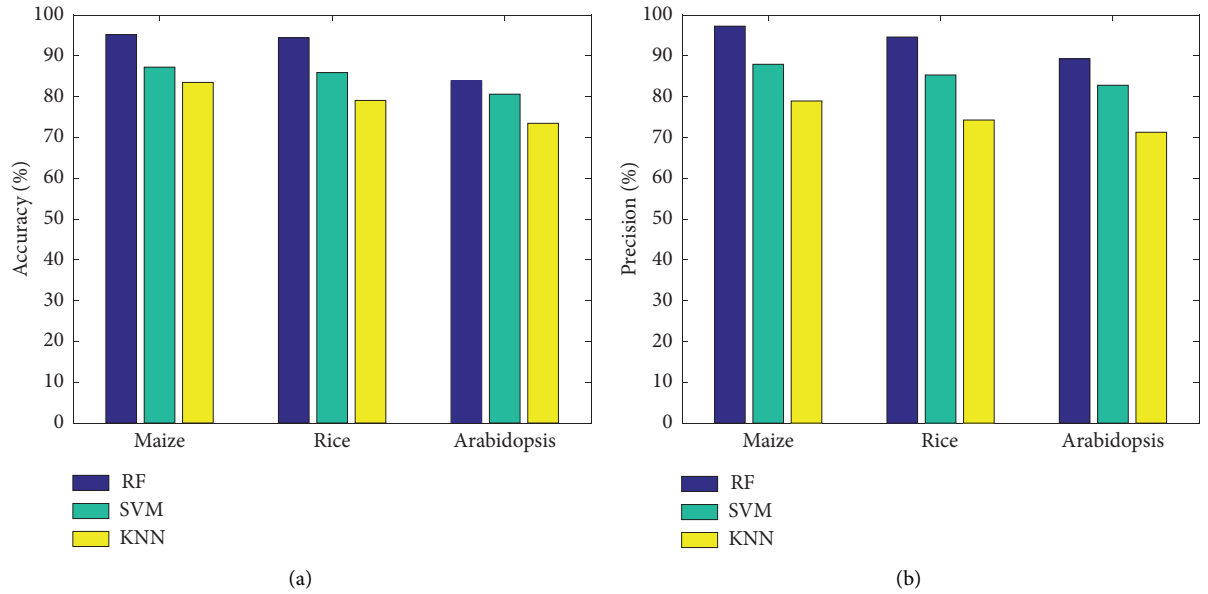

Figure 5: Continued.

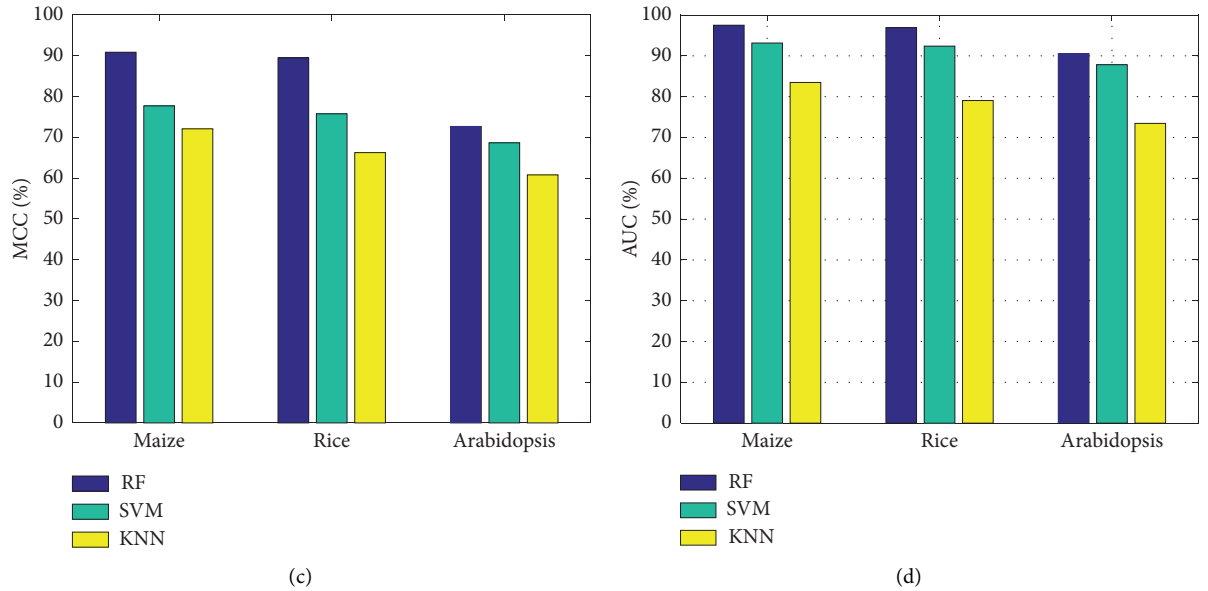

Figure 5: Performance comparisons of four validation metrics for three classifiers: RF (blue bar), SVM (green bar) and KNN (yellow bar). (a) Accuracy. (b) Precision. (c) MCC. (d) AUC.

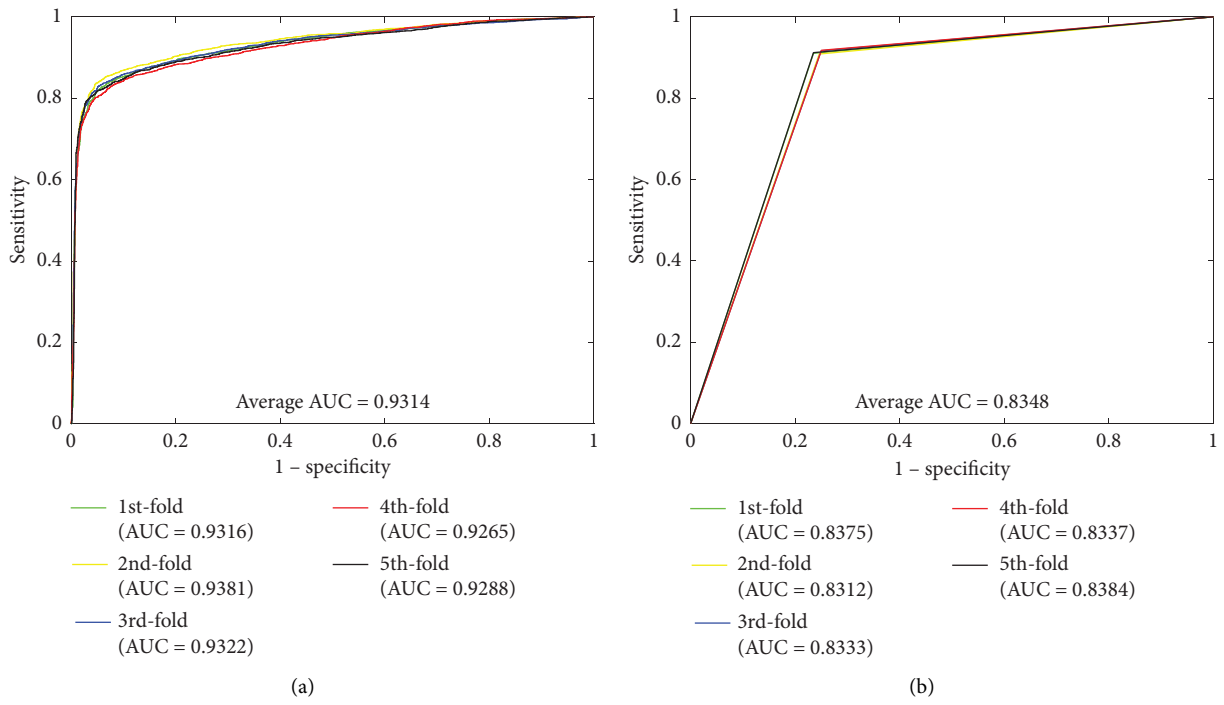

Figure 6: ROC curves performed on *Maize* dataset (5-fold cross validation). (a) is the ROC curves of SVM method. (b) is the ROC curves of KNN classifier.

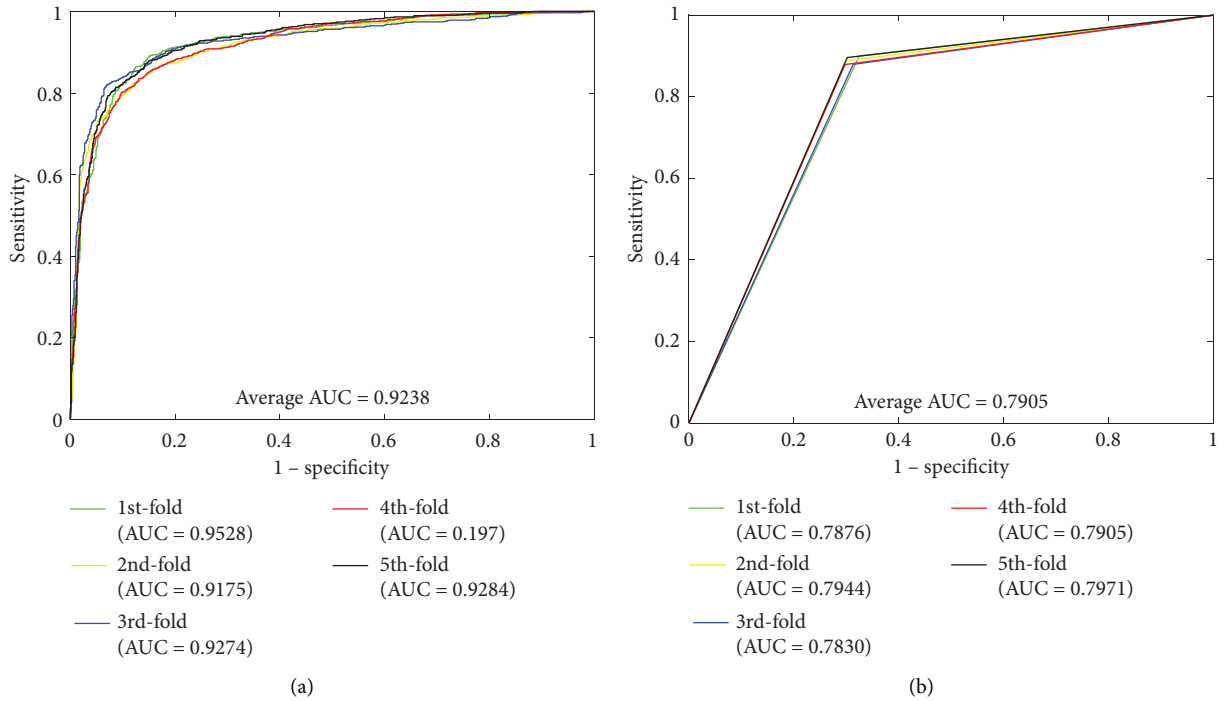

Figure 7: ROC curves performed on *Rice* dataset (5-fold cross validation). (a) is the ROC curves of SVM method. (b) is the ROC curves of KNN classifier.

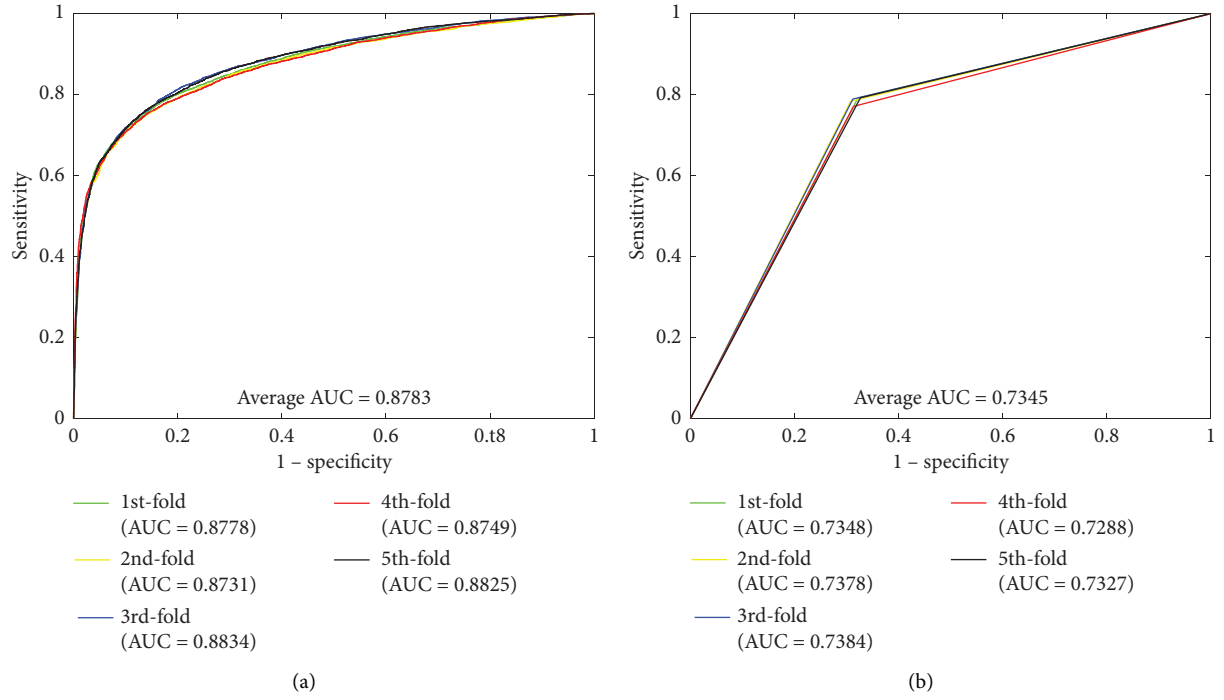

Figure 8: ROC curves performed on *Arabidopsis* dataset (5-fold cross validation). (a) is the ROC curves of SVM method. (b) is the ROC curves of KNN classifier.

Table 4: Comparing results of RF with SVM and KNN model on three plants PPIs dataset.

Dataset	Classifier	Acc. (%)	Sen. (%)	Prec. (%)	MCC (%)	AUC (%)
Maize	RF	95.20 ± 0.38	92.99 ± 0.62	97.29 ± 0.26	90.85 ± 0.69	97.50 ± 0.33
	SVM	87.22 ± 0.41	86.26 ± 0.89	87.95 ± 0.71	77.70 ± 0.62	93.14 ± 0.44
	KNN	83.48 ± 0.38	91.29 ± 0.40	78.96 ± 0.87	72.08 ± 0.51	83.48 ± 0.30
Rice	RF	**94.42 ± 0.56**	**94.17 ± 0.72**	**94.63 ± 0.84**	**89.46 ± 0.99**	**96.90 ± 0.37**
	SVM	85.89 ± 0.91	86.65 ± 1.76	85.33 ± 0.55	75.76 ± 1.29	92.38 ± 0.49
	KNN	79.06 ± 0.65	88.86 ± 0.96	74.29 ± 0.91	66.25 ± 0.74	79.05 ± 0.55
Arabidopsis	RF	**83.85 ± 0.35**	76.95 ± 1.16	**89.29 ± 0.62**	**72.66 ± 0.52**	**90.55 ± 0.41**
	SVM	80.59 ± 0.37	77.22 ± 0.85	82.81 ± 0.41	68.65 ± 0.46	87.83 ± 0.45
	KNN	73.45 ± 0.41	**78.53 ± 0.88**	71.29 ± 0.72	60.79 ± 0.38	73.45 ± 0.40

4. Discussion and Conclusions

In this study, we presented an effective sequence-based method called FWHT-RF to predict potential PPIs in plants. This method combined position-specific scoring matrix (PSSM) with fast Walsh–Hadamard transform (FWHT) and rotation forest (RF) classifier. First, we transformed the plant protein sequences into PSSM to obtain the evolutionary information of plants' protein sequences. Then, the FWHT algorithm was used to extract as much hidden information as possible from the plant protein sequences. At last, the RF classifier was trained for predicting PPIs in plants. When performed FWHT-RF on three plants' PPI datasets *Maize*, *Rice*, and *Arabidopsis*, it achieved a high prediction accuracy of 95.20%, 94.42%, and 83.85%, respectively. Moreover, we compared FWHT-RF with the state-of-art SVM and KNN classifier by adopting the same feature extraction method. The comprehensive experiments demonstrated that FWHT-RF is an effective tool to predict PPIs in plants. In the future work, we will consider applying FWHT-RF to other bioinformatics problems.

Data Availability

The data are original, and the data source is restricted.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China, Grant nos. 61722212 and 62002297.

References

[1] R. Subramaniam, D. Desveaux, C. Spickler, S. W. Michnick, and N. Brisson, "Direct visualization of protein interactions in plant cells," *Nature Biotechnology*, vol. 19, no. 8, pp. 769–772, 2001.

[2] T. Pawson and P. Nash, "Protein-protein interactions define specificity in signal transduction," *Genes & Development*, vol. 14, no. 9, pp. 1027–1047, 2000.

[3] M. Alves, S. Dadalto, A. Gonçalves, G. De Souza, V. Barros, and L. Fietto, "Transcription factor functional protein-protein interactions in plant defense responses," *Proteomes*, vol. 2, no. 1, pp. 85–106, 2014.

[4] C. C. Matiolli and M. Melotto, "A comprehensive Arabidopsis yeast two-hybrid library for protein-protein interaction studies: a resource to the plant research community," *Molecular Plant-Microbe Interactions*, vol. 31, no. 9, pp. 899–902, 2018.

[5] N. Ohad and S. Yalovsky, "Utilizing bimolecular fluorescence complementation (BiFC) to assay protein-protein interaction in plants," in *Plant Developmental Biology*, pp. 347–358, Springer, Berlin, Germany, 2010.

[6] J. S. Rohila, M. Chen, S. Chen et al., "Protein-protein interactions of tandem affinity purified protein kinases from rice," *PLoS One*, vol. 4, no. 8, Article ID e6685, 2009.

[7] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[8] P. Wang, R. Pleskot, J. Zang et al., "Plant AtEH/Pan1 proteins drive autophagosome formation at ER-PM contact sites with actin and endocytic machinery," *Nature Communications*, vol. 10, no. 1, pp. 5132–16, 2019.

[9] K. Knox, P. Wang, V. Kriechbaumer et al., "Putting the squeeze on plasmodesmata: a role for reticulons in primary plasmodesmata formation," *Plant Physiology*, vol. 168, no. 4, pp. 1563–1572, 2015.

[10] A. Onan, "On the performance of ensemble learning for automated diagnosis of breast cancer," in *Artificial Intelligence Perspectives and Applications*, pp. 119–129, Springer, Berlin, Germany, 2015.

[11] A. Onan, "Biomedical text categorization based on ensemble pruning and optimized topic modelling," *Computational and Mathematical Methods in Medicine*, vol. 2018, Article ID 2497471, 22 pages, 2018.

[12] A. Onan, "Ensemble learning based feature selection with an application to text classification," in *Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU)*, 2018.

[13] H.-C. Yi, Z.-H. You, D.-S. Huang, X. Li, T.-H. Jiang, and L.-P. Li, "A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information," *Molecular Therapy—Nucleic Acids*, vol. 11, pp. 337–344, 2018.

[14] S. Hashemifar, B. Neyshabur, A. A. Khan, and J. Xu, "Predicting protein-protein interactions through sequence-based deep learning," *Bioinformatics*, vol. 34, no. 17, pp. i802–i810, 2018.

[15] L. Zhang, G. Yu, D. Xia, and J. Wang, "Protein-protein interactions prediction based on ensemble deep neural networks," *Neurocomputing*, vol. 324, pp. 10–19, 2019.

[16] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel

negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine*, vol. 83, pp. 67–74, 2017.

[17] T. Sun, B. Zhou, L. Lai, and J. Pei, "Sequence-based prediction of protein protein interaction using a deep-learning algorithm," *BMC Bioinformatics*, vol. 18, no. 1, pp. 277–8, 2017.

[18] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics*, vol. 34, no. 4, pp. 660–668, 2018.

[19] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, pp. 1–16, 2016.

[20] K. Bracha-Drori, "Detection of protein-protein interactions in plants using bimolecular fluorescence complementation," *The Plant Journal*, vol. 40, no. 3, pp. 419–427, 2004.

[21] P. Zhu, H. Gu, Y. Jiao, D. Huang, and M. Chen, "Computational identification of protein-protein interactions in rice based on the predicted rice interactome network," *Genomics, Proteomics & Bioinformatics*, vol. 9, no. 4-5, pp. 128–37, 2011.

[22] G. Zhu, A. Wu, X.-J. Xu et al., "PPIM: a protein-protein interaction database for maize," *Plant Physiology*, vol. 170, no. 2, pp. 618–626, 2016.

[23] T. Tian, Y. Liu, H. Yan et al., "agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update," *Nucleic Acids Research*, vol. 45, no. W1, pp. W122–W129, 2017.

[24] H. Gu, P. Zhu, Y. Jiao, Y. Meng, and M. Chen, "PRIN: a predicted rice interactome network," *BMC Bioinformatics*, vol. 12, no. 1, p. 161, 2011.

[25] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. D1, pp. D841–D846, 2012.

[26] R. Oughtred, C. Stark, B.-J. Breitkreutz et al., "The BioGRID interaction database: 2019 update," *Nucleic Acids Research*, vol. 47, no. D1, pp. D529–D541, 2019.

[27] S. Y. Rhee, "The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community," *Nucleic Acids Research*, vol. 31, no. 1, pp. 224–228, 2003.

[28] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences*, vol. 84, no. 13, pp. 4355–4358, 1987.

[29] G. Raicar, H. Saini, A. Dehzangi, S. Lal, and A. Sharma, "Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids," *Journal of Theoretical Biology*, vol. 402, pp. 117–128, 2016.

[30] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases," *Trends in biochemical sciences*, vol. 23, no. 11, pp. 444–447, 1998.

[31] B. J. Fino and V. R. Algazi, "Unified matrix treatment of the fast Walsh-Hadamard transform," *IEEE Transactions on Computers*, vol. 25, no. 11, pp. 1142–1146, 1976.

[32] P. Zheng and J. Huang, "Walsh-Hadamard transform in the homomorphic encrypted domain and its application in image watermarking," in *Proceedings of the 2012 International Workshop on Information Hiding*, pp. 240–254, Springer, Berkeley, CA, USA, 2012.

[33] A. Thompson, "The cascading Haar wavelet algorithm for computing the Walsh-Hadamard transform," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1020–1023, 2017.

[34] C.-S. Park, "Recursive algorithm for sliding Walsh Hadamard transform," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2827–2836, 2014.

[35] S. Weinstein and P. Ebert, "Data transmission by frequency-division multiplexing using the discrete Fourier transform," *IEEE Transactions on Communication Technology*, vol. 19, no. 5, pp. 628–634, 1971.

[36] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974.

[37] Y. A. Geadah and M. Corinthios, "Natural, dyadic, and sequency order algorithms and processors for the Walsh-Hadamard transform," *IEEE Computer Architecture Letters*, vol. 26, no. 5, pp. 435–442, 1977.

[38] M. T. Hamood and S. Boussakta, "Fast Walsh-Hadamard-Fourier transform algorithm," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5627–5631, 2011.

[39] P. Zheng and J. Huang, "Efficient encrypted images filtering and transform coding with Walsh-Hadamard transform and parallelization," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2541–2556, 2018.

[40] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[41] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[42] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.