Sequence-Based Prediction of Plant Protein-Protein Interactions by Combining Discrete Sine Transformation With Rotation Forest

Evolutionary Bioinformatics
Volume 17: 1–9

DOI: 10.1177/11769343211050067
SAGE

Jie Pan, Li-Ping Li, Chang-Qing Yu, Zhu-Hong You, Yong-Jian Guan and Zhong-Hao Ren

College of Information Engineering, Xijing University, Xi'an, China.

ABSTRACT: Protein-protein interactions (PPIs) in plants are essential for understanding the regulation of biological processes. Although high-throughput technologies have been widely used to identify PPIs, they are usually laborious, expensive, and suffer from high false-positive rates. Therefore, it is imperative to develop novel computational approaches as a supplement tool to detect PPIs in plants. In this work, we presented a method, namely DST-RoF, to identify PPIs in plants by combining an ensemble learning classifier-Rotation Forest (RoF) with discrete sine transformation (DST). Specifically, plant protein sequence is firstly converted into Position-Specific Scoring Matrix (PSSM). Then, the discrete sine transformation was employed to extract effective features for obtaining the evolutionary information of proteins. Finally, these optimal features were fed into the RoF classifier for training and prediction. When performed on the plant datasets Arabidopsis, Rice, and Maize, DST-RoF yielded high prediction accuracy of 82.95%, 88.82%, and 93.70%, respectively. To further evaluate the prediction ability of our approach, we compared it with 4 state-of-the-art classifiers and 3 different feature extraction methods. Comprehensive experimental results anticipated that our method is feasible and robust for predicting potential plant-protein interacted pairs.

KEYWORDS: Plant, protein-protein interactions, discrete sine transformation, position-specific scoring matrix, rotation forest

RECEIVED: June 26, 2021. **ACCEPTED:** September 13, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is funded by the National Natural Science Foundation of China, under Grant 61722212 and 62002297.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Li-Ping Li, College of Information Engineering, Xijing University, Xi'an 710123, China. Email: Lipingli@gmail.com

Introduction

Protein-protein interactions (PPIs) in plants are an important aspect of systems biology.[1] It is very important for the investigation of biological processes, including signal transduction,[2] homeostasis control,[3] stress responses,[4] and plant defense.[5] Many traditional biological methods have been presented for exploring the functions and relationships between plant genes and proteins, such as yeast 2-hybrid system,[6,7] PPIs mapping,[8] tandem affinity purification (TAP),[9] and regulatory interaction.[10] However, these experimental approaches are time-consuming and cost a lot, and the plant PPI pairs collected from experiments only cover a small part of the Genome-wide protein interaction data. Due to this limitation, it is now believed that the problem of identifying unknown PPIs on a large scale is difficult to be solved entirely by traditional experimental methods.[11-13]

In recent years, various computational approaches have been developed to detect protein-protein interactions in plants.[14-17] These approaches can broadly fall into several categories: methods based on protein structure, protein domain, genomic information, evolutionary relationships, and protein primary sequence. Generally, the first 4 methods have a higher prediction accuracy. However, these approaches always require the structural details of proteins, such as 3D structural details. When the prior knowledge is not available, their drawbacks will be exposed. Theoretically, the amino acid sequence of proteins contains all the necessary information for identifying PPIs. In addition, with the development of sequencing technologies, more sequences information has been discovered. Therefore, sequence-based methods have attracted extensive attention.[18]

To date, numerous computational studies have been reported to predict PPIs from amino acid sequences. For example, Chen et al[19] developed a predictive framework named StackPPI. It is a stacked ensemble classifier constructed by extremely randomized trees, random forest, and logistic regression algorithms. Li et al[20] proposed an approach to predict PPIs only using sequence information. They converted sequences into Position Weight Matrix (PWM) and used Scale-Invariant Feature Transform (SIFT) method to extract features. Then PCA algorithm is employed to reduce the dimensionality of features. Finally, using the Weighted Extreme Learning Machine (WELM) classifier to detect PPIs. Khorsand et al[21] extracted several features from protein sequences and combined them with the human PPI network (HPPIN) to detect PPIs between Alphainfluenzavirus proteins and human proteins (HI-PPIs). Hashemifar et al[22] introduced a new framework called DPPI. It utilized a deep, Siamese-like convolutional neural network combined with data augmentation and random projection to identify PPIs from sequence information. Zhang et al[23] present a neural network-based method named EnsDNN, which used local descriptor, autocovariance descriptor, discontinuous local descriptor, and multi-scale continuous to represent amino acid sequence and detect PPIs. Kulmanov et al[24] presented an approach named DeepGO, which employed a deep ontology-aware classifier to predict protein functions and interactions

from protein sequence. Sun et al[25] used stacked autoencoder (SAE) to predict PPIs. Ding et al[26] employed a new multivariate mutual information (MMI) feature representation scheme and combined it with normalized Moreau-Broto Autocorrelation to extract features from protein sequence. Lastly, these features will be fed to Random Forest for training and predicting. Hu and Chan[27] present a novel coevolutionary feature extraction method, called CoFex, to predict PPIs. The coevolutionary features detect by this method are the covariations found at coevolving positions. Despite these achievements, there remains significant room for further improvement in terms of accuracy.

In this article, we present a novel computational model, called DST-RoF, to predict PPIs in plants that only adopting protein sequences information. It combined discrete sine transformation (DST), position-specific scoring matrix (PSSM), and rotation forest (RoF) classifier. More specifically, we first converted the protein primary sequences into PSSM to obtain the biological information. Then, the discrete sine transformation (DST) was performed on PSSM to extract primary features of different dimensions. Finally, these feature vectors were trained by the RoF classifier for prediction. When performed DST-RoF on the Arabidopsis, Rice, and Maize PPIs datasets, it yielded promising results of average accuracy of 82.95%, 88.82%, and 93.70%, respectively. To further verify the prediction performance of our approach, we compared DST with some popular feature extraction methods. We also compared RoF with *k*-nearest neighbor (KNN), support vector machine (SVM), deep neural network (DNN), and LightGBM classifier by using the same DST descriptors. The comprehensive results indicated that DST-RoF is effective and reliable for predicting potential PPIs in plants.

Materials and Methods

Data source

To evaluate the predictive ability of our method, we applied our method on 3 plant PPIs datasets. The first dataset is Arabidopsis. We collected it from public PPIs databases TAIR,[28] BioGRID,[29] and IntAct.[30] After removing redundant datasets, we selected the remaining 28 110 protein pairs as the positive dataset, which contained 7437 Arabidopsis proteins.[31] For the construction of the negative dataset, we used a bipartite graph to formulate a network of plant PPIs,[32] where the nodes denote the plants' proteins, and the links represent the interactions between them. Here, we set the Arabidopsis dataset as an example. The whole interactions of their connections are 55 308 969 (7437×7437) in the corresponding bipartite. However, only 28 110 protein pairs had been indicated to have the associations. Thus, the possible number of negative pairs is 55 280 859 (55 308 969-28 110), which is significantly more than the positive samples. To deal with this bias problem, we randomly collectedly 28 110 non-interacting PPIs pairs as the negative samples. Although in theory, these negative samples may contain a small count of positive pairs. However, given the size of whole PPIs dataset, the probability of this situation is very small. In this way, the whole Arabidopsis dataset is made up of 56 220 protein pairs.

Rice and maize are the 2 most important foods in the world.[33] To further validate the generality of the proposed approach, we also performed DST-RoF on the Rice and Maize dataset. We collected the 4800 Rice protein pairs from the rice protein reference database PRIN[34] and agriGO.[35] Similarly, we assumed that the proteins in different subcellular work compartments have no interactions, and finally yielded 4800 non-interacting protein pairs. Lastly, the Rice dataset consists of 9600 rice protein pairs. The Maize dataset was gathered from PPIM.[36] The whole Maize dataset consists of 29 600 maize protein pairs (14 800 positive protein pairs and 14 800 negative protein pairs).

Representation of plant protein sequence

The position-specific scoring matrix (PSSM)[37] was developed for detecting distantly related proteins. In this work, we employed PSSM to encode the plant protein sequences. Let $K = \left\{ \lambda_{i,j} : i = 1 \cdots L \text{ and } j = 1 \cdots 20 \right\}$, and each matrix can be defined as follows:

$$PSSM = \begin{bmatrix} \lambda_{1,1}, & \lambda_{1,2} & \cdots & \lambda_{1,20} \\ \lambda_{2,1}, & \lambda_{2,2} & \cdots & \lambda_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{L,1} & \lambda_{L,2} & \cdots & \lambda_{L,20} \end{bmatrix} \quad (1)$$

where $\lambda_{i,j}$ represent the probability that the ith residue changed to the jth amino acid.

In our experiment, we used the PSI-BLAST[38] to convert the Arabidopsis, Rice, and Maize sequence as a matrix. The PSI-BLAST is an accurate tool, which was against the database of *SwissProt* to generate the PSSM. To obtain a highly and widely homologous sequence, we select 3 iterations and assigned the *e*-value of PSI-BLAST to be 0.001. Finally, each plant protein sequence can be represented as a $L \times 20$ matrix, L represents the length of an amino acid sequence and 20 represents twenty different kinds of amino acids. The *SwissProt* database and PSI-BLAST can be freely obtained from http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Feature extraction method

Discrete Sine Transformation (DST)[39] is a kind of sinusoidal unitary and separable Transform. It plays a key role in the field of signal and image processing, not only because of its transforming coding capabilities but also for some other applications, including adaptive beamforming, signal interpolation, and image resizing.[40] As it is a separable transform, the

2D-DST can be constructed by two 1D-DST. The first 1D-DST is applied column-wise and the obtaining results will be adopted as the input for the second 1D-DST, which is then calculated by row-wise. The most common DST definitions for 1D sequence of length T can be described as:

$$P(v)=\alpha(v)\sum_{x=0}^{T-1}f(x)\sin\left[\frac{\pi(2x+1)v}{2T}\right] \quad (2)$$

for $v=0,1,\ldots,T-1$. Similarly, the inverse transformation is defined as:

$$f(x)=\sum_{v=0}^{T-1}\alpha(v)\,p(v)\sin\left[\frac{\pi(2x+1)v}{2T}\right] \quad (3)$$

for $x=0,1,\ldots,T-1$. For the both equations (2) and (3), the $\alpha(v)$ can be described as:

$$\alpha(v)=\begin{cases}\sqrt{\frac{1}{T}} \text{ for } u=0\\ \sqrt{\frac{2}{T}} \text{ for } u\neq 0\end{cases} \quad (4)$$

Thus, the 2D-DST can be described as:

$$P(v,u)=\alpha(v)\alpha(u)\sum_{a=0}^{T-1}\sum_{b=0}^{T-1}f(a,b)\sin\left[\frac{\pi(2a+1)v}{2T}\right]\sin\left[\frac{\pi(2b+1)u}{2T}\right] \quad (5)$$

for $v,u=0,1,2,\cdots,T-1$. $\alpha(v)$ and $\alpha(u)$ is defined in equation (4). Where x represents the length of 1D sequence, u and v denotes the length and width of input images in 2D-DST. In this study, $f(a,b)$ represents the input signal matrix and here is the $L\times 20$ PSSM. In this way, plant protein sequences can be represented by the DST feature descriptors.

Rotation Forest classifier

Rotation Forest (RoF)[41] is a popular ensemble classification method, which uses the concept of feature transformation to improve the diversity and accuracy of the classifier in the ensemble system.[42] It applies the Principal component analysis (PCA)[43] algorithm to construct a rotational matrix and transforms initial variables into new variables. In this way, RoF can build independent decision trees. At the same time, the PCA algorithm maintains the integrity of the information contained in the data while ensuring the diversity of the classifiers. The framework of the Rotation Forest can be described as follows.

Let a training set $\eta=\left\{\left[F_i,G_j\right]\right\}_{i=1}^{R}$ consisting of R training samples, in which F_i represents the input feature vector and G_i denotes the corresponding class label. Assuming that the feature set is randomly split into K subsets with the same size, and RoF has L decision trees denoted by $T_1,T_2,\ldots,T_L$, respectively. L and K are the 2 parameters that need to be optimized in advance. The training process for a base classifier T_i is shown as follows:

(1) The feature set F will be randomly split into K disjoint subsets. As a result, each subset has $M=n/K$ number of features.
(2) Let β_{ij} represents the jth subsets of features for training classifier T_i, and φ_{ij} denotes the dataset X for the features in β_{ij}. Employing a new training set φ'_{ij}, which is a non-empty subset of classes randomly extracted from φ_{ij}, and it accounts for 75% of the dataset X. After using the PCA technique on the T_i, the coefficients in a matrix Q_{ij} can be generated.
(3) Build a sparse rotation matrix R_i with the achieved coefficients in matrix Q_{ij} as follows:

$$R_i=\begin{bmatrix}\gamma_{i,1}^{(1)},\cdots,\gamma_{i,1}^{(M_2)} & 0 & \cdots & 0\\ 0 & \gamma_{i,2}^{(1)},\cdots\gamma_{i,2}^{(M_2)} & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & \gamma_{i,s}^{(1)},\cdots,\gamma_{i,s}^{(M_s)}\end{bmatrix} \quad (6)$$

The columns R_i will be rearranged from the new rotation matrix R_i^a. Accordingly, the transformed classifier sample T_i is XR_i^a. In this way, the classifiers can be trained in parallel.

During the prediction process, given a test sample x, let the probability of this test sample detected by classifier T_i into class y_i, which is expressed as $d_{i,j}(xR_i^a)$. Assign x is split into a class with the largest confidence $\omega_j(x)$. Thus, the class of confidence can be calculated according to formula (7).

$$\omega_j(x)=\frac{1}{L}\sum_{i=1}^{L}d_{i,j}(XR_i^a),\ j=1,2 \quad (7)$$

Experimental and Results

Evaluation metrics

In this work, we used the following 4 metrics to access the performance of the prediction method, including accuracy (ACC.), precision (PR.), sensitivity (Sen.), and Matthews Correlation Coefficient (MCC). They can be calculated as:

$$ACC.=\frac{TP+TN}{TP+TN+FN+FP} \quad (8)$$

$$PR.=\frac{TP}{TP+FP} \quad (9)$$

Table 1. Prediction results of different dimensions on 3 plants dataset.

DIMENSIONS	DATASETS	ACC. (%)	PR. (%)	SEN. (%)	MCC. (%)	AUC
40	Arabidopsis	81.36 ± 0.40	86.71 ± 0.74	74.07 ± 0.76	69.34 ± 0.53	0.8756 ± 0.0028
	Rice	84.06 ± 1.09	89.90 ± 1.52	76.74 ± 1.25	72.92 ± 1.51	0.8706 ± 0.0096
	Maize	91.82 ± 0.31	94.79 ± 0.38	88.51 ± 0.52	84.95 ± 0.54	0.9546 ± 0.0025
60	Arabidopsis	82.37 ± 0.50	87.80 ± 0.68	75.19 ± 0.71	70.66 ± 0.66	0.8847 ± 0.0026
	Rice	85.04 ± 1.06	90.72 ± 0.69	78.07 ± 1.80	74.32 ± 1.53	0.8839 ± 0.0094
	Maize	92.43 ± 0.59	95.43 ± 0.39	89.13 ± 1.10	85.98 ± 1.02	0.9583 ± 0.0041
80	Arabidopsis	82.95 ± 0.13	88.21 ± 0.36	76.06 ± 0.34	71.44 ± 0.19	0.8897 ± 0.0028
	Rice	87.21 ± 0.56	91.69 ± 1.00	81.83 ± 0.55	77.56 ± 0.83	0.8999 ± 0.0064
	Maize	92.93 ± 0.42	95.67 ± 0.33	89.94 ± 0.76	86.84 ± 0.73	0.9621 ± 0.0024
100	Arabidopsis	81.97 ± 0.54	88.87 ± 0.78	73.09 ± 0.62	69.98 ± 0.70	0.8830 ± 0.0035
	Rice	87.41 ± 0.92	92.03 ± 1.27	81.88 ± 1.31	77.85 ± 1.41	0.9058 ± 0.0103
	Maize	93.38 ± 0.43	95.92 ± 0.43	90.63 ± 1.03	87.62 ± 0.74	0.9630 ± 0.0035
120	Arabidopsis	80.31 ± 0.41	87.36 ± 0.17	70.87 ± 0.90	67.81 ± 0.56	0.8645 ± 0.0036
	Rice	88.82 ± 0.58	92.91 ± 0.78	84.08 ± 1.41	80.05 ± 0.92	0.9194 ± 0.0035
	Maize	93.60 ± 0.44	96.20 ± 0.58	90.79 ± 0.61	88.00 ± 0.76	0.9647 ± 0.0019
140	Arabidopsis	80.65 ± 0.27	87.62 ± 0.42	71.37 ± 0.46	68.25 ± 0.34	0.8679 ± 0.0026
	Rice	87.71 ± 0.95	92.57 ± 1.08	82.01 ± 1.38	78.31 ± 1.45	0.9057 ± 0.0070
	Maize	93.70 ± 0.43	96.09 ± 0.31	91.11 ± 0.79	88.18 ± 0.75	0.9666 ± 0.0039

$$Sen. = \frac{TP}{FN + TP} \tag{10}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{11}$$

where TP is a true positive, standing for the count of true samples that identified correctly; FP represents false positive, indicating the number of true non-interacting pairs that correctly predicted; TN denotes the true negative, standing for the number of negative samples that are determined has no interactions; FN is the false negative, indicating the number of true samples predicted to be non-interacting pairs incorrectly. Moreover, the receiver operating characteristic (ROC)[44] curve is employed as a measure, and the area under the ROC curve (AUC)[45] is also calculated to visually demonstrate the predictive capacity of the proposed model.

Selecting the best dimensions

In order to obtain the best prediction results, we tested the accuracy of the proposed method in different dimensions. From Table 1, it can clearly see that the best dimensions for Arabidopsis and Rice are 80 and 120, and the best dimension of Maize is 140. We also implemented a series of experiments to optimize the parameters of the RoF classifier. As a result, on the Arabidopsis dataset, the parameters L and K are set to be 35 and 22; on the Rice dataset, the parameters L and K are set to be 2 and 3, the parameters L and K for the Maize dataset were set to be 17 and 15, respectively. Here, L represents the number of decision trees and the count of feature subsets is denoted by K.

Prediction performance of proposed method

To avoid overfitting of the proposed method, 5-fold cross-validation (CV)[46] was applied to verify the predictive ability of DST-RoF on the Arabidopsis, Rice, and Maize datasets. Specifically, the whole dataset was randomly split into 5 equal subsets, where 4 of them were used as training sets and the remaining 1 for testing, so we can conduct 5 experiments in 1 dataset. The prediction results obtained from the proposed approach on the Arabidopsis, Rice, and Maize datasets are shown in Tables 2 to 4.

Table 2. The 5-fold CV results yielded from the Arabidopsis dataset by the DST-RoF.

TEST SET	ACC. (%)	PR. (%)	SEN. (%)	MCC. (%)	AUC
1	82.87	88.02	76.09	71.35	0.8868
2	82.86	87.67	76.01	71.31	0.8878
3	83.18	88.37	76.54	71.78	0.8891
4	82.91	88.58	75.57	71.35	0.8898
5	82.92	88.40	76.01	71.42	0.8923
Average	82.95 ± 0.13	88.21 ± 0.36	76.06 ± 0.34	71.44 ± 0.19	0.8897 ± 0.0028

Table 3. The 5-fold CV results yielded from the Rice dataset by the DST-RoF.

TEST SET	ACC. (%)	PR. (%)	SEN. (%)	MCC. (%)	AUC
1	88.54	93.69	82.13	79.50	0.9196
2	88.28	92.74	83.54	79.24	0.9158
3	88.80	93.19	83.82	80.02	0.9168
4	88.70	91.66	85.30	79.91	0.9200
5	89.79	93.25	85.61	81.59	0.9248
Average	88.82 ± 0.58	92.91 ± 0.78	84.08 ± 1.41	80.05 ± 0.92	0.9194 ± 0.0035

Table 4. The 5-fold CV results yielded from the Maize dataset by the DST-RoF.

TEST SET	ACC. (%)	PR. (%)	SEN. (%)	MCC. (%)	AUC
1	93.78	96.30	91.12	88.32	0.9645
2	93.66	95.87	91.05	88.09	0.9650
3	94.03	95.72	92.14	88.76	0.9705
4	94.04	96.47	91.30	88.77	0.9707
5	92.99	96.11	89.94	86.95	0.9623
Average	93.70 ± 0.43	96.09 ± 0.31	91.11 ± 0.79	88.18 ± 0.75	0.9666 ± 0.0039

When applying DST-RoF to the Arabidopsis dataset, we achieved high average prediction accuracy (ACC.), precision (PR.), sensitivity (Sen.), and MCC of 82.95%, 88.21%,76.06%, and 71.44%, with the standard deviation of 0.13%, 0.36%, 0.34%, and 0.19%, respectively. The ROC curves achieved by the proposed approach on the Arabidopsis dataset are shown in Figure 1, with the average AUC value and standard deviation of 0.8897 and 0.0028, respectively. On the Rice dataset, DST-RoF obtained average ACC., PR., Sen. and MCC of 88.82%, 92.91%, 84.08%, and 80.05%, with standard deviation of 0.58%, 0.78%, 1.41%, and 0.92%, respectively. The ROC curves obtained by DST-RoF on the Rice dataset are shown in Figure 2, with the average value of AUC and its standard deviation are 0.9194 and 0.0035, respectively. When applying DST-RoF on the Maize dataset, the average ACC., PR., Sen., and MCC were 93.70%, 96.09%, 91.11%, and 88.18%, with the standard deviation of 0.43%,0.31%,0.79%, and 0.75%, respectively. The ROC curves yielded by DST-RoF on the Maize dataset are shown in Figure 3, with the average value of AUC and standard deviation are 0.9666 and 0.0039, respectively.

Comparison with previous studies on the maize dataset

Various kinds of computational approaches have been developed for predicting PPIs in plants. To further verify the predictive power of DST-RoF, we compared it with some existing methods on the Maize dataset. Table 5 lists the prediction results of the other 4 different methods. It can be observed that DST-ROF obtained the best results in accuracy, MCC, and AUC values. Although the precision and sensitivity were lower

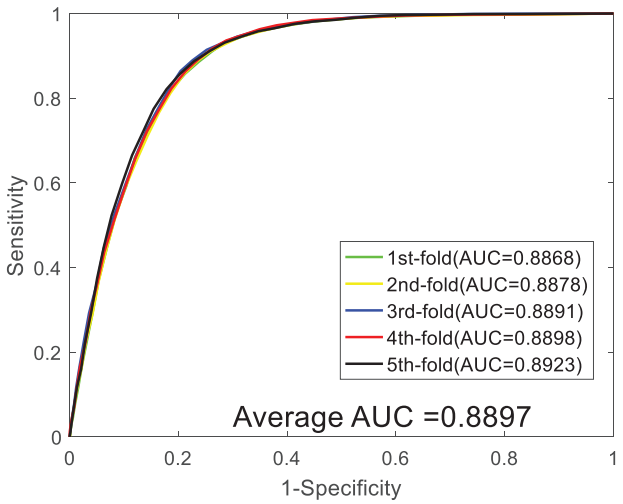

Figure 1. The ROC curves achieved by DST-RoF on Arabidopsis dataset.

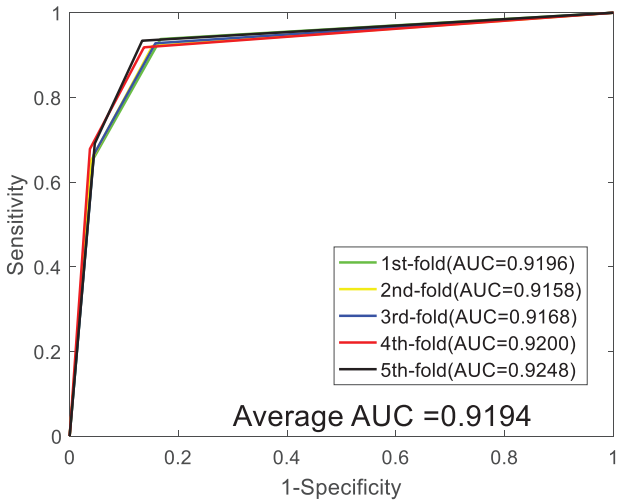

Figure 2. The ROC curves achieved by DST-RoF on Rice dataset.

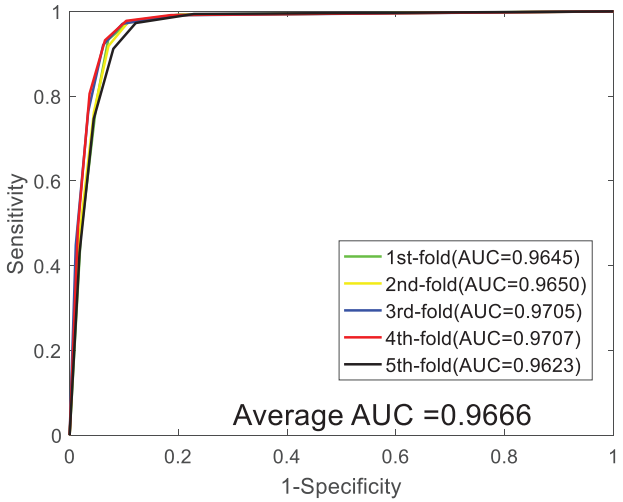

Figure 3. The ROC curves achieved by DST-RoF on Maize dataset.

than some previous methods, it still attained promising results of 96.09% and 91.11%. The ACC values yielded by these methods are between 79.58% and 89.9%, lower than 93.7%, which was achieved by the proposed method. In terms of MCC and AUC values, the average increase of our method over the best results of the 4 existing methods is 7.59% and 0.26%, respectively. These comparison results indicated that DST-RoF can improve predictive ability. This improvement may attribute to the novel feature extraction method and the use of the Rotation Forest algorithm which has been indicated to be powerful and effective for PPIs prediction.

Comparison with different feature descriptors on the rice dataset

In order to verify the superiority of the DST feature extraction method, we compared it with different feature extraction methods in the same RoF classifier. In this part, we employed DCT (Discrete Cosine Transform),[49] FFT (Fast Fourier Transform),[50] and HHT (Hilbert-Huang transform)[51] to further evaluate the prediction performance of DST-RoF. DCT is a linear and invertible transformation using in image transformation. FFT has been widely performed in digital signal processing. HHT is a signal decomposition method that employed empirical mode decomposition (EMD) to decompose a real-world signal into pseudo monochromatic waves. The comparison results of different feature extraction methods on the Rice dataset are summarized in Table 6. We can indicate that DST descriptor is better than the other 3 feature extraction methods. The detailed 5-fold CV results performed by DCT, FFT, and HHT algorithm on the Rice dataset are summarized in Supplemental Tables S1 to S3.

Comparison with the KNN, SVM, DNN, and LightGBM-based method

There are many machine learning algorithms that have been to used detect PPIs. In order to further evaluate the prediction performance of DST-RoF, we combined the same DST feature descriptors with *k*-nearest neighbor (KNN),[52] support vector machine (SVM),[53] deep neural network (DNN),[54,55] and LightGBM[56,57] classifier. *k*-nearest neighbor (KNN) is a supervised machine learning method and it is simple and effective for classification tasks. The main idea of SVM classifier is to find a high-dimensional decision plane to solve the classification prediction problems. DNN is a deep-learning-based method, which is composed of an input layer, multiple hidden layers, and an output layer. Recently, it has been widely applied to predict PPIs.[58-60] LightGBM was introduced by Ke et al[57] that combined the exclusive feature bundling (EFB) and gradient-based 1-side sampling (GOSS) algorithm.

In this part, we employed the LIBSVM[61] tool to train the SVM model. In addition, 2 parameters need to be optimized when applying the SVM classifier (the penalty c of the model and the gamma g of the kernel function). In the experiments of Arabidopsis and Rice datasets, we set $c=17$, $g=5$ and $c=11$, $g=0.09$, respectively. For the Maize dataset, we set $c=7$, $g=0.4$. The KNN classifier needs to choose the number of neighbors

Table 5. Comparing DST-RoF with other approaches on the Maize dataset.

MODEL	ACC (%)	PR (%)	SEN (%)	MCC (%)	AUC
SIPMA[47]	89.9	N/A	62.0	68.0	0.964
PPIM[36]	79.58	96.44	61.44	N/A	0.8636
WSRC + IFFT[48]	89.12	87.49	91.32	80.59	0.9376
Our method	93.70	96.09	91.11	88.18	0.9666

Abbreviation: N/A, not applicable.

Table 6. The results obtained by RoF classifier based on different feature extraction methods on the Rice dataset.

FEATURE EXTRACTION METHODS	ACC. (%)	PR. (%)	SEN. (%)	MCC (%)
DCT	63.40 ± 1.84	91.88 ± 1.00	29.41 ± 3.66	41.53 ± 2.77
FFT	62.92 ± 1.55	90.82 ± 1.78	28.69 ± 2.42	41.02 ± 1.77
HHT	63.68 ± 1.89	91.43 ± 2.01	30.17 ± 4.08	42.07 ± 3.11
Our method	88.82 ± 0.58	92.91 ± 0.78	84.08 ± 1.41	80.05 ± 0.92

Table 7. Comparing results of RoF with 4 different classifiers on 3 PPIs dataset.

DATASET	MODEL	ACC (%)	PR (%)	SEN (%)	MCC (%)	AUC
Arabidopsis	KNN	74.77 ± 0.96	73.65 ± 4.55	78.16 ± 5.62	61.94 ± 0.89	0.7459 ± 0.0055
	SVM	75.09 ± 0.39	77.24 ± 0.69	71.15 ± 0.65	62.48 ± 0.40	0.8252 ± 0.0050
	DNN	64.89 ± 2.15	60.39 ± 2.14	87.49 ± 3.54	33.55 ± 2.62	0.7901 ± 0.0044
	LightGBM	79.95 ± 0.28	82.32 ± 0.26	76.29 ± 0.40	60.07 ± 0.56	0.8725 ± 0.0024
	RoF	82.95 ± 0.13	88.21 ± 0.36	76.06 ± 0.34	71.44 ± 0.19	0.8891 ± 0.0021
Rice	KNN	79.11 ± 1.30	72.63 ± 0.94	87.10 ± 1.56	64.03 ± 1.54	0.7713 ± 0.0143
	SVM	77.46 ± 1.53	77.86 ± 1.14	76.79 ± 1.83	65.10 ± 1.65	0.8557 ± 0.0134
	DNN	72.55 ± 1.77	66.24 ± 2.21	92.39 ± 2.33	49.23 ± 2.34	0.8695 ± 0.0065
	LightGBM	84.34 ± 0.89	84.53 ± 0.93	83.21 ± 1.47	68.70 ± 1.78	0.8935 ± 0.0083
	RoF	88.82 ± 0.58	92.91 ± 0.78	84.08 ± 1.41	80.05 ± 0.92	0.9194 ± 0.0025
Maize	KNN	86.07 ± 0.59	83.45 ± 2.92	90.26 ± 3.71	75.89 ± 0.91	0.8605 ± 0.0060
	SVM	84.24 ± 0.49	86.20 ± 0.56	81.55 ± 1.21	73.41 ± 0.68	0.9136 ± 0.0042
	DNN	82.00 ± 1.07	75.16 ± 1.65	93.21 ± 1.02	65.71 ± 1.84	0.9353 ± 0.0051
	LightGBM	85.56 ± 0.33	89.02 ± 0.63	81.12 ± 0.81	71.40 ± 0.66	0.9105 ± 0.0087
	RoF	93.70 ± 0.43	96.09 ± 0.31	91.11 ± 0.79	88.18 ± 0.75	0.9641 ± 0.0039

k and distance measuring function. Here, we set k to 1 and the distance measuring function is set to be $L1$ for the 3 datasets. The DNN classifier that used in this paper consists of 2 hidden layers with 48 and 30 neurons. Table 7 list the experimental results of KNN, SVM, DNN, LightGBM, and RoF classifiers on 3 plant PPIs datasets.

It can be seen from Table 7 that when DST-RoF is used to predict the Arabidopsis dataset, high accuracy (82.95%) is obtained, which is 8.18%, 7.86%, 18.06%, and 3% higher than those of KNN, SVM, DNN, and LightGBM, respectively. On the Rice dataset, the accuracy of DST-RoF is 88.82%, which is much better than that of the other 4 methods. The

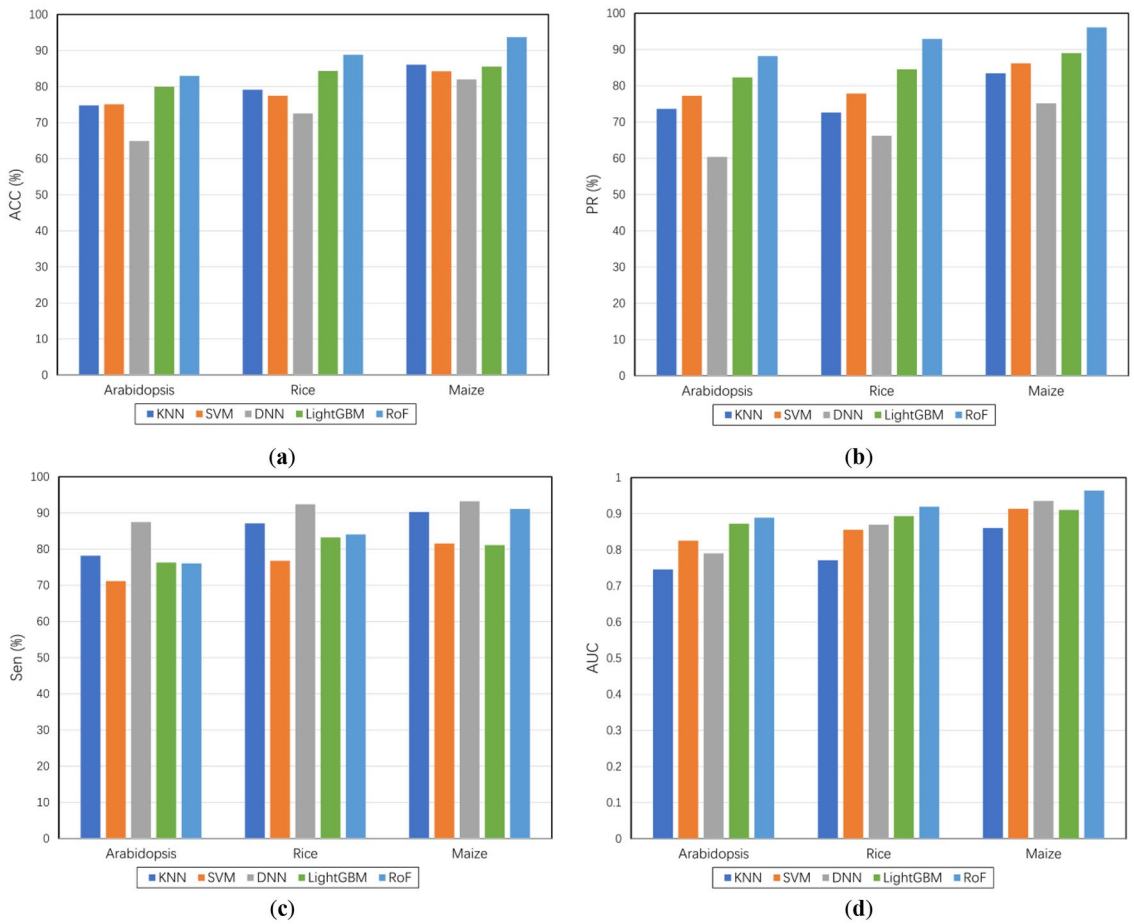

Figure 4. Performance comparisons of 4 validation metrics of the 5 classifiers: (a) accuracy, (b) precision, (c) sensitivity, and (d) AUC.

accuracy of KNN, SVM, DNN, and LightGBM on the Rice dataset is 9.71%, 11.36%, 16.27%, and 4.48% lower than that of the proposed method, respectively. When DST-RoF is applied to identify the Maize dataset, the accuracy of the proposed approach is 93.70%, which is 7.63%, 9.46%, 11.7%, and 8.14% higher than our approach, respectively. When employing DST-RoF on the Arabidopsis dataset, its AUC value is 0.8891, which is 14.32%, 6.39%, 9.9%, and 1.66% higher than KNN, SVM, DNN, and LightGBM, respectively. On the Rice dataset, the AUC value of RoF is 0.9194, which is better than the other 4 algorithms. The AUC values of KNN, SVM, DNN, and LightGBM classifier on the Rice dataset are 14.81%, 6.37%, 4.99%, and 2.59% lower than our method. When performed DST-RoF on the Maize dataset, its AUC value is 0.9641, which is 10.36%, 5.05%, 2.88%, and 5.36% higher than the other 4 classifiers. In addition, the higher accuracies and low standard deviations further indicated that the combination of RoF classifier and DST descriptors can significantly improve the performance in plant PPIs prediction. Figure 4a to d reports the results yielded by the 5 classifiers on the 3 plant PPIs datasets.

Conclusions

In this paper, we present a novel sequence-based approach called DST-RoF, to predict protein-protein interactions (PPIs) in plants by combining discrete sine transformation (DST) with Rotation Forest (RoF). For obtaining rich evolutionary information, we first convert the plant protein sequence into Position-Specific Scoring Matrix (PSSM) and then extract feature vectors using the DST algorithm. Finally, these features are fed into the RoF classifier to determine whether there is an interaction between these protein pairs. When performed on 3 benchmark datasets (Arabidopsis, Rice, and Maize), DST-RoF obtained high average accuracies of 82.95%, 88.82%, and 93.70%, respectively. In order to verify the predictive ability of rotation forest, we compared it to state-of-the-art KNN, SVM, DNN, and LightGBM classifiers. In addition, we also compared DST with some popular feature descriptors. These results demonstrated that the presented approach is feasible and accurate for predicting potential PPIs in plants. In future work, we aim to find more efficient feature descriptors and develop a better model to explore the functions of plant proteins.

ORCID iD

Jie Pan https://orcid.org/0000-0002-4993-298X

Dataset

The source codes and datasets explored in this work are available at https://github.com/jie-pan111/Prediction-of-PPIs-in-plants.

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*. 2002;417:399-403.
2. McDowell JM, Dangl JL. Signal transduction in the plant immune response. *Trends Biochem Sci*. 2000;25:79-82.
3. May MJ, Vernoux T, Leaver C, Montagu MV, Inze D. Glutathione homeostasis in plants: implications for environmental sensing and plant development. *J Exp Bot*. 1998;49:649-667.
4. Chinnusamy V, Zhu J-K. Epigenetic regulation of stress responses in plants. *Curr Opin Plant Biol*. 2009;12:133-139.
5. Hammond-Kosack KE, Jones JD. Resistance gene-dependent plant defense responses. *Plant Cell*. 1996;8:1773-1791.
6. Ehlert A, Weltmeier F, Wang X, et al. Two-hybrid protein–protein interaction analysis in Arabidopsis protoplasts: establishment of a heterodimerization map of group C and group S bZIP transcription factors. *Plant J*. 2006;46:890-900.
7. Fang Y, Macool DJ, Xue Z, et al. Development of a high-throughput yeast two-hybrid screening system to study protein-protein interactions in plants. *Mol Genet Genomics*. 2002;267:142-153.
8. Struk S, Jacobs A, Sánchez Martín-Fontecha E, Gevaert K, Cubas P, Goormachtig S. Exploring the protein–protein interaction landscape in plants. *Plant Cell Environ*. 2019;42:387-409.
9. Van Leene J, Eeckhout D, Persiau G, et al. Isolation of transcription factor complexes from Arabidopsis cell suspension cultures by tandem affinity purification. In: Yuan L, Perry S, eds. *Plant Transcription Factors*. Springer; 2011;195-218.
10. Chow C-N, Zheng H-Q, Wu NY, et al. PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res*. 2016;44:D1154-D1160.
11. Zhang Y, Gao P, Yuan JS. Plant protein-protein interaction network and interactome. *Curr Genomics*. 2010;11:40-46.
12. Gookin TE, Kim J, Assmann SM. Whole proteome identification of plant candidate G-protein coupled receptors in Arabidopsis, rice, and poplar: computational prediction and in-vivo protein coupling. *Genome Biol*. 2008;9:R120-R126.
13. Haque S, Ahmad JS, Clark NM, Williams CM, Sozzani R. Computational prediction of gene regulatory networks in plant growth and development. *Curr Opin Plant Biol*. 2019;47:96-105.
14. Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN Jr. Plant systems biology comes of age. *Trends Plant Sci*. 2008;13:165-171.
15. Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M. A predicted interactome for Arabidopsis. *Plant Physiol*. 2007;145:317-329.
16. Kumari S, Ware D. Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. *PLoS One*. 2013;8:e79011.
17. Adai A, Johnson C, Mlotshwa S, et al. Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*. 2005;15:78-91.
18. Zhang Y, Natale R, Domingues AP, et al. Rapid identification of protein-protein interactions in plants. *Curr Protoc Plant Biol*. 2019;4:e20099.
19. Chen C, Zhang Q, Yu B, et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Comput Biol Med*. 2020;123:103899.
20. Li J, Shi X, You Z-H, et al. Using weighted extreme learning machine combined with scale-invariant feature transform to predict protein-protein interactions from protein evolutionary information. *IEEE/ACM Trans Comput Biol Bioinform*. 2020;17:1546-1554.
21. Khorsand B, Savadi A, Zahiri J, Naghibzadeh M. Alpha influenza virus infiltration prediction using virus-human protein-protein interaction network. *Math Biosci Eng*. 2020;17:3109-3129.
22. Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*. 2018;34:i802-i810.
23. Zhang L, Yu G, Xia D, Wang J. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing*. 2019;324:10-19.
24. Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*. 2018;34:660-668.
25. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*. 2017;18:277.
26. Ding Y, Tang J, Guo F. Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics*. 2016;17:398.
27. Hu L, Chan KCC. Extracting coevolutionary features from protein sequences for predicting protein-protein interactions. *IEEE/ACM Trans Comput Biol Bioinform*. 2017;14:155-166.
28. Lamesch P, Berardini TZ, Li D, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*. 2012;40:D1202-D1210.
29. Oughtred R, Stark C, Breitkreutz B-J, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019;47:D529-D541.
30. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40:D841-D846.
31. Yang S, Li H, He H, Zhou Y, Zhang Z. Critical assessment and performance improvement of plant-pathogen protein-protein interaction prediction methods. *Brief Bioinform*. 2019;20:274-287.
32. Pavlopoulos GA, Kontou PI, Pavlopoulou A, Bouyioukos C, Markou E, Bagos PG. Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience*. 2018;7:giy014.
33. Zhang H, Xu F, Wu Y, Hu HH, Dai XF. Progress of potato staple food research and industry development in China. *J Integr Agric*. 2017;16:2924-2932.
34. Gu H, Zhu P, Jiao Y, Meng Y, Chen M. PRIN: a predicted rice interactome network. *BMC Bioinformatics*. 2011;12:161.
35. Tian T, Liu Y, Yan H, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res*. 2017;45:W122-W129.
36. Zhu G, Wu A, Xu XJ, et al. PPIM: a protein-protein interaction database for maize. *Plant Physiol*. 2016;170:618-626.
37. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA*. 1987;84:4355-4358.
38. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389-3402.
39. Jain A. A fast Karhunen-Loeve transform for a class of random processes. *IEEE Trans Commun*. 1976;24:1023-1029.
40. Ramadan K, Fiky AS, Dessouky MI, Abd El-Samie FE. Equalization and carrier frequency offset compensation for UWA-OFDM communication systems based on the discrete sine transform. *Digit Signal Process*. 2019;90:142-149.
41. Rodríguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: a new classifier ensemble method. *IEEE Trans Pattern Anal Mach Intell*. 2006;28:1619-1630.
42. Guo Z-H, You Z-H, Wang Y-B, Yi HC, Chen ZH. A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest. *iScience*. 2019;19:786-795.
43. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab Syst*. 1987;2:37-52.
44. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39:561-577.
45. Qu K, Guo F, Liu X, Lin Y, Zou Q. Application of machine learning in microbiology. *Front Microbiol*. 2019;10:827.
46. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011;21:137-146.
47. Khatun MS, Hasan MM, Mollah MNH, et al. SIPMA: A systematic identification of protein-protein interactions in Zea mays using autocorrelation features in a machine-learning framework. In: *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, Taichung, Taiwan, 29-31 October, 2018. IEEE; 2018:122-125.
48. Pan J, Yu C, Li L, et al. *Computational Prediction of Protein-Protein Interactions in Plants Using Only Sequence Information*. Springer International Publishing; 2021:115-125.
49. Ahmed N, Natarajan T, Rao KR. Discrete cosine transform. *IEEE Trans Comput*. 1974;100:90-93.
50. Nussbaumer HJ. The fast Fourier transform. In: Nussbaumer HJ, ed. *Fast Fourier Transform and Convolution Algorithms*. Springer; 1981;80-111.
51. Huang NE. *Hilbert-Huang Transform and Its Applications*. World Scientific; 2014.
52. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. *IEEE Trans Syst Man Cybern*. 1985;4:580-585.
53. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273-297.
54. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006;313:504-507.
55. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. *Neural Comput*. 2006;18:1527-1554.
56. Chen C, Zhang Q, Ma Q, Yu B. LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemometr Intell Lab Syst*. 2019;191:54-64.
57. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.
58. Gui Y, Wang R, Wei Y, Wang X. DNN-PPI: a large-scale prediction of protein–protein interactions based on deep neural networks. *J Biol Syst*. 2019;27:1-18.
59. Patel S, Tripathi R, Kumari V, Varadwaj P. DeepInteract: deep neural network based protein-protein interaction prediction tool. *Curr Bioinform*. 2017;12:551-557.
60. Li H, Gong X-J, Yu H, Zhou C. Deep neural network based predictions of protein interactions using primary sequences. *Molecules*. 2018;23:1923.
61. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2:1-27.