SIPEC: Systematic identification of self-interacting proteins with ensemble classifiers using evolutionary information

Lei Wang[1,*], Z.-H. You[1,*], Shan-Wen Zhang[1], Tao Wang[1], Li-Ping Li[2] and Ya-Ping Wu[1]

[1]College of Information Engineering, Xijing University, Xi'an 710123, China

[2]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China

[*]Joint First Authors

Correspondence to: *Z.-H. You,* ***email:*** *zhuhongyou@gmail.com*
Shan-Wen Zhang, ***email:*** *wjdw716@163.com*

Keywords: *self-interacting proteins (sips); disease; pernicious anemia; amino acids; evolutionary information*

Received: *August 05, 2017* ***Accepted:*** *January 01, 2018* ***Published:*** *January 11, 2018*

ABSTRACT

As the center of most biological processes, Protein-Protein Interactions (PPIs) constitute the basis of the formation of biological mechanisms. Deregulation of PPIs results in many diseases including cancer and pernicious anemia. As a special type of PPIs, the Self-interacting Proteins (SIPs) occupy an important position in them. Although a large number of SIPs data have been generated by experimental methods, currently-detected self-interacting proteins cover only a small part of the complete network. Therefore, there is a great need for computational methods to efficiently and accurately predict SIPs. In the present study, we introduce a novel computational method based on protein sequence information to predict SIPs. More specifically, each protein sequence is converted to Position-Specific Scoring Matrix (PSSM) containing the evolutionary information. And then an effective feature extraction approach, namely, Auto Covariance (AC) is employed to construct a feature set. Finally, the improved Rotation Forest (RF) model is used to remove the noise of the feature set and give prediction results. When performed on *yeast* and *human* SIPs data sets, the proposed method can achieve high accuracies of 80.50% and 93.70%, respectively. Our method also shows a good performance when compared with the SVM classifier and other existing methods. Consequently, the proposed method can be considered to be a promising model to predict SIPs. In addition, for the purpose of further research in the future, the user-friendly web server is freely available to academic use at http://www.proteininteraction.cn/sip/.

INTRODUCTION

As both the material base of life and the main bearer of life activities, proteins affect the cells through interaction with other components. In these interactions, Protein-Protein Interactions (PPIs) has attracted more attention of researchers because of their critical roles in living organisms. Deregulation of PPIs results in many diseases including cancer and pernicious anemia. The PPI data accumulated from the previous numerous small-scale experiments and some recent large-scale experiments allow us to establish the proteome-wide PPI networks [1–4], which will help us to deepen the understanding of cell structure and function from the perspective of the system and provide theoretical basis for the discovery of new drug targets and drug design.

One special type of PPIs is Self-interacting proteins (SIPs). They represent those with more than two copies that can interact with each other. Two interaction partners of SIPs are two identical copies represented by the same gene, which can result in the formation of homodimer. More than two copies of a protein interact with each other

to form a homotrimer or a higher order homo-oligomer. Recent research have shown that homo-oligomerization plays important roles in a variety of vital biological processes, such as immune response, enzyme activation, signal transduction and gene expression regulation [5–9]. Ispolatov *et al*. [10] noted that SIPs occupy a significant position in the protein interaction networks (PINs), meaning that there are great possibilities that the SIPs can interact with a large number of other proteins. At the same time, it also shows its functional importance for cellular systems. Pereira-Leal and their collaborators proposed a genome-wide, cross-species analysis of the origins and evolution of protein complexes. Their conclusion indicates that the evolution of many protein complexes was first established through self-interactions and then through the duplication of these self-interacting proteins [11]. In addition, one of the key factors that regulate protein function is self-interaction. Without increasing the size of the genome, through self-interactions, the functional diversity of proteins can be greatly expanded [12].

Recently, some computational methods for the prediction of PPIs have been developed [13, 14]. By analyzing the relationship between codon pair usage and PPIs in yeast, Zhou *et al*. drew a conclusion that codon pair usage of interacting protein pairs has great difference on the random expectations. And it is used as a motivation by proposing a novel method named CCPPI to predict PPIs by using codon pair frequency difference as Support Vector Machine input [15]. Based on pairwise similarity theory, Zaki *et al*. used only the protein primary structure before proposing a simple and efficient method for predicting PPIs [16]. You *et al*. used only the protein sequence information to predict PPI, in which a kind of method called PCA-EELM (Principal Component Analysis-Ensemble Extreme Learning Machine) is designed. When performed on the PPIs data of *Saccharomyces cerevisiae*, this model yields 87.00% prediction accuracy, 86.15% sensitivity and 87.59% precision [17]. These methods generally take into account the correlational information between protein pairs, such as coevolution, co-localization and co-expression. However, such information is not available when predicting protein self-interacting. Furthermore, the data sets used in these methods do not contain protein interactions among the same partners, making them unsuitable for SIP prediction. Therefore, there is a strong motivation to design efficient and reliable computation methods for large-scale prediction SIPs.

Based on the Rotation Forest (RF) algorithm [18, 19], in this study, we designed an improved RF-based approach (ImRF) [20, 21] for predicting SIPs by only using protein amino acids sequences. First, the candidate self-interacting protein sequence is converted into Position-Specific Scoring Matrix (PSSM) [22]. Second, an effective feature extraction method called Auto Covariance (AC) [23] is used to extract feature vector from PSSM. Finally, the features of weighted selection are fed into the RF classifier to predict SIPs. In the experiments, the proposed model was evaluated on *yeast* and *human* SIPs data sets. The experiment result shows that our model achieved 80.50% and 93.70% prediction accuracy with 85.30% and 94.70% specificity on these two datasets, respectively. In order to further evaluate the performance of our model, we compared it with other existing methods and the state-of-the-art support vector machine (SVM) classifiers on *yeast* and *human* data sets. Excellent results indicate that our model can effectively extract useful information from large amounts of data and produce better prediction accuracy.

RESULTS AND DISCUSSION

Performance of the proposed method

In order to avoid over-fitting to affect the performance of our model, we divided the data set into training set and independent test set. Taking the *human* data set as an example, we randomly selected about 1/6 of the samples from the whole *human* data set as the independent test set. Since the number of negative instances is much larger than that of the positive ones in *human* data set, we randomly selected negative samples from the remaining *human* negative data set to set up the training set with the ratio of about 1:1. To ensure the reliability of the results, the independent test set and training set were constructed for 5 times and so were the experiments . The final results were expressed in the form of mean and standard deviation. The same strategy was also used to apply to the *yeast* dataset. For the sake of guaranteeing the fair outcome, there are several parameters that should be optimized for our model. Through the grid search method, in this experiment, the parameter *lg* of the feature extraction method AC is set to 5. In the improved rotation forest algorithm, feature selection rate $r = 0.7$, the number of sub sets $K = 5$, and the number of decision trees $L = 7$.

The results of our method on *yeast* and *human* datasets are shown in Tables 1, 2. It can be seen from Table 1 that the overall accuracies of five experiments are all above 79.09% for *yeast* dataset. Specifically, the accuracies of each experiment are 79.89%, 79.09%, 80.91%, 82.95% and 79.55%, respectively. We can see that the average accuracy, specificity, sensitivity, and MCC are 80.50%, 85.30%, 42.60%, and 23.20%, respectively. The standard deviations of them are 1.50%, 2.10%, 3.40%, and 1.60%, respectively. Table 2 lists the experimental results of our method on the *human* data set. Accuracies of the five experiments are 93.88%, 92.72%, 93.56%, 94.44%, and 94.40%, respectively. The good results of average accuracy, specificity, sensitivity, and MCC of 93.70%, 94.70%, 34.00%, and 15.40%, respectively. The standard deviations of them are 0.60%, 0.70%, 3.80%, and 0.90%, respectively. The ROC curves performed on *yeast* and *human* datasets was shown in Figures 1, 2. In those

Table 1: The experimental results obtained by using proposed method on *yeast* SIPs data set

Testing set	Accu. (%)	Spe. (%)	Sen. (%)	MCC (%)	AUC (%)
1	79.89	84.10	47.00	24.95	64.82
2	79.09	83.46	45.00	22.68	63.01
3	80.91	86.15	40.00	22.19	61.25
4	82.95	88.59	39.00	24.84	64.36
5	79.55	84.36	42.00	21.48	65.33
Average	80.50 ± 1.50	85.30 ± 2.10	42.60 ± 3.40	23.20 ± 1.60	63.75 ± 1.64

Table 2: The experimental results obtained by using proposed method on *human* data set

Testing set	Accu. (%)	Spe. (%)	Sen. (%)	MCC (%)	AUC (%)
1	93.88	94.84	35.00	16.23	69.84
2	92.72	93.58	40.00	16.56	75.59
3	93.56	94.55	32.50	14.43	75.44
4	94.44	95.49	30.00	14.79	71.64
5	94.40	95.00	32.50	15.21	80.64
Average	93.70 ± 0.60	94.70 ± 0.70	34.00 ± 3.80	15.40 ± 0.90	74.63 ± 4.17

figures, x-ray depicts False Positive Rate (FPR) while y-ray delineates True Positive Rate (TPR).

Thanks to choosing the appropriate classifier and feature extraction method, we can see from Table 1 and Table 2 that our method has achieved good results when predicting SIPs. Our method plays an important role in improving the accuracy of prediction, which may be attributed to the following three reasons: (1) PSSM has the advantage of resisting background noise and reducing the redundancy of prediction results. It can retain enough prior information of protein sequences, thus helping to improve the prediction accuracy. (2) Feature extraction method AC takes neighboring effect into account, which makes it possible to discover patterns of the entire sequences. (3) High-dimensional data not only increases the computational cost but also is likely to contain

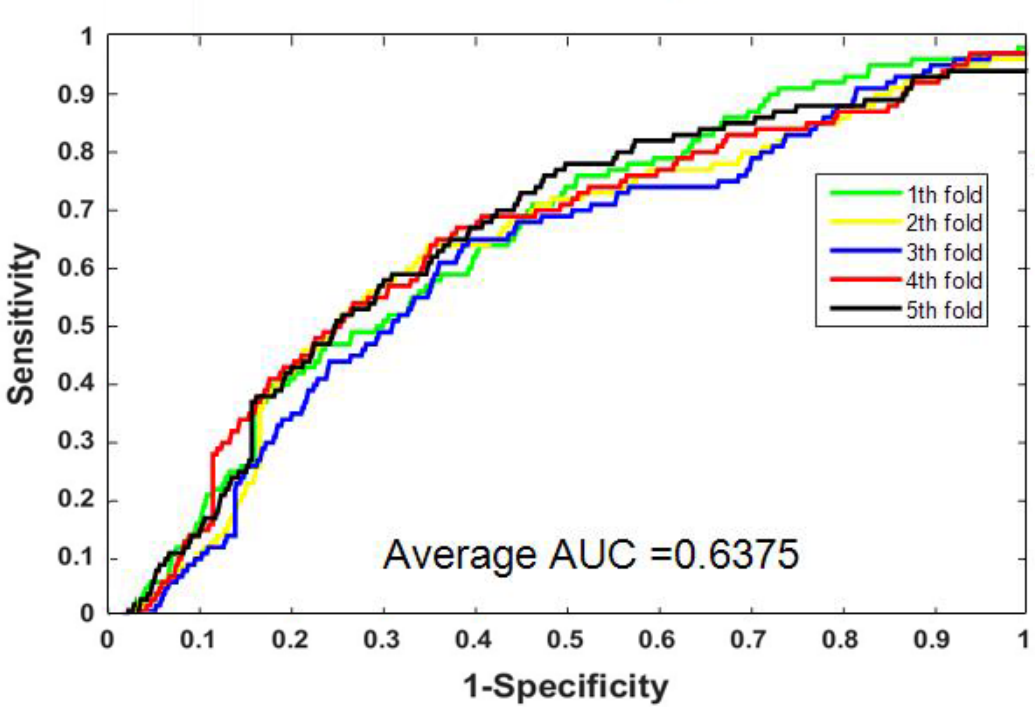

Figure 1: Performance comparison performed by our proposed model on *Yeast* SIPs data set in terms of ROC curves and AUCs. As a result, SIP-ECEI yielded high performance with the AUC of 0.6375.

redundant information. In this experiment, we use the improved rotation forest method to calculate the weight of the feature and remove the features of small weight. This increases the proportion of the useful information and helps to improve the performance of the classifier. The experimental results show that these powerful factors can provide help for the prediction of SIPs.

Comparison with the SVM Classifier

Although the experimental results show that the performance of our proposed prediction model is good, in order to have a clearer understanding of our classifier, we compare it with the state-of-the-art support vector machine (SVM) classifier. In the experiment, we have taken the same feature extraction method and implemented it in the *yeast* and *human* data sets, respectively. We use the LIBSVM tool [24] to execute the SVM classifier. The SVM parameters determined by the grid search method are $c = 10$ and $g = 10$, and other parameters use the default value.

Tables 3, 4 list the prediction results of SVM classifier on *yeast* and *human* datasets respectively. It can be seen from Table 3 that the average accuracy of SVM on *yeast* dataset is 78.10%, while the results of five experiments are 78.64%, 77.27%, 78.07%, 78.07%, and 78.30%. However, the improved rotation forest classifier achieved 80.50% average accuracy. Similarly as displayed in Table 4, the average accuracy of SVM on *human* dataset is 91.30%, while the results of five experiments are 91.72%, 92.16%, 90.16%, 91.48%, and 91.12%. At the same time, the accuracy of the improved rotation forest classifier is 93.70%. The ROC curves performed on *yeast* and *human* data sets were shown in Figures 3, 4.

Comparison with other methods

In order to further evaluate the performance of the proposed method, we also compared our final model with three existing SIPs predictor SLIPPER [25], CRS [26], SPAR [26] and three PPI predictors DXECPPI [27], PPIevo [28] and LocFuse [29] based on the *yeast* and *human* datasets. Tables 5, 6 list the results of the above-mentioned methods on *yeast* and *human* data sets. We can observe from Table 5 that the proposed method performs well and the accuracy is only next to the highest, 6.84% higher than the average accuracy of other six methods on *yeast* data set. Similarity, as shown in Table 6, the prediction results of the proposed method are obviously higher those of the other six different methods on *human* dataset. Accuracy is 1.61% higher than the highest method, and 16.31% higher than that the average of the other six methods. The prediction results show that the proposed method can more effectively improve the accuracy than the current existing methods and suitable for predicting SIPs.

Web server

For the convenience of using the proposed model, a user-friendly web server has been made available at http://www.proteininteraction.cn/sip/. Web server mainly

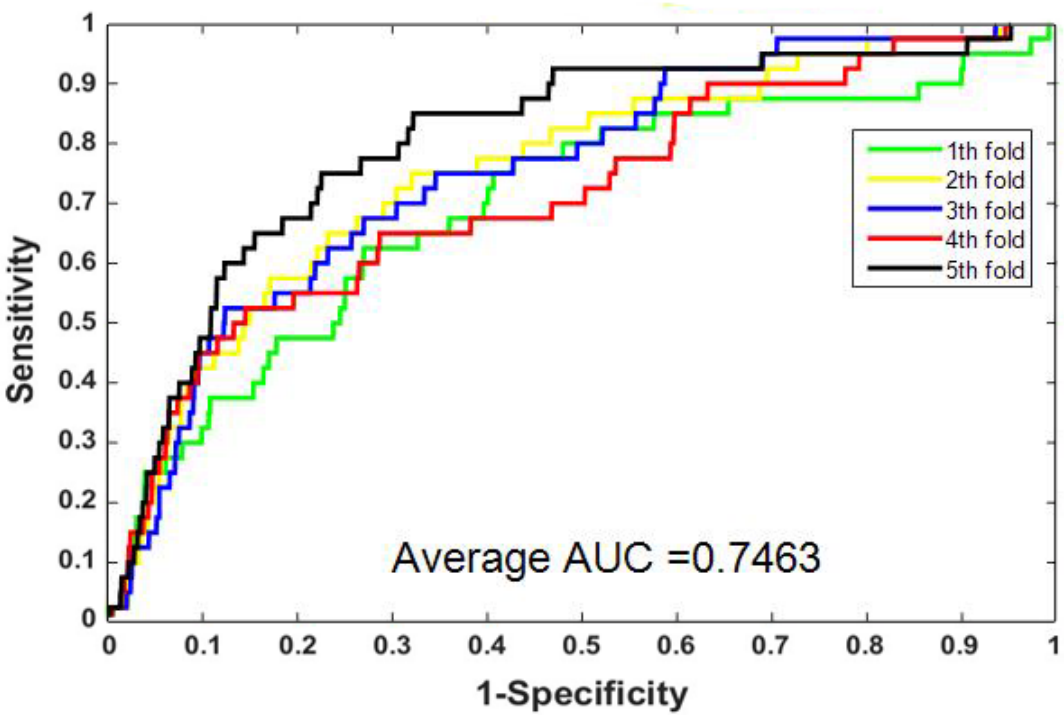

Figure 2: Performance comparison performed by our proposed model on *Human* SIPs data set in terms of ROC curves and AUCs. As a result, SIP-ECEI yielded high performance with the AUC of 0.7463.

Table 3: The experimental results obtained by using SVM classifier on yeast data set

Testing set	Accu. (%)	Spe. (%)	Sen. (%)	MCC (%)	AUC (%)
1	78.64	83.97	37.00	17.18	64.02
2	77.27	81.67	43.00	19.17	62.69
3	78.07	83.72	34.00	14.54	60.29
4	78.07	82.31	45.00	21.35	64.18
5	78.30	82.82	43.00	20.44	65.24
Average	78.10 ± 0.50	82.90 ± 1.00	40.40 ± 4.70	18.50 ± 2.70	63.28 ± 1.90
Our method	**80.50 ± 1.50**	**85.30 ± 2.10**	**42.60** ± 3.40	**23.20 ± 1.60**	**63.75 ± 1.64**

Table 4: The experimental results obtained by using SVM classifier on *human* data set

Testing set	Accu. (%)	Spe. (%)	Sen. (%)	MCC (%)	AUC (%)
1	91.72	92.72	30.00	10.73	69.77
2	92.16	93.05	37.50	14.61	75.56
3	90.16	90.89	45.00	15.23	75.49
4	91.48	92.32	40.00	14.78	71.66
5	91.12	91.91	42.50	15.37	80.68
Average	91.30 ± 0.80	92.20 ± 0.80	39.00 ± 5.80	14.10 ± 1.90	74.63 ± 4.20
Our method	**93.70 ± 0.60**	**94.70 ± 0.70**	**34.00** ± 3.80	**15.40 ± 0.90**	**74.63 ± 4.17**

provides predictive proteins self-interacting on *yeast* data set. Users input *yeast* protein sequences in the web page and enter the received email address. After pressing the submit button, the server will automatically predict whether the proteins can interact with each other based on our proposed method. After the completion of the server computing, users can check E-mail in the mailbox, which shows the predicted results.

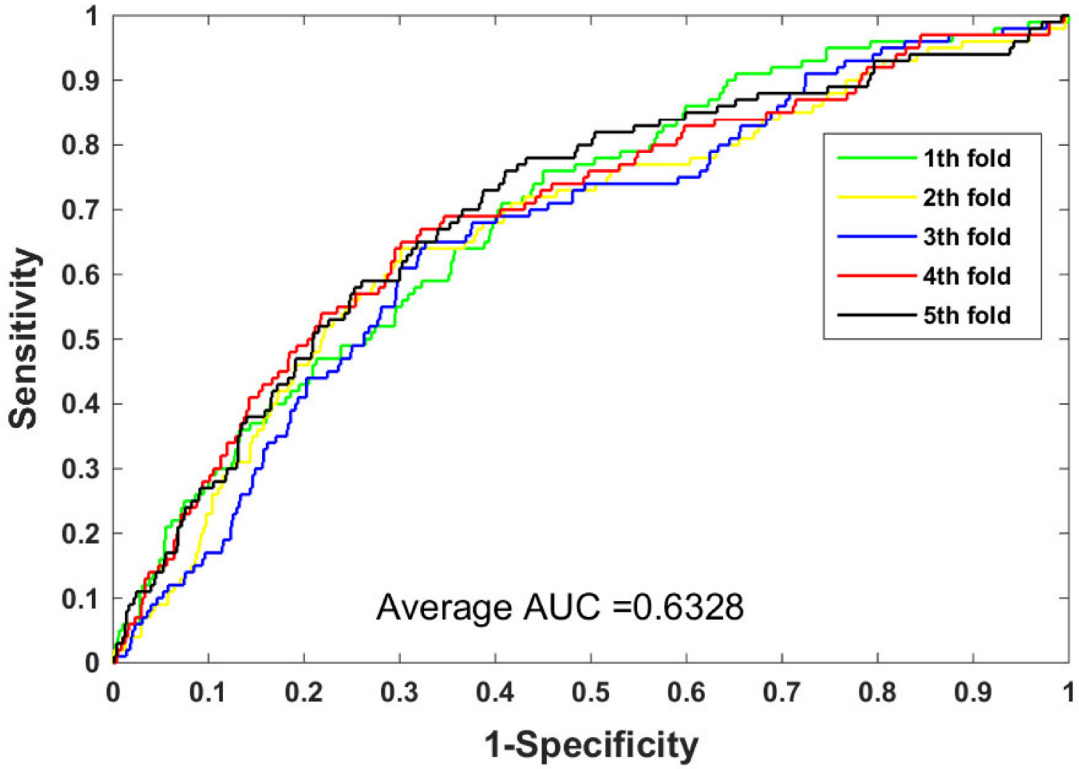

Figure 3: Performance comparison performed by SVM model on *Yeast* SIPs data set in terms of ROC curves and AUCs. As a result, SIP-ECEI yielded high performance with the AUC of 0.6328.

CONCLUSIONS

In this paper, we proposed a novel computational method based on protein sequence information to large-scale and efficient prediction protein self-interaction, which combines the feature extraction method AC and improved rotation forest classifier. In order to evaluate the performance of the proposed method, we implemented it on the *yeast* and *human* data sets. We also compared the state-of-the-art support vector machine classifier with other popular methods commonly used for PPIs prediction. In these comparisons, we achieved good performance. The experimental results on *yeast* and *human* data sets show that the prediction accuracy achieved by our method has been significantly improved. In addition, for the convenience of researchers, we construct a user-friendly web server based on the proposed method. It can provide users with the predicted result of whether proteins could interact with each other. In future research, we will focus on more effective feature extraction methods and machine learning algorithms to improve the prediction accuracy.

MATERIALS AND METHODS

Dataset

We can obtain 20,199 curated *human* protein sequences from the *UniProt* database [30]. These PPIs data come from various resources, including *MatrixDB* [31], *InnateDB* [32], *IntAct* [33], *BioGRID* [34] and *DIP* [35]. In this experiments, we only extract protein sequences in which two interaction partners are exactly the same and interactive type is the 'direct interaction' in relevant databases. Eventually, the number of *human* protein self-interaction instances we have obtained is 2,994.

We construct datasets through the following steps in order to achieve the purpose of evaluating the performance of our model [26]. Firstly, we only preserve the number of residues in proteins ranging from 50 to 5,000.The rest of the proteins were removed from the whole *human* proteome. Secondly, to ensure the quality of the protein self-interaction data, each sample in positive data set must satisfy one of the following conditions: (1) At least two publications reported the protein self-interaction; (2) The protein is defined as homo-oligomer (including homodimer and homotrimer) in *UniProt*; (3) At least two large-scale experiments or one small-scale experiment detected the self-interaction. Finally, to construct the negative dataset, we removed the predicted SIPs annotated in *UniProt* and all types of SIPs from the whole *human* proteome (including proteins annotated as more extensive 'physical association' and 'direct interaction'). As a result, 1,441 positive samples and 15,938 negative samples were constructed as *human* SIPs data set. In addition, we used the same strategy in the construction of *yeast* data set, which contained 710 positive samples and 5,511 negative samples.

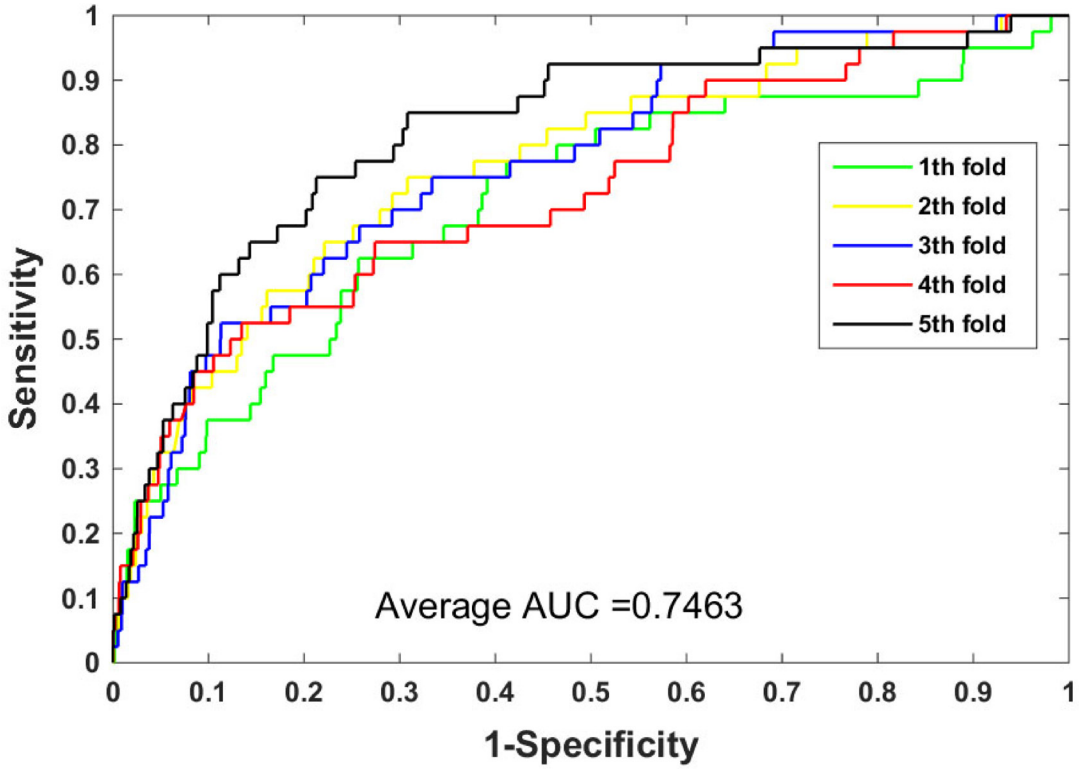

Figure 4: Performance comparison performed by our proposed model on *Human* SIPs data set in terms of ROC curves and AUCs. As a result, SIP-ECEI yielded high performance with the AUC of 0.7463.

Table 5: Performance comparisons between SIP-ECEI and six existing computational models (SLIPPER, DXECPPI, PPIevo, LocFuse, CRS and SPAR) on yeast SIPs data set for predicting SIPs in terms of Accuracy, Specificity, Sensitivity, MCC based on cross validations

Model	Accu. (%)	Spe. (%)	Sen. (%)	MCC (%)
SLIPPER [25]	71.90	72.18	69.72	28.42
DXECPPI [45]	87.46	94.93	29.44	28.25
PPIevo [28]	66.28	87.46	60.14	18.01
LocFuse [29]	66.66	68.10	55.49	15.77
CRS [26]	72.69	74.37	59.58	23.68
SPAR [26]	76.96	80.02	53.24	24.84
Our method	**80.50 ± 1.50**	**85.30 ± 2.10**	**42.60 ± 3.40**	**23.20 ± 1.60**

Table 6: Performance comparisons between SIP-ECEI and six existing computational models (SLIPPER, DXECPPI, PPIevo, LocFuse, CRS and SPAR) on *Human* SIPs data set for predicting SIPs in terms of Accuracy, Specificity, Sensitivity, MCC based on cross validations

Model	Accu. (%)	Spe. (%)	Sen. (%)	MCC (%)
SLIPPER [25]	91.10	95.06	47.26	41.97
DXECPPI [45]	30.90	25.83	87.08	8.25
PPIevo [28]	78.04	25.82	87.83	20.82
LocFuse [29]	80.66	80.50	50.83	20.26
CRS [26]	91.54	96.72	34.17	36.33
SPAR [26]	92.09	97.40	33.33	38.36
Our method	**93.70 ± 0.60**	**94.70 ± 0.70**	**34.00** ± 3.80	**15.40 ± 0.90**

Position-specific scoring matrix

Position-specific scoring matrix (PSSM) is generated by a set of sequences which has the structure or sequence similarity. Initially introduced by Gribskov *et al.* [22], it is used for detecting distantly related protein. PSSM has made outstanding achievements in areas such as protein secondary structure prediction [36], protein binding site prediction [37], and prediction of disordered regions [38]. A PSSM is a matrix of $N\times 20$, which can be denoted as $M=\{e_{i,j}:i=1\cdots N \text{ and } j=1\cdots 20\}$, where N represents the length of the protein sequence and 20 the number of the amino acids. Each matrix $M(i,j)$ is defined as follows:

$$M=\begin{bmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,20} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ e_{N,1} & e_{N,2} & \cdots & e_{N,20} \end{bmatrix} \quad (1)$$

where $e_{i,j}$ represents the probability that the *ith* residue being mutated into the *jth* naive amino acid during the evolutionary process of protein multiple sequence alignment.

In order to generate the PSSM matrix of evolutionary information, we implement the Position-Specific Iterated BLAST (PSI-BLAST) tool [39] on each protein sequence. PSI-BLAST will return a 20-dimensional vector which indicates the probabilities of conservation against mutations to 20 different amino acids including its own. To get broad and high homologous sequences, in this study, we decide that the value of *e*-value which is *0.001*, the value of iterations is *3*, and matching database is *SwissProt*, respectively. Applications of PSI-BLAST and *SwissProt* database can be downloaded from http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Auto covariance

As one of the most efficient methods for analyzing the sequence of vector statistics, the Auto Covariance (AC) has been widely used in the prediction of secondary structure content [40, 41], protein family classification by researchers [42, 43], and protein interaction prediction [23]. AC variable indicates that in a given protein sequence of two residues average correlation, the expression is:

$$AC(i,lg)=\frac{1}{L-lg}\sum_{i=1}^{L-lg}\left(M_{i,j}-\frac{1}{L}\sum_{i=1}^{L}M_{i,j}\right)\times\left(M_{(i+lg),j}-\frac{1}{L}\sum_{i=1}^{L}M_{i,j}\right) \quad (2)$$

where lg is the distance between residues, i represents the i *th* amino acid, L denotes the length of the protein sequence, $M_{i,j}$ indicates the matrix score of amino acid i at position j.

Using the above expression, the value of AC variable M can be figured out as $M = lg \times N$, where N is the number of descriptors. When all the data in the database complete the operation, each protein sequence was represented as a vector of AC variables and a protein pair was characterized by concatenating the vectors of two proteins in this protein pair.

Feature weighted rotation forest

In this paper, an improved rotation forest algorithm is proposed, which adds the weight selection on the basis of the original rotation forest. This will remove the features of small weight, namely noise, increase the proportion of useful information and improve the accuracy of the classifier. We use χ^2 statistical method to calculate the weight of the features. A feature F against the class feature is calculated as follows:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{2}\frac{(p_{ij} - q_{ij})^2}{q_{i,j}} \tag{3}$$

where n is the number of values in feature F and p_{ij} is the count of the value d_i in feature F belonging to class c_j, defined as:

$$p_{ij} = count\left(F = d_i \; and \; C = c_j\right) \tag{4}$$

$q_{i,j}$ is the expected` value of d_i and c_j, defined as:

$$q_{i,j} = \frac{count(F = d_i) \times count(C = c_j)}{N} \tag{5}$$

where $count(F = d_i)$ is the number of samples in the feature F value d_i, $count(C = c_j)$ is the number of samples in the class C value c_j, and N is the total number of samples in the training set.

In order to make full use of the useful information, we perform the following steps. First, use formula (3) to calculate the weight of each feature; second, descend sort features according to the weight value; finally, select new features from the full feature set in accordance with a given feature selection rate r. After executing these steps, we construct a new data set and use it as the input of the rotation forest.

Rotation forest is a popular ensemble classifier. In order to generate the training samples of the base classifier, the feature set is randomly divided into K subsets. The linear transformation method is applied to each subset and retains all the principal components to maintain the precision of data. The rotation formed the training sample of new features to ensure the diversity of data. Therefore, the rotation forest can enhance the accuracy for individual classifier and the diversity in the ensemble at the same time.

Assuming that $\{x_i, y_i\}$ contains T training samples in which $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ is an n-dimensional feature vector. Let X be the training sample set, Y the corresponding labels and F the feature set. Then X is $T \times$ n matrix, which is composed of n observation feature vector composition. The feature set is randomly divided into K equal subsets by a suitable factor. Let the number of decision trees be L, then the decision trees in the forest can be represented as $G_1, G_2, \ldots, G_L$. The implementation process of the algorithm is as follows.

(1) Select the suitable parameter K which is a factor of n randomly dividing F into K parts of the disjoint subsets and each subset containing a number of features is n/k.

(2) From the training data set X, select the corresponding column of the feature in the subset $G_{i,j}$ to form a new matrix $X_{i,j}$ followed by a bootstrap subset of objects extracted *75%* of X constituting a new training set $X'_{i,j}$.

(3) Matrix $X'_{i,j}$ is used as the feature transform for producing the coefficients in a matrix $M_{i,j}$, in which the jth column coefficient is considered as the characteristic component jth.

(4) The coefficients obtained in the matrix $M_{i,j}$ are constructed a sparse rotation matrix P_i, which is expressed as follows:

$$P_i = \begin{bmatrix} r_{i,1}^{(1)}, \ldots, r_{i,1}^{(N_1)} & 0 & \cdots & 0 \\ 0 & r_{i,2}^{(1)}, \cdots, r_{i,2}^{(N_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{i,k}^{(1)}, \ldots, r_{i,k}^{(N_k)} \end{bmatrix} \tag{6}$$

In the prediction period, the test sample x is provided and generated by the classifier G_i of $d_{i,j}(XP_i^r)$ to determine x belonging to class y_i. Next, the class of confidence is calculated by means of the average combination, and the formula is as follows:

$$\alpha_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{i,j}(XP_i^r) \tag{7}$$

Then assign the category with the largest $\alpha_j(x)$ value to x.

Performance evaluation

In this experiment, we use the prediction accuracy (Accu.), sensitivity (Sen.), Specificity (Spe.), and Matthews Correlation Coefficient (MCC) as the evaluation criterion to assess the performance of our method, they are defined as:

$$\text{Accu.} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Sen.} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{Spe.} = \frac{TN}{TN + FP} \quad (10)$$

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

where TP, TN, FP, FN represent the number of true positives, true negatives, false positives and false negatives, respectively. Moreover, the receiver operating characteristic (ROC) curve [44] is used to visually display the performance of the classifier. The area under the ROC curves (AUC) is also calculated as an indicator of assessment.

ACKNOWLEDGMENTS AND FUNDING

ZHY was supported by National Natural Science Foundation of China under Grant No. 61572506, Pioneer Hundred Talents Program of Chinese Academy of Sciences. LW was supported by National Natural Science Foundation of China under Grant No. 61702444.

CONFLICTS OF INTEREST

None.

REFERENCES

1. Xenarios I, Salwinski L, Duan XQJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research. 2002; 30:303–305.
2. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, et al. Human protein reference database - 2006 update. Nucleic Acids Research. 2006; 34:D411–D414.
3. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. MINT: the molecular INTeraction database. Nucleic Acids Research. 2007; 35:D572–D574.
4. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, et al. IntAct - open source resource for molecular interaction data. Nucleic Acids Research. 2007; 35:D561–D565.
5. Katsamba P, Carroll K, Ahlsena G, Bahna F, Vendome J, Posy S, Rajebhosale M, Price S, Jessell TM, Ben-Shaul A, Shapiro L, Honig BH. Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:11594–11599.
6. Hattori T, Ohoka N, Inoue Y, Hayashi H, Onozaki K. C/EBP family transcription factors are degraded by the proteasome but stabilized by forming dimer. Oncogene. 2003; 22:1273–1280.
7. Baisamy L, Jurisch N, Diviani D. Leucine zipper-mediated homo-oligomerization regulates the Rho-GEF activity of AKAP-Lbc. Journal of Biological Chemistry. 2005; 280:15405–15412.
8. Koike R, Kidera A, Ota M. Alteration of oligomeric state and domain architecture is essential for functional transformation between transferase and hydrolase with the same scaffold. Protein Science. 2009; 18:2060–2066.
9. Woodcock JM, Murphy J, Stomski FC, Berndt MC, Lopez AF. The dimeric versus monomeric status of 14-3-3 zeta is controlled by phosphorylation of Ser(58) at the dimer interface. Journal of Biological Chemistry. 2003; 278:36323–36327.
10. Ispolatov I, Yuryev A, Mazo I, Maslov S. Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Research. 2005; 33:3629–3635.
11. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. Evolution of protein complexes by duplication of homomeric interactions. Genome Biology. 2007; 8.
12. Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. Trends in Biochemical Sciences. 2004; 29:618–625.
13. Kuncheva LI. Combining Pattern Classifiers: Methods and Algorithms. Technometrics. 2005; 47:517–518.
14. Gao ZG, Wang L, Xia SX, You ZH, Yan X, Zhou Y. Ens-PPI: A Novel Ensemble Classifier for Predicting the Interactions of Proteins using Auto Covariance Transformation from PSSM. BioMed Research International,2016,(2016-6-29). 2016; 2016:1–8.
15. Zhou Y, Zhou YS, He F, Song J, Zhang Z. Can simple codon pair usage predict protein-protein interaction? Molecular Biosystems. 2012; 8:1396–1404.
16. Zaki N, Lazarovamolnar S, Elhajj W, Campbell P. Protein-protein interaction based on pairwise similarity. Bmc Bioinformatics. 2009; 10:1–12.
17. You ZH, Lei YK, Zhu L, Xia J, Wang B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. Bmc Bioinformatics. 2013; 14:1–11.
18. Rodriguez JJ, Kuncheva LI. Rotation forest: A new classifier ensemble method. Ieee Transactions on Pattern Analysis and Machine Intelligence. 2006; 28:1619–1630.
19. Lei W, You ZH, Xia SX, Feng L, Xing C, Xin Y, Yong Z. Advancing the Prediction Accuracy of Protein-Protein Interactions by Utilizing Evolutionary Information from Position-Specific Scoring Matrix and Ensemble Classifier. Journal of Theoretical Biology. 2017; 418:105–110.
20. Nanni L, Lumini A. Ensemble generation and feature selection for the identification of students with learning disabilities. Expert Systems with Applications. 2009; 36:3896–3900.
21. Wang L, You ZH, Xia SX, Chen X, Yan X, Zhou Y, Liu F. An improved efficient rotation forest algorithm to predict

the interactions among proteins. Soft Computing. 2017; 1–9.

22. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences of the United States of America. 1987; 84:4355–4358.
23. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. Nucleic Acids Research. 2008; 36:3025–3030.
24. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. Acm Transactions on Intelligent Systems and Technology. 2011; 2.
25. Liu Z, Guo F, Zhang J, Wang J, Lu L, Li D, He F. Proteome-wide prediction of self-interacting proteins based on multiple properties. Molecular & cellular proteomics: MCP. 2013; 12:1689–1700.
26. Liu X, Yang S, Li C, Zhang Z, Song J. SPAR: a random forest-based predictor for self-interacting proteins with fine-grained domain information. Amino Acids. 2016; 48:1655–1665.
27. Du XJ, Zheng T, Zheng D, Qian F. A Novel Feature Extraction Scheme with Ensemble Coding for Protein–Protein Interaction Prediction. International Journal of Molecular Sciences. 2014; 15:12731–12749.
28. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPIevo: Protein-protein interaction prediction from PSSM based evolutionary information. Genomics. 2013; 102:237–242.
29. Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, Masoudi-Nejad A. LocFuse: Human protein–protein interaction prediction via classifier fusion using protein localization information. Qrevchemsoc. 2014; 104:496–503.
30. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, Antunes R, Arganiska J, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, et al. UniProt: a hub for protein information. Nucleic Acids Research. 2015; 43:D204–D212.
31. Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Research. 2015; 43:D321–D327.
32. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock REW, Brinkman FSL, Lynn DJ. InnateDB: systems biology of innate immunity and beyond-recent updates and continuing curation. Nucleic Acids Research. 2013; 41:D1228–D1233.
33. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, et al. The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research. 2014; 42:D358–D363.
34. Chatr-aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Research. 2015; 43:D470–D478.
35. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Research. 2004; 32:D449–D451.
36. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology. 1999; 292:195–202.
37. Chen XW, Jeong JC. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics. 2009; 25:585–591.
38. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins-Structure Function and Bioinformatics. 2003; 53:573–578.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; 25:3389–3402.
40. Lin Z, Pan XM. Accurate prediction of protein secondary structural content. Journal of Protein Chemistry. 2001; 20:217–220.
41. Zhang CT, Lin ZS, Zhang Z, Yan M. Prediction of the helix/strand content of globular proteins based on their primary sequences. Protein engineering. 1998; 11:971–979.
42. Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Science. 2002; 11:795–805.
43. Guo Y, Li M, Lu M, Wen Z, Huang Z. Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins-Structure Function and Bioinformatics. 2006; 65:55–60.
44. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical chemistry. 1993; 39:561–577.
45. Du X, Cheng J, Zheng T, Duan Z, Qian F. A Novel Feature Extraction Scheme with Ensemble Coding for Protein-Protein Interaction Prediction. International Journal of Molecular Sciences. 2014; 15:12731–12749.