Research Article

Self-Interacting Proteins Prediction from PSSM Based on Evolutionary Information

Zheng Wang , Yang Li, Li-Ping Li , Zhu-Hong You , and Wen-Zhun Huang

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

Correspondence should be addressed to Li-Ping Li; cs2bioinformatics@gmail.com and Zhu-Hong You; zhuhongyou@gmail.com

Received 16 October 2020; Revised 8 January 2021; Accepted 17 February 2021; Published 10 March 2021

Academic Editor: Jianping Gou

Copyright © 2021 Zheng Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Self-interacting proteins (SIPs) play an influential role in regulating cell structure and function. Thus, it is critically important to identify whether proteins themselves interact with each other. Although there are some existing experimental methods for self-interaction recognition, the limitations of these methods are both expensive and time-consuming. Therefore, it is very necessary to develop an efficient and stable computational method for predicting SIPs. In this study, we develop an effective computational method for predicting SIPs based on rotation forest (RF) classifier, combined with histogram of oriented gradients (HOG) and synthetic minority oversampling technique (SMOTE). When performing SIPs prediction on yeast and human datasets, the proposed method achieves superior accuracies of 97.28% and 89.41%, respectively. In addition, the proposed approach was compared with the state-of-the-art support vector machine (SVM) classifiers and other different methods on the same datasets. The experimental results demonstrate that our method has good robustness and effectiveness and can be regarded as a useful tool for SIPs prediction.

1. Introduction

Protein is the material basis of life and an important part of all cells and tissues. The study of protein-protein interactions has become a fundamental problem. These interactions can elucidate the mechanism of biological reactions and also play a crucial role in living organisms. In recent years, protein-protein interactions (PPIs) have become a hot spot for many researchers because they are the key to further understanding of protein functions. [1] Self-interacting proteins (SIPs) are a special problem of PPIs, referring to the interactions between two or more copies of the protein. At the same time, copies of proteins from the same gene will likely lead to the formation of homo-oligomers. Among them, a large number of studies have shown that homo-oligomerization has a wide range of promoting effects on biological development, such as enzyme activation, gene expression regulation, signal transduction, and immune response. SIPs also play an important role in the study of protein interaction networks (PINs). Specifically, they can interact with other protein partners and thus play an important role in the study of biological cell systems [2]. On the contrary, the

genes of SIPs have a good superiority, which is reflected in the high duplicability of coding genes compared with other genes. At the same time, these genes are often displayed on a large scale at the level of the whole genome [3]. Although there is a lot of theoretical knowledge about SIPs in the field of biological proteomics, few machine learning methods have been used to study this problem [4, 5]. Due to the high cost and time consumption of traditional SIPs prediction experiments, it is necessary to develop a reliable and efficient computational method for detecting SIPs.

So far, a large number of different computational methods based on biological sequences have been proposed for the prediction of PPIs. For instance, Zhang et al. [6] proposed an ensemble deep neural networks approach which used different configurations based on each descriptor to train deep neural networks. Jha et al. [7] extracted the features from the 3D voxel representation of proteins and the amino acid sequences, respectively, and then concatenated these features in pairs. Finally, the bidirectional GRU-based classifier used the concatenated features to predict PPIs. Jia et al. [8] used a discrete wavelet transform (DWT) to extract the corresponding protein information and

obtained a variety of physicochemical descriptors. The random forests model was used to classify the prediction results. The experimental results fully demonstrate that this method can effectively predict PPIs. Sumonja et al. [9] developed a stacking algorithm GA-STACK for automatic ensembling of different machine learning algorithms, and this algorithm achieves a good prediction performance. Tain et al. [10] proposed an efficient prediction method based on multi-information fusion, in which the pseudoamino acid composition (PseAAC), autocovariance (AC), and encoding based on grouped weight (EBGW) extracted the features of protein sequences. By two-dimensional wavelet denoising, the denoised features are input to the support vector machine (SVM) for classification. Xu et al. [11] proposed a prediction method based on protein sequences, in which the method used graph energy to extract the effective information between proteins, and then sent the information to the weighted sparse representation-based classification (WSRC) classifier for predicting PPIs. Wang et al. [12] proposed a method which called novel stochastic block model. This method can capture the latent structural features of proteins from the perspective of forming protein complexes by simulating the generative process of the protein interaction network. In summary, these methods are mainly used to study the interaction between protein pairs with related information such as co-localization, co-expression, and co-evolution, and they cannot be used to predict SIPs. Among them, the most important reason is that the datasets used by these methods do not contain protein information between the same partners. Given the impact of such limitations, these methods are not desirable in predicting SIPs. Therefore, it is very important that an efficient and practical computing method be developed for the study of large-scale SIPs.

In this study, a method for predicting SIPs from a PSSM of protein sequences is proposed, which combines the HOG feature extraction method with SMOTE and an RF classifier. In this experiment, two SIPs datasets, yeast and human, were used to evaluate the predictive performance of the proposed method. These experimental results further indicate that the proposed method can be effectively used to identify self-interacting proteins. Among them, the proposed method achieved high prediction accuracy on yeast and human datasets, which were 97.28% and 89.41%, respectively. At the same time, our method was compared to the most popular support vector machine (SVM) classifiers and other different methods on the same yeast and human datasets. All the experimental results show that the proposed method has good robustness and effectiveness and can be used to predict SIPs.

2. Materials and Methodology

2.1. Datasets. First, there are 20,199 reliable curated human protein sequences available from the published UniProt database [13]. Second, these protein sequences were also obtained from other datasets in different ways, including BioGRID, [14] DIP, [15] InnateDB, [16] IntAct, [17] and MatrixDB [18]. In this experiment, those PPIs derived from the same PPIs datasets and containing only two interacting protein partners were extracted, and the PPIs obtained in

this way were considered as “direct interaction.” Finally, a total of 2,994 human self-interaction protein sequences were calculated. In order to further better test the accuracy of the proposed method, these human self-interaction proteins were processed by using the following measures. (1) Protein sequences with a length of more than 50 and less than 5,000 residues in the entire human proteome are retained. (2) In order to construct an accurate and reliable SIPs positive dataset, the selected data must meet one of the following requirements: (a) the data should be at least reliable experimental data obtained from experiments that were applied to a small-scale or two large-scale tests for self-interactions; (b) the annotation of proteins in the UniProt protein database is a homo-oligomer, which also includes homodimers and homotrimers; (c) the SIPs positive dataset should be at least from the experimental data reported in the 2 publications. (3) The annotated SIPs data in the entire human proteome were deleted. Its main purpose was to construct a negative dataset. Thus, after such processing, a total of 17,379 proteins were used to construct human SIPs datasets, of which 1441 human SIPs were defined as positive datasets and 15,938 human non-SIPs were defined as negative datasets. In addition, in order to further verify the validity of our method on other datasets, we also used the same data processing measures to construct a yeast dataset with 710 positive SIPs and 5511 negative non-SIPs. Table 1 shows the summary of the datasets used in this experiment, and the positive samples were experimental identified samples.

2.2. Position Specific Scoring Matrix (PSSM). The PSSM can detect distantly related proteins by sequence comparison. At the same time, this PSSM information is established based on a previous set of protein sequences that have been arranged by the structure or sequence similarity. [19] In addition, the most commonly used homologous search Position-Specific Iterated BLAST (PSI-BLAST) tools [20] have also been applied to this experiment, and its purpose is to convert each protein sequence into a PSSM. Here, each PSSM representation can be given. Specifically, it is an $R \times 20$ matrix $Q = \{Q_{ij}: i = 1, \dots, R, j = 1, \dots, 20\}$, where the number of Q matrix row vectors is used to represent the size of the protein sequence and the number of Q matrix column vectors is used to refer to 20 amino acids. The position specific score Q_{ij} is used to specify the score of the j th amino acid at the i th position, which is expressed as $Q_{ij} = \sum_{k=1}^{20} m(i, k) \times n(j, k)$, where $m(i, k)$ represents the value of the frequency at which the k th amino acid in the protein appears at position i , and $n(j, k)$ is Dayhoff's matrix consisting of mutation values between the j th amino acid and the k th amino acid [21].

In order to obtain biological evolutionary information of higher homologous proteins in this study, the PSI-BLAST tool was used to search for important information in each protein sequence and convert it into PSSM for further use in the prediction of SIPs. Among them, the parameter e -value of PSI-BLAST is set to 0.001 and the number of iterations is set to 3. Each PSSM is eventually converted to a matrix of

TABLE 1: Summary of the datasets used in this experiment.

Dataset	Protein pairs	Positive samples	Negative samples
Yeast	6221	710	5511
Human	17379	1441	15938

$S \times 20$, where the value of row vector S represents the total number of residues in each protein and 20 represents the number of amino acids. The website address of the PSI-BLAST tool is <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

2.3. Histogram of Oriented Gradients (HOG). The HOG feature descriptor was first applied to human detection by Dalal and Triggs [22]. Among them, the HOG feature descriptor describes the local appearance and shape of the image according to the gradient density distribution. The descriptor is created by dividing the original image into small cell regions. Each cell builds a histogram of the gradient direction for each pixel inside it. [23] At the same time, the important features are obtained by calculating the HOG of the image local area and then the HOG algorithm is used to identify the target object to be detected. So far, HOG descriptors have been widely used for various computer vision problems. In most cases, they are combined with machine learning methods to solve specific practical problems. For example, face recognition, facial expression recognition, vehicle detection, and pedestrian detection in video. In addition, because of its good performance in some complex problems, HOG as an attribute extraction method has attracted the attention of many researchers and is widely used in a large number of literature. Therefore, HOG is used to extract information of the target edge in a local region of the original image, and this has become an important means of referencing certain target shapes. [24] The description of HOG feature extraction is as follows.

Using the HOG algorithm to perform gradient calculation on the original image is a crucial step. The gradient direction of each pixel in the image can be further determined based on the obtained gradient values of the abscissa and the ordinate. Its main purpose is to obtain the contour information of the image so as to further avoid the influence of light on the experimental result. [25] The gradient calculation formula is as follows:

$$G_a(a, b) = H(a + 1, b) - H(a - 1, b), \quad (1)$$

$$G_b(a, b) = H(a, b + 1) - H(a, b - 1), \quad (2)$$

where $G_a(a, b)$ represents the horizontal gradient of the position (a, b) in the original image, $G_b(a, b)$ indicates the vertical gradient of the position (a, b) in the original image, and $H(a, b)$ refers to the pixel value of the position (a, b) in the original image. Then, the convolution kernel $[-1, 0, 1]$ is used to perform a convolution operation with the original image to obtain the gradient component $G_a(a, b)$ in the horizontal direction, and then, the convolution operation is performed on the original image through another convolution kernel $[1, 0, -1]^T$ to further obtain the gradient component $G_b(a, b)$ in the vertical direction.

The next step is to calculate the gradient magnitude and orientation through the image's horizontal gradient $G_a(a, b)$ and vertical gradient $G_b(a, b)$. They are expressed as follows:

$$m(a, b) = G_a(a, b)^2 + G_b(a, b)^2, \quad (3)$$

$$\theta(a, b) = \frac{180}{\pi} \left(\tan^{-1} \left(\frac{G_b(a, b)}{G_a(a, b)} \right) \right) + \pi, \quad (4)$$

where $m(a, b)$ represents the gradient magnitude of point (a, b) in the image and $\theta(a, b)$ denotes the gradient orientation of point (a, b) in the image. In this experiment, each PSSM of the protein sequence was divided into 8×8 size cells with an overlap ratio of 1/2. Each of the cells has a gradient orientation of 0–180° and is divided into 9 bins; specifically, a direction bin is 20° so that the pixel magnitude in each direction can be obtained.

Since the illumination and contrast factors have a large influence on the gradient magnitude, the gradient magnitude can be normalized. The HOG feature in each cell is further extracted by normalizing all cell values in the block. The block normalization formula is as follows:

$$u = \frac{u}{\sqrt{\|u\|_2^2 + \epsilon^2}}, \quad (5)$$

where u is the unnormalized vector of the obtained histogram, $\|u\|_2$ is used to refer to the L2-norm of the vector u , and ϵ is a particularly small constant value used for zero division.

Therefore, we use the HOG algorithm to extract the HOG features on each PSSM for a given protein sequence. In order to avoid the influence of noise on the experimental results, we further use the singular value decomposition (SVD) method to process the extracted HOG features to obtain significant features. Finally, the protein sequences of the yeast dataset and the human dataset were converted into 180-dimensional vectors for prediction of SIPs experiments.

2.4. Synthetic Minority Oversampling Technique (SMOTE).

SMOTE is a technique for oversampling the minority classes to balance datasets. This method was proposed by Chawla et al. [26]. The main idea is to form new minority samples by interpolating and selecting a neighbor of a minority class. This technique is capable of generating synthetic sample data rather than simply copying minority class samples. Thus, this method can avoid overfitting problem. The SMOTE algorithm is described as follows. The first step randomly selects a point x in the minority class sample and obtains its k nearest neighbors from other minority samples. Then, get one minority class sample $\tilde{x}$ among the k neighbors. Finally, an interpolation method is used to generate a new synthetic sample x_{new} in the vector between x and $\tilde{x}$. It can be concluded that the mathematical expression of SMOTE is as follows:

$$x_{\text{new}} = x + (\tilde{x} - x) \times \text{rand}(0, 1), \quad (6)$$

where $\text{rand}(0, 1)$ represents a random number between 0 and 1. The above process is then repeated based on the beginning of the set of oversampling rates, and a new

minority class sample is added to the original training samples using the SMOTE algorithm to increase the number of minority class samples. This method can be seen as an interpolation between two minority class samples from a geometric perspective. It expands the decision space for the minority class, which allows the classifier to make higher predictions for unknown minority class samples. The SMOTE method is efficient and simple in generating synthetic samples, which avoids overfitting problems. In our datasets, the positive samples are few. In order to prove the effectiveness of the SMOTE method in Section 3, we performed the SMOTE method after the five-fold cross-division to expand the positive samples of training dataset. Finally, the new training dataset is entered into the classifier for classification, and the prediction results are obtained. The process of SMOTE is shown in Figure 1.

2.5. Rotation Forest (RF). Rotation forest, an ensemble classifier based on the bagging algorithm, was originally introduced by Rodriguez et al. [27]. The idea of rotation forest classifier comes from random forest, but the main difference is that it can build accurate and diverse base classifiers. RF has been widely used in machine learning to perform classification on practical problems. For example, cancer classification [28], sequence-based protein-protein interactions' prediction [29], and diagnosis of neuromuscular disorders [30]. The rotation forest framework is described as follows. Assuming N training samples are composed of $\{x_i, y_i\}$, where $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ is a D -dimensional feature vector. Suppose the training sample set consists of $X = (x_1, x_2, \dots, x_n)^T$ ($n \times D$ matrix), which contains n observation feature vectors, S be the feature set, and $Y = (y_1, y_2, \dots, y_n)^T$ be the corresponding class labels. Here, we assume that the number of decision trees in the RF algorithm is L , denoted as $R_1, R_2, \dots, R_L$, respectively. The processing steps for a single classifier R_i can be described as follows. The appropriate parameters K is selected and the feature set S is randomly divided into K disjoint subsets, wherein the feature number contained in each feature subset is $C = n/K$. Next, the corresponding column of the feature in the subset $R_{i,j}$ is selected from the training sample set X to form a new matrix $X_{i,j}$. Using the bootstrap method, we can extract three-quarters of the dataset from X to form a new nonempty training set $X'_{i,j}$. After that, a principal component analysis method is applied to $X'_{i,j}$ to obtain coefficients in a matrix $M_{i,j}$, where the j th column coefficient is used as the characteristic component j th. Finally, a sparse rotation matrix R_m is constructed, and its coefficients stored in the matrix $M_{i,j}$ are represented as follows:

$$R_m = \begin{bmatrix} \lambda_{i,1}^{(1)}, \dots, \lambda_{i,1}^{(C_1)} & 0 & \dots & 0 \\ 0 & \lambda_{i,1}^{(1)}, \dots, \lambda_{i,1}^{(C_2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_{i,K}^{(1)}, \dots, \lambda_{i,K}^{(C_K)} \end{bmatrix}. \quad (7)$$

FIGURE 1: The schematic of SMOTE algorithm.

In the classification problem, for a given test sample x , $d_{i,j}(xR_m^{\lambda})$ is produced by the classifier R_i to determine that x belongs to class y_i . Then, the average combination method is used to calculate the confidence of a class:

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^L d_{i,j}(xR_m^{\lambda}). \quad (8)$$

Finally, x is assigned to the class with the greatest confidence. The pseudocodes of rotation forest algorithm (Algorithm 1) can be summarized as follows.

3. Results and Discussion

3.1. Performance Evaluation. In order to further test the accuracy and reliability of the proposed method, the values of the five evaluation indicators were calculated: Accuracy (Ac), Sensitivity (Sn), Specificity (Sp), and Matthew correlation coefficient (MCC). They are defined as follows:

$$Ac = \frac{TP + TN}{TP + TN + FP + FN},$$

$$S_n = \frac{TP}{FN + TP},$$

$$S_p = \frac{TN}{FP + TN},$$

$$MCC = \frac{TP + TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \quad (9)$$

where true positive TP represents the number of true SIPs correctly predicted by the model, false positive FP refers to the number of non-self-interacting proteins predicted to be self-interacting by the model, true negative TN replaces the number of true non-self-interacting proteins correctly

predicted by the model, and false negative FN was used to represent the number of self-interacting proteins predicted to be non-self-interacting by the model. In addition, to better demonstrate the predictive results of our model, Receiver Operating Characteristics (ROC) [31] curves were also used to evaluate the performance of the proposed method.

3.2. Performance of the Proposed Method. In this experiment, the proposed method was applied to yeast and human datasets for detection of SIPs. In order to avoid overfitting of the prediction results to better test the validity and stability of our model, the datasets used in the experiment were divided into training datasets and independent test datasets. At the same time, a five-fold crossvalidation method was also used in this experiment to further evaluate the performance of the proposed method. The workflow diagram of the five-fold crossvalidation is shown in Figure 2.

In addition, in order to ensure the fairness of the experimental results, the two main parameters of the RF classifier were optimized, and the same parameter settings were used for SIPs experiments of yeast and human datasets, where the parameter K was set to 5 and L was set to 13. Here, K is the number of feature subsets and L is the number of decision trees. The predicted results of the original method and the proposed method for SIPs experiments on yeast datasets are shown in Tables 2 and 3.

From Table 2, we can clearly see that the RF model has achieved excellent prediction results on the yeast dataset, and the accuracy of the five experiments is higher than 97%. Specifically, their accuracy is 97.99%, 97.19%, 97.83%, 97.35%, and 97.59%, respectively. Among them, the average accuracy, specificity, sensitivity, and MCC were 97.59%, 99.47%, 82.87%, and 87.78%, respectively. Their standard deviations are 0.33%, 0.28%, 2.42%, and 1.92%, respectively. From Table 3, when using the SMOTE and RF models to predict SIPs on the yeast dataset, the average accuracy, specificity, sensitivity, and MCC were 97.28%, 98.02%, 91.50%, and 87.34%, respectively. Although the SMOTE and RF method also achieved 97% accuracy, the average sensitivity increased by 8.63%. At the same time, to better demonstrate and visualize the performance of the model, ROC curves on the yeast dataset are shown in Figures 3 and 4. X-rays on both figures represent the false positive rate (FPR) and y-rays indicate the true positive rate (TPR).

Similarly, the same model is applied to the human dataset to perform SIPs prediction. The predicted results of the original method and the proposed method for SIPs experiments on human datasets are shown in Tables 4 and 5. It is worth mentioning that the proposed method also achieved an average accuracy of more than 89% on the human dataset. As shown in Table 4, when exploring SIPs on the human dataset, the average accuracy, specificity, sensitivity, and MCC of 91.68%, 99.92%, 0.56%, and 5.08% were obtained by using the RF model. The standard deviations of these evaluation indicators were 0.59%, 0.04%, 0.21%, and 1.14%, respectively. As shown in Table 5, when predicting SIPs on the human dataset, the average accuracy, specificity, sensitivity, and MCC of 89.41%, 96.44%, 11.69%, and 19.76%

were achieved by using the SMOTE and RF models. From this comparison, we can see that although the average accuracy decreased by 0.0227, the average sensitivity and average MCC increased by 0.1113 and 0.1468, respectively. This is benefited from the combination of the SMOTE method and the rotation forest model.

It can be seen that the superior feature extraction methods and the choice of classifier can effectively help us to improve the prediction accuracy of SIPs experiments. In addition, our method can obtain such good prediction results mainly depending on the following aspects. (1) PSSM not only can obtain effective information of protein sequences but also retains important prior information, which plays a crucial role in predicting SIPs experiments. (2) Since the Histogram of Oriented Gradients (HOG) can extract effective local texture features and maintain good invariance in both geometric and optical deformation of the image, therefore, the HOG feature extraction method can be used to acquire important protein information from the PSSM. (3) It is very necessary to choose a reliable and robust rotation forest classifier to classify the sample features. (4) The SMOTE method can effectively solve the problem of low sensitivity caused by unbalanced samples. Finally, these good experimental results fully demonstrate that the proposed method can predict SIPs efficiently and accurately. Figures 5 and 6 show the ROC curves performed by the proposed method on the human dataset.

3.3. Comparison with SVM-Based Method. A large number of machine learning models have been proposed for detecting interactions between proteins. There is no doubt in saying that one of the most popular classifiers in the field of machine learning is the support vector machine (SVM). In order to better test the performance of the proposed method, we used the state-of-the-art SVM classifier to perform SIPs experiments on yeast and human datasets to better compare the superiority of the two classifiers. Among them, the same protein data and feature extraction methods were applied to this experiment. In addition, the LibSVM toolbox was also used in this experiment to complete the SVM classification task, which can be downloaded from <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>. [32] Here, the radial basis function (RBF) is used for the kernel function of the SVM and the grid search method is also used to determine the values of the main parameters c and g , where c is set to 0.1 and g is set to 0.9. The predicted results of SIPs experiments on yeast and human datasets by using the SMOTE and SVM method are shown in Tables 6 and 7.

From Table 6, we can easily see that the accuracy of the five experiments predicted on the yeast dataset using the SVM classifier was 93.09%, 92.60%, 94.53%, 94.29%, and 94.22%, respectively. The preformed averages of accuracy, specificity, sensitivity, and MCC are 93.75%, 94.14%, 90.72%, and 75.48%, slightly lower than those performed by the proposed method, which are 97.28%, 98.02%, 91.50%, and 87.34%, respectively. Similarly, from Table 7, we can also see that the average accuracy, specificity, sensitivity, and MCC predicted by the SVM classifier on the human dataset are 85.49%, 91.51%, 18.89%, and 22.88%, respectively. However,

Training Phase

Input

X : the training samples.

Y : the labels of training samples.

L : the ensemble size of classifiers.

K : the number of subsets.

R : the proportion of resampling new samples from original samples ($R=0.75$).
for $i = 1, \dots, L$

*construct sparse rotation matrix R_m .

*divided the total samples into K disjoint subsets randomly.

for $j = 1, \dots, K$

form a new matrix $X_{i,j}$.

using bootstrap algorithm to obtain R proportion subset $X'_{i,j}$.

using PCA on $X'_{i,j}$ to obtain coefficients in a matrix $M_{i,j}$.

build decision tree R_i .

Classification Phase

Input

x : the test samples.

for $n = 1, \dots, L$

*calculate the probability $d_{i,j}(xR_m^l)$ of each classes.

Finally, using the largest average confidence $\mu(x)$ to classification.

ALGORITHM 1: The pseudocodes of rotation forest algorithm.

FIGURE 2: The workflow diagram of the five-fold crossvalidation.

TABLE 2: Five-fold crossvalidation prediction results by using the RF method on the yeast dataset.

Testing set	Ac (%)	Sp (%)	Sn (%)	MCC (%)
1	97.99	99.63	86.00	90.35
2	97.19	99.09	82.76	86.19
3	97.83	99.82	81.88	88.66
4	97.35	99.46	79.55	85.59
5	97.59	99.36	84.14	88.10
Average	97.59 ± 0.33	99.47 ± 0.28	82.87 ± 2.42	87.78 ± 1.92

TABLE 3: Five-fold crossvalidation prediction results by using the SMOTE and RF methods on the yeast dataset.

Testing set	Ac (%)	Sp (%)	Sn (%)	MCC (%)
1	97.67	98.35	92.67	89.49
2	96.62	97.27	91.72	85.06
3	97.59	98.37	91.30	88.31
4	97.19	98.11	89.39	85.93
5	97.35	98.00	92.41	87.90
Average	97.28 ± 0.42	98.02 ± 0.45	91.50 ± 1.30	87.34 ± 1.81

FIGURE 3: ROC curves' performance of the RF method on the yeast SIPs dataset.

FIGURE 4: ROC curves' performance of the SMOTE and RF methods on the yeast SIPs dataset.

TABLE 4: Five-fold crossvalidation prediction results by using the RF method on the human dataset.

Testing set	Ac (%)	Sp (%)	Sn (%)	MCC (%)
1	90.68	99.90	0.31	3.46
2	92.17	99.91	0.74	5.87
3	92.00	99.94	0.72	6.25
4	91.91	99.97	0.36	4.40
5	91.64	99.87	0.69	5.43
Average	91.68 ± 0.59	99.92 ± 0.04	0.56 ± 0.21	5.08 ± 1.14

the RF classifier achieved an average accuracy of 89.41% on the human dataset. Figures 7 and 8 show the ROC curves obtained on yeast and human datasets for the SVM-based

TABLE 5: Five-fold crossvalidation prediction results by using the SMOTE and RF methods on the human dataset.

Testing set	Ac (%)	Sp (%)	Sn (%)	MCC (%)
1	88.43	96.45	9.94	18.60
2	89.93	96.57	11.44	19.14
3	89.04	95.96	9.35	16.60
4	89.58	96.21	14.23	21.98
5	90.08	97.02	13.49	22.47
Average	89.41 ± 0.68	96.44 ± 0.40	11.69 ± 2.14	19.76 ± 2.45

FIGURE 5: ROC curves' performance of the RF method on the human SIPs dataset.

FIGURE 6: ROC curves' performance of the SMOTE and RF methods on the human SIPs dataset.

methods. Comparing these experimental results, we can conclude that the RF classifier is significantly better than the SVM classifier in predicting SIPs experiments. Therefore, the

TABLE 6: Five-fold crossvalidation prediction results by using the SMOTE and SVM method on the yeast dataset.

Testing set	Ac (%)	Sp (%)	Sn (%)	MCC (%)
1	93.09	93.69	88.67	73.98
2	92.60	92.90	90.34	72.63
3	94.53	94.85	92.03	77.58
4	94.29	94.78	90.15	75.78
5	94.22	94.45	92.41	77.41
Average	93.75 ± 0.85	94.14 ± 0.83	90.72 ± 1.52	75.48 ± 2.16

TABLE 7: Five-fold crossvalidation prediction results by using the SMOTE and SVM method on the human dataset.

Testing set	Ac (%)	Sp (%)	Sn (%)	MCC (%)
1	85.58	92.39	18.94	24.40
2	85.04	90.54	19.93	22.62
3	85.53	91.37	18.35	22.12
4	85.27	91.01	19.93	23.16
5	86.03	92.26	17.30	22.09
Average	85.49 ± 0.37	91.51 ± 0.80	18.89 ± 1.12	22.88 ± 0.96

FIGURE 7: ROC curves' performance of the SMOTE and SVM method on the yeast SIPs dataset.

proposed prediction model can be well applied to the prediction of SIPs, which may promote the research of other bioinformatics tasks.

3.4. Comparison with Other Methods. So far, an increasing number of calculation methods have been developed for predicting PPIs. Furthermore, in order to test the reliability and effectiveness of the proposed model, the following methods SLIPPER, CRS, SPAR [33], DXECPPI [34], PPIevo [35], and LocFuse [36] were used for comparison, and these methods were proposed based on the same yeast and human datasets. The prediction results obtained on the yeast and human datasets by using these seven methods are shown in Tables 8 and 9, respectively. It can be seen from Table 8 that the average prediction accuracy obtained by the six different

FIGURE 8: ROC curves' performance of the SMOTE and SVM method on the human SIPs dataset.

TABLE 8: Predictive performance of seven different methods on yeast dataset.

Model	Ac (%)	Sp (%)	Sn (%)	MCC (%)
SLIPPER	71.90	72.18	69.72	28.42
DXECPPI	87.46	94.93	29.44	28.25
PPIevo	66.28	87.46	60.14	18.01
LocFuse	66.66	68.10	55.49	15.77
CRS	72.69	74.37	59.58	23.68
SPAR	76.96	80.02	53.24	24.84
Proposed method	97.28	98.02	91.50	87.34

TABLE 9: Predictive performance of seven different methods on human dataset.

Model	Ac (%)	Sp (%)	Sn (%)	MCC (%)
SLIPPER	91.10	95.06	47.26	41.97
DXECPPI	30.90	25.83	87.08	8.25
PPIevo	78.04	25.82	87.83	20.82
LocFuse	80.66	80.50	50.83	20.26
CRS	91.54	96.72	34.17	36.33
SPAR	92.09	97.40	33.33	38.36
Proposed method	89.41	96.44	11.69	19.76

methods on the yeast dataset is between 66.28% and 87.46%, and these accuracies are lower than the prediction accuracy of the proposed method. Similarly, the specificity, sensitivity, and MCC values of these methods are also lower than the predicted values obtained by the proposed method. At the same time, Table 9 shows the average accuracy, specificity, sensitivity, and MCC values obtained by seven different methods on the human dataset. Among them, the proposed method has a prediction accuracy of 89.41% on the human dataset, which also achieves a better prediction performance. The experimental results obtained by further comparing these methods can fully demonstrate that the proposed method can effectively predict the interaction between SIPs.

4. Conclusions

In this study, an effective computational method was developed for predicting SIPs by using sequence information of proteins. Among them, the proposed method is based on an RF classifier combined with a position specific scoring matrix (PSSM), HOG, and SMOTE methods. In the first step, all protein sequences are transformed into PSSM, which contains evolutionary information about the protein sequence. Secondly, by using the HOG feature extraction method to obtain valuable protein features from PSSM, the final feature is further processed by the SVD method to reduce the interference of noise in the experimental results. To address the sample class imbalance problem of self-interacting proteins, we used the SMOTE method to process the HOG features of the obtained proteins. Finally, a reliable RF classifier is used to predict whether a given protein sequence can interact with itself. In addition, in order to better evaluate the performance of the proposed model, our method was compared to the most popular support vector machine (SVM) classifiers and other different methods on the same yeast and human datasets. All the experimental results show that the proposed method has good robustness and effectiveness and can be used to predict SIPs, which will be beneficial to the future research of bioinformatics tasks.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

Zheng Wang and Yang Li are the co-first authors.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Zheng Wang and Yang Li contributed equally to this work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China under Grant nos. 62072378 and 61873212.

References

- [1] Y. Katoh, K. Nakamura, and K. Nakayama, "Visible immunoprecipitation (VIP) assay: a simple and versatile method for visual detection of protein-protein interactions," *Bio-protocol*, vol. 8e2687, pages, 2018.
- [2] C. Chen, Q. Zhang, Q. Ma, B. Yu, and I. L. Systems, "LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion," *Chemometrics and Intelligent Laboratory Systems*, vol. 191, pp. 54–64, 2019.
- [3] A. Perez-Bercoff, T. Makino, and A. McLysaght, "Duplicability of self-interacting human genes," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 160, 2010.
- [4] J. Qu, Y. Zhao, L. Zhang, S.-B. Cai, Z. Ming, and C.-C. Wang, "Computational models for self-interacting proteins prediction," *Protein & Peptide Letters*, vol. 27, no. 5, pp. 392–399, 2020.
- [5] J.-Y. An, Y. Zhou, Z.-J. Yan, and Y. Zhao, "Predicting self-interacting proteins using a recurrent neural network and protein evolutionary information," *Evolutionary Bioinformatics*, vol. 16, Article ID 1176934320924674, 2020.
- [6] 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.
- [7] K. Jha, S. Saha, and M. Khushi, "Protein-protein interactions prediction based on Bi-directional gated recurrent unit and multimodal representation," in *Proceedings of the International Conference on Neural Information Processing*, pp. 164–171, Springer, Berlin, Germany, December 2020.
- [8] J. Jia, X. Xiao, and B. Liu, "Prediction of protein-protein interactions with physicochemical descriptors and wavelet transform via random forests," *Journal of Laboratory Automation*, vol. 21, no. 3, pp. 368–377, 2016.
- [9] N. Sumonja, B. Gemovic, N. Veljkovic, and V. Perovic, "Automated feature engineering improves prediction of protein-protein interactions," *Amino Acids*, vol. 51, no. 8, pp. 1187–1200, 2019.
- [10] B. Tian, X. Wu, C. Chen, W. Qiu, Q. Ma, and B. Yu, "Predicting protein-protein interactions by fusing various chou's pseudo components and using wavelet denoising approach," *Journal of Theoretical Biology*, vol. 462, pp. 329–346, 2019.
- [11] D. Xu, H. Xu, Y. Zhang, W. Chen, and R. Gao, "Protein-protein interactions prediction based on graph energy and protein sequence information," *Molecules*, vol. 25, no. 8, p. 1841, 2020.
- [12] X. Wang, P. Hu, and L. Hu, "A novel stochastic block model for network-based prediction of protein-protein interactions," in *Proceedings of the International Conference on Intelligent Computing*, pp. 621–632, Springer, Berlin, Germany, December 2020.
- [13] U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.
- [14] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred et al., "The BioGRID interaction database: 2015 update," *Nucleic Acids Research*, vol. 43, no. D1, pp. D470–D478, 2015.
- [15] L. Salwinski, C. S. Miller, A. J. Smith et al., "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, no. 1, pp. D449–D451, 2004.
- [16] K. Breuer, A. K. Foroushani, M. R. Laird et al., "InnateDB: systems biology of innate immunity and beyond-recent updates and continuing curation," *Nucleic Acids Research*, vol. 41, no. D1, pp. D1228–D1233, 2013.
- [17] S. Orchard, M. Ammari, B. Aranda et al., "The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research*, vol. 42, no. D1, pp. D358–D363, 2014.
- [18] G. Launay, R. Salza, D. Multedo, N. Thierry-Mieg, and S. Ricard-Blum, "MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities," *Nucleic Acids Research*, vol. 43, no. D1, pp. D321–D327, 2015.
- [19] 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.

- [20] 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.
- [21] K. Tomii and M. Kanehisa, "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins," *Protein Engineering, Design and Selection*, vol. 9, no. 1, pp. 27–36, 1996.
- [22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection IEEE computer society conference on computer vision and pattern recognition (CVPR'05)," *IEEE*, vol. 1, pp. 886–893, 2005.
- [23] B. Sugiarto, E. Prakasa, R. Wardoyo et al., "Wood identification based on histogram of oriented gradient (HOG) feature and support vector machine (SVM) classifier," in *Proceedings of the 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 337–341, IEEE, Yogyakarta, Indonesia, November 2017.
- [24] S. A. Korkmaz, A. Akçiçek, H. Bınoğlu, and M. F. Korkmaz, "Recognition of the stomach cancer images with probabilistic HOG feature vector histograms by using HOG features," in *Proceedings of the 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 339–342, IEEE, Subotica, Serbia, September 2017.
- [25] J.-Q. Gui and Z.-M. Lu, "Pedestrian detection based on modified dynamic background using Gaussian mixture models and HOG-SVM detection," *International Journal of Innovative Computing, Information Control*, vol. 14, no. 1, pp. 279–295, 2018.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [27] J. J. Rodríguez, 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.
- [28] K.-H. Liu and D.-S. Huang, "Cancer classification using rotation forest," *Computers in Biology and Medicine*, vol. 38, no. 5, pp. 601–610, 2008.
- [29] J.-F. Xia, K. Han, and D.-S. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor," *Protein & Peptide Letters*, vol. 17, no. 1, pp. 137–145, 2010.
- [30] A. Subasi, "Diagnosis of neuromuscular disorders using DT-CWT and rotation forest ensemble classifier," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 1940–1947, 2019.
- [31] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [32] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.
- [33] X. Liu, S. Yang, C. Li, Z. Zhang, and J. Song, "SPAR: a random forest-based predictor for self-interacting proteins with fine-grained domain information," *Amino Acids*, vol. 48, no. 7, pp. 1655–1665, 2016.
- [34] X. Du, J. Cheng, T. Zheng, Z. Duan, and F. Qian, "A novel feature extraction scheme with ensemble coding for protein-protein interaction prediction," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 12731–12749, 2014.
- [35] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, "PPIevo: protein-protein interaction prediction from PSSM based evolutionary information," *Genomics*, vol. 102, no. 4, pp. 237–242, 2013.
- [36] J. Zahiri, M. Mohammad-Noori, R. Ebrahimpour et al., "LocFuse: human protein-protein interaction prediction via classifier fusion using protein localization information," *Genomics*, vol. 104, no. 6, pp. 496–503, 2014.