International Journal of
Molecular Sciences

Article

An Ensemble Classifier to Predict Protein–Protein Interactions by Combining PSSM-based Evolutionary Information with Local Binary Pattern Model

Yang Li [1], Li-Ping Li [1,*], Lei Wang [2,*], Chang-Qing Yu [1,*], Zheng Wang [1] and Zhu-Hong You [1]

[1] School of Information Engineering, Xijing University, Xi'an 710123, China
[2] College of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China
[*] Correspondence: Lipingli@gmail.com (L.-P.L.); leiwang@ms.xjb.ac.cn (L.W.); 20160082@xijing.edu.cn (C.-Q.Y.); Tel.: +86-29-6189-0087 (L.-P.L.); +86-632-378-6721 (L.W.); +86-29-6368-1238 (C.-Q.Y.)

Received: 2 June 2019; Accepted: 15 July 2019; Published: 17 July 2019

Abstract: Protein plays a critical role in the regulation of biological cell functions. Among them, whether proteins interact with each other has become a fundamental problem, because proteins usually perform their functions by interacting with other proteins. Although a large amount of protein–protein interactions (PPIs) data has been produced by high-throughput biotechnology, the disadvantage of biological experimental technique is time-consuming and costly. Thus, computational methods for predicting protein interactions have become a research hot spot. In this research, we propose an efficient computational method that combines Rotation Forest (RF) classifier with Local Binary Pattern (LBP) feature extraction method to predict PPIs from the perspective of Position-Specific Scoring Matrix (PSSM). The proposed method has achieved superior performance in predicting *Yeast*, *Human*, and *H. pylori* datasets with average accuracies of 92.12%, 96.21%, and 86.59%, respectively. In addition, we also evaluated the performance of the proposed method on the four independent datasets of *C. elegans*, *H. pylori*, *H. sapiens*, and *M. musculus* datasets. These obtained experimental results fully prove that our model has good feasibility and robustness in predicting PPIs.

Keywords: protein–protein interactions; position-specific scoring matrix; rotation forest; protein sequence

1. Introduction

Protein is the essential part of the life activities of cells and organisms [1], and its function is usually performed by interacting with other proteins [2]. With the development of high-throughput biotechnology, experimental methods such as mass spectrometry, microarray analysis, and *Yeast* two-hybrid system have been widely used to detect protein–protein interactions (PPIs) [3–8]. However, these biological experimental methods are not only expensive and time-consuming, but also have a high false positive rate. In addition, the experimentally identified PPI can only cover a small portion of the entire PPIS network. Therefore, it is particularly important to design an accurate and effective computational method to predict PPIs.

At present, many computational methods have been proposed for predicting PPIs. These methods are usually based on the information of gene co-expression, phylogenetic relationship, and three-dimensional structural and so on [9–21]. Although these methods have achieved excellent results, they need to rely on prior knowledge of proteins [22]. Therefore, in order to overcome this drawback, many researchers have proposed the PPIs prediction method based on protein amino acid sequence information in recent years [23–28]. This kind of method can use the machine learning algorithm to extract important information from protein sequence data, and extract key features through feature extraction methods, so as to accurately and effectively predict the relationship among proteins [29]. For example, Shen et al. [30] rely on the properties of amino acids to extract the features of protein sequences by adopting the method of the conjoint triad. In order to reduce the dimension of the feature vector space, they divide 20 amino acids into 7 groups, which is determined by the volume of the side chain and the dipole. Zhou and Yang [31] separated the entire protein sequence into different local regions of different lengths, and then obtained three local descriptors of each local region, so as to further study the overlapping continuous and discontinuous interactions in the protein sequence [32]. Nakashima et al. [33] used the method of amino acid composition (AAC) to detect PPIs. The final experimental results show that this method can effectively predict PPIs. Guo et al. [34] proposed a combination of auto covariance (AC) and support vector machine (SVM) to predict PPIs. AC can efficiently obtain the interaction between a certain number of amino acids and amino acids in the sequence. Under the classification of SVM, the model achieved 87.36% accuracy on *Yeast* dataset. Zhou et al. [32] used a combination of local descriptors (LD) and support vector machines (SVM) to predict PPIs. The model achieved a prediction accuracy of 88.56% on the *Yeast* dataset. Wang et al. presented a computational model called PCVMZM to detect PPIs from protein amino acid sequences based on Zernike moments descriptor and probabilistic classification vector machines. This method yielded excellent performance on the *Yeast* dataset, and an average prediction accuracy of 94.48% indicates that the method is reliable for predicting protein–protein interactions.

In this study, we propose a novel sequence-based method to predict protein–protein interactions by combining Local Binary Pattern (LBP) feature extraction method and Rotation Forest (RF) classifier. More specifically, the method first converts the protein sequence information into a numerically represented Position-Specific Scoring Matrix (PSSM), then uses LBP to extract the effective features of the protein, and finally sends them into the RF classifier for accurate prediction. In the experiment, we used PPIs datasets of *Yeast*, *Human*, and *H. pylori* to evaluate the performance of the proposed model. The evaluation results show that our model achieved an average accuracy of 92.12%, 96.21%, and 86.59% on the three datasets, respectively. For the sake of verifying the reliability of our method, we have also predicted the protein–protein interactions on four independent datasets of *C. elegans, H. pylori, H. sapiens*, and *M. musculus* datasets and their accuracies are 94.82%, 94.79%, 95.11%, and 93.93%, respectively.

2. Results and Discussion

2.1. Performance Evaluation

To make the experimental results more reliable, we implemented the 5-fold cross-validation on all data to evaluate the performance of the proposed method. The evaluation index of the model includes overall prediction accuracy (ACC), sensitivity (SN), precision (PE), and Matthews correlation coefficient (MCC). The calculation formula for the evaluation criteria are as follows:

$$ACC = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{1}$$

$$SN = \frac{T_P}{T_P + F_N} \tag{2}$$

$$PE = \frac{T_P}{T_P + F_P} \tag{3}$$

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P) \times (T_P + F_N) \times (T_N + F_P) \times (T_N + F_N)}} \tag{4}$$

where True Positive (T_P) indicates the number of positive samples that are correctly predicted. False Positive (F_P) refers to the number of positive samples that are incorrectly predicted. True Negative (T_N) indicates the number of negative samples that are correctly predicted. False Negative (F_N) represents the number of negative samples that are incorrectly predicted. At the same time, the

Receiver Operating Characteristic (ROC) curves and the Area Under a Curve (AUC) are also used as an evaluation index to assess the performance of the model [35]. The workflow of the proposed model is shown in Figure 1.

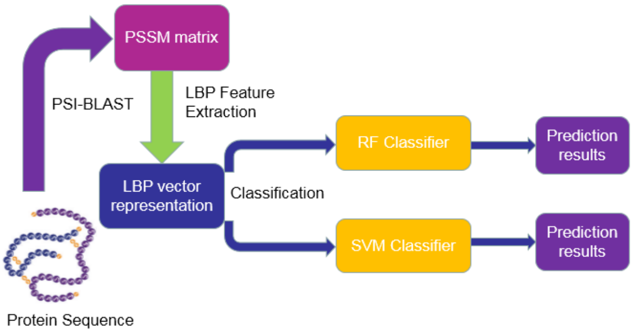

Figure 1. The workflow of the proposed method.

2.2. *Assessment of Prediction Ability*

In order to obtain more accurate and reliable experimental results, we optimized two important parameters of the rotation forest classifier on three different datasets of *Yeast*, *Human*, and *H. pylori*. Through the grid search method, we get the number of the optimal feature subset K of RF classifier is 10, and the number of the optimal decision trees L is 21. Meanwhile, we utilized a 5-fold cross-validation method to avoid over-fitting of the results. Specifically, we divide the total dataset into five roughly equal subsets, four of which are used as a training set and the rest one as a test set. This process is executed five times until all subsets are used as a test set once and only once. Finally, we take the average and standard deviation of the five experiments as the experimental results of the model. The prediction results of the three datasets are shown in Table 1. Additional materials are available online, Tables S1–S3.

Table 1. 5-fold cross-validation results obtained using the proposed method on three datasets.

Data Sets	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
Yeast	92.12 ± 0.54	94.20 ± 0.78	89.76 ± 0.96	85.46 ± 0.92	96.11 ± 0.77
Human	96.21 ± 0.76	97.23 ± 1.19	94.77 ± 1.09	92.70 ± 1.42	98.62 ± 0.48
H. pylori	86.59 ± 0.48	87.70 ± 1.89	85.17 ± 2.20	76.73 ± 0.74	92.69 ± 0.48

ACC = accuracy, PE = precision, SN = sensitivity, MCC = Matthews correlation coefficient, AUC = Area Under the Curve.

When our method is used to predict the PPIs of the *Yeast* dataset, the average accuracy, precision, sensitivity, and MCC of the prediction results are well displayed, which are 92.12%, 94.20%, 89.76%, and 85.46%, respectively. The standard deviations of these predicted results are 0.54%, 0.78%, 0.96%, and 0.92%, respectively. When our method is adopted to predict the PPIs of the *Human* dataset, our method also obtains good prediction results of average accuracy, precision, sensitivity, and MCC, which are 96.21%, 97.23%, 94.77%, and 92.70%, respectively. The standard deviations of these predicted results are 0.76%, 1.19%, 1.09%, and 1.42%, respectively. When our method was utilized to predict the PPIs of the *H. pylori* dataset, the average accuracy, precision, sensitivity, and MCC were predicted to be 86.59%, 87.70%, 85.17%, and 76.73%, respectively. The standard deviations of these predicted results are 0.48%, 1.89%, 2.20%, and 0.74%, respectively. The ROC curves of the proposed model on three datasets are Figures 2–4. Here, the X-axis indicates the false positive rate, while the Y-axis denotes the true positive rate. In order to better verify the feasibility of our method, AUC values are calculated on *Yeast*, *Human*, and *H. pylori* datasets and their average AUC values are 96.11%, 98.62%, and 92.69%, respectively.

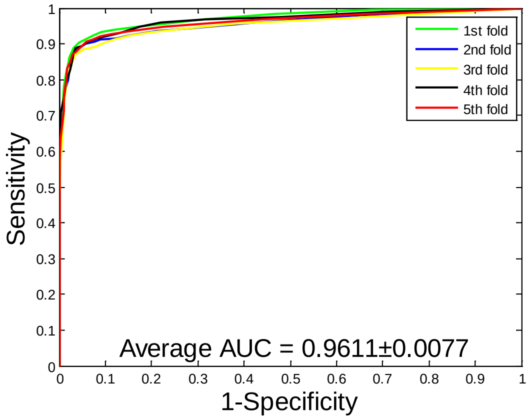

Figure 2. Receiver Operating Characteristic (ROC) curves are performed by the proposed method on *Yeast* protein–protein interactions (PPIs) dataset.

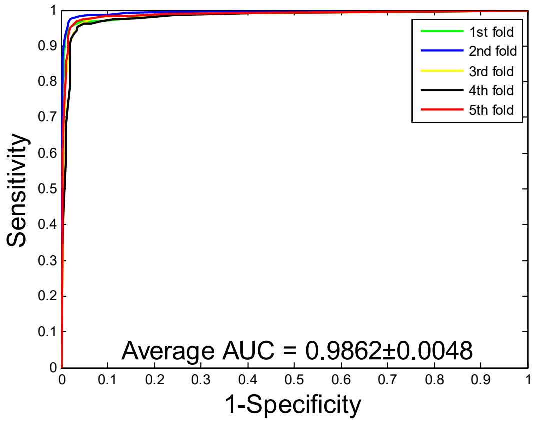

Figure 3. Receiver Operating Characteristic (ROC) curves are performed by the proposed method on *Human* protein–protein interactions (PPIs) dataset.

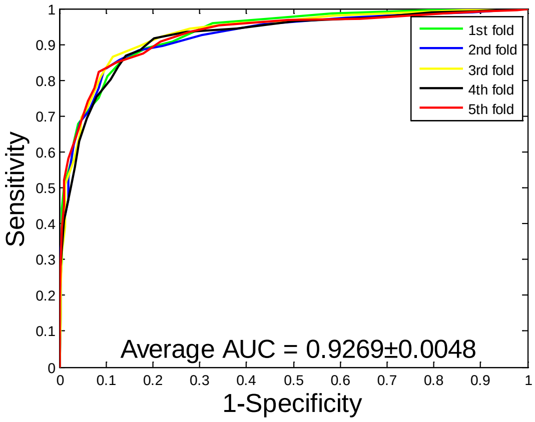

Figure 4. Receiver Operating Characteristic (ROC) curves are performed by the proposed method on *H. pylori* protein–protein interactions (PPIs) dataset.

2.3. Comparison with Support Vector Machine (SVM) Classifier

To more clearly assess the impact of the RF classifier on model performance, we compare the results of RF classifier model with those of Support Vector Machine (SVM) classifier model on the same dataset. To be fair, the data fed into the two classifier models are identical, both of which have undergone numerical transformation and feature extraction. The LIBSVM tool package used by SVM can be downloaded from its official website https://www.csie.ntu.edu.tw/~{}cjlin/libsvm/. When using

SVM, the regularization parameter *c* and the kernel parameter *g* are optimized by taking a grid search method. Eventually, we set *c* as 10 and *g* as 60 on the *Yeast*, *Human*, and *H. pylori* datasets, respectively.

The experimental results generated by the proposed model and the SVM model on the three datasets are summarized in Table 2. From the table, we can see that the average accuracy, precision, sensitivity, and MCC of the SVM model generated on the *Yeast* dataset are 86.99%, 88.05%, 85.62%, and 77.36%, respectively. When exploring the PPIs of the *Human* dataset through SVM model, the average accuracy, precision, sensitivity, and MCC obtained are 92.56%, 93.71%, 90.47%, and 86.18%, respectively. When the SVM is used to predict the PPIs of *H. pylori* dataset, the average accuracy is 81.62%. By comparing the results of two classifier models on the three datasets, we can see that the accuracy of the classifier based on SVM is lower than that of RF classifier. The results of the ROC curves on the three datasets predicted by the SVM classifier are reflected in Figures 5–7. Through observing and analyzing the results in the table, we can see that the model based on RF classifier has better performance than SVM classifier model in predicting PPIs.

Table 2. Comparison of the results of the proposed model and Support Vector Machine (SVM) model in three datasets.

Dataset	Classifier	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
Yeast	RF	92.12 ± 0.54	94.20 ± 0.78	89.76 ± 0.96	85.46 ± 0.92	96.11 ± 0.77
	SVM	86.99 ± 0.43	88.05 ± 0.88	85.62 ± 1.23	77.36 ± 0.64	93.66 ± 0.64
Human	RF	96.21 ± 0.76	97.23 ± 1.19	94.77 ± 1.09	92.70 ± 1.42	98.62 ± 0.48
	SVM	92.56 ± 0.70	93.71 ± 1.06	90.47 ± 0.82	86.18 ± 1.23	97.36 ± 0.65
H. pylori	RF	86.59 ± 0.48	87.70 ± 1.89	85.17 ± 2.20	76.73 ± 0.74	92.69 ± 0.48
	SVM	81.62 ± 1.22	80.73 ± 3.79	83.40 ± 3.56	69.93 ± 1.56	89.52 ± 0.53

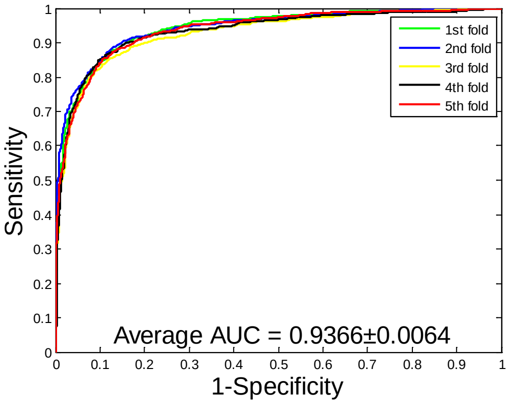

Figure 5. Receiver Operating Characteristics (ROC) curves are performed by the Support Vector Machine (SVM) method on *Yeast* protein–protein interactions (PPIs) dataset.

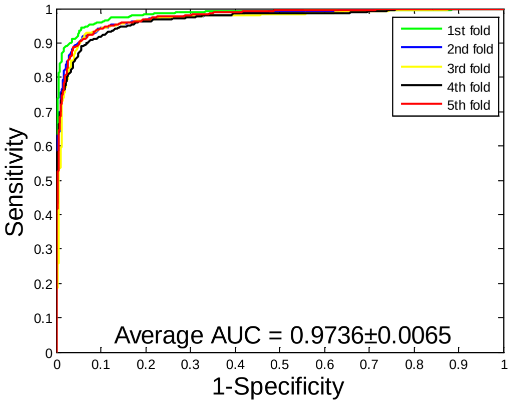

Figure 6. Receiver Operating Characteristics (ROC) curves are performed by the Support Vector Machine (SVM) method on *Human* protein–protein interactions (PPIs) dataset.

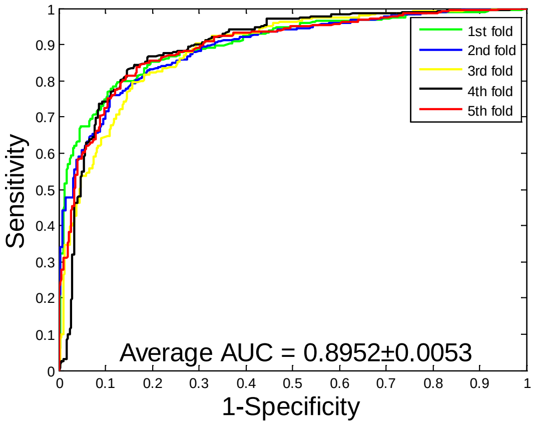

Figure 7. Receiver Operating Characteristics (ROC) curves are performed by the Support Vector Machine (SVM) method on *H. pylori* protein–protein interactions (PPIs) dataset.

2.4. Comparison with Existing Methods

In order to better evaluate the performance of the proposed method, we compare it with other existing methods on the same dataset. Tables 3 and 4 show the results obtained by different methods on the *Yeast* and *Human* datasets. As can be seen from Table 3, there are six methods applied to the *Yeast* dataset. Among them, our method shows a good average accuracy, which is 92.12%. In addition, the standard deviation obtained by the proposed model is also low. It can be seen from Table 4 that the proposed method also achieves better overall performance on the *Human* dataset. These results indicate that the proposed model has better performance and robustness than other methods on the *Yeast* and *Human* dataset.

There are two main reasons for this result: The first is that we use a sequence-based approach to predict PPIs. The discriminative information contained in the protein sequence combined with the effective LBP feature extraction method can contribute to the improvement of model performance. The second is that we use the ensemble classifier RF, which can synthesize the results of each sub-classifier and effectively improve the accuracy of prediction.

Table 3. Performance comparison of different methods on *Yeast* dataset.

Author	Model	ACC (%)	PE (%)	SN (%)	MCC (%)
Guos' work [34]	ACC	89.33 ± 2.67	88.87 ± 6.16	89.93 ± 3.68	N/A
	AC	87.36 ± 1.38	87.82 ± 4.33	87.30 ± 4.68	N/A
You et al.'s work [17]	PCA-EELM	87.00 ± 0.29	87.59 ± 0.32	86.15 ± 0.43	77.36 ± 0.44
Yang et al.'s work [31]	Cod1	75.08 ± 1.13	74.75 ± 1.23	75.81 ± 1.20	N/A
	Cod2	80.04 ± 1.06	82.17 ± 1.35	76.77 ± 0.69	N/A
	Cod3	80.41 ± 0.47	81.86 ± 0.99	78.14 ± 0.90	N/A
	Cod4	86.15 ± 1.17	90.24 ± 1.34	81.03 ± 1.74	N/A
Zhou et al.'s work [32]	SVM + LD	88.56 ± 0.33	89.50 ± 0.60	87.37 ± 0.22	77.15 ± 0.68
Wang et al.'s work [36]	PCVM + ZM	94.48 ± 1.2	93.92 ± 2.4	95.13 ± 2.0	89.58 ± 2.2
Our method	SVM + PSSM	86.99 ± 0.43	88.05 ± 0.88	85.62 ± 1.23	77.36 ± 0.64
	RF + PSSM	92.12 ± 0.54	94.20 ± 0.78	89.76 ± 0.96	85.46 ± 0.92

ACC: Auto Cross Covariance; AC: Auto Covariance; PCA-EELM: Principal component analysis-ensemble extreme learning machine; LD: Local description; PCVM + ZM: Probabilistic Classification Vector Machines+ Zernike Moments.

Table 4. Performance comparison of different methods on *Human* dataset.

Model	ACC (%)	SN (%)	MCC (%)
LDA + RF [37]	96.4	94.2	92.8
LDA + RoF	95.7	97.6	91.8
LDA + SVM	90.7	89.7	81.3
AC + RF	95.5	94.0	91.4
AC + RoF	95.1	93.3	91.0
AC + SVM	89.3	94.0	79.2
Our method	96.21	94.77	92.70

LDA: Linear discriminant analysis; RoF: Rotation forest; RF: Random forest.

2.5. Performance on Independent Datasets

By analyzing the results obtained from previous experiments, it is no exaggeration to say that our method gives superior performance in predicting PPIs on three datasets. In this part of the experiment, we validated the performance of the proposed method using an independent dataset, which were selected from the Database of Interacting Proteins(DIP) database, namely *C. elegans*, *H. pylori*, *H. sapiens*, and *M. musculus* datasets. In the experiment, we train the model with all of the 11,188 protein pairs in the *Yeast* dataset, and then predict the PPIs of the four independent datasets. The experimental results are listed in Table 5. From table we can see that the accuracies of the proposed model on *C. elegans*, *H. pylori*, *H. sapiens*, and *M. musculus* datasets were 94.82%, 94.79%, 95.11%, and 93.93%, respectively. The proposed model achieves high accuracy in all four independent datasets, which indicates that the proposed model has strong competitiveness in predicting the PPIs of different species.

Table 5. Predicted results on four independent datasets.

Species	Test Pairs	ACC (%)
C. elegans	4013	94.82
H. pylori	1420	94.79
H. sapiens	1412	95.11
M. musculus	313	93.93

3. Materials and Methodology

3.1. Dataset and Data Collection

In this paper, we employed a highly credible PPIs dataset of *Saccharomyces cerevisiae*, which comes from the open Database of Interacting Proteins (DIP) [38]. Since this dataset contains a large number of homologous proteins, in order to eliminate the differences, we deleted more than 40% sequence identities in these homologous sequences. At the same time, lower than 50 residues of protein pairs will also be removed, because they may be only a small fragment. After this treatment, the remaining 5594 protein pairs are established, which are used as positive datasets. In addition, 5594 additional protein pairs in different subcellular localization are also constructed, which are considered as negative datasets [12]. Eventually, we built a total *Yeast* dataset consisting of 11,188 protein pairs, half of which came from the positive dataset, and the other half from the negative dataset. Similarly, we constructed *Human* and *Helicobacter pylori* (*H. pylori*) datasets. The *Human* dataset contains 8161 protein pairs, of which 4262 negative protein pairs were used to construct the negative dataset and 3899 positive protein pairs were used to construct the positive dataset. The *H. pylori* dataset contains 2916 protein pairs, half of which are positive datasets and the other half are negative datasets.

3.2. Position-Specific Scoring Matrix (PSSM)

Protein sequences have undergone various changes in the process of biological evolution. With these constant changes, one or more amino acid residues are displaced, inserted, or deleted in the protein sequence, and the comparability between proteins has also decreased gradually. However, these homologous proteins may still have similar structures. Therefore, in order to demonstrate this characteristic of proteins, we introduce the Position-Specific Scoring Matrix (PSSM) which can fully acquire the evolutionary information of protein sequences. In the experiment, we make use of the Position-Specific Iterated BLAST (PSI-BLAST) search tool to generate PSSMs on the local machine [39]. In order to obtain reliable homologous sequence data, we optimize its main parameters, in which E-value is set to 0.001 and the number of interactions is set to 3, respectively. PSI-BLAST toolkit can be downloaded from http://blast.ncbi.nlm.nih.gov/Blast.cgi. PSI-BLAST will return a PSSM where each PSSM is R rows and 20 columns. The PSSM can be defined as:

$$\mathrm{PSSM} = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,20} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{\mathrm{R},1} & \rho_{\mathrm{R},2} & \cdots & \rho_{\mathrm{R},20} \end{bmatrix}, \tag{5}$$

where R expresses the length of the amino acid sequence and 20 stands for 20 amino acids. The value of $\rho_{i,j}$ in the PSSM indicates that the ith amino acid residue is mutated into the type j amino acids among the 20 native amino acids.

3.3. Local Binary Pattern (LBP)

Local Binary Pattern (LBP) is an effective algorithm for describing the local texture features of an image [40]. It has significant features of rotation invariance and grayscale invariance. At present, LBP has been widely used in image processing, including facial expression recognition, image recovery and scene analysis [41,42]. The original LBP operator is defined as the window of 3×3, which uses the gray value in the fixed neighborhood. As a result, the texture information around the image pixels is unlikely to be obtained correctly. Ojala et al. [43] proposed the original LBP operator, which uses the central pixel value of the window as a threshold and gives the 8-bit codes through the eight pixel values around the center pixel. For the sake of adapting the texture features of different scales, researchers improved the original LBP operator, in which the operator was extended to any radius and neighborhood, while the original square neighborhood was replaced with a circular neighborhood. The LBP operator can have any number of pixels in a circular neighborhood of radius R. Therefore, the circular LBP operator with radius R can be obtained.

In this experiment, LBP features of all PSSM matrices can be calculated. Where N is used to indicate the number of neighboring pixels around the center pixel, and R is employed to represent the radius of a circle around N equidistant neighborhoods of the center pixel. Here, we set the corresponding parameters of the LBP. $R = 1$ represents a circular neighborhood with a radius of 1, and $N = 8$ represents an LBP operator with eight sample points in the circular neighborhood. The i_c is used to represent the luminance value of the center pixel and i_i to represent the intensity value of the circular neighborhood. The central pixel is regarded as the threshold of the window, and then the gray values of the eight neighboring pixels are compared with them. If the surrounding pixel value is greater than the central pixel value, its pixel position is marked as 1. Otherwise, it is marked as 0. The formula calculation of Local Binary Pattern can be defined as follows:

$$B = b(s(i_0 - i_c), s(i_1 - i_c), \ldots, s(i_{N-1} - i_c)), \tag{6}$$

where

$$s(x) = \begin{cases} 1 & ,x \geq 0 \\ 0 & ,x < 0 \end{cases} \tag{7}$$

$$LBP_{N,R}(x_c, y_c) = \sum_{i=0}^{N-1} s(i_i - i_c)2^i \tag{8}$$

here, the appropriate gray value in a circular neighborhood is calculated as [43].

$$i_i = I\left(x + R\sin\frac{2\pi i}{N}, y - R\cos\frac{2\pi i}{N}\right) \tag{9}$$

where (x_c, y_c) represents the gray value i_c of the center pixel in the LBP. The rotation invariance problem can be solved by selecting the smallest binary number of all LBPs. There are 256 kinds of LBP features in the experiment when all possible outcomes are considered among neighborhoods. Finally, we extract the LBP features of PSSM, each of which is the feature matrix of the 1×256.

3.4. Rotation Forest (RF)

Rotation forest (RF) is an ensemble classifier consisting of a set of decision trees. It was proposed by Rodriguez et al. [44]. For each decision tree in the RF, the bootstrap sample is derived from the original training set to be used to form a new training set. The feature set of the new training set is randomly divided into several subsets and transformed using a linear transformation method. Thus, a complete feature set can be reconstructed by transforming all the features of each tree during the ensemble process. Since a small rotation of axis can construct completely different trees, the transformation method can guarantee the diversity of the ensemble system. Finally, we can use the main voting rules to fuse the output of all trees.

Let the training sample set X be an $N \times n$ matrix, which contains N training samples and n features. Let F be the feature set and the corresponding label vector be $Y = (y_1, y_2, \ldots, y_n)^T$ with size $N \times 1$. Suppose that the feature set of the sample set is randomly partitioned into K subsets with the same size. In this case, the decision tree L in the RF can be represented as $T_1, \ldots, T_L$, respectively. Here, we need to determine the two parameters L and K in advance. The implementation of the rotation forest classifier is as follows:

(1) The feature set F is randomly divided into K disjoint subsets, and each subset contains $M = n/K$ features.

(2) Assuming that F_{ij} be the jth subset of features, which is used to train the classifier T_i. Let X_{ij} be the dataset for X. For each subset, a nonempty random subset is selected for X_{ij}. Then, a bootstrap resampling is selected from X_{ij} with a size of 75% of the dataset to generate a new training set X'_{ij}.

(3) Apply principal component analysis to X'_{ij} to produce the coefficients in matrix C_{ij}. The size of each X'_{ij} is $M \times 1$ with the coefficients of $\lambda_{ij}^{(1)}, \ldots, \lambda_{ij}^{(M_j)}$.

(4) The coefficients obtained in the matrix C_{ij} are used to generate a sparse rotation matrix R_i, which is given as follows:

$$R_i = \begin{bmatrix} \lambda_{i1}^{(1)}, \ldots, \lambda_{i1}^{(M_1)} & \{0\} & \cdots & \{0\} \\ \{0\} & \lambda_{i2}^{(1)}, \ldots, \lambda_{i2}^{(M_2)} & \cdots & \{0\} \\ \vdots & \vdots & \ddots & \vdots \\ \{0\} & \{0\} & \cdots & \lambda_{iK}^{(1)}, \ldots, \lambda_{iK}^{(M_K)} \end{bmatrix}. \tag{10}$$

In the classification process, let $d_{ij}(xR_i^\lambda)$ be the probability generated by the classifier T_i, which is used to determine whether x belongs to class y_i. Next, the average combination method is used to calculate the confidence of each class in a given test sample, and the formula is as follows:

$$\omega_j(x) = \frac{1}{L}\sum_{i=1}^{L} d_{ij}(xR_i^\lambda). \tag{11}$$

Finally, the test sample x will be assigned to the class with the greatest confidence.

4. Conclusions

In this paper, we proposed a computational method using only protein sequence information to predict PPIs. The proposed method can accurately predict the interaction among proteins by combining the local binary pattern algorithm and rotation forest classifier. In the experiment, we validated the proposed model on the *Yeast, Human*, and *H. pylori* datasets using the 5-fold cross-validation method. To evaluate the performance of the proposed model, we compared it with the SVM model and the existing methods in the same dataset. Among them, the proposed method obtained average prediction accuracy of 92.12%, 96.21%, and 86.59% on the *Yeast*, *Human*, and *H. pylori* datasets, respectively. Comparing these good experimental results, it can be seen that the proposed method is reliable and feasible for predicting PPIs. In addition, we also evaluate the proposed model in four independent datasets, including *C. elegans, H. pylori, H. sapiens*, and *M. musculus*. In the above experiments, the proposed models have achieved excellent results. This demonstrated that the proposed model is highly competitive and can be used as an effective tool for PPIs prediction. In future research, we will introduce a deep learning algorithm into the model to help the model achieve better prediction performance.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/20/14/3511/s1, Tables S1–S3.

Author Contributions: Y.L., C.-Q.Y. and L.W. conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript. L.-P.L., Z.W. and Z.-H.Y. analyzed experiments. All authors read and approved the final manuscript.

Funding: This work is supported in part by the National Science Foundation of China, under Grant 61572506 and 61702444, in part by the West Light Foundation of the Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008.

Acknowledgments: The authors would like to thank all the editors and anonymous reviewers for their constructive advices.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Várnai, C.; Burkoff, N.S.; Wild, D.L. Improving protein-protein interaction prediction using evolutionary information from low-quality MSAs. *PLoS ONE* **2017**, *12*, 0169356.
2. Lei, H.; Li, L.; Wu, C.H. Protein-protein interaction prediction based on multiple kernels and partial network with linear programming. *BMC Syst. Biol.* **2016**, *10*, 45.
3. Yi, H.-C.; You, Z.-H.; Huang, D.-S.; Li, X.; Jiang, T.-H.; Li, L.-P. A Deep Learning Framework for Robust and Accurate prediction of ncRNA-Protein Interactions using Evolutionary Information. *Mol. Ther. Nucleic Acids* **2018**, *1*, 1–11. [CrossRef] [PubMed]
4. Li, Z.; Ivanov, A.A.; Su, R.; Gonzalez-Pecchi, V.; Qi, Q.; Liu, S.; Webber, P.; McMillan, E.; Rusnak, L.; Pham, C.; et al. The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat. Commun.* **2017**, *8*, 14356.
5. Yang, B.; Tang, S.; Ma, C.; Li, S.T.; Shao, G.C.; Dang, B.; Degrado, W.F.; Dong, M.Q.; Wang, P.G.; Ding, S. Spontaneous and specific chemical cross-linking in live cells to capture and identify protein interactions. *Nat. Commun.* **2017**, *8*, 2240. [CrossRef] [PubMed]

6. Schlecht, U.; Liu, Z.; Blundell, J.R.; St Onge, R.P.; Levy, S.F. A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. *Nat. Commun.* **2017**, *8*, 15586. [CrossRef]
7. Li, J.; Bonkowski, M.S.; Moniot, S.; Zhang, D.; Hubbard, B.P.; Ling, A.J.; Rajman, L.A.; Qin, B.; Lou, Z.; Gorbunova, V. A conserved NAD+ binding pocket that regulates protein-protein interactions during aging. *Science* **2017**, *355*, 1312. [CrossRef]
8. Gierer, A. Model for DNA and Protein Interactions and the Function of the Operator. *Nature* **2017**, *212*, 1480–1481. [CrossRef]
9. An, J.Y.; Meng, F.R.; You, Z.H.; Fang, Y.H.; Zhao, Y.J.; Zhang, M. Using the Relevance Vector Machine Model Combined with Local Phase Quantization to Predict Protein-Protein Interactions from Protein Sequences. *BioMed Res. Int.* **2016**, *2016*, 1–9. [CrossRef]
10. Huang, Q.; You, Z.; Zhang, X.; Zhou, Y. Prediction of Protein–Protein Interactions with Clustered Amino Acids and Weighted Sparse Representation. *Int. J. Mol. Sci.* **2015**, *16*, 10855–10869. [CrossRef]
11. Huang, Y.; Chen, X.; You, Z.; Huang, D.; Chan, K. ILNCSIM: Improved lncRNA functional similarity calculation model. *Oncotarget* **2016**, *7*, 25902–25914. [CrossRef] [PubMed]
12. Huang, Y.-A.; You, Z.-H.; Gao, X.; Wong, L.; Wang, L. Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. *BioMed Res. Int.* **2015**, *2015*, 902198. [CrossRef] [PubMed]
13. Luo, X.; Ming, Z.; You, Z.; Li, S.; Xia, Y.; Leung, H. Improving network topology-based protein interactome mapping via collaborative filtering. *Knowl. Based Syst.* **2015**, *90*, 23–32. [CrossRef]
14. Wong, L.; You, Z.-H.; Ming, Z.; Li, J.; Chen, X.; Huang, Y.-A. Detection of Interactions between Proteins through Rotation Forest and Local Phase Quantization Descriptors. *Int. J. Mol. Sci.* **2015**, *17*, 21. [CrossRef] [PubMed]
15. You, Z.H.; Li, S.; Gao, X.; Luo, X.; Ji, Z. Large-Scale Protein-Protein Interactions Detection by Integrating Big Biosensing Data with Computational Model. *BioMed Res. Int.* **2014**, *2014*, 598129. [CrossRef] [PubMed]
16. You, Z.H.; Zhou, M.; Luo, X.; Li, S. Highly Efficient Framework for Predicting Interactions Between Proteins. *IEEE Tran. Cybern.* **2016**, *47*, 731–743. [CrossRef] [PubMed]
17. You, Z.-H.; Lei, Y.-K.; 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 Bioinform.* **2013**, *14* (Suppl. 8), S10. [CrossRef] [PubMed]
18. You, Z.-H.; Li, J.; Gao, X.; He, Z.; Zhu, L.; Lei, Y.-K.; Ji, Z. Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. *BioMed Res. Int.* **2015**, *2015*, 867516. [CrossRef]
19. You, Z.-H.; Yin, Z.; Han, K.; Huang, D.-S.; Zhou, X. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *BMC Bioinform.* **2010**, *11*, 343. [CrossRef]
20. Zhu, L.; You, Z.-H.; Huang, D.-S. Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding. *Neurocomputing* **2013**, *121*, 99–107. [CrossRef]
21. Zhu, L.; You, Z.-H.; Huang, D.-S. Identifying Spurious Interactions in the Protein-Protein Interaction Networks Using Local Similarity Preserving Embedding. In *Bioinformatics Research and Applications*; Springer International Publishing: Basel, Switzerland, 2014; pp. 138–148.
22. Atashin, A.A.; Bagherzadeh, P.; Ghiasishirazi, K. A two-stage learning method for protein-protein interaction prediction. *arXiv* **2016**, arXiv:1606.04561.
23. Kotlyar, M.; Pastrello, C.; Pivetta, F.; Sardo, A.L.; Cumbaa, C.; Li, H.; Naranian, T.; Niu, Y.; Ding, Z.; Vafaee, F. In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat. Methods* **2015**, *12*, 79–84. [CrossRef] [PubMed]
24. Schoenrock, A.; Samanfar, B.; Pitre, S.; Hooshyar, M.; Jin, K.; Phillips, C.A.; Wang, H.; Phanse, S.; Omidi, K.; Gui, Y. Efficient prediction of human protein-protein interactions at a global scale. *BMC Bioinform.* **2014**, *15*, 383. [CrossRef] [PubMed]
25. Huang, D.-S.; Zhang, L.; Han, K.; Deng, S.; Yang, K.; Zhang, H. Prediction of Protein-Protein Interactions Based on Protein-Protein Correlation Using Least Squares Regression. *Curr. Protein Pept. Sci.* **2014**, *15*, 553–560. [CrossRef] [PubMed]

26. Zhang, Q.C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C.A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **2012**, *490*, 556–560. [CrossRef] [PubMed]
27. Muppirala, U.K.; Honavar, V.G.; Dobbs, D. Predicting RNA-Protein Interactions Using Only Sequence Information. *BMC Bioinform.* **2011**, *12*, 489. [CrossRef]
28. Wang, L.; You, Z.H.; Chen, X.; Li, J.Q.; Yan, X.; Zhang, W.; Huang, Y.A. An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences. *Oncotarget* **2017**, *8*, 5149. [CrossRef]
29. Zhou, C.; Yu, H.; Ding, Y.; Guo, F.; Gong, X.J. Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. *PLoS ONE* **2017**, *12*, e0181426. [CrossRef]
30. Juwen, S.; Jian, Z.; Xiaomin, L.; Weiliang, Z.; Kunqian, Y.; Kaixian, C.; Yixue, L.; Hualiang, J. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4337–4341.
31. Yang, L.; Xia, J.-F.; Gui, J. Prediction of Protein-Protein Interactions from Protein Sequence Using Local Descriptors. *Protein Pept. Lett.* **2010**, *17*, 1085–1090. [CrossRef]
32. Zhou, Y.Z.; Gao, Y.; Zheng, Y.Y. Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence. In *Advances in Computer Science and Education Applications, Pt Ii*; Zhou, M., Tan, H.H., Eds.; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2011; Volume 202, pp. 254–262.
33. Nakashima, H.; Nishikawa, K.; Ooi, T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* **1986**, *99*, 153–162. [CrossRef] [PubMed]
34. Guo, Y.; Yu, L.; Wen, Z.; Li, M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.* **2008**, *36*, 3025–3030. [CrossRef] [PubMed]
35. Zweig, M.H.; Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561–577. [PubMed]
36. Wang, Y.; You, Z.; Li, X.; Chen, X.; Jiang, T.; Zhang, J. PCVMZM: Using the Probabilistic Classification Vector Machines Model Combined with a Zernike Moments Descriptor to Predict Protein–Protein Interactions from Protein Sequences. *Int. J. Mol. Sci.* **2017**, *18*, 1029. [CrossRef] [PubMed]
37. Liu, B.; Liu, F.; Fang, L.; Wang, X.; Chou, K.C. repDNA: A Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics* **2015**, *31*, 1307. [CrossRef] [PubMed]
38. Ioannis, X.; Lukasz, S.; Xiaoqun Joyce, D.; Patrick, H.; Sul-Min, K.; David, E. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **2002**, *30*, 303–305.
39. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
40. Bhatti, M.N.A.; Jung, S.K. Local binary pattern variants-based adaptive texture features analysis for posed and nonposed facial expression recognition. *J. Electron. Imaging* **2017**, *26*, 053017.
41. Ahonen, T.; Hadid, A.; Pietikainen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef]
42. Huynh, T.; Min, R.; Dugelay, J.L. *An Efficient LBP-Based Descriptor for Facial Depth Images Applied to Gender Recognition Using RGB-D Face Data*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 133–145.
43. Ojala, T.; Harwood, I. A Comparative Study of Texture Measures with Classification Based on Feature Distributions. *Pattern Recogn.* **1996**, *29*, 51–59. [CrossRef]
44. Rodriguez, J.J.; Kuncheva, L.I. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1619–1630. [CrossRef] [PubMed]