

Research Article

Predicting Protein-Protein Interactions via Random Ferns with Evolutionary Matrix Representation

Yang Li ¹, Zheng Wang ², Zhu-Hong You ³, Li-Ping Li ⁴, and Xuegang Hu ¹

¹School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China

²School of Information Engineering, Xijing University, Xi'an 710123, China

³School of Computer Science, Northwestern Polytechnical University, Xi'an Shaanxi 710129, China

⁴College of Grassland and Environment Sciences, Xinjiang Agricultural University, Urumqi 830052, China

Correspondence should be addressed to Zhu-Hong You; zhuhongyou@gmail.com and Xuegang Hu; jsjxhuxg@hfut.edu.cn

Received 3 August 2021; Revised 15 January 2022; Accepted 18 January 2022; Published 22 February 2022

Academic Editor: Prem Chapagain

Copyright © 2022 Yang Li 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.

Protein-protein interactions (PPIs) play a crucial role in understanding disease pathogenesis, genetic mechanisms, guiding drug design, and other biochemical processes, thus, the identification of PPIs is of great importance. With the rapid development of high-throughput sequencing technology, a large amount of PPIs sequence data has been accumulated. Researchers have designed many experimental methods to detect PPIs by using these sequence data, hence, the prediction of PPIs has become a research hotspot in proteomics. However, since traditional experimental methods are both time-consuming and costly, it is difficult to analyze and predict the massive amount of PPI data quickly and accurately. To address these issues, many computational systems employing machine learning knowledge were widely applied to PPIs prediction, thereby improving the overall recognition rate. In this paper, a novel and efficient computational technology is presented to implement a protein interaction prediction system using only protein sequence information. First, the Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) was employed to generate a position-specific scoring matrix (PSSM) containing protein evolutionary information from the initial protein sequence. Second, we used a novel data processing feature representation scheme, MatFLDA, to extract the essential information of PSSM for protein sequences and obtained five training and five testing datasets by adopting a five-fold cross-validation method. Finally, the random fern (RFs) classifier was employed to infer the interactions among proteins, and a model called MatFLDA_RF was developed. The proposed MatFLDA_RF model achieved good prediction performance with 95.03% average accuracy on *Yeast* dataset and 85.35% average accuracy on *H. pylori* dataset, which effectively outperformed other existing computational methods. The experimental results indicate that the proposed method is capable of yielding better prediction results of PPIs, which provides an effective tool for the detection of new PPIs and the in-depth study of proteomics. Finally, we also developed a web server for the proposed model to predict protein-protein interactions, which is freely accessible online at http://120.77.11.78:5001/webserver/MatFLDA_RF.

1. Introduction

Recognition of protein-protein interactions (PPIs) is distinctly important for understanding various cellular biological activities [1]. The knowledge of PPIs can help us to explore and elucidate the functions of proteins, drug design, new drug development, and the mechanisms of biological activity and related proteins in cells [2]. Additionally, it can also provide new ideas for other studies, such as the ranking of disease genes [3], functional module identification [4], and human

disease prevention and treatment. In general, the research approaches for PPIs mainly include two categories: computational-based methods and biological experimental-based methods. In the last decades, many different experimental techniques have been used for large-scale PPIs validation, such as yeast two-hybrid (Y2H) screens [5], coimmunoprecipitation (Co-IP) [6], nuclear magnetic resonance (NMR) [7], protein chip [8], and other high-throughput biological techniques. However, there are some inevitable disadvantages of these methods: they are not only time-consuming and

expensive but also suffer from high false-positive rates and weak generalization ability. Thus, it has great practical significance to develop a new effective machine learning approach for PPIs prediction in order to save cost and time, thereby ultimately improving the prediction accuracy of protein interactions. To date, numerous computational approaches have been suggested to detect PPIs based on different data types, including protein domains, genomic information, evolutionary knowledge, structure information, gene fusion, and phylogenetic profiles [9–14]. Although these methods can be used to detect PPIs, the abovementioned methods are not universally applicable unless prior knowledge of the protein is known. Although amino acid sequence information is readily available for a large number of proteins, the 3D structural information of many proteins is still unclear, and the known and available PPIs for most species are still incomplete or very sparse. Consequently, it is particularly important to design novel computational methods for PPI prediction utilizing only protein amino acid sequence information, so as to better employ these abundant protein sequence data.

Numerous previous works have shown that using protein amino acid sequence information alone is sufficient to predict PPIs. So far, many different computational methods based on sequence information have been presented to implement this pattern in PPI prediction, such as combining average blocks with relevance vector machine [15], combining principal component analysis with ensemble extreme learning machine [16], combining conventional auto covariance with support vector machine [17], local descriptors using k-nearest neighbor [18], discrete cosine transformation using weighted sparse representation model [19], and so on. In 2017, Wang et al. [20] proposed a PCVMZM method based on protein sequence. The Zernike moments (ZM) are used as the feature extraction method. ZM can capture multiangle useful and representative information. Probabilistic classification vector machines (PCVM) are a sparse classification model that optimizes the kernel parameters by the expectation-maximization (EM) algorithm, which not only improves the prediction performance of PPIs but also reduces the computational time in the testing phase. The average prediction accuracy achieved by the PCVMZM method was 94.48% on the *Yeast* dataset. In the same year, Du et al. [21] proposed a method called DeepPPI from the angle of deep learning technology by using amphiphilic pseudo amino acid composition feature extraction algorithm to extract features from amino acid sequences, which opens a new way for studying PPIs. This DeepPPI method reached a prediction accuracy of 94.43% on the *Saccharomyces cerevisiae* dataset. In 2018, Göktepe and Kodaz [22] applied a new technique called weighted skip-sequential conjoint triads to predict PPIs. The method adopts principal component analysis (PCA) to remove noise information, captures protein sequence information by combining Bi-gram representation and Pseudo-amino acid composition, and finally uses support vector machine (SVM) as a prediction classifier to identify interactions between proteins. In the same year, Song et al. [23] presented a novel feature fusion scheme based on random projection ensemble method, which separately used three algorithms (fast fourier transform, discrete cosine transform, and singular value decomposition) to

explore and denote the patterns of interactions between amino acids. In 2019, Chen et al. [1] developed an end-to-end framework, called PIPR, to predict PPIs using only the protein sequences. They capture effectively the local significant features and sequential features from protein sequence pairs by using a deep residual recurrent convolutional neural network. Experimental results demonstrate that the framework has good scalability on different datasets. In the same year, Beltran et al. [24] used five feature extraction methods, namely, dipeptide composition, tripeptide composition, autocovariance, amino acid composition, and pseudo-amino-acid composition to represent amino acid sequences. They then employed SVM, random forest (RF), and extreme gradient boosting (XGBoost) to predict PPIs, respectively, and finally achieved good prediction performance. More recently, Jha and Saha [25] presented a deep-learning-based predictor to identify PPIs. They introduced two deep learning algorithms, ResNet50 and stacked autoencoder, to extract features from the autocovariance and conjoint triad representations of protein sequences. Then, LSTM-based classifier model was constructed for each feature encoding scheme. The experimental results show that the introduced deep learning scheme can learn valuable features from multimodal information of proteins. Although a number of computational-based methods have achieved good progress and application prospects, the accuracy and efficiency of PPIs prediction still need to be further enhanced so as to provide a supplementary tool for proteomics research and other bioinformatics tasks.

In this paper, an efficient computational method for detecting PPIs from amino acid sequences is presented by using the evolutionary matrix representation of protein sequences and combining with an ensemble classifier. Among them, an important improvement of the proposed model is to develop a more accurate numerical representation of protein sequences. Specifically, we applied the MatFLDA feature extraction algorithm to a position-specific scoring matrix (PSSM) to extract the evolutionary information of protein sequences and utilized a random ferns classifier to predict the PPIs. More specifically, each protein sequence is denoted as a PSSM numerical matrix. Subsequently, for the purpose of obtaining more representative information, we utilize the MatFLDA descriptor to extract the feature information in each PSSM, so as to obtain a 400-dimensional feature vector from the model and thus obtain an 800-dimensional feature vector representation of each protein pair. Finally, we employ the feature vector of protein pairs as the input of the model and combine the RF ensemble model in machine learning to accomplish the classification task of PPIs. The proposed method is estimated on the PPI datasets of *Yeast* and *H. pylori* with prediction accuracy of 95.03% and 85.35%, respectively. By comparing with a series of previous computational methods, we clearly found that the proposed model has good generalization performance in predicting PPIs.

2. Materials and Methodology

2.1. Datasets. So far, a number of PPIs databases have been created, including HAPPI database [26], Molecular Interaction Database (MINT) [27], APID database [28], Biomolecular Interaction Network Database (BIND) [29], and

Database of Interacting Proteins (DIP) [30]. In this section, we use two high-quality benchmark datasets, which are extracted from DIP, to test the generality of the model and assess the performance of the proposed method. The first dataset is the *yeast* dataset collected by Guo et al. [17]. To evaluate our method, a data preprocessing procedure that deleted protein pairs of greater than 40% sequence identity and less than 50 residues was used in this experiment to avoid the bias introduced by these homologous sequence pairs. By performing this process, we extracted 5594 protein pairs which formed the golden standard positive dataset. The additional 5594 protein pairs were retained to construct the golden standard negative dataset by removing interaction pairs with the same subcellular localization information. The second dataset is the *H. pylori* dataset, which was validated by the yeast two-hybrid technology [31] and collected by Martin et al. [32]. The PPI dataset of *H. pylori* contains 1458 positive protein pairs and 1458 negative protein pairs, which are regarded as positive and negative datasets, respectively. Consequently, *yeast* and *H. pylori* datasets are composed of a total of 11,188 and 2916 protein pairs, respectively.

2.2. Numerical Characterization of Protein Sequences. Position-Specific Scoring Matrix (PSSM) serves as a very useful scoring matrix that can contain evolutionary information of protein sequences, which is crucial in proteomics. PSSM was originally introduced by Gribskov et al. [33] in 1987 and is commonly used to detect distantly related proteins and protein folding patterns [34]. Currently, some researchers have done a lot of related work using PSSM encoding information in many fields of bioinformatics such as identification of DNA binding proteins [35], the identification of drug-target interaction [36], prediction of membrane protein types [37], and protein-protein interaction site prediction [38]. In this experiment, we employed the Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) [39] to convert each protein sequence into a PSSM, which is widely adopted for the numerical representation of protein sequences for further use in PPI detection tasks. PSSM is a matrix composed of T rows and 20 columns, where the row represents the length of the protein sequence and 20 columns are attributed to the 20 naive amino acids. Suppose that $M = \{\partial_{i,j} : i = 1, \dots, T \text{ and } j = 1, \dots, 20\}$, PSSM can be described as follows:

$$M = \begin{bmatrix} \partial_{1,1} & \partial_{1,2} & \cdots & \partial_{1,20} \\ \partial_{2,1} & \partial_{2,2} & \cdots & \partial_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ \partial_{T,1} & \partial_{T,2} & \cdots & \partial_{T,20} \end{bmatrix}. \quad (1)$$

The elements in this matrix usually contain positive or negative integers, where the element $\partial_{i,j}$ is the probability that the i th amino acid mutates into the j th amino acid in the process of biological evolution. Here, positive scores in this matrix mean that amino acid substitutions occur more frequently in the alignment, whereas negative scores mean that the substitution occurs less frequently.

In our study, we set the e -value and iteration times of PSI-BLAST, which are 0.001 and 3, respectively, to obtain highly and broadly homologous protein sequences. Consequently, each protein sequence is denoted as a 20-dimensional matrix containing $T \times 20$ elements, where T is the length of a given protein sequence and 20 indicates the number of amino acids. The application information of PSI-BLAST can be downloaded at <http://blast.ncbi.nlm.nih.gov/Blast.cgi> [40, 41].

2.3. Matrix Fisher Linear Discriminant Analysis (MatFLDA). Fisher linear discriminant analysis (FLDA), as a popular feature extraction method [42], has recently gained considerable attention in the areas of data mining and pattern recognition, such as software fault prediction [43], Arabic text classification [44], and face recognition [45]. In Section 2.2, each PSSM can be denoted as $M = \{\partial_{i,j} : i = 1, \dots, T \text{ and } j = 1, \dots, 20\}$, which is a $T \times 20$ matrix. To construct the FLDA of the matrix pattern, we give the matrix pattern A_{ij} for the i th class containing N_i samples, which can be denoted as $A_{ij} = M^T \times M (i = 1, 2, \dots, C, j = 1, 2, \dots, N_i)$, where represents the number of PSSMs, and the total sample mean is defined as \bar{A} . For Matrix Fisher Linear Discriminant Analysis (MatFLDA), assume that a class matrix pattern $A_i, i = 1, 2, \dots, C$ containing C classes is given, where $C = 20$ represents the 20 classes of amino acids, and their class mean is \bar{A}_i . Let x be a vector with m components. MatFLDA aims to project a matrix pattern A onto the x satisfying the constraint that $x^T x = 1$, and then a $1 \times n$ dimensional feature matrix can be generated by using the following linear transformation.

$$y = x^T A, \quad (2)$$

where y is an extracted feature matrix or projected value. Hence, for each matrix pattern A_{ij} , all their feature matrices are projected as follows:

$$y_{ij} = x^T A_{ij}, i = 1, 2, \dots, C; j = 1, 2, \dots, N_i. \quad (3)$$

To find the optimal projection vector x , we use the following criterion function and maximize it:

$$J_{Mat}(x) = \frac{\text{tr}(x^T S_b^{Mat} x)}{\text{tr}(x^T S_w^{Mat} x)}, \quad (4)$$

where S_b^{Mat} is the total between-class scatter matrix, which is defined as

$$S_b^{Mat} = \sum_{i=1}^C N_i (\bar{A}_i - \bar{A})(\bar{A}_i - \bar{A})^T, \quad (5)$$

where S_w^{Mat} is the total within-class scatter matrix, which is defined as

$$S_w^{Mat} = \sum_{i=1}^C \sum_{j=1}^{N_i} (A_{ij} - \bar{A}_i)(A_{ij} - \bar{A}_i)^T. \quad (6)$$

In the MatFLDA algorithm, by maximizing $J_{Mat}(x)$, we want to keep the between-class scatter matrix as large as possible and the within-class scatter matrix as small as possible in the projection space. Furthermore, under the constraint $x^T x = 1$, this optimization problem can be further equated to solve the following eigenvalue-eigenvector matrix equation:

$$S_b^{Mat} x = \lambda S_w^{Mat} x. \quad (7)$$

At last, the completely new features are obtained by determining the appropriate x , which will be used in the subsequent classification task. In this experiment, the PSSM of N protein sequences of size $T \times 20$ was used as input to the MatFLDA algorithm on the *yeast* and *H. pylori* datasets, where the MatFLDA algorithm was only used for feature extraction. In this way, we obtained the output of a 20×20 dimensional feature matrix by using the MatFLDA algorithm on an original PSSM of protein sequence. In other words, we obtained a feature vector of 1×400 dimensions from each PSSM. Consequently, the output of N PSSMs is N fixed size 20×20 dimensional feature matrices. Thus, each protein pair contains 800 features. Here, in order to clearly understand how to use the MatFLDA algorithm for feature extraction of protein sequences, we give a schematic diagram of MatFLDA feature extraction for a protein pair namely Histone H4 and Regulatory protein SIR3 in the *Saccharomyces cerevisiae* dataset, as shown in Figure 1.

2.4. Random Ferns (RFs). Random fern classifier is developed based on random forests, but it is different from the random forest [46, 47]. Here, by giving a PSSM in a protein sequence, our main task is to assign it to the most likely class. Let $c_i, i = 1, 2$, be the set of classes, where 1 indicates an interacting protein and 2 is a noninteracting protein. Let $x_j, j = 1, 2, \dots, N$, be the set of normalized 20×20 dimensional features that will be calculated by using the MatFLDA algorithm on the PSSM that we are trying to classify. Formally, we are looking for [48]

$$c_i' = \arg \max_{c_i} P(C = c_i | x_1, x_2, \dots, x_N), \quad (8)$$

where C , a random variable, represents the class of protein. The aim of this paper is to model the posterior interacting protein class probability by giving a set of N features. This can be expressed in terms of the Bayesian formula, as

$$P(C = c_i | x_1, x_2, \dots, x_N) = \frac{P(x_1, x_2, \dots, x_N | C = c_i) \times P(C = c_i)}{P(x_1, x_2, \dots, x_N)}. \quad (9)$$

Assuming a uniform prior $P(C)$, since the denominator is just a scale factor, it is independent and is common for all the classes. Thus, by removing the priors $P(x_1, x_2, \dots, x_N)$, the problem reduces to finding

$$c_i' = \arg \max_{c_i} P(x_1, x_2, \dots, x_N | C = c_i). \quad (10)$$

But learning the complete representation of the joint probability of all features is very intractable. According to the Naive Bayes theory, it is assumed that all features are completely independent, that is,

$$P(x_1, x_2, \dots, x_N | C = c_i) = \prod_{j=1}^N P(x_j | C = c_i). \quad (11)$$

However, this independence assumption is usually wrong because it completely ignores the correlation between features in practice. To account for the dependencies between these features while making the problem tractable, a better compromise is to divide our features into M groups of size $S = N/M$. These groups are what we define as ferns, and we calculate the joint probability for features in each fern. The conditional probability is expressed as follows:

$$P(x_1, x_2, \dots, x_N | C = c_i) = \prod_{k=1}^M P(F_k | C = c_i). \quad (12)$$

where $F_k = \{x_{\vartheta(k,1)}, x_{\vartheta(k,2)}, \dots, x_{\vartheta(k,S)}\}$, $k = 1, \dots, M$, refers to the k th fern, and $\vartheta(k, j)$ is a random permutation function. Therefore, we follow a seminaive Bayesian method by modeling only some of the dependencies between features. In addition, the class conditional probabilities $P(F_m | C = c_i)$ are estimated for each fern F_m and class c_i in the training phase. For each fern F_m , these terms can be described as

$$p_{k,c_i} = P(F_m = k | C = c_i) = \frac{N_{k,c_i}}{N_{c_i}}, \quad (13)$$

where N_{k,c_i} represents the number of training samples of class c_i that evaluates to fern value $k, k = 1, 2, \dots, K$. Here, $K = 2^S$, and N_{c_i} represents the total number of samples for class c_i . However, when the number of samples given is not infinitely large, both N_{k,c_i} and p_{k,c_i} will be zero. To overcome this problem, p_{k,c_i} is rewritten as

$$p_{k,c_i} = \frac{N_{k,c_i} + N_r}{N_{c_i} + K \times N_r}, \quad (14)$$

where N_r is a regularization term, which behaves as a uniform Dirichlet prior over feature values. $N_r = 1$ is used to guarantee the results above zero. In this experiment, we set two important parameters of the random ferns classifier, where S (the depth of ferns) was set to 20 and M (the number of ferns) was set to 140. Finally, the features extracted by the MatFLDA algorithm are normalized and then fed into the random ferns classifier to predict whether each protein pair interacts with each other.

3. Results and Discussion

3.1. Evaluation Criteria. In this paper, to ensure the robustness of the proposed model and avoid overfitting and data dependency, we adopted five-fold cross-validation to assess

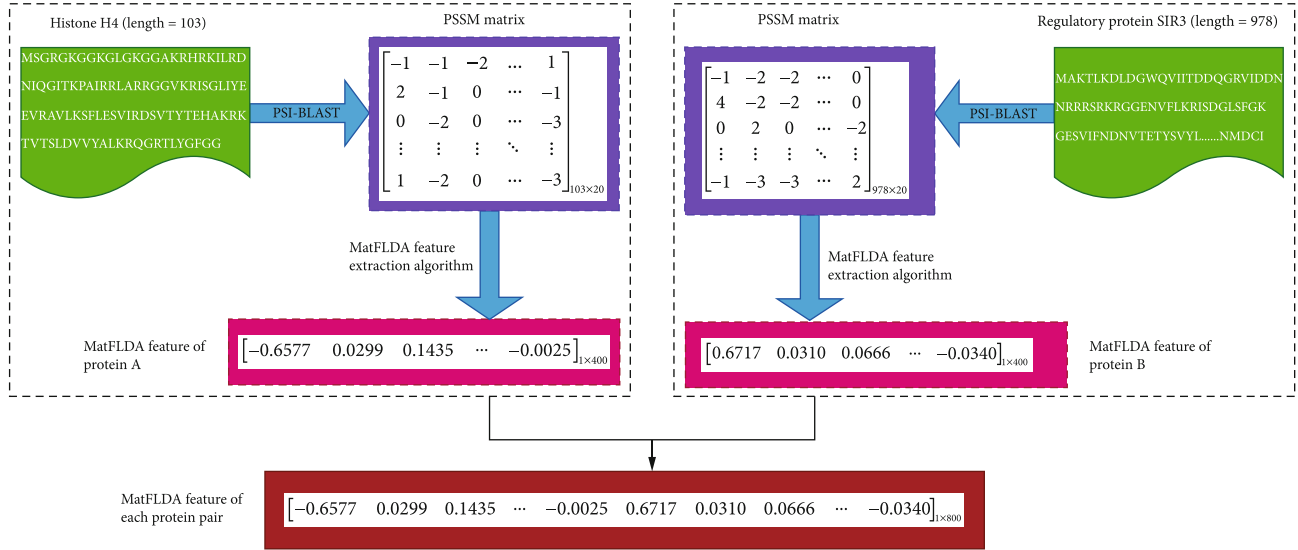


FIGURE 1: Flow chart of MatFLDA feature extraction for each protein pair.

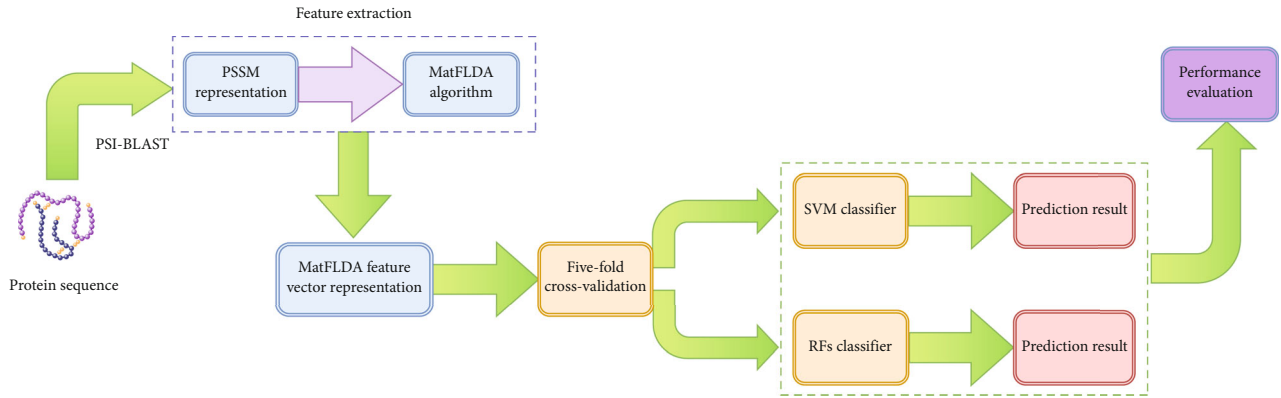


FIGURE 2: The flow of the proposed scheme.

 TABLE 1: Five-fold cross-validation prediction results achieved in predicting *Yeast* PPI dataset.

Testing set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
1	95.26	99.41	91.06	90.94	94.79
2	94.99	99.33	90.85	90.47	93.44
3	94.81	98.81	90.55	90.12	94.11
4	94.77	99.21	90.27	90.05	94.00
5	95.31	98.92	91.49	91.02	94.99
Average	95.03 ± 0.25	99.14 ± 0.26	90.84 ± 0.47	90.52 ± 0.45	94.27 ± 0.63

the effectiveness of this method in predicting PPIs. Specifically, we first divide the experimental dataset into five parts and then select four of them as the training dataset and the additional one as the testing dataset. Finally, the average values of the five independent experiments are used as prediction results. Here, the following assessments are used, including overall prediction accuracy (ACC), precision (PE), sensitivity (SN), and Matthews correlation coefficient (MCC), which are defined as follows

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (15)$$

$$\text{PE} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (16)$$

$$\text{SN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (17)$$

TABLE 2: Five-fold cross-validation prediction results achieved in predicting *H. pylori* PPI dataset.

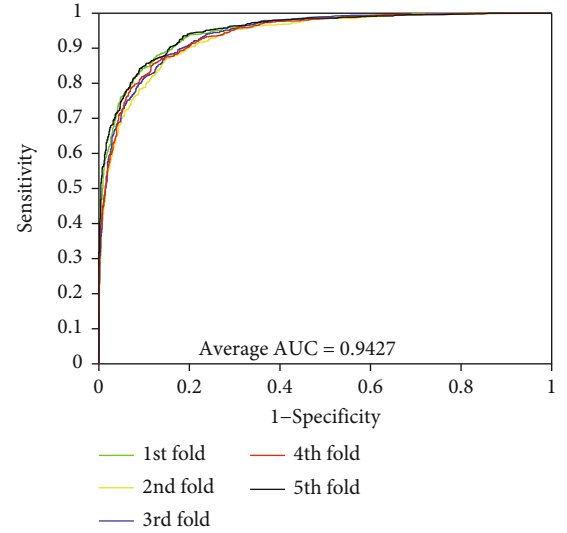
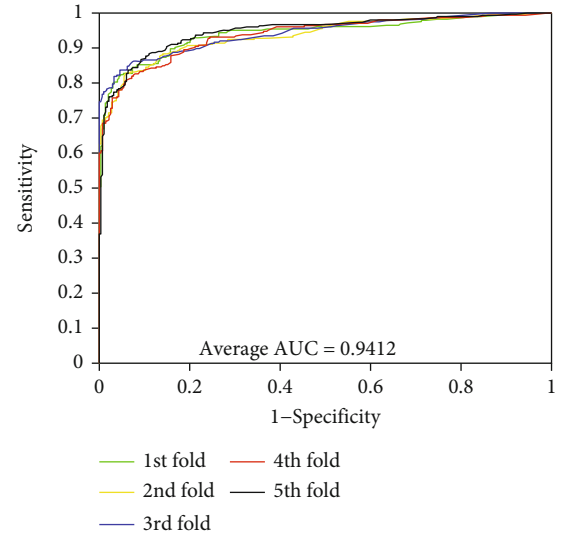
Testing set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
1	85.76	79.30	95.77	75.19	94.16
2	85.59	79.15	96.56	74.76	93.63
3	85.59	79.27	94.20	75.11	94.28
4	85.59	80.44	95.74	74.58	93.78
5	84.22	78.17	96.35	72.43	94.78
Average	85.35 ± 0.64	79.27 ± 0.81	95.72 ± 0.92	74.41 ± 1.14	94.12 ± 0.45

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \quad (18)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Among them, TP indicates the number of true PPIs that are predicted correctly, TN represents the number of true noninteracting pairs that are predicted correctly. FP indicates the number of true interacting pairs not found in positive dataset, and FN represents the number of true interacting pairs not found in negative dataset. MCC is used as a balance indicator to measure the quality of binary classification in data mining, which value ranges between -1 and +1 representing the correlation coefficient between the observed results and the predicted results. In this experiment, the receiver operating characteristic (ROC) curve [49] and the area under the ROC curve (AUC) [50] are employed to evaluate the prediction performance of the proposed model. The AUC value of the classifier is larger, the prediction performance of the method is superior, and the model constructed is more stable. The flow of the proposed scheme is shown in Figure 2.

3.2. Prediction Performance of Proposed Model. In order to assess the effectiveness and stability of the model combining MatFLDA and RFs to predict PPIs, we tested the model on *Yeast* and *H. pylori* datasets in this section. In addition, for reducing deviations of the proposed method and avoiding overfitting, five-fold cross-validation was performed in the experiment. Specifically, the whole dataset was divided into five parts, including five training and five testing datasets, respectively, and then we obtained five models by carrying out separate experiments for each dataset. Finally, the average values of the five models were selected as the prediction results of our experiments. In order to obtain more accurate and reliable experimental results, the fern size S and fern number M of the random ferns classifier were set to be the same on *Yeast* and *H. pylori* datasets. Here, $S = 20$ and $M = 140$. The five-fold cross-validation prediction results of the RFs prediction model employing the MatFLDA feature extraction algorithm of protein sequence on two benchmark datasets are shown in Tables 1 and 2.

As can be seen from Table 1, the accuracies of the five experiments were 95.26%, 94.99%, 94.81%, 94.77%, and 95.31% when PPIs were performed on the *Yeast* dataset. The precisions are $\geq 98.81\%$, the sensitivities are $\geq 90.27\%$,

FIGURE 3: ROC curves performed using the proposed method on *Yeast* dataset.FIGURE 4: ROC curves performed using the proposed method on *H. pylori* dataset.

and the MCCs are $\geq 90.05\%$. The standard deviations corresponding to these four assessment values are 0.25%, 0.26%, 0.47%, and 0.45%, respectively. At the same time, we can see that these standard deviations are relatively low.

TABLE 3: Five-fold cross-validation results by using two models on the *Yeast* dataset.

Classifier	Testing set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
SVM	1	81.63	84.29	77.73	69.91	87.06
	2	80.02	83.86	75.61	67.92	86.23
	3	79.44	80.79	76.39	67.25	84.55
	4	80.20	83.28	75.63	68.11	84.74
	5	80.69	82.83	76.83	68.72	86.34
	Average	80.39 ± 0.82	83.01 ± 1.36	76.44 ± 0.89	68.38 ± 1.00	85.78 ± 1.09
RFs	1	95.26	99.41	91.06	90.94	94.79
	2	94.99	99.33	90.85	90.47	93.44
	3	94.81	98.81	90.55	90.12	94.11
	4	94.77	99.21	90.27	90.05	94.00
	5	95.31	98.92	91.49	91.02	94.99
	Average	95.03 ± 0.25	99.14 ± 0.26	90.84 ± 0.47	90.52 ± 0.45	94.27 ± 0.63
Random Forest	Average	95.48 ± 0.29	97.71 ± 0.38	93.14 ± 0.71	91.35 ± 0.53	95.48 ± 0.28
XGBoost	Average	94.08 ± 1.08	96.43 ± 0.92	91.54 ± 1.52	88.86 ± 1.91	98.59 ± 0.34

TABLE 4: Five-fold cross-validation results by using two models on the *H. pylori* dataset.

Classifier	Testing set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
SVM	1	82.85	81.72	83.45	71.57	89.26
	2	82.33	80.52	85.22	70.86	89.87
	3	79.42	76.17	82.25	67.25	86.20
	4	82.33	83.22	82.95	70.85	89.16
	5	83.53	84.75	83.06	72.47	90.22
	Average	82.09 ± 1.57	81.28 ± 3.26	83.39 ± 1.12	70.60 ± 1.99	88.94 ± 1.60
RFs	1	85.76	79.30	95.77	75.19	94.16
	2	85.59	79.15	96.56	74.76	93.63
	3	85.59	79.27	94.20	75.11	94.28
	4	85.59	80.44	95.74	74.58	93.78
	5	84.22	78.17	96.35	72.43	94.78
	Average	85.35 ± 0.64	79.27 ± 0.81	95.72 ± 0.92	74.41 ± 1.14	94.12 ± 0.45
Random Forest	Average	87.27 ± 0.82	85.90 ± 0.72	89.09 ± 2.45	77.73 ± 1.21	93.28 ± 0.69
XGBoost	Average	85.11 ± 1.22	84.28 ± 3.10	86.49 ± 3.25	74.64 ± 1.72	91.59 ± 0.82

Similarly, the average values of accuracy, precision, sensitivity, and MCC were 85.35%, 79.27%, 95.72%, and 74.41% when exploring PPIs of *H. pylori* dataset. We can see from Table 2 that the standard deviations corresponding to these four evaluation values are 0.64%, 0.81%, 0.92%, and 1.14%, respectively. In order to better visualize the performance of combining RFs and MatFLDA to predict PPIs, we plot the ROC curves on two benchmark datasets. In addition, MCC and AUC values were also calculated to better quantify the predictive performance of the proposed model. The ROC curves performed on the two benchmark datasets are shown in Figures 3 and 4.

From Figures 1 and 2, we can see that the average AUC values obtained by the proposed method were 94.27% and 94.12% for the experiments on *Yeast* and *H. pylori* datasets, respectively. The promising results show that the proposed

method is feasible, effective, and practical for detecting PPIs. The excellent prediction performance mainly depends on the selection of the feature extraction algorithm and classification model of the proposed method. It can be seen that the MatFLDA feature extraction descriptor can effectively retain useful information from the original protein sequences. Moreover, the high prediction accuracies and low standard deviations further indicate that the proposed method is robust for predicting PPIs.

3.3. Comparison of the Four Methods Using the Same Feature Representation. Generally, the same feature extraction approach by combining different classifiers will yield different prediction results when using machine-learning-based methods to predict PPIs. In this section, we performed PPI experiments using the same feature extraction method on

the state-of-the-art individual classifier support vector machine (SVM) and the proposed ensemble learning classifier random ferns in order to further evaluate the prediction performance of the proposed model. It should be noted that the LIBSVM toolbox, which was downloaded from <https://www.csie.ntu.edu.tw/~cjlin/libsvm/> [51], was employed in this experiment to carry out the PPI classification task. In our experiment, a polynomial function is used as the kernel function and the initial values of SVM are $c = 0.1$, $g = 0.2$ and $c = 0.01$, $g = 0.1$ when predicting PPIs using five-fold cross-validation on *Yeast* and *H. pylori* datasets, respectively. For SVM and RF classifiers, all input feature vectors are normalized by the zero-mean normalization method.

The experimental results of PPIs based on RFs and SVM-based classifiers are presented in Tables 3 and 4 on *Yeast* and *H. pylori* datasets, respectively. From Table 3, the average values of accuracy, precision, sensitivity, and MCC of the RF method on *Yeast* dataset are as high as 95.03%, 99.14%, 90.84%, and 90.52%, respectively. However, when employing the SVM classifier, we yielded relatively poor prediction results with the average values of accuracy, precision, sensitivity, and MCC of 80.39%, 83.01%, 76.44%, and 68.38%, respectively. It can be observed that the maximum accuracy obtained by the SVM classifier is 81.63%, which is 13% lower than the minimum accuracy obtained by the RF method. Similarly, as presented in Table 4, the average accuracy by utilizing SVM method in *H. pylori* dataset is 82.09%, among which the results of five models are 82.85%, 82.33%, 79.42%, 82.33%, and 83.53%, respectively. Additionally, for further evaluation, the ROC (receiver operating characteristic) curves and AUC values based on the SVM method are also calculated (see Figures 5 and 6). The average AUC values obtained by the same feature extraction method on *Yeast* and *H. pylori* datasets were 85.78% and 88.94%, respectively. In addition, we also evaluate the prediction performance of the proposed model using Random Forest and XGBoost classifiers by employing the same features. Comparing the proposed model with these three models, we can clearly see the proposed model achieves good performance in the prediction of PPIs. Thus, the proposed model can provide a useful tool for detecting PPIs and other bioinformatics tasks.

3.4. Comparison with other PPI Prediction Methods. Currently, many computational methods that are based on data mining knowledge have been presented for predicting sequence-based PPIs. In this section, to verify the performance of the proposed model, we measure the proposed method by comparing with several other state-of-the-art methods on the *Yeast* and *H. pylori* datasets. Specifically, we compared the proposed method with previous work on PPI prediction presented by Guo et al., Yang et al., Zhou et al., You et al., Du et al., and Wong et al. on the *Yeast* dataset. Table 5 lists the PPI prediction results of the above methods on the same *Yeast* dataset.

As shown in Table 5, the accuracy, sensitivity, precision, and MCC of the MatFLDA_RFs method are 95.03%, 90.84%, 99.14%, and 90.52%, respectively. Compared with other existing methods listed, the accuracy of the proposed

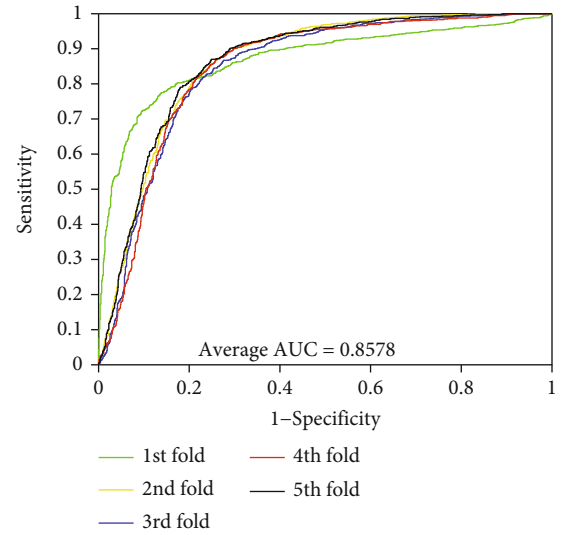


FIGURE 5: ROC curves performed using the SVM method on *Yeast* dataset.

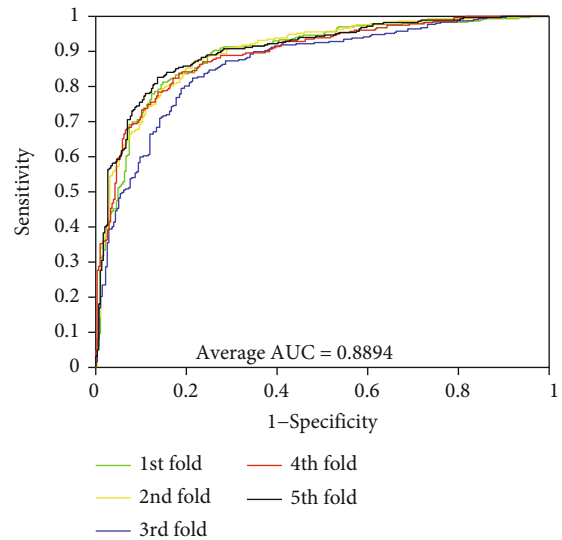


FIGURE 6: ROC curves performed using the SVM method on *H. pylori* dataset.

method increased by about 0.1% to 9%. The ACC of MatFLDA_RFs method is 7.67% higher than the AC method, 8.88% higher than the Cod4 + KNN method, 6.47% higher than the SVM + LD method, 3.67% higher than the MCD + SVM method, 0.89% higher than the LRA + RF method, 0.60% higher than the DeepPPI method, and 1.11% higher than the PR - LPQ + RF method. The PE of MatFLDA_RF method is 11.32% higher than the AC method, 8.90% higher than the Cod4 + KNN method, 9.64% higher than the SVM + LD method, 7.20% higher than the MCD + SVM method, 2.04% higher than the LRA + RF method, 2.49% higher than the DeepPPI method, and 2.69% higher than the PR - LPQ + RF method. The MCC of MatFLDA_RFs method is 13.37% higher than the SVM + LD method, 6.31% higher than the MCD + SVM

TABLE 5: The prediction ability of the other methods on the *Yeast* dataset.

Related work	Method	ACC (%)	SN (%)	PE (%)	MCC (%)	AUC (%)
Guo et al.'s work [17]	AC	87.36 ± 1.38	87.30 ± 4.68	87.82 ± 4.33	N/A	N/A
	ACC	89.33 ± 2.67	89.93 ± 3.68	88.87 ± 6.16	N/A	N/A
Yang et al.'s work [18]	Cod4 + KNN	86.15 ± 1.17	81.03 ± 1.74	90.24 ± 1.34	N/A	N/A
Zhou et al.'s work [52]	SVM + LD	88.56 ± 0.33	87.37 ± 0.22	89.50 ± 0.60	77.15 ± 0.68	95.07 ± 0.39
You et al.'s work [53]	MCD + SVM	91.36 ± 0.36	90.67 ± 0.69	91.94 ± 0.62	84.21 ± 0.59	97.07 ± 0.12
You et al.'s work [54]	LRA + RF	94.14 ± 1.8	91.22 ± 1.6	97.10 ± 2.1	88.96 ± 2.6	94.20 ± 1.7
Du et al.'s work [21]	DeepPPI	94.43 ± 0.30	92.06 ± 0.36	96.65 ± 0.59	88.97 ± 0.62	N/A
Wong et al.'s work [55]	PR – LPQ + RF	93.92 ± 0.36	91.10 ± 0.31	96.45 ± 0.45	88.56 ± 0.63	N/A
Proposed method	MatFLDA_RFs	95.03 ± 0.25	90.84 ± 0.47	99.14 ± 0.26	90.52 ± 0.45	94.27 ± 0.63

Note: N/A means not available.

TABLE 6: The prediction ability of the different methods on the *H. pylori* PPI dataset.

Related work	Method	ACC (%)	SN (%)	PE (%)	MCC (%)
Martin et al.'s work [32]	Signature products + SVM	83.40	79.90	85.70	N/A
You et al.'s work [53]	MCD + SVM	84.91	83.24	86.12	74.40
Nanni's work [56]	WSR	83.70	79.00	87.00	N/A
Bock and Gough's work [57]	Phylogenetic Booststrap	75.80	69.80	80.20	N/A
Nanni's work [56]	LDC	83.00	80.60	85.10	N/A
Shi et al.'s work [58]	Boosting	79.52	80.37	81.69	70.64
Proposed method	MatFLDA_RFs	85.35	95.72	79.27	74.41

Note: N/A means not available.

method, 1.56% higher than the LRA + RF method, 1.55% higher than the DeepPPI method, and 1.96% higher than the PR – LPQ + RF method.

Similarly, Table 6 presents the PPI prediction results of other existing methods on the same *H. pylori* dataset. As shown in Table 6, the prediction performance of the proposed method is better than other existing methods. The obtained values of ACC, SN, PE, and MCC are 85.35%, 95.72%, 79.27%, and 74.41%, respectively. In terms of ACC, the MatFLDA_RFs method is 0.44%-9.55% higher than other methods, 1.95% higher than the Signature Products + SVM method, 0.44% higher than the MCD + SVM method, 1.65% higher than the WSR method, 9.55% higher than the Phylogenetic Booststrap method, 2.35% higher than the LDC method, and 5.83% higher than the Boosting method. These excellent results prove that the proposed method is an effective computational tool suitable for predicting PPIs.

4. Conclusion

The study of proteins and their interactions is essential to understand most biological activities in living cells, such as development, signal transduction, and apoptosis. Therefore, the successful prediction of PPIs will facilitate the study of other related problems in biomedical science. In this work, we present a novel computational approach to detect PPIs,

using the MatFLDA algorithm, the RF classifier, and the PSSM matrix that can preserve protein evolutionary information. More specifically, MatFLDA is used to obtain the feature representation from the PSSM, an evolutionary matrix of protein sequences. This PSSM contains a great deal of valuable and important knowledge for PPI prediction. The RF classifier is then applied to detect novel PPIs. Finally, to measure the PPI identification ability of the developed method, we conducted extensive computational experiments on several benchmark PPI datasets. These excellent experimental results have indicated that the proposed MatFLDA_RF method has a higher identification rate of PPIs than other existing methods and SVM-based approaches. Consequently, the proposed method to identify PPIs is reliable and effective, which can be used as a practical tool for experimental methods, thus, facilitating further research on related problems in the field of bioinformatics.

Data Availability

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

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Yang Li and Zheng Wang contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61873212).

References

- [1] M. Chen, C. J.-T. Ju, G. Zhou et al., "Multifaceted protein-protein interaction prediction based on Siamese residual RCNN," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, 2019.
- [2] J. Y. An, F. R. Meng, Z. H. You, X. Chen, G. Y. Yan, and J. P. Hu, "Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model," *Protein Science*, vol. 25, no. 10, pp. 1825–1833, 2016.
- [3] D.-H. Le and Y.-K. Kwon, "Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization," *Computational Biology and Chemistry*, vol. 44, pp. 1–8, 2013.
- [4] S. Navlakha, A. Gitter, and Z. Bar-Joseph, "A network-based approach for predicting missing pathway interactions," *PLoS Computational Biology*, vol. 8, no. 8, article e1002640, 2012.
- [5] F. Pazos and A. Valencia, "In silico two-hybrid system for the selection of physically interacting protein pairs," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 2, pp. 219–227, 2002.
- [6] M. Foltman and A. Sanchez-Diaz, "Studying protein-protein interactions in budding yeast using co-immunoprecipitation," in *Yeast Cytokinesis*, pp. 239–256, Springer, 2016.
- [7] M. R. O'Connell, R. Gamsjaeger, and J. P. Mackay, "The structural analysis of protein-protein interactions by NMR spectroscopy," *Proteomics*, vol. 9, no. 23, pp. 5224–5232, 2009.
- [8] H. Zhu and M. Snyder, "Protein chip technology," *Current Opinion in Chemical Biology*, vol. 7, no. 1, pp. 55–63, 2003.
- [9] C. Huang, F. Morcos, S. P. Kanaan, S. Wuchty, D. Z. Chen, and J. A. Izaguirre, "Predicting protein-protein interactions from protein domains using a set cover approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 78–87, 2007.
- [10] R. Jansen, H. Yu, D. Greenbaum et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.
- [11] B. Wang, P. Chen, D. S. Huang, J. Li, T. M. Lok, and M. R. Lyu, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate," *FEBS Letters*, vol. 580, no. 2, pp. 380–384, 2006.
- [12] Q. Sheng and C. Lu, "Predicting protein-protein interaction based on protein secondary structure information using Bayesian classifier," *Journal of Inner Mongolia University of Science and Technology*, vol. 1, 2010.
- [13] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events," *Nature*, vol. 402, no. 6757, pp. 86–90, 1999.
- [14] T. Sato, Y. Yamanishi, M. Kanehisa, H. Toh, U. Jp, and T. Kyoto, "Prediction of protein-protein interactions based on real-valued phylogenetic profiles using partial correlation coefficient," *BMC Genomics*, vol. 10, p. 288, 2004.
- [15] J.-Y. An, Z.-H. You, F.-R. Meng, S.-J. Xu, and Y. Wang, "RVMA: using the relevance vector machine model combined with average blocks to predict the interactions of proteins from protein sequences," *International Journal of Molecular Sciences*, vol. 17, no. 5, p. 757, 2016.
- [16] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC bioinformatics*, vol. 14, no. S10, pp. 1–11, 2013.
- [17] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.
- [18] L. Yang, J.-F. Xia, and J. Gui, "Prediction of protein-protein interactions from protein sequence using local descriptors," *Protein and Peptide Letters*, vol. 17, no. 9, pp. 1085–1090, 2010.
- [19] Y.-A. Huang, Z.-H. You, X. Gao, L. Wong, and L. Wang, "Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence," *BioMed Research International*, vol. 2015, Article ID 902198, 10 pages, 2015.
- [20] Y. Wang, Z. You, X. Li, X. Chen, T. Jiang, and J. Zhang, "PCVMZM: using the probabilistic classification vector machines model combined with a Zernike moments descriptor to predict protein-protein interactions from protein sequences," *International Journal of Molecular Sciences*, vol. 18, no. 5, p. 1029, 2017.
- [21] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "DeepPPI: boosting prediction of protein-protein interactions with deep neural networks," *Journal of Chemical Information and Modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.
- [22] Y. E. Göktepe and H. Kodaz, "Prediction of protein-protein interactions using an effective sequence based combined method," *Neurocomputing*, vol. 303, pp. 68–74, 2018.
- [23] X.-Y. Song, Z.-H. Chen, X.-Y. Sun, Z.-H. You, L.-P. Li, and Y. Zhao, "An ensemble classifier with random projection for predicting protein-protein interactions using sequence and evolutionary information," *Applied Sciences*, vol. 8, no. 1, p. 89, 2018.
- [24] J. C. Beltran, P. Valdez, and P. Naval, "Predicting protein-protein interactions based on biological information using extreme gradient boosting," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–6, Siena, Italy, 2019.
- [25] K. Jha and S. Saha, "Amalgamation of 3D structure and sequence information for protein-protein interaction prediction," *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020.
- [26] J. Y. Chen, S. Mamidipalli, and T. Huan, "HAPPI: an online database of comprehensive human annotated and predicted protein interactions," *BMC Genomics*, vol. 10, no. S1, p. S16, 2009.
- [27] L. Licata, L. Briganti, D. Peluso et al., "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, no. D1, pp. D857–D861, 2012.
- [28] D. Alonso-López, F. J. Campos-Laborie, M. A. Gutiérrez et al., "APID database: redefining protein-protein interaction experimental evidences and binary interactomes," *Database*, vol. 2019, article baz005, p. 2019, 2019.

- [29] G. D. Bader, D. Betel, and C. W. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.
- [30] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.
- [31] J.-C. Rain, L. Selig, H. De Reuse et al., "The protein-protein interaction map of *Helicobacter pylori*," *Nature*, vol. 409, no. 6817, pp. 211–215, 2001.
- [32] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.
- [33] 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.
- [34] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "A trigram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition," *IEEE Transactions on Nanobioscience*, vol. 13, no. 1, pp. 44–50, 2014.
- [35] M. Waris, K. Ahmad, M. Kabir, and M. Hayat, "Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix," *Neurocomputing*, vol. 199, pp. 154–162, 2016.
- [36] Y. Li, X. Z. Liu, Z. H. You, L. P. Li, J. X. Guo, and Z. Wang, "A computational approach for predicting drug–target interactions from protein sequence and drug substructure fingerprint information," *International Journal of Intelligent Systems*, vol. 36, no. 1, pp. 593–609, 2021.
- [37] X. Pu, J. Guo, H. Leung, and Y. Lin, "Prediction of membrane protein types from sequences and position-specific scoring matrices," *Journal of Theoretical Biology*, vol. 247, no. 2, pp. 259–265, 2007.
- [38] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, and Q. Ma, "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique," *Bioinformatics*, vol. 35, no. 14, pp. 2395–2402, 2019.
- [39] 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.
- [40] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of Theoretical Biology*, vol. 320, pp. 41–46, 2013.
- [41] Y. Li, L.-P. Li, L. Wang, C.-Q. Yu, Z. Wang, and Z.-H. You, "An ensemble classifier to predict protein–protein interactions by combining PSSM-based evolutionary information with local binary pattern model," *International Journal of Molecular Sciences*, vol. 20, no. 14, p. 3511, 2019.
- [42] S. Chen, Y. Zhu, D. Zhang, and J.-Y. Yang, "Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA," *Pattern Recognition Letters*, vol. 26, no. 8, pp. 1157–1167, 2005.
- [43] A. Kalsoom, M. Maqsood, M. A. Ghazanfar, F. Aadil, and S. Rho, "A dimensionality reduction-based efficient software fault prediction using fisher linear discriminant analysis (FLDA)," *The Journal of Supercomputing*, vol. 74, no. 9, pp. 4568–4602, 2018.
- [44] D. AbuZeina and F. S. Al-Anzi, "Employing fisher discriminant analysis for Arabic text classification," *Computers & Electrical Engineering*, vol. 66, pp. 474–486, 2018.
- [45] T. Wenjing, G. Fei, D. Renren, S. Yujuan, and L. Ping, "Face recognition based on the fusion of wavelet packet sub-images and fisher linear discriminant," *Multimedia Tools and Applications*, vol. 76, no. 21, pp. 22725–22740, 2017.
- [46] Y. Dong, Y. Zhang, J. Yue, and Z. Hu, "Comparison of random forest, random ferns and support vector machine for eye state classification," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11763–11783, 2016.
- [47] M. Zhang and M. Xin, "Human detection using random color similarity feature and random ferns classifier," *PLoS One*, vol. 11, no. 9, article e0162830, 2016.
- [48] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.
- [49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [50] L.-P. Li, Y.-B. Wang, Z.-H. You, Y. Li, and J.-Y. An, "PCLPred: a bioinformatics method for predicting protein–protein interactions by combining relevance vector machine model with low-rank matrix approximation," *International Journal of Molecular Sciences*, vol. 19, no. 4, p. 1029, 2018.
- [51] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.
- [52] Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, "Prediction of protein-protein interactions using local description of amino acid sequence," in *Advances in computer science and education applications*, pp. 254–262, Springer, 2011.
- [53] Z.-H. You, L. Zhu, C.-H. Zheng, H.-J. Yu, S.-P. Deng, and Z. Ji, "Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set," *BMC Bioinformatics*, vol. 15, no. S9, pp. 1–9, 2014.
- [54] Z.-H. You, X. Li, and K. C. Chan, "An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers," *Neurocomputing*, vol. 228, pp. 277–282, 2017.
- [55] L. Wong, Z.-H. You, S. Li, Y.-A. Huang, and G. Liu, "Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor," in *International Conference on Intelligent Computing*, pp. 713–720, Fuzhou, China, 2015.
- [56] L. Nanni, "Fusion of classifiers for predicting protein-protein interactions," *Neurocomputing*, vol. 68, pp. 289–296, 2005.
- [57] J. R. Bock and D. A. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–134, 2003.
- [58] M.-G. Shi, J.-F. Xia, X.-L. Li, and D.-S. Huang, "Predicting protein–protein interactions from sequence using correlation coefficient and high-quality interaction dataset," *Amino Acids*, vol. 38, no. 3, pp. 891–899, 2010.