Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5764060, 11 pages
https://doi.org/10.1155/2020/5764060

Research Article

Prediction of Protein-Protein Interactions from Protein Sequences by Combining MatPCA Feature Extraction Algorithms and Weighted Sparse Representation Models

Zheng Wang, Yang Li, Zhu-Hong You, Li-Ping Li, Xin-Ke Zhan, and Jie Pan

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

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

Received 26 February 2020; Revised 25 July 2020; Accepted 12 August 2020; Published 24 September 2020

Academic Editor: Elena Zaitseva

Identifying protein-protein interactions (PPIs) plays a vital role in a number of biological activities such as signal transduction, transcriptional regulation, and apoptosis. Although advances in high-throughput technologies have generated large amounts of PPI data for different species, they only cover a small part of the entire PPI network. Furthermore, traditional experimental methods are generally expensive, time-consuming, tedious, and prone to high false-positive rates. Therefore, to overcome this problem, it is necessary to develop a novel computational method for predicting PPIs. In this article, we propose an efficient computational method to detect protein-protein interactions using only protein sequence information, which integrates the MatPCA feature extraction algorithm and the weighted sparse representation classifier. As a result, when predicting PPIs on yeast, human, and *H. pylori* datasets, the proposed method achieves superior prediction performance with an average accuracy of 94.55%, 97.48%, and 83.64%, respectively. These experimental results further illustrate that the proposed method is reliable and robust in predicting PPIs, which can be regarded as a useful complement to the experimental method.

1. Introduction

Proteins are an important part of all organisms and are also one of the most versatile organic macromolecules in living systems. Generally, proteins interact with each other to perform their functions [1]. Therefore, predicting protein-protein interactions (PPIs) is critical to elucidating the function of proteins and exploring the pathogenesis of diseases. Nowadays, researchers have developed many biologically based experimental methods to identify protein interactions by employing high-throughput technologies, for example, mass spectrometry [2], immunoprecipitation [3], yeast two-hybrid systems [4], and protein chips [5]. Although traditional experimental methods have also achieved some results in detecting PPIs, they only account for a small part of the entire PPI network. Furthermore, these methods have weak generalization performance and high false-negative and false-positive rates, which are both costly and time-consuming [6, 7]. Thus, as a complement to experimental methods to reduce costs, it is indispensable to develop reliable computational methods to predict PPIs [8].

To date, a series of PPI detection methods based on different data types have been suggested, including gene fusion [9], protein domains [10], amino acid index distribution [11], phylogenetic profile [12], and protein structure information [13]. Nevertheless, the disadvantage of these methods is that they need to consider pre-existing knowledge about the protein pairs, such as protein functional domains and 3D structure information of proteins. In practice, there are relatively few known proteins with a 3D structure, which will limit the extensive use of these methods [14]. Moreover, with the development of genomic technologies, the sequence data of proteins have also shown explosive growth and are readily available. Compared to these data types, researchers have developed many methods based on the amino acid sequence of proteins to infer potential PPIs. Experimental results confirm that using only protein amino acid sequences is feasible in predicting PPIs [15–17].

A growing number of computational approaches have been developed to predict PPIs based on protein sequence information [18, 19]. For example, Martin et al. [20] suggested a new signature descriptor to predict PPIs by combining the sequence-based protein description with the experimental information collected by the protein-protein interaction screen. The method achieves 70% to 80% accuracy on publicly available yeast and *H. pylori* datasets. Qi et al. [21] proposed six different classifiers such as decision tree, random forest (RF), naive Bayes, logistic regression, RF similarity-based k-nearest neighbors, and support vector machine to compare their prediction results in detecting PPIs. Guo et al. [22] used autocovariance to characterize protein sequences and combined support vector machine classifiers to predict PPIs. The method reached 88.09% prediction accuracy on the *Saccharomyces cerevisiae* dataset. You et al. [23] designed a computational method named PCA-EELM (principal component analysis-ensemble extreme learning machine) to detect PPIs, which used only protein sequence information. The prediction accuracy, precision, and sensitivity of the method on the yeast dataset were 87.00%, 87.59%, and 86.15%, respectively. Huang et al. [24] applied a global encoding feature extraction algorithm and a weighted sparse representation classifier to protein sequences to predict PPIs, in which good prediction accuracy was obtained. Du et al. [25] adopted a deep neural network method to predict PPIs. This method uses the integrated multiple feature descriptors to represent the feature information of the amino acid sequence. When performing PPIs on the yeast dataset, this model yields 92.5% prediction accuracy. Although these machine learning methods have achieved good results in predicting PPIs, the prediction accuracy of the algorithms still needs to be further improved to develop efficient and accurate prediction models.

In this study, we present a novel protein sequence-based computational method to detect protein-protein interactions by employing a matrix representation of protein sequences. Specifically, we first converted each protein sequence into a numerical substitution matrix representation (SMR). Secondly, we adopt MatPCA to extract features from the SMR to characterize protein amino acid sequences. By doing this, we can represent each protein pair as an 800-dimensional feature vector. Finally, we feed the obtained feature vectors into a weighted sparse representation-based classifier (WSRC) model, which is used to perform classification tasks for protein-protein interaction prediction. In this work, the proposed method was applied to three different biological datasets, namely, yeast, *H. pylori*, and human, for identifying protein-protein interactions. To further evaluate the prediction performance of our model, we compared the proposed method with the SVM-based method and other existing methods, respectively. The experimental results show that the proposed method achieves superior prediction accuracy in predicting PPIs as compared with the existing methods.

2. Materials and Methods

2.1. Dataset Construction. In this experiment, we used three available PPI datasets including yeast, human, and *H. pylori* to evaluate the predictive performance of the proposed method. Table 1 shows the summary of the datasets used in this experiment. The first dataset is a high-confidence *Saccharomyces cerevisiae* PPI dataset, which was collected from a freely available database of interacting proteins (DIP) [26] provided by Guo et al. [22]. To better perform the proposed method, we preprocessed the protein data, namely, protein pairs with sequence lengths less than fifty residues were deleted because these could be fragments. In addition, more than 40% sequence identity in sequence pairs is considered homologous. In order to eliminate the bias of homologous sequence pairs, we also removed these protein sequence pairs. In this way, we obtained the remaining 5594 protein pairs, which were formed into the positive dataset. Besides, to construct negative samples, we used 5594 additional protein pairs as the negative dataset, which came from different subcellular localizations. At last, the yeast PPI dataset contained 11,188 protein pairs, half of which were negative samples, and the other half were positive samples. The second dataset is the human dataset, which is derived from the Human Protein Reference Database (HPRD) [27]. We obtained 3899 interacting pairs and 4262 noninteracting pairs after removing more than 25% sequence identity from those protein pairs. Specifically, the interacting protein pairs were screened from 2502 different human proteins. Considering that proteins in diverse subcellular compartments do not interact with each other, noninteracting protein pairs were obtained from 661 different human proteins [28]. Finally, the human dataset consisted of 8161 protein pairs (3899 positive samples and 4262 negative samples). The third dataset is the *Helicobacter pylori* dataset introduced by Martin et al. [20]. The dataset contains a total of 2916 samples, including 1458 interacting pairs and 1458 noninteracting pairs.

2.2. Substitution Matrix Representation. It is vital to efficiently represent the intrinsic information of protein sequences when using computational methods to predict protein-protein interactions because an accurate and reliable protein sequence feature representation method will directly influence the prediction results of PPI predictors. Previous studies have shown those interacting or functionally related proteins tend to exhibit similarities in molecular phylogenetic trees during coevolution [29]. In this section, we present a novel feature representation method for predicting PPIs by transforming the evolutionary information of protein sequences into a matrix representation.

The proposed substitution matrix representation (SMR) method is a variant of the method described by Yu et al. [30], which retains the evolutionary information of protein sequences. For any given N-length protein sequence, we can use this novel protein matrix representation method to convert it into an $N \times 20$ matrix. In this experiment, we used a matrix representation method called BLOSUM62, which is

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

Dataset	Protein pairs	Positive samples	Negative samples
Yeast	11,188	5594	5594
Human	8161	3899	4262
H. pylori	2916	1458	1458

a very popular sequence alignment substitution scoring matrix. In this transformation, SMR can be expressed as follows:

$$\mathrm{SMR}(i, j) = M(V(i), j), \quad i = 1, \ldots, N, j = 1, \ldots, 20, \tag{1}$$

where M stands for BLOSUM62, which is a substitution matrix with 20 rows and 20 columns and $M(i, j)$ denotes the probability value of the ith amino acid mutation to the jth amino acid in the evolutionary process. Here, $V = (v1, v2, \ldots, vN)$ is a given protein sequence consisting of N amino acids.

2.3. Matrix Principal Component Analysis (MatPCA). As an effective feature extraction algorithm, MatPCA can deal with both vector pattern and matrix pattern, which was originally introduced by Chen et al. [31]. The idea of MatPCA is derived from the 2-dimensional principal component analysis or image principal component analysis, which mainly performs operations on the image matrix. The description of MatPCA is as follows.

Suppose a training sample set $A = (A_1, A_2, \ldots, A_M)$ consisting of M samples, and their mean is $\overline{A}$, where $A_j (j = 1, 2, \ldots, M)$ is an $m \times n$ matrix and represents the jth training sample. Let $X = (x_1, x_2, \ldots, x_d)$ be a $d \times n$ projection matrix and x be a projection vector with m components. MatPCA tries to use this projection matrix X for feature extraction by performing the following linear transformation on arbitrary A.

$$Y = X^T (A - \overline{A}), \tag{2}$$

where $Y = [y_1^T, y_2^T, \ldots, y_d^T]^T$ is the $d \times n$ feature matrix and d is the number of projection directions with each row vector $y_i (i = 1, 2, \ldots, d)$ satisfying the equation $y_i = x_i^T (A - \overline{A})$. Hence, for each $A_i (i = 1, \ldots, M)$ in the training set, we have

$$Y_i = X^T (A_i - \overline{A}). \tag{3}$$

In order to obtain the optimal projection vector X and to retain more original information of the training set in the projected space, next, we want to construct the reconstructed error (Re) criterion of all training samples, specifically, minimizing the following criterion:

$$\mathrm{Re}(X) = \frac{1}{M} \sum_{i=1}^{M} \left\| A_i - \widehat{A}_i \right\|^2. \tag{4}$$

Here, $\widehat{A}_i = XY_i + \overline{A}$ is the reconstructed representation for A_i and $\|A_i - \widehat{A}_i\|$ is the matrix $2-$norm, and the equation can also be written as $\|A\|^2 = tr(AA^T)$, where $tr()$ denotes a matrix trace operation. In this way, Re can be simplified to

$$\mathrm{Re}(X) = \frac{1}{M} \sum_{i=1}^{M} \left\| A_i - \overline{A} \right\|^2 - tr\left(X^T S_t^{mat} X\right), \tag{5}$$

and specifically,

$$S_t^{mat} = \frac{1}{M} \sum_{i=1}^{M} (A_i - \overline{A})(A_i - \overline{A})^T, \tag{6}$$

where S_t^{mat} is called the total covariance matrix composed of given sample matrices, and it is easy to prove that S_t^{mat} is a positive semidefinite matrix without negative eigenvalues. It can be seen that the first term of $\mathrm{Re}(X)$ is a constant. Therefore, minimizing $\mathrm{Re}(X)$ is equivalent to maximizing $J(X)$ as follows:

$$J(X) = tr\left(X^T S_t^{mat} X\right). \tag{7}$$

Additionally, under the constraint condition $X^T X = I$, $J(X)$ is optimized to obtain the following eigenvalue-eigenvector matrix equation:

$$S_t^{mat} X = X\Lambda, \tag{8}$$

where I is an identity matrix and T is the transpose of the matrix. Here, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$ is a diagonal matrix, in which diagonal elements correspond to all nonnegative eigenvalues of S_t^{mat}, and X is a matrix composed of eigenvectors. Then, the following formula is used to select appropriate θ:

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geq \theta. \tag{9}$$

Finally, the corresponding fusion features are obtained by determining appropriate d. In this experiment, we obtained 400 features by using MatPCA to analyze the substitution matrix representation of a given protein sequence.

2.4. Weighted Sparse Representation-Based Classification (WSRC). Sparse representation classifier (SRC) is a popular nonparametric data mining method, which was originally proposed by Wright et al. [32]. The main purpose of this method is to sparsely represent the test set by linearly combining the training set in the original sample data. Finally, the new test set is assigned to the class with minimal residue. So far, the sparse representation classifier has been widely used in various fields to solve different practical problems, such as face recognition [33], speaker recognition system [34], text classification [35], and diabetes detection based on facial block color features [36]. The SRC algorithm is described as follows.

Given a training sample matrix $X \in \mathbf{R}^{d \times n}$, consisting of n samples and d-dimensional feature vectors. The SRC algorithm assumes that the lth sample of X is c_l and that there are enough samples corresponding to the K object classes. Then, the entire dataset can be expressed as $X_i = [c_{i1}, c_{i2}, \ldots, c_{in_i}]$, where n_i represents the number of samples that belong to the

ith class. Here, the training sample matrix X can be further written as $X = [X_1, X_2, \ldots, X_K]$. For a new test sample $y \in R^d$ that belongs to the ith class, SRC is used to find such a column vector $\alpha = [\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ini}]$ such that

$$y = \alpha_{i1}c_{i1} + \alpha_{i2}c_{i2} + \cdots + \alpha_{in_i}c_{in_i}. \tag{10}$$

When using a linear combination to denote all training samples, y can be described as follows:

$$y = X\alpha_0. \tag{11}$$

It should be noted that nonzero entries in α_0 are only relevant for the ith class. Then, we have the following:

$$\alpha_0 = \left[0, \ldots, 0, \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in_i}, 0, \ldots, 0\right]^T. \tag{12}$$

Next, to solve the l_0 − norm minimization problem, this can be written as

$$\begin{aligned} \widehat{\alpha}_0 &= \arg\min \|\alpha\|_0 \\ s.t. \quad y &= X\alpha. \end{aligned} \tag{13}$$

However, this is an NP-hard problem to solve (13). Based on the theory of compressive sensing [37], we know that when α is sufficiently sparse, the problem can be transformed into solving the l_1 − norm minimization:

$$\begin{aligned} \widehat{\alpha}_1 &= \arg\min \|\alpha\|_1 \\ s.t. \quad y &= X\alpha. \end{aligned} \tag{14}$$

In order to avoid occlusion, this l_1 − norm minimization should be further transformed into the following stable l_1 − norm minimization problem:

$$\begin{aligned} \widehat{\alpha}_1 &= \arg\min \|\alpha\|_1 \\ s.t. \quad \|y - X\alpha\| &\le \varepsilon, \end{aligned} \tag{15}$$

where $\varepsilon > 0$ denotes the threshold of the residue. Here, we can adopt standard linear programming approaches to solve this minimization problem [38]. Subsequently, the SRC algorithm classifies the given test sample y based on the following rule:

$$g_k(y) = \left\|y - X\widehat{\alpha}_k^1\right\|, \quad k = 1, \ldots, K, \tag{16}$$

where $X\widehat{\alpha}_k^1$ is the reconstructed value, which is formed by the training samples of class k, and g_k represents the residual. When it is satisfied that y belongs to class k, then those entries of $\widehat{\alpha}_1$ related to class k have nonzero values. Finally, the obtained result y will be assigned to the smallest residual.

However, previous research results have demonstrated the fact that, in some cases, the local structure is more crucial than the sparsity of the data [39, 40]. Therefore, in this paper, we propose the weighted sparse representation-based classification model by combining the local structure of the data with the sparse representation. The WSRC algorithm is mainly used to find an appropriate method to assess the relationship between testing and training samples. Here, the distance based on Gaussian kernel is applied to the WSRC algorithm to evaluate the similarity between the two samples because the distance can capture the nonlinear relationship in the dataset. The Gaussian-based distance can be formulated as follows:

$$d_g(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right), \tag{17}$$

where $x, y \in R^d$ represent training and testing samples, respectively, and σ denotes the Gaussian kernel width. Unlike the SRC method, WSRC is used to solve the following weighted l_1 − norm minimization problem:

$$\begin{aligned} \widehat{\alpha}_1 &= \arg\min \|W\alpha\|_1 \\ s.t. \quad y &= X\alpha, \end{aligned} \tag{18}$$

and specifically,

$$diag(W) = \left[d_g\left(y, x_1^1\right), \ldots, d_g\left(y, x_{n_k}^k\right)\right]^T, \tag{19}$$

where W is a diagonal-weighted matrix and n_k represents the number of training samples in class k. Similarly, when dealing with occlusion, the WSRC algorithm needs to solve the stable l_1 − norm minimization problem:

$$\begin{aligned} \widehat{\alpha}_1 &= \arg\min \|W\alpha\|_1 \\ s.t. \quad \|y - X\alpha\| &\le \varepsilon, \end{aligned} \tag{20}$$

where $\varepsilon > 0$ is the threshold of the residue.

2.5. Procedure of the Proposed Method. In this experiment, the workflow diagram of the proposed method is presented in Figure 1. More specifically, the positive samples of the datasets used by this study were experimentally identified PPI data. Each protein sequence is converted into a numerical substitution matrix representation, respectively. Following this, MatPCA algorithm is employed to obtain numerous valuable feature vectors for PPI prediction. Finally, we feed the obtained feature vectors into a WSRC model, and then five-fold cross-validation is performed to evaluate the proposed method.

3. Results and Discussion

3.1. Evaluation Measures. In order to assess the reliability and robustness of the proposed method from different perspectives, we used the following four criteria, overall prediction accuracy (ACC), sensitivity (SN), precision (PE), and Matthews correlation coefficient (MCC), to comprehensively evaluate the model. The definitions of these evaluation indexes are as follows:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}}, \tag{21}$$

$$\text{SN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{22}$$

$$\text{PE} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{23}$$

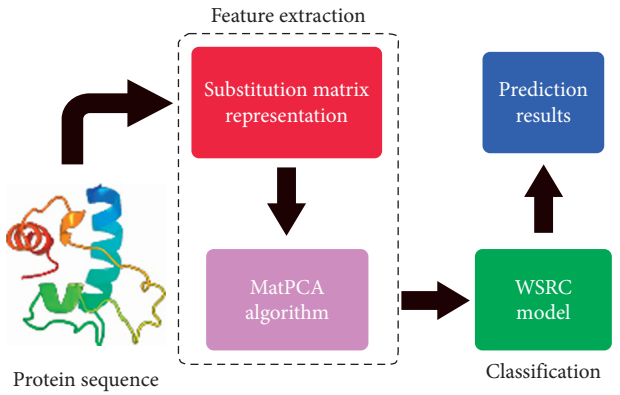

Figure 1: The workflow diagram of the proposed method.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN})}}, \tag{24}$$

where true positive (TP) is the count of samples that are correctly detected as positive by the model, false negative (FN) is the number of samples that are wrongly predicted as negative by the model, true negative (TN) is the count of samples that are correctly detected as negative by the model, and false positive (FP) is the number of samples that are wrongly predicted as positive by the model. Meanwhile, we also constructed the receiver operating characteristic (ROC) curves [41] and calculated the area under the ROC curve (AUC) to show the predictive performance of the proposed model. In general, the stability of the proposed model can be evaluated by comparing the AUC values of different predictors. A larger AUC shows a better predictor.

3.2. Assessment of Prediction Ability. For the sake of fairness, we set the two main parameters σ and ε of the weighted sparse representation-based classifier in this experiment, which correspond to 1.5 and 0.00005, respectively, when performing PPI prediction on three benchmark datasets. In addition, to avoid overfitting, we used a five-fold cross-validation method to verify the stability of the proposed model by dividing the entire dataset into a training set and an independent test set. More specifically, each dataset is divided into five parts, one of which is selected as the test set, and the remaining parts are used as the training set. In this way, we obtained five models, and the final prediction result depends on the average value of five separate experiments. Tables 2–4 list the prediction results on the three datasets based on the proposed model combined with the five-fold cross-validation method.

As shown in Table 2, when using the proposed method to predict the PPIs of the yeast dataset, the average accuracy, precision, sensitivity, and MCC were 94.55%, 92.33%, 97.15%, and 89.68%, respectively, with the corresponding standard deviations as 0.63%, 0.91%, 0.42%, and 1.12%, respectively. From Table 3, which gives the five-fold cross-validation PPI prediction results of the proposed method on the human dataset, we can see that the average accuracy, precision, sensitivity, and MCC of the proposed method are 97.48%, 96.27%, 99.01%, and 95.06%, respectively, with the corresponding standard deviations of 0.31%, 0.55%, 0.32%, and 0.60%, respectively. It can be seen from Table 4 that when detecting the PPIs of the *H. pylori* dataset by using a five-fold cross-validation method, the proposed method achieved an average accuracy, precision, sensitivity, and MCC of 83.64%, 89.71%, 75.98%, and 72.26%, with the corresponding standard deviations of 1.15%, 0.79%, 1.63%, and 1.49%, respectively. Although the proposed model obtained good prediction accuracy, to further demonstrate the reliability of our model, we also calculated the AUC value, which denotes the area under the ROC curve and plotted the ROC curves of the three datasets. Figures 2–4 show the prediction results of ROC curves for performing PPIs on the yeast, human, and *H. pylori* datasets by adopting the proposed method. Finally, we can find that the average AUC values of the proposed model on these three datasets including yeast, human, and *H. pylori* are 97.04%, 99.05%, and 86.35%, respectively. In summary, these excellent experimental results show that our method is effective and reasonable in predicting protein-protein interactions.

3.3. Comparison with the Support Vector Machine Classification Model. To predict the interactions between proteins by using computational methods, researchers have proposed a variety of different machine learning models to detect PPIs. Among them, using the support vector machine (SVM) to predict protein interactions has become one of the most popular approaches in this field. In this section, we use the same feature extraction method combined with the SVM to perform PPI classification experiments on the same dataset to further evaluate the predictive performance of the proposed method. Here, we adopt the radial basis function as the kernel function, and we also optimize the two parameters BoxConstraint and KernelScale of the support vector machine, where BoxConstraint and KernelScale are 10 and 2ˆ1.5 respectively. Finally, the prediction results obtained by combining the five-fold cross-validation method and the SVM classifier are shown in Table 5.

As can be seen from Table 5, when we detected the PPIs by using the support vector machine model on the yeast dataset, the average values of accuracy, sensitivity, precision, MCC, and AUC were 87.25%, 88.22%, 86.58%, 77.75%, and 93.96%, respectively. The average accuracy of the method is reduced by about 7.30% as compared to the WSRC classifier. When we predicted the PPIs by applying the support vector machine model on the human dataset, the average values of accuracy, sensitivity, precision, MCC, and AUC were 92.88%, 95.97%, 90.91%, 86.70%, and 97.56%, respectively. The average accuracy of the method is reduced by about 4.60% as compared to the WSRC classifier. When we detected the PPIs by utilizing the support vector machine model on the *H. pylori* dataset, the average values of accuracy, sensitivity, precision, MCC, and AUC were 82.64%, 94.33%, 76.51%, 70.53%, and 91.60%,

Table 2: Five-fold cross-validation result obtained in predicting the yeast PPI dataset.

Test set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
1	94.90	92.86	97.21	90.32	97.33
2	93.61	90.86	96.61	88.02	96.70
3	94.19	92.03	96.91	89.03	96.96
4	94.90	92.85	97.32	90.31	97.11
5	95.13	93.05	97.72	90.71	97.13
Average	**94.55 ± 0.63**	**92.33 ± 0.91**	**97.15 ± 0.42**	**89.68 ± 1.12**	**97.04 ± 0.24**

Table 3: Five-fold cross-validation result obtained in predicting the human PPI dataset.

Test set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
1	97.00	95.56	98.82	94.15	99.13
2	97.49	96.59	98.72	95.08	98.89
3	97.43	95.82	99.54	94.94	99.23
4	97.86	96.83	99.04	95.80	99.22
5	97.61	96.53	98.93	95.32	98.77
Average	**97.48 ± 0.31**	**96.27 ± 0.55**	**99.01 ± 0.32**	**95.06 ± 0.60**	**99.05 ± 0.21**

Table 4: Five-fold cross-validation result obtained in predicting the *H. pylori* PPI dataset.

Test set	ACC (%)	PE (%)	SN (%)	MCC (%)	AUC (%)
1	84.56	90.11	78.74	73.75	87.85
2	84.05	90.34	75.44	72.73	88.10
3	83.70	90.17	74.56	72.18	86.87
4	81.65	89.51	75.16	69.80	81.19
5	84.22	88.41	76.01	72.82	87.73
Average	**83.64 ± 1.15**	**89.71 ± 0.79**	**75.98 ± 1.63**	**72.26 ± 1.49**	**86.35 ± 2.92**

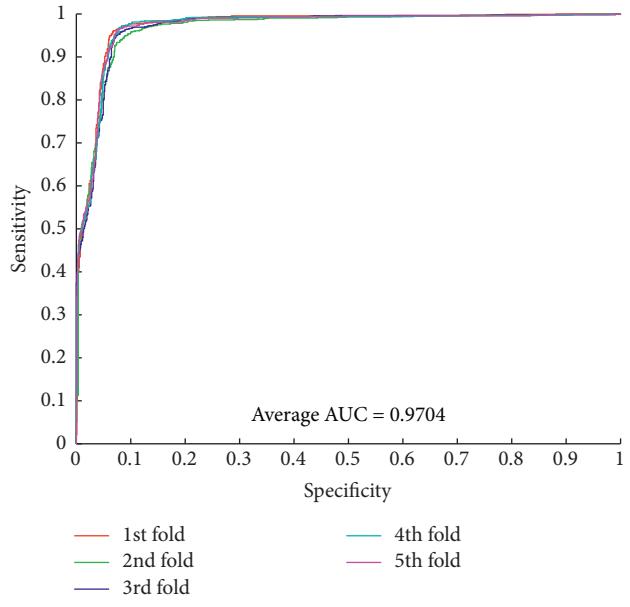

Figure 2: ROC curve prediction results for the yeast PPI dataset.

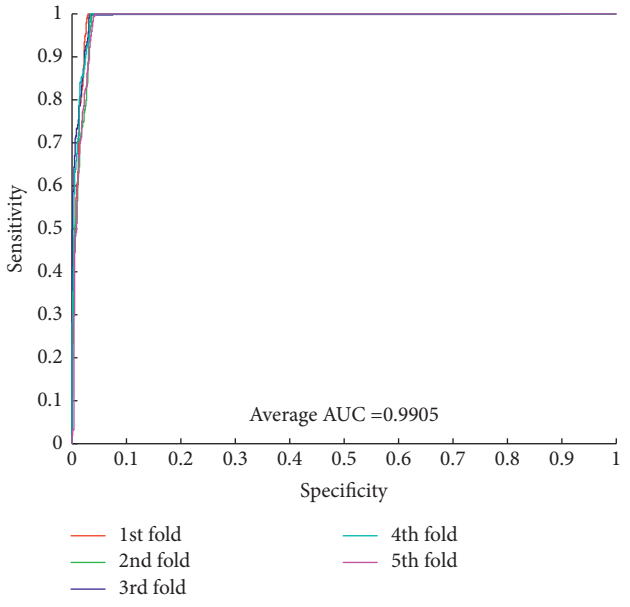

Figure 3: ROC curve prediction results for the human PPI dataset.

respectively. The average accuracy of the method is reduced by about 1.00% as compared to the WSRC classifier. In addition, we also plotted ROC curves based on SVM classifiers on three datasets, which are shown in Figures 5–7, respectively. The comparison results show that the WSRC-based method is superior to the SVM-based method in predicting PPIs on three different datasets.

3.4. Prediction Performance on the Independent Dataset. Despite the proposed method achieved good prediction accuracy on the yeast, human, and *H. pylori* datasets, it is worth considering whether the trained model can be adapted to predicting PPIs from other species. Therefore, 4 independent datasets were constructed from the DIP database to validate the performance of the trained model for predicting protein-

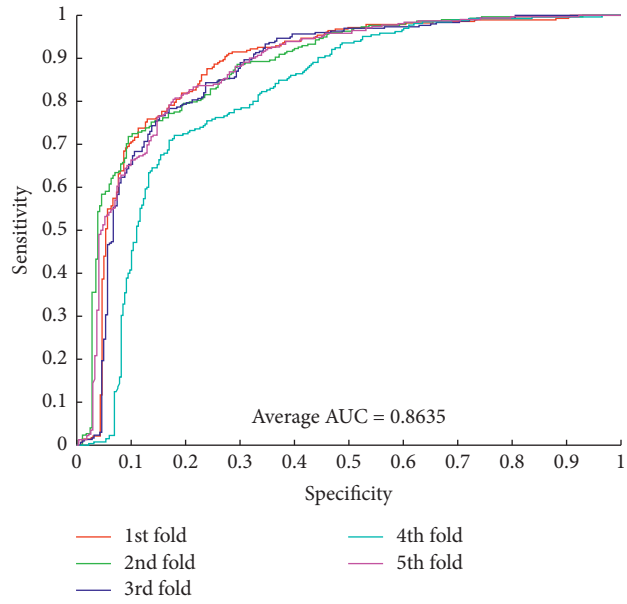

Figure 4: ROC curve prediction results for the *H. pylori* PPI dataset.

Table 5: Comparison with the SVM-based method on three datasets.

Dataset	Classifier	ACC (%)	SN (%)	PE (%)	MCC (%)	AUC (%)
Yeast	WSRC	94.55 ± 0.63	97.15 ± 0.42	92.33 ± 0.91	89.68 ± 1.12	97.04 ± 0.24
	SVM	87.25 ± 0.66	88.22 ± 1.61	86.58 ± 1.71	77.75 ± 0.98	93.96 ± 0.41
Human	WSRC	97.48 ± 0.31	99.01 ± 0.32	96.27 ± 0.55	95.06 ± 0.60	99.05 ± 0.21
	SVM	92.88 ± 0.94	95.97 ± 0.74	90.91 ± 1.27	86.70 ± 1.63	97.56 ± 0.44
H. pylori	WSRC	83.64 ± 1.15	75.98 ± 1.63	89.71 ± 0.79	72.26 ± 1.49	86.35 ± 2.92
	SVM	82.64 ± 1.61	94.33 ± 1.87	76.51 ± 2.40	70.53 ± 2.38	91.60 ± 1.11

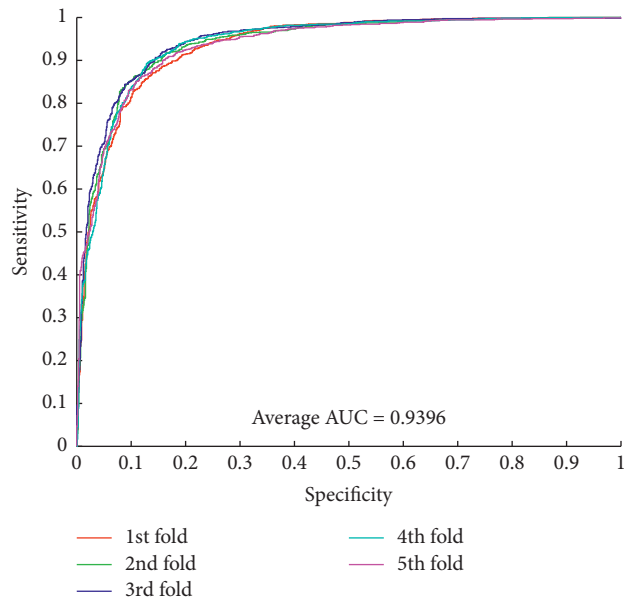

Figure 5: ROC curve prediction results based on the support vector machine for the yeast PPI dataset.

protein interactions from other species, in which all 11,188 samples of the yeast dataset were used as training sets, while *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus* datasets were treated as test sets separately. All of the test sets are positive. Here, we utilize the same SMR method and MatPCA feature extraction algorithm to convert the protein sequences of the four datasets into feature vectors and send them to the WSRC to perform PPI classification. The experimental results of our

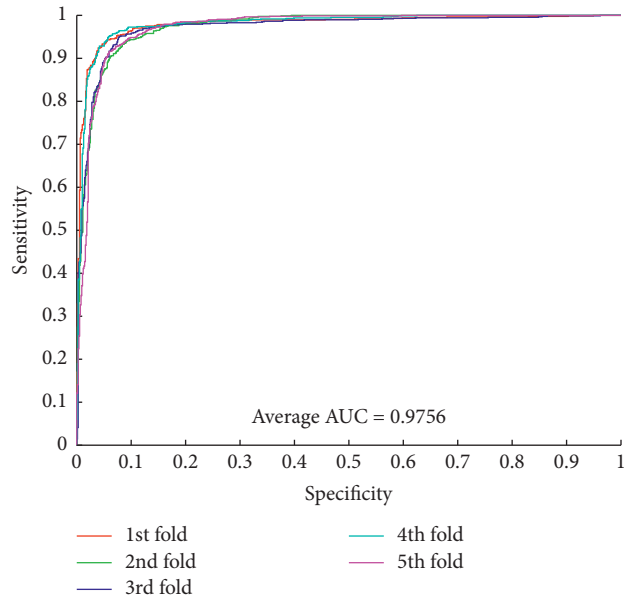

Figure 6: ROC curve prediction results based on the support vector machine for the human PPI dataset.

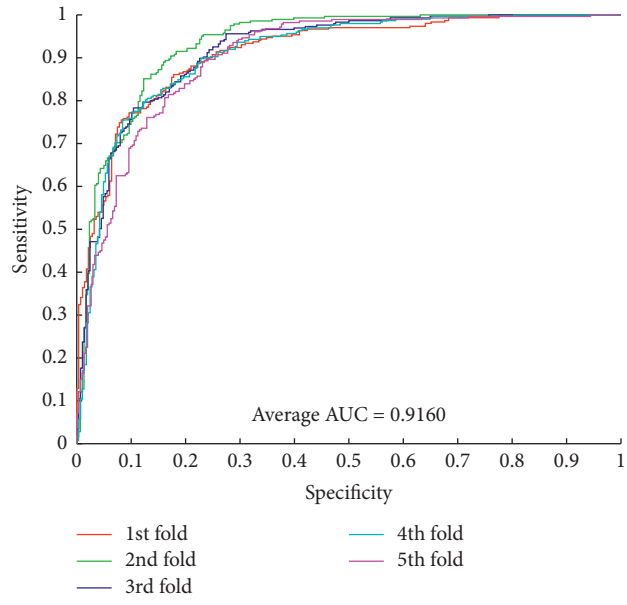

Figure 7: ROC curve prediction results based on the support vector machine for the *H. pylori* PPI dataset.

method are listed in Table 6. It can be seen that when predicting the PPIs of the *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus* datasets by using the proposed method, the accuracy is 96.84%, 90.14%, 97.10%, and 95.21%, respectively. These results indicate that our trained model also has excellent performance in detecting PPIs from other species.

Table 6: Prediction results on independent datasets.

Species	Test pairs	ACC (%)
C. elegans	4013	96.84
E. coli	6954	90.14
H. sapiens	1412	97.10
M. musculus	313	95.21

3.5. Comparison with the Previous Methods. Currently, a large number of computational methods have been proposed for predicting PPIs. Here, to further validate the effectiveness of the proposed method, we compared it with several state-of-the-art approaches on two benchmark datasets, namely, yeast and *H. pylori*, respectively. Table 7 shows the prediction results performed by using five different methods on the yeast dataset. We can clearly observe that the accuracy achieved by the other four existing

Table 7: Performance comparison of other methods on the yeast dataset.

Model	Test set	ACC (%)	PE (%)	SN (%)	MCC (%)
Zhou et al.'s work [42]	LD + SVM	88.56 ± 0.33	89.50 ± 0.60	87.37 ± 0.22	77.15 ± 0.68
Guo et al.'s work [22]	AC	87.36 ± 1.38	87.82 ± 4.33	87.30 ± 4.68	N/A
	ACC	89.33 ± 2.67	88.87 ± 6.16	89.93 ± 3.68	N/A
You et al.'s work [23]	PCA-EELM	87.00 ± 0.29	87.59 ± 0.32	86.15 ± 0.43	77.36 ± 0.44
Yang et al.'s work [43]	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
Proposed method	MatPCA + WSRC	**94.55 ± 0.63**	**92.33 ± 0.91**	**97.15 ± 0.42**	**89.68 ± 1.12**

Table 8: Performance comparison of other methods on the *H. pylori* dataset.

Model	ACC (%)	PE (%)	SN (%)	MCC (%)
Signature products [20]	83.40	85.70	79.90	N/A
Boosting [44]	79.52	81.69	80.37	70.64
Phylogenetic bootstrap [45]	75.80	80.20	69.80	N/A
HKNN [46]	84.00	84.00	86.00	N/A
Ensemble of HKNN [47]	86.60	85.00	86.70	N/A
Proposed method	**83.64**	**89.71**	**75.98**	**72.26**

methods is between 75.08% and 89.33%, which is significantly lower than the average accuracy of 94.55% by the proposed method. At the same time, the proposed method also obtains relatively low standard deviations, which further shows the stability and robustness of our model.

Similarly, we also compared the experimental results of the proposed method with five other existing methods on the *H. pylori* dataset. From Table 8, we can see that the proposed method achieves good prediction results, including accuracy, precision, and sensitivity. Specifically, signature products, boosting, and phylogenetic bootstrap methods have obtained a relatively high precision, which is 85.70%, 81.69%, and 80.20%, respectively. The HKNN and ensemble of HKNN methods have obtained a relatively high sensitivity, which is 86.00% and 86.70%, respectively. However, compared with other methods, the proposed model achieved 83.64% accuracy, 89.71% precision, 75.98% sensitivity, and 72.26% MCC. The above results show that using MatPCA and weighted sparse representation-based classification can effectively predict protein-protein interactions with good prediction performance.

4. Conclusions

In the postgenomic era, it is essential to employ computational methods to predict the interactions between protein pairs as this is important for explaining the molecular basis of complex cellular processes. In this paper, a novel computational model using solely protein sequence information was proposed for protein-protein interaction prediction. The proposed model first transforms the original protein sequence into a substitution matrix representation. Secondly, MatPCA, as a feature extraction algorithm, is applied to the SMR to capture evolutionary information of protein sequences. Finally, a weighted sparse representation-based classifier was used in this experiment to detect whether there are interactions between protein pairs. At the same time, we also used a five-fold cross-validation method to detect PPIs on three highly credible benchmark datasets including yeast, human, and *H. pylori*. Furthermore, we compared the experimental results of the proposed method with the SVM as well as different existing models to further demonstrate the prediction performance of different models in predicting PPIs. As expected, the experimental results confirm that our method has obtained good prediction results. These excellent predictive indicator values further demonstrate the effectiveness and feasibility of the proposed model, which will be regarded as a powerful tool for detecting potential PPIs.

Data Availability

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

Disclosure

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

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Zheng Wang and Yang Li contributed equally to this work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 61572506).

References

[1] Q. C. Zhang, D. Petrey, L. Deng et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[2] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, "A statistical model for identifying proteins by tandem mass

spectrometry," *Analytical Chemistry*, vol. 75, no. 17, pp. 4646–4658, 2003.

[3] N. E. Williams, "Immunoprecipitation procedures," *Methods in Cell Biology*, vol. 62, p. 449, 2000.

[4] M. Koegl and P. Uetz, "Improving yeast two-hybrid screening systems," *Briefings in Functional Genomics & Proteomics*, vol. 6, no. 4, pp. 302–312, 2007.

[5] H. Zhu, M. Bilgin, R. Bangham et al., "Global analysis of protein activities using proteome chips," *Science*, vol. 293, no. 5537, pp. 2101–2105, 2001.

[6] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal, "Effect of sampling on topology predictions of protein-protein interaction networks," *Nature Biotechnology*, vol. 23, no. 7, pp. 839–844, 2005.

[7] Z.-H. You, Z. Yin, K. Han, D.-S. Huang, and X. Zhou, "A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network," *Bmc Bioinformatics*, vol. 11, no. 1, p. 343, 2010.

[8] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright, "Computational prediction of protein-protein interactions," *Molecular Biotechnology*, vol. 38, no. 1, pp. 1–17, 2008.

[9] 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.

[10] C. Huang, F. Morcos, S. Kanaan, S. Wuchty, D. Chen, and J. 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.

[11] S.-W. Zhang, L.-Y. Hao, and T.-H. Zhang, "Prediction of protein-protein interaction with pairwise kernel support vector machine," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 3220–3233, 2014.

[12] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein-protein interaction," *Protein Engineering, Design and Selection*, vol. 14, no. 9, pp. 609–614, 2001.

[13] Q. Sheng and C. Lu, "Predicting protein-protein interaction based on protein secondary structure information using Bayesian classifier," *Journal on Applying Mathematics and Mathematical Applications*, vol. 1, p. 021, 2010.

[14] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[15] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein and Peptide Science*, vol. 15, no. 6, pp. 553–560, 2014.

[16] 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," 2014.

[17] Z.-H. You, J. Li, X. Gao et al., "Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines," *BioMed Research International*, vol. 2015, 2015.

[18] C. Zhou, H. Yu, Y. Ding, F. Guo, and X.-J. Gong, "Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree," *PLoS One*, vol. 12, no. 8, 2017.

[19] 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.

[20] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[21] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 490–500, 2006.

[22] 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.

[23] 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," 2008.

[24] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, vol. 17, no. 1, p. 184, 2016.

[25] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "Deep PPI: boosting prediction of protein-protein interactions with deep neural networks," *Journal of Chemical Information and Modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

[26] 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.

[27] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database-2009 update," *Nucleic Acids Research*, vol. 37, pp. D767–D772, 2009.

[28] Z.-H. You, J.-Z. Yu, L. Zhu, S. Li, and Z.-K. Wen, "A MapReduce based parallel SVM for large-scale predicting protein-protein interactions," *Neurocomputing*, vol. 145, pp. 37–43, 2014.

[29] F. Pazos, D. Juan, J. M. Izarzugaza, E. Leon, and A. Valencia, "Prediction of protein interaction based on similarity of phylogenetic trees," *Functional Proteomics*, vol. 145, pp. 523–535, 2008.

[30] X. Yu, X. Zheng, T. Liu, Y. Dou, and J. Wang, "Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: approach from amino acid substitution matrix and auto covariance transformation," *Amino Acids*, vol. 42, no. 5, pp. 1619–1625, 2012.

[31] 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.

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[33] J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, "Sparse representation classifier steered discriminative projection with applications to face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1023–1035, 2013.

[34] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," *SIAM Review*, vol. 24, pp. 4548–4551, 2013.
[35] N. Sharma, A. Sharma, V. Thenkanidiyoor, and A. Dileep, "Text classification using combined sparse representation classifiers and support vector machines," *SIAM Review*, vol. 24, pp. 181–185, 2008.
[36] B. Zhang and D. Zhang, "Noninvasive diabetes mellitus detection using facial block color with a sparse representation classifier," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1027–1033, 2013.
[37] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.
[38] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
[39] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," pp. 3360–3367 2011.
[40] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.
[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[42] Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, "Prediction of protein-protein interactions using local description of amino acid sequence," *Advances in Computer Science and Education Applications*, vol. 27, pp. 254–262, 2011.
[43] L. Yang, J.-F. Xia, and J. Gui, "Prediction of protein-protein interactions from protein sequence using local descriptors," *Protein & Peptide Letters*, vol. 17, no. 9, pp. 1085–1090, 2010.
[44] 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.
[45] J. R. Bock and D. A. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–134, 2003.
[46] L. Nanni, "Hyperplanes for predicting protein–protein interactions," *Neurocomputing*, vol. 69, no. 1–3, pp. 257–263, 2005.
[47] L. Nanni and A. Lumini, "An ensemble of *K*-local hyperplanes for predicting protein-protein interactions," *Bioinformatics*, vol. 22, no. 10, pp. 1207–1210, 2006.