Hindawi
Complexity
Volume 2018, Article ID 4216813, 12 pages
https://doi.org/10.1155/2018/4216813

WILEY | Hindawi

Research Article

Predicting Protein Interactions Using a Deep Learning Method-Stacked Sparse Autoencoder Combined with a Probabilistic Classification Vector Machine

Yanbin Wang,[1,2] **Zhuhong You**,[1] **Liping Li**,[1] **Li Cheng**,[1] **Xi Zhou**,[1] **Libo Zhang**,[3] **Xiao Li**,[1] **and Tonghai Jiang**[1]

[1]*Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*Institute of Software, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Zhuhong You; zhuhongyou@hotmail.com and Liping Li; lipingli@ms.xjb.ac.cn

Received 1 February 2018; Revised 10 May 2018; Accepted 13 June 2018; Published 10 December 2018

Academic Editor: Panayiotis Vlamos

Protein-protein interactions (PPIs), as an important molecular process within cells, are of pivotal importance in the biochemical function of cells. Although high-throughput experimental techniques have matured, enabling researchers to detect large amounts of PPIs, it has unavoidable disadvantages, such as having a high cost and being time consuming. Recent studies have demonstrated that PPIs can be efficiently detected by computational methods. Therefore, in this study, we propose a novel computational method to predict PPIs using only protein sequence information. This method was developed based on a deep learning algorithm-stacked sparse autoencoder (SSAE) combined with a Legendre moment (LM) feature extraction technique. Finally, a probabilistic classification vector machine (PCVM) classifier is used to implement PPI prediction. The proposed method was performed on human, unbalanced-human, *H. pylori*, and *S. cerevisiae* datasets with 5-fold cross-validation and yielded very high predictive accuracies of 98.58%, 97.71%, 93.76%, and 96.55%, respectively. To further evaluate the performance of our method, we compare it with the support vector machine- (SVM-) based method. The experimental results indicate that the PCVM-based method is obviously preferable to the SVM-based method. Our results have proven that the proposed method is practical, effective, and robust.

1. Introduction

Most important molecular processes in cells are performed by different types of protein interactions. Thus, one of the main objectives of functional proteomics is to determine the protein-protein interactions of organisms. With continuous research and the development of technique, it is now possible to detect protein interactions on a large scale by using high-throughput experimental techniques. Such research is obviously very important, because the research of PPIs is closely related to many functions of complex life systems, and these functions are not determined by the characteristics of the individual components. For example, molecular cell signaling is carried out through protein interactions. This process is not only the basis of many life functions, but it is also related to many diseases. In addition, the study of protein interactions has been of great value in the development of new drugs and in the prevention and diagnosis of disease.

As some high-throughput experimental techniques have been successfully applied to postgenomic era PPI research tasks, a large number of different species of PPI data have been collected, and some databases have been created to systematically collect and store experimentally determined PPIs [1–3]. Even though experimentally validated PPI data drives research and development of proteomics, they often have high false positives and false negatives [4–7]. In addition, because the experimental method has some unavoidable defects, such as having a high cost and being time consuming, the researchers have only verified a small part of the whole

PPI network even after a long period of effort. With advances in mathematical and computational methods [8–12], computer technology has been applied in more and more fields. Vlachakis et al. proposed computational methods to simulate catalytic mechanisms, complete drug design, and model protein three-dimensional structures [13–17]. Vlamos et al. developed several intelligent disease diagnosis applications and hybrid models for vulnerability detection [18–25]. Some researchers also introduced computational methods into the medical field and developed several automated diagnostic models [26, 27]. Therefore, using a machine learning algorithm to develop an efficient and accurate automatic discriminative system to predict new protein interactions has important practical significance.

To date, a variety of protein information has been used to build PPI prediction models based on machine learning algorithms. Protein information that can be used includes, but is not limited to, physicochemical information, structural information, evolutionary information, and protein domains. However, these methods have some limitations when they are used. For example, some computational methods using genomic information predict protein interactions by calculating a set of gene presence or absence patterns. The main factor limiting these methods is that they can only be applied to fully sequenced genomic data [28, 29]. Recently, methods that directly extract information from a protein primary sequence have attracted much attention. Methods that use only protein sequence information are more general than methods that rely on some additional information about proteins. Many researchers are working on the development of sequence-based computational models to predict new PPIs. Hamp and Rost developed a computational method for predicting PPIs based on profile-kernel support vector machines combined with evolutionary profiles [30]. An et al. proposed a PPI prediction method that combines the local phase quantization and relevance vector machine [31]. Yang et al. used a new local descriptor to describe the interaction between the contiguous and discontinuous regions of the protein sequence, which is able to obtain more protein interaction information from sequences [32]. Zhang et al. introduced two ensemble methods to predict PPIs. These ensemble methods are based on undersampling techniques and fusion classifiers [33]. You et al. proposed a prediction framework for detecting PPIs using a low-rank approximation-kernel extreme learning machine [34]. Several other sequence-based computational methods have been reported in previous work [35–38]. These sequence-based methods show that the individual information of the amino acid sequence is sufficient to determine the interaction of the protein. However, these methods usually use physical, chemical, or structural information, and even the fusion of all of these types of information as features of the protein sequence. Therefore, the feature extraction steps of these methods are not efficient. In addition, the above information can only represent each specific protein sequence but does not contain knowledge related to protein interactions. Therefore, even these methods combined with advanced classification algorithms have a difficulty in producing enough accuracy.

Compared with the physicochemical information, the evolutionary information of proteins can reflect the potential interactions between proteins. Therefore, we consider the evolutionary information of the protein as a feature of the protein sequence. Extracting the evolutionary information of a protein is challenging as there is currently no strategy that can efficiently obtain the evolutionary information of a protein. We hypothesize that there is a potential relationship between the conservation of amino acid residues during evolution and the interaction of proteins. Based on this hypothesis, we propose an efficient protein evolution feature extraction scheme, which used a deep learning algorithm combined with Legendre moments (LMs) and position weight matrix (PWM). Specifically, we first convert the protein sequence into a PWM containing the amino acid residue conservative score. Then, we use LMs to extract important evolutionary information from the PWM and generate the feature vector $\overrightarrow{\mathbf{F}}$. Last but not least, this feature $\overrightarrow{\mathbf{F}}$ was further optimized by using SSAE deep neural networks to eliminate noise, obtain primary information, and reduce feature dimensions. In addition, in response to the challenges posed by big data and imbalanced datasets, a sparse model, PCVM, was used to perform classification. Our contributions can be summarized as follows:

(1) We propose a method to predict PPIs quickly, efficiently, and accurately.

(2) We have abandoned the traditional materialized information and structural information, considered the evolutionary information associated with PPIs as a feature of the protein sequence, and proposed a feature extraction strategy to quickly and efficiently extract the evolutionary information of the protein and improve the prediction performance.

(3) We confirm that sparse classification algorithms can greatly benefit prediction of PPIs and present results showing that they can provide a benefit in dealing with large-scale data and unbalanced data (as is the case with PCVM).

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the datasets and methods used in this paper. Section 4 shows the results of the experiment. Section 5 concludes the paper.

2. Related Work

The study of the PPI prediction model is mainly divided into two parts. One is the development of protein sequence feature extraction strategies, and the other is the application of classification algorithms. This section briefly reviews related research.

2.1. Sequence-Based Feature Extraction Algorithm. Previous methods of extracting sequence features were mainly the direct use of physicochemical information or amino acid sequence structure information or evolutionary information of proteins. Since the amino acid composition model has

been proposed, many subsequent works have been carried out for the composition model. Chou [39] proposed a feature extraction method called pseudoamino acid composition. This feature extraction method greatly increases the information content of the amino acid sequences contained in the features. It does not only consider the composition of amino acids, but also processes the amino acid position information. Another excellent research was done by Shen et al. [40]. In that study, 20 amino acids were clustered into 7 classes based on their dipole and side chain volumes, and then the features of the protein pairs were extracted based on the amino acid class. Combined with the SVM classifier, this method has a prediction accuracy of 83.9% on human PPIs. In a study by Guo et al. [41], an autocovariance-based method was developed to extract the interaction information of discontinuous amino acid fragments in a sequence. The method replaces the protein sequence with a digital sequence based on the physicochemical properties, and the replaced digital sequence is regarded as a group of information for analysis.

Different from the previous classical computational method, we did not use the traditional sequence-coding scheme and did not consider the physicochemical information of the protein sequence. Our method uses the evolutionary information of the protein sequence indirectly (using Legendre moments to extract feature vectors on the PSSM matrix containing evolutionary information), trying to use image-processing ideas to complete the task of PPI prediction; this is a direction in which only a few people are exploring. The introduction of our method and the satisfactory results produced on several gold standard datasets have greatly encouraged the scholars who explored on this direction. The advantage of this method is that the feature extraction strategy is simple and efficient, does not require complicated sequence coding, and does not need to consider the physicochemical information of the protein. Compared with traditional feature extraction methods, this method greatly improves the accuracy of PPI prediction and saves time and computational overhead.

In addition, the deep learning algorithm has shown extraordinary performance in many fields, but its ability has not been effectively verified in the PPI prediction task. A deep learning algorithm-stacked sparse autoencoder was used to reconstruct a protein feature vector in our work. This algorithm uses sparse network structures and adds sparseness restrictions on neurons. This not only allows us to obtain low-dimensional, low-noise protein feature vectors, but also improves the efficiency of the network. The results of our method applied to the test set demonstrate once again that deep learning algorithms can be used to assist in solving bioinformatics problems.

2.2. Classifier. The support vector machine (SVM) is one of the most commonly used classification algorithms in the PPI prediction model [42–44]. However, the SVM approach has some obvious drawbacks: (1) As the dataset becomes larger, the support vector increases rapidly. (2) Cross-validation-based kernel parameter optimization strategy consumes a large amount of computing resources. Another widely used classifier is the relevance vector machine (RVM) [45–47], which effectively avoids the disadvantages of SVM. It was developed to take advantage of the Bayesian inference and the prior weight following a zero-mean Gaussian distribution. However, the RVM has the potential to produce some unreliable vectors that lead to system error decisions. Because the weights of the negative class and the positive class are given by the zero-mean Gaussian prior, partial training samples that do not interact might be assigned confident weights or vice versa.

In order to avoid the problems of the above classifiers, we used the probability classification vector machine (PCVM) method to perform PPI classification, which provides different priors for different types of samples. The positive class is associated with a right-truncated Gaussian and the negative class is associated with a left-truncated Gaussian. The PCVM method has the following advantages: (1) PCVM produces sparse predictive models and has better efficiency in the testing phase. (2) PCVM provides probabilistic results for each output. (3) PCVM uses the EM algorithm to automatically find the optimal initial point, which saves time and improves the performance of the system.

3. Materials and Methodology

3.1. Datasets. To evaluate the performance of the proposed method, there are a total of 4 different PPI datasets used in our experiments, two of which are human, one is *S. cerevisiae*, and one is *H. pylori*.

The first human PPI dataset we used was from Pan et al. [48], which was downloaded from the Human Protein Reference Database (HPRD). After the self-interaction and repetitive interactions were removed, the remaining 36,630 PPI pairs formed the final gold standard positive (GSP) dataset. For the selection of gold standard negative (GSN) datasets, we followed the previous work [48] and generated GSN datasets from the Swiss-Prot version 57.3 database according to the following criteria: (1) Protein sequences annotated by uncertain terms are removed. (2) Multiple unlocalized protein sequences are deleted. (3) Protein sequences that may be only "fragments" or containing "fragments" are deleted.

After strictly following the above steps, 1773 human proteins were screened out. Noninteracting protein pairs are then constructed by randomly pairing proteins from different subcellular compartments. In addition, another golden negative dataset was downloaded, which was used in the study by Smialowski et al. [49]. The final GSN dataset was constructed by combining the above two negative datasets, which consisted of 36,480 noninteracting protein pairs. Therefore, the entire gold standard dataset (GSD) consists of 73,110 protein pairs, of which almost half is from the positive dataset and half is from the negative dataset.

Due to the fact that there are serious imbalances in the dataset in real-world tasks, this can lead to a failure of the PPI prediction model. Considering this issue, we have constructed another set of human datasets with an unbalanced number of positive and negative samples to evaluate the stability and robustness of our proposed method. This unbalanced human PPI dataset consists of 3899 positive samples and 13,000 negative samples.

The third PPI gold standard dataset we used was from downloaded datasets from the *S. cerevisiae* core subset of the database of interacting proteins (DIP). We strictly followed the work of Guo et al. [41] to construct the *S. cerevisiae* dataset. Finally, we obtained a gold standard dataset containing 11,188 protein pairs, of which 5594 positive protein pairs form a GSP dataset and 5594 negative protein pairs form a GSN dataset.

The last PPI dataset uses the pair of *H. pylori* proteins described by Martin et al. [50], which includes 1458 positive sample pairs and 1458 negative sample pairs.

3.2. Position Weight Matrix. In this article, we use the position weight matrix (PWM) to derive evolutionary information from protein sequences. A PWM for a query protein is a $Y \times 20$ matrix $\mathbf{M} = \{m_{ij},\ i = 1, \dots, Y, j = 1, \dots, 20\}$, where Y represents the size of the protein sequence and the number of columns of the $\mathbf{M}$ matrix denotes 20 amino acids. In order to construct PWM, a position frequency matrix is first created by calculating the presence of each nucleotide on each position. This frequency matrix can be represented as $\mathbf{p}(a, c)$, where u means position and k is the kth nucleotide. The PWM can be expressed as $M_{ij} = \sum_{k=1}^{20} \mathbf{p}(a, c) \times \mathbf{w}(b, c)$, where $\mathbf{w}(b, c)$ is a matrix whose elements represent the mutation value between two different amino acids. Consequently, high scores represent highly conservative positions, and low points represent a weak conservative position. It's an extremely useful tool for predicting protein disulfide connectivity, protein structural classes, subnuclear localization, and DNA or RNA binding sites. Here, we also employ PWMs to detect PPIs. In this paper, each protein is interpreted as PWMs using the position-specific iterated BLAST (PSI-BLAST). The PSI-BLAST has two important parameters, e value and iteration number, which were set at 0.001 and 3, respectively [51–53].

3.3. Legendre Moments. Legendre moments (LMs) are typical orthogonal moments, whose kernel function is the Legendre polynomial. It has been widely involved in a lot of applications, such as image analysis, computer vision, and remote sensing [54–58]. Here, we use the Legendre moment to extract the evolutionary information of the protein indirectly from the PWM and generate a 961-dimensional eigenvector. The two-dimensional discrete form of the LM is represented as

$$L_{mn} = \mu_{mn} \sum_{i=1}^{K} \sum_{j=1}^{L} h_{mn}(x, y) g(x_i, y_i), \tag{1}$$

where $g(x, y)$ is defined as a set of discrete points (x_i, y_i), $-1 \le x_i,\ y_i \le +1$. K represents the number of columns of the PWM matrix, L represents the sum of each column of PWM matrix.

$$h_{mn}(x, y) = \int_{x_i - \Delta x/2}^{x_i + \Delta x/2} \int_{y_i - \Delta y/2}^{y_i + \Delta y/2} R_m(x) R_n(y) dx dy. \tag{2}$$

The integral terms in (2) are frequently estimated by zeroth-order approximation; in other words, the values of Legendre polynomials are assumed to be constant over the intervals $[x_i - \Delta x/2, x_i + \Delta x/2]$ and $[y_i - \Delta y/2, y_i + \Delta y/2]$. In this case, the set of approximated LMs is defined as:

$$L'_{mn} = \frac{(2m + 1)(2n + 1)}{KL} \sum_{i=1}^{K} \sum_{j=1}^{L} R_m(x_i) R_n(y_i) g(x_i, y_i). \tag{3}$$

3.4. Stacked Sparse Autoencoder. Deep learning is a new field in machine learning research. Its motivation lies in building and simulating the neural network of the human brain for analytical learning. It imitates the mechanism of the human brain to interpret data. In this paper, the deep structure stacked sparse autoencoder (SSAE) is adopted for feature reduction and reconstruction [59–62]. SSAE forms a more abstract high-level representation feature by combining low-level features to discover the distributed feature representation of protein feature data.

The SSAE is an unsupervised network that is a large-scale nonlinear system composed by multilayer neuron cells in which the outputs of the current layer neuron are fed to the connectivity layer neuron. In this work, the aim of SSAE is to learn a distinctive representation for the Legendre moment (LM) feature. The underlying purposes are noise elimination and dimensionality reduction. The process of feature reconstruction is layer by layer in SSAE. The first layer is in charge of rough integration original input. The second layer is responsible for extracting and integrating the features learned earlier. Higher successive layers will be inclined to produce low-dimensional, low-noise, and high-cohesion features. In this paper, the SSAE was used to reduce the LM feature to 200 dimensional.

SSAE or Sparse autoencoder network is mainly made up of two parts, the encoding part and the decoding part [63], where the encode network compresses high-dimensional into low-dimensional attributes. The decoding network is responsible for restoring the original input layer by layer, and the network structure is symmetrical with the structure of the encoding network. In the coding stage, the primary data x is mapped onto a hidden layer. This process can be represented as

$$z = \sigma_1(w_1 x + b_1). \tag{4}$$

Here, σ_1 is a nonlinear function, w_1 is the weight of the encoding part and b_1 is the bias. After that, the original data is reconstructed by the decoding network:

$$x' = \sigma_2(w_2 z + b_2), \tag{5}$$

where w_2 is the weight of decoding network and b_2 is the bias. The purpose of SAE is to make the output as close as possible to the input by minimizing loss function:

$$\theta = \arg\min \left[\frac{1}{n} \sum_{i=1}^{n} L\left(x_i, x'_i\right) + \beta \sum_{j=1}^{S_2} KL\left(\rho \| \widehat{\rho}_j\right) \right], \tag{6}$$

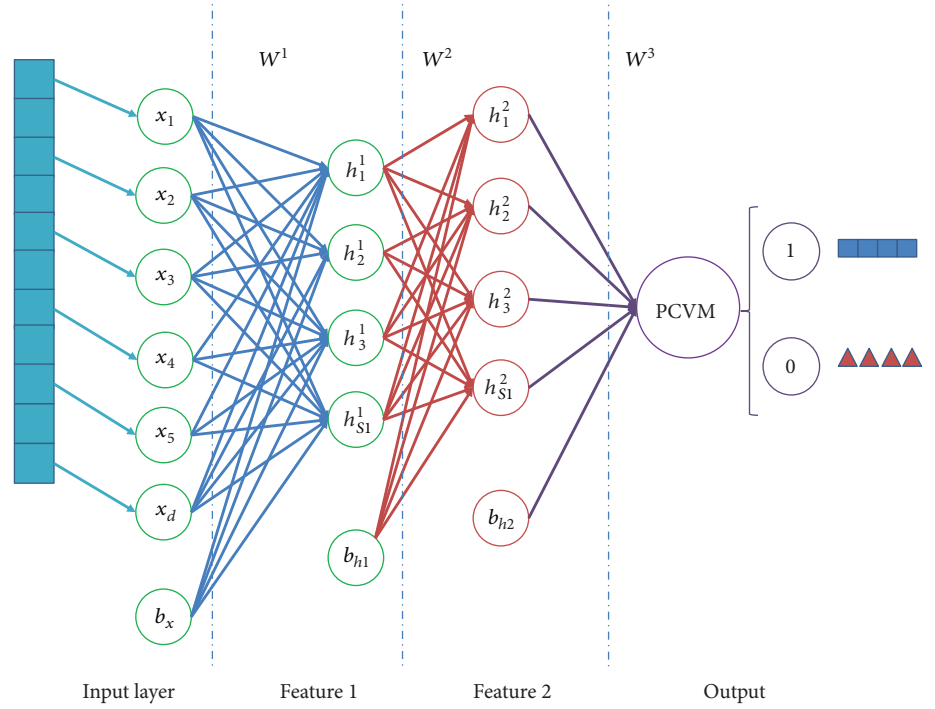

Figure 1: Stacked sparse autoencoder with two hidden layer structures.

$$\theta = \frac{1}{2N}\sum_{i=1}^{N}\left\|x_i' - x_i\right\|^2 + \beta\sum_{j=1}^{S_2} KL\left(\rho\|\widehat{\rho}_j\right), \tag{7}$$

where N is the number of hidden layer nodes, β is the weight of the sparse penalty item, $\widehat{\rho}_j$ represents the average activation value of the hidden layer element, and ρ is the sparse parameter.

Figure 1 shows a SSAE network with two hidden layers, of which the decoding part has not been shown, in order to highlight the feature reduction function of the network. Similar to the sparse autoencoder (SAE), the key to the training model is to learn the parameters $\theta = (W, b)$, which allows the model to have minimum input and output deviation. Once the optimal parameters θ are obtained, the SSAE yield function $R^{d_x} \to R^{d_{h(2)}}$ that transforms original data to a low-dimensional space.

3.5. Probabilistic Classification Vector Machines. The design of feature extraction strategies and the selection of classifiers are two crucial parts in developing an excellent PPI prediction model. In the previous description, we developed a new deep learning-based amino acid sequence feature extraction method. Here, we use the stronger PCVM classifier to replace the Softmax layer of the stacked sparse autoencoder to achieve the output of our model. Like most classification models, the goal of a PCVM [64–66] is to generate a model $f(X; W)$ by learning a set of labeled data $\{X, Y\}$. The model is determined by parameters W learned and expressed as

$$f(X; W) = \sum_{i=1}^{N} w_i \varnothing_{i,\theta}(x) + b, \tag{8}$$

where the $W = (w_1, \ldots, w_N)^T$ denotes the parameter of the model, $\varnothing_{i,\theta}(x)$ is a set of primary functions, and b represents the bias. A Gaussian cumulative distribution function $\beth(x)$ is used for obtaining the binary outputs. The function is defined as

$$\beth(d) = \int_{-\infty}^{d} N(r \mid 0, 1). \tag{9}$$

After incorporating (7) with (8), the model becomes

$$K(X; W, b) = \beth\left(\sum_{i=1}^{N} w_i \varnothing_{i,\theta}(x) + b\right) = \beth(\Phi_\theta(X)W + b). \tag{10}$$

Each weight w_i is assigned a prior by a truncated Gaussian distribution, as follows:

$$p(W \mid \boldsymbol{\alpha}) = \prod_{i=1}^{N} p(w_i \mid \alpha_i) = \prod_{i=1}^{N} N_t\left(w_i \mid 0, \alpha_i^{-1}\right), \tag{11}$$

where the bias b is assigned a zero-mean Gaussian prior, as follows:

$$p(b \mid \beta) = N\left(b \mid 0, \beta^{-1}\right), \tag{12}$$

where the $N_t(w_i \mid 0, \alpha_i^{-1})$ is a truncated Gaussian, and α_i denotes the inverse of the variance. The EM algorithm is used for obtaining all parameters of a PCVM model [67].

4. Results

4.1. Evaluation Criteria. In this work, the following criteria, such as accuracy (Accu), precision (Prec), sensitivity (Sens), and Matthews's correlation coefficient (Mcc), are used to assess the proposed method. Accuracy is used to describe the overall system error. Since the key task of PPI prediction is to correctly predict the interacting protein pairs, the sensitivity and accuracy indicators are used to assess the model's ability to predict positive data. In addition, data imbalance exists in real PPI prediction tasks. In view of this situation, we used an unbalanced PPI dataset in this paper. Therefore, Mcc is used to evaluate the reliability and stability of the model when dealing with unbalanced data. When the model appears "preference prediction" (i.e., the dataset is very unbalanced, the model can only correctly predict negative data), the Mcc score is lower. When the model is strong and robust, the indicator score is high. These indicators are defined as

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

$$\text{Sens} = \frac{\text{TP}}{\text{TN} + \text{TP}},$$

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Mcc} = \frac{(\text{TN} \times \text{TP}) - (\text{FN} \times \text{FP})}{\sqrt{(FN + TP) \times (FP + TN) \times (FP + TP) \times (FN + TN)}}, \quad (13)$$

where TP means those samples, true interacting with each other, are predicted correctly. FP represents those samples, true noninteracting with each other, are judged to be interacting. TN represents those samples, true noninteracting with each other, are predicted correctly. FN represents those samples, true interacting with each other, are judged to be noninteracting. Furthermore, the ROC (receiver operating characteristic) is portrayed to appraise the performance of a set of classification results [68] and the AUC (area under ROC) is computed as an important evaluation indicator.

4.2. Assessment of Prediction. In this paper, the proposed sequence-based PPI predictor is implemented using a MATLAB platform. All the simulations are carried out on a computer with a 3.1 GHz 8-core CPU, 16 GB memory, and a Windows operating system. In order to make the prediction system independent of the training data, each PPI dataset is segmented into five parts by the five-fold cross-validation method. The performance of the PCVM-based method on human, unbalanced-human, *H. pylori*, and *S. cerevisiae* datasets are exposed in Tables 1–4. The corresponding ROC curves are depicted in Figures 2–6, respectively.

Analyzing Table 1 allows drawing the conclusion that the PCVM-based method yields a satisfactory result on the human dataset, where the accuracy of each fold is above 98% and the accuracy standard deviations of five experiments are only 0.2%. The corresponding average sensitivity, precision, and Mcc are 98.47%, 98.67%, and 97.19%, respectively. Their standard deviations are 0.3%, 0.5%, and 0.3%, respectively. The average AUC (Figure 2) of the five experiments reached 0.9984. The high accuracies and AUC show that the PCVM-based approach has a strong classification ability in identifying PPIs. The low standard deviations illustrate that this model is robust and stable.

When predicting PPIs on the unbalanced-human dataset (Table 2), the method produced an average accuracy of 97.71%, sensitivity of 91.87%, precision of 98.10%, and AUC of 0.9971, respectively.

When applied on the *H. pylori* dataset with the smallest training set, the PCVM-based methods also yielded a high average prediction accuracy of 93.76%, high precision of 92.10%, high sensitivity of 95.77%, and high Mcc of 88.26%, respectively (Table 3). The standard deviations of Accu, Sens, Prec, and Mcc in the five experiments are 0.1%, 2.0%, 1.9%, and 0.2%, respectively. Moreover, the average AUC on the *H. pylori* dataset reached 0.9860.

Table 1: 5-Fold cross-validation results using the proposed method on the human dataset.

Testing set	Accu (%)	Sens (%)	Prec (%)	Mcc (%)
1	98.50	98.87	98.13	97.04
2	98.69	98.53	98.89	97.41
3	98.31	98.35	98.22	96.68
4	98.69	98.51	98.88	97.41
5	98.69	98.11	99.23	97.41
Average	98.58 ± 0.2	98.47 ± 0.3	98.67 ± 0.5	97.19 ± 0.3

Table 2: 5-Fold cross-validation results using the proposed method on the unbalanced-human dataset.

Testing set	Accu (%)	Sens (%)	Prec (%)	Mcc (%)
1	97.57	91.71	97.67	93.23
2	97.78	92.44	98.00	93.86
3	97.72	92.20	97.12	93.32
4	97.75	91.26	99.19	93.78
5	97.75	91.76	98.50	93.74
Average	97.71 ± 0.1	91.87 ± 0.5	98.10 ± 0.8	93.59 ± 0.3

Table 3: 5-Fold cross-validation results using the proposed method on the *H. pylori* dataset.

Testing set	Accu (%)	Sens (%)	Prec (%)	Mcc (%)
1	94.00	96.76	92.28	88.62
2	93.65	95.73	91.50	88.10
3	93.65	92.52	94.77	88.11
4	93.83	95.67	92.58	88.38
5	93.66	98.18	89.37	88.10
Average	93.76 ± 0.1	95.77 ± 2.0	92.10 ± 1.9	88.26 ± 0.2

Table 4: The prediction performance comparison of PCVM with SVM.

Model	Testing set	Accu (%)	Sens (%)	Prec (%)	MCC (%)
PCVM	1	96.83	97.37	96.44	93.85
	2	96.33	97.33	95.22	92.93
	3	96.33	96.86	96.02	92.93
	4	96.60	96.85	96.33	93.44
	5	96.64	97.75	95.19	93.11
	Average	96.55 ± 0.2	97.23 ± 0.3	95.84 ± 0.5	93.25 ± 0.3
SVM	1	94.46	93.68	95.36	89.53
	2	93.70	90.32	96.46	88.13
	3	93.92	92.49	95.49	88.58
	4	92.76	91.99	93.33	86.56
	5	93.53	92.99	93.92	87.89
	Average	93.67 ± 0.6	92.29 ± 1.2	94.91 ± 1.2	88.13 ± 1.0

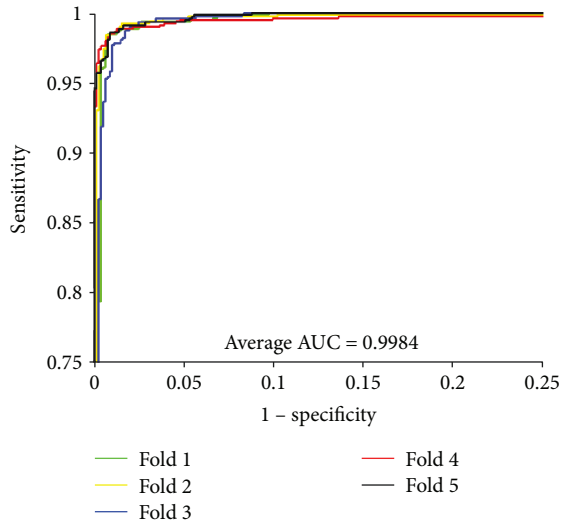

Figure 2: ROC curves performed by the proposed approach on the human dataset.

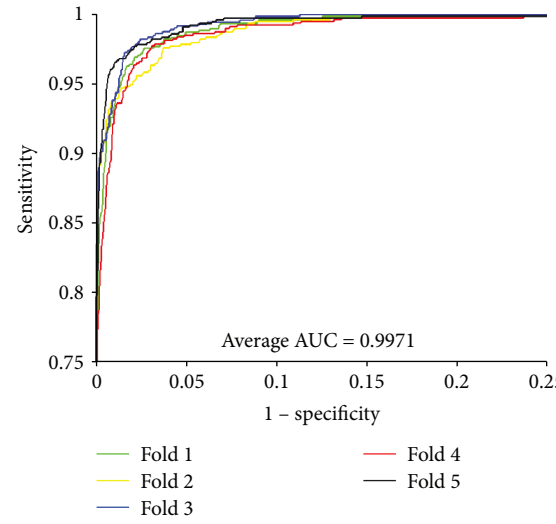

Figure 3: ROC curves performed by the proposed approach on the unbalanced-human dataset.

4.3. Comparison with the SVM-Based Approach. In order to highlight the feasibility of our classifier, the state-of-the-art SVM classifier was used to compare with PCVM. To make it fair, the same feature extraction scheme and the same *S. cerevisiae* dataset were used in this experience. The LIBSVM tool [69] is available for SVM classification, and the grid search approach was adopted for optimizing SVM model parameters c and g.

The classification results of the PCVM and SVM classifiers on the *S. cerevisiae* dataset are listed in Table 4, and the ROC curves of SVM are displayed in Figures 5 and 6. As we have seen, the average result of the PCVM method achieved 96.55% Accu, 97.23% Sens, 95.84% Prec, and 93.25% Mcc. The standard deviation of these indicators in five experiments are 0.2%, 0.3%, 0.5%, and 0.3%, respectively. The average results of the SVM method yielded 93.67% Accu, 92.29% Sens, 94.91% Prec, and 88.13% Mcc. The standard deviations are 0.6%, 1.2%, 1.2%, and 1.0%, respectively. In comparison with SVM, the PCVM classifier achieves significantly better results on this gold standard dataset. From Figures 5 and 6, the average AUC of the SVM classifier is 0.9856, which is significantly lower than those of PCVM of 0.9963. Higher AUC values clearly illustrate that the PCVM method is more accurate and more reliable for detection PPIs. The improved classification performance of the PCVM classifier compared with the SVM classifier can be explained by two reasons: (1) The number of PCVM basis functions is less than the number of training points, resulting in a reduction in the computational effort involved. (2) PCVM uses truncated Gauss priors to flexibly assign a priori information about weights, thus ensuring the generation of reliable support vectors.

4.4. Compare with Previous Studies. Some other computational approaches for predicting PPI have been reported in previous studies. These highlight the advantages of the

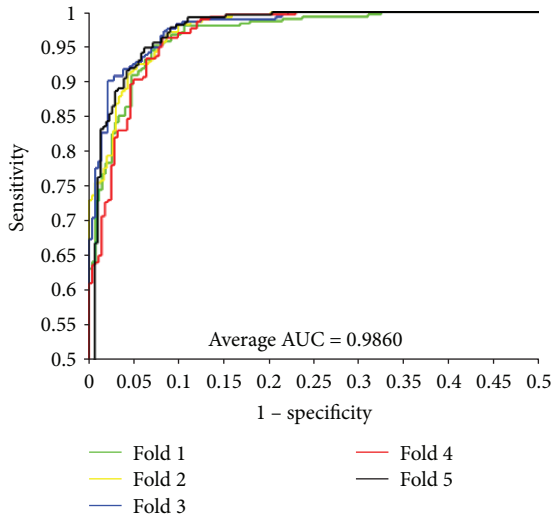

Figure 4: ROC curves performed by the proposed approach on the *H. pylori* dataset.

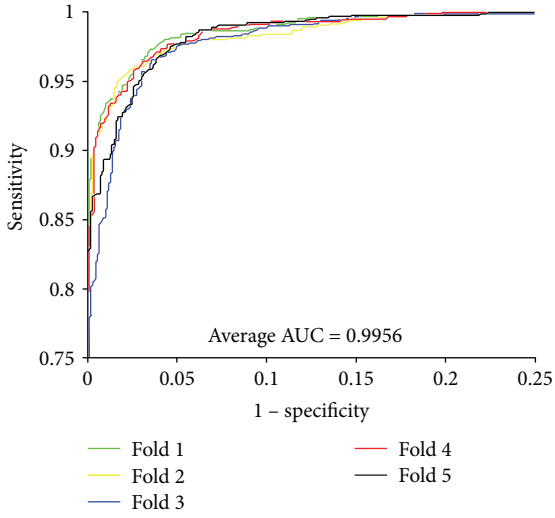

Figure 5: ROC curves performed by the proposed approach on the *S. cerevisiae* dataset.

proposed approach, which was compared with the existing approaches that attract wide attention on the same PPI datasets, respectively. We can see from Table 5 that our method also produces better results than other existing methods. The performance of the several different approaches on the *H. pylori* dataset is presented in Table 6. As seen from the Table 6, our proposed approach produces better performances than the four other main methods. The 93.76% prediction accuracy is much higher than any of the several other methods. Table 7 shows the results of comparing with several other different methods that achieved an average prediction accuracy of less than 93.92% on the *S. cerevisiae* dataset, while our PCVM-based approach obtained an average prediction accuracy of 96.55% with the lowest standard deviation of 0.2%. Meanwhile, the sensitivity of 97.23% is also far better than those of the other methods.

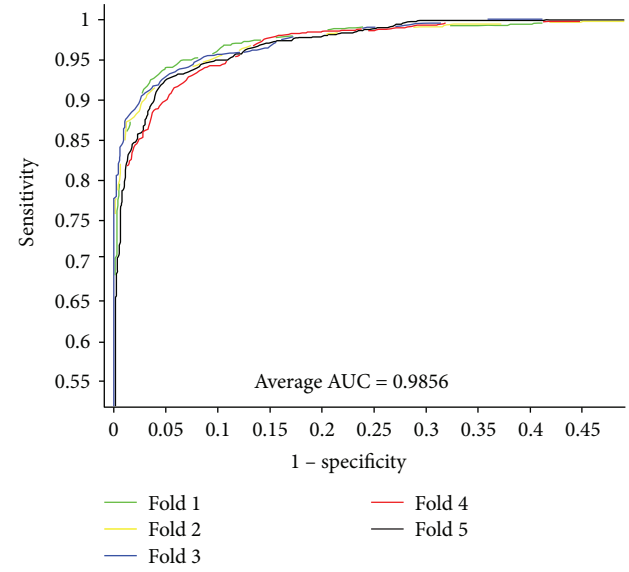

Figure 6: ROC curves performed by the SVM-based approach on the *S. cerevisiae* dataset.

Table 5: Performance comparison of different methods on the human dataset.

Model	Accu (%)	Sens (%)	Prec (%)	MCC (%)
LDA + RF [70]	96.40	94.20	N/A	92.80
LDA + RoF [70]	95.70	97.60	N/A	91.80
LDA + SVM [70]	90.70	89.70	N/A	81.30
AC + RF [70]	95.50	94.00	N/A	91.40
AC + RoF [70]	95.10	93.30	N/A	91.10
AC + SVM [70]	89.30	94.00	N/A	79.20
Proposed method	97.71	91.87	98.10	93.59

Extensive experiments indicate that the method we employ can sufficiently meet the needs of large-scale protein detection and can be used as a meaningful adjunct application for proteomics investigation.

5. Conclusion

The function and activity of proteins are usually regulated by other proteins that interact with it. In order to understand biological processes, we need to develop a tool that gives us an insight into the knowledge of protein interactions. Although many efforts have been taken to develop the method for detecting PPIs, the accuracy and robustness of most existing methods still have potential room to be improved. Hence, we explore a fresh and efficient computational system based on protein sequences using a PCVM classifier combined with Legendre moments and a stacked sparse autoencoder. Four strictly screened PPI datasets are used to assess the prediction ability of our devised approach and the prediction outcomes display that the approach provides practical predictive capability for PPI detection. In

Table 6: Performance comparison of different methods on the *H. pylori* dataset.

Model	Accu (%)	Sens (%)	Prec (%)	MCC (%)
Phylogenetic bootstrap [71]	75.80	69.80	80.20	N/A
Boosting [71]	79.52	80.30	81.69	70.64
Signature products [72]	83.40	79.90	85.70	N/A
HKNN [73]	84.00	86.00	84.00	N/A
Proposed method	93.76	95.77	92.10	88.26

Table 7: Performance comparison of different methods on the *S. cerevisiae* dataset.

Model	Testing set	Accu (%)	Sens (%)	Prec (%)	MCC (%)
Guo [41]	ACC	89.33 ± 2.67	89.93 ± 3.68	88.87 ± 6.16	N/A
	AC	87.36 ± 1.38	87.30 ± 4.68	87.82 ± 4.33	N/A
Yang [32]	Code1	75.08 ± 1.13	75.81 ± 1.20	74.75 ± 1.23	N/A
	Code2	80.04 ± 1.06	76.77 ± 0.69	82.17 ± 1.35	N/A
	Code3	80.41 ± 0.47	78.14 ± 0.90	81.66 ± 0.99	N/A
	Code4	86.15 ± 1.17	81.03 ± 1.74	90.24 ± 1.34	N/A
You [74]	PCA-EELM	87.00 ± 0.29	86.15 ± 0.43	87.59 ± 0.32	77.36 ± 0.44
Wong [75]	PR-LPQ + RF	93.92 ± 0.36	91.10 ± 0.31	96.45 ± 0.45	88.56 ± 0.63
Proposed method	PCVM	96.55 ± 0.2	97.23 ± 0.3	95.84 ± 0.5	93.25 ± 0.3

a subsequent comparative experiment, the prediction performance by our approach is obviously better than that of an SVM-based method and previous methods. We also found that prediction quality continues to improve with increasing dataset size. This finding underscores the value of this model to train and apply very large datasets, and suggests that further performance gains may be had by increasing the data size. Therefore, this proposed method is a reliable, efficient, and powerful PPI prediction model. It can be adopted to guide the validation of relevant experiments and to be an auxiliary tool for proteomics research.

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 state no conflict of interest.

Authors' Contributions

Yanbin Wang, Zhuhong You, Liping Li, and Li Cheng considered the algorithm, arranged the datasets, and performed the analyses. Xi Zhou, Libo Zhang, Xiao Li, and Tonghai Jiang wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work is supported in part by the National Science Foundation of China (Grant nos. 61722212 and 61572506) and in part by the Pioneer Hundred Talents Program of the Chinese Academy of Sciences. The authors would like to thank the editors and anonymous reviewers for their constructive advice.

References

[1] 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.

[2] G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue, "BIND—the biomolecular interaction network database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 242–245, 2001.

[3] I. Xenarios, E. Fernandez, L. Salwinski et al., "DIP: the database of interacting proteins: 2001 update," *Nucleic Acids Research*, vol. 29, no. 1, pp. 239–241, 2001.

[4] O. Puig, F. Caspary, G. Rigaut et al., "The tandem affinity purification (TAP) method: a general procedure of protein complex purification," *Methods*, vol. 24, no. 3, pp. 218–229, 2001.

[5] M. Koegl and P. Uetz, "Improving yeast two-hybrid screening systems," *Briefings in Functional Genomics and Proteomics*, vol. 6, no. 4, pp. 302–312, 2008.

[6] U. Rüetschi, A. Rosén, G. Karlsson et al., "Proteomic analysis using protein chips to detect biomarkers in cervical and amniotic fluid in women with intra-amniotic inflammation," *Journal of Proteome Research*, vol. 4, no. 6, pp. 2236–2242, 2005.

[7] J. Sun, J. Xu, Z. Liu et al., "Refined phylogenetic profiles method for predicting protein-protein interactions," *Bioinformatics*, vol. 21, no. 16, pp. 3409–3415, 2005.

[8] I. Kotsireas, R. Melnik, and B. West, "Advances in mathematical and computational methods: addressing modern

challenges of science, technology, and society," in *AIP Conference Proceedings*, p. 1, Melville, NY, USA, 2011.

[9] I. Kotsireas, E. Lau, and R. Voino, "Exact implicitization of polynomial curves and surfaces," *ACM SIGSAM Bulletin*, vol. 37, no. 3, p. 78, 2003.

[10] I. Kotsireas and E. Zima, "Abstracts of WWCA 2011 in honor of Herb Wilf's 80th birthday," *ACM Communications in Computer Algebra*, vol. 45, no. 1/2, pp. 92–99, 2011.

[11] I. Kotsireas and E. Volcheck, "ANTS VI: algorithmic number theory symposium poster abstracts," *ACM SIGSAM Bulletin*, vol. 38, no. 3, pp. 93–107, 2004.

[12] I. Kotsireas, "Proceedings of the 2011 International Workshop on Symbolic-Numeric Computation," in *ISSAC '11 International Symposium on Symbolic and Algebraic Computation (Co-located with FCRC 2011)*, p. 18, San Jose, CA, USA, 2011.

[13] D. Vlachakis, A. Pavlopoulou, G. Tsiliki et al., "An integrated in silico approach to design specific inhibitors targeting human poly(A)-specific ribonuclease," *PLoS One*, vol. 7, no. 12, article e51113, 2012.

[14] D. Vlachakis, G. Tsiliki, A. Pavlopoulou, M. G. Roubelakis, S. C. Tsaniras, and S. Kossida, "Antiviral stratagems against HIV-1 using RNA interference (RNAi) technology," *Evolutionary Bioinformatics*, vol. 9, article EBO.S11412, 2013.

[15] D. Vlachakis, D. Tsagrasoulis, V. Megalooikonomou, and S. Kossida, "Introducing Drugster: a comprehensive and fully integrated drug design, lead and structure optimization toolkit," *Bioinformatics*, vol. 29, no. 1, pp. 126–128, 2013.

[16] D. Vlachakis, V. L. Koumandou, and S. Kossida, "A holistic evolutionary and structural study of *Flaviviridae* provides insights into the function and inhibition of HCV helicase," *PeerJ*, vol. 1, article e74, 2013.

[17] D. Vlachakis, D. G. Kontopoulos, and S. Kossida, "Space constrained homology modelling: the paradigm of the RNA-dependent RNA polymerase of dengue (type II) virus," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 108910, 9 pages, 2013.

[18] P. Vlamos, K. Lefkimmiatis, C. Cocianu, L. State, and Z. Luo, "Artificial intelligence applications in biomedicine," *Advances in Artificial Intelligence*, vol. 2013, Article ID 219137, 2 pages, 2013.

[19] P. Vlamos, V. Chrissikopoulos, and M. Psiha, "Building vulnerability: an interdisciplinary concept," *Key Engineering Materials*, vol. 628, pp. 193–197, 2014.

[20] P. Vlamos, A. Pateli, and M. Psiha, "Hybrid model for measurement of building vulnerability," *Key Engineering Materials*, vol. 628, pp. 237–242, 2014.

[21] P. Vlamos, "On the monotony of certain sequences," *Octogon Mathematical Magazine*, vol. 10, pp. 370-371, 2002.

[22] P. Vlamos and S. Tefarikis, "Numerical solution of partial differential equations," *The Mathematical Gazette*, vol. 50, pp. 179–449, 2005.

[23] A. Alexiou, M. Psiha, and P. Vlamos, "An integrated ontology-based model for the early diagnosis of Parkinson's disease," in *IFIP Advances in Information and Communication Technology*, pp. 442–450, Springer, Berlin, Heidelberg, 2012.

[24] A. Alexiou, M. Psiha, and P. Vlamos, *Towards an Expert System for Accurate Diagnosis and Progress Monitoring of Parkinson's Disease*, Springer International Publishing, 2015.

[25] A. T. Alexiou, P. Maria, J. Rekkas, and P. Vlamos, "A stochastic approach of mitochondrial dynamics," *World Academy of Science, Engineering and Technology*, vol. 55, pp. 77–80, 2011.

[26] A. Athanasios, P. Maria, T. Georgia, and V. Panayiotis, "Automated prediction procedure for Charcot-Marie-Tooth disease," in *13th IEEE International Conference on BioInformatics and BioEngineering*, pp. 1–4, Chania, Greece, 2013.

[27] M. Psiha and P. Vlamos, "Modeling neural circuits in Parkinson's disease," *Advances in Experimental Medicine and Biology*, vol. 822, pp. 139–147, 2015.

[28] 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.

[29] T. N. Tran, K. Satou, and B. H. Tu, "Using inductive logic programming for predicting protein-protein interactions from multiple genomic data," in *Knowledge Discovery in Databases: PKDD 2005*, vol. 3721 of Lecture Notes in Computer Science, pp. 321–330, Springer, Berlin, Heidelberg, 2005.

[30] T. Hamp and B. Rost, "Evolutionary profiles improve protein-protein interaction prediction from sequence," *Bioinformatics*, vol. 31, no. 12, pp. 1945–1950, 2015.

[31] H. C. Yi, Z. H. You, D. S. Huang, X. Li, T. H. Jiang, and L. P. Li, "A deep learning framework for robust and accurate prediction of ncrna-protein interactions using evolutionary information," *Molecular Therapy - Nucleic Acids*, vol. 11, pp. 337–344, 2018.

[32] 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.

[33] Y. Zhang, D. Zhang, G. Mi et al., "Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions," *Computational Biology and Chemistry*, vol. 36, pp. 36–41, 2012.

[34] Z.-H. You, M. C. Zhou, X. Luo, and S. Li, "Highly efficient framework for predicting interactions between proteins," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 731–743, 2017.

[35] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions," *Journal of Molecular Biology*, vol. 268, no. 1, pp. 209–225, 1997.

[36] 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, 2017.

[37] C. von Mering, R. Krause, B. Snel et al., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[38] T. Berggard, S. Linse, and P. James, "Methods for the detection and analysis of protein-protein interactions," *Proteomics*, vol. 7, no. 16, pp. 2833–2842, 2007.

[39] K. C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Current Proteomics*, vol. 6, no. 4, pp. 262–274, 2009.

[40] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[41] 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.

[42] W. Zhou, H. Yan, X. Fan, and Q. Hao, "Prediction of protein-protein interactions based on molecular interface features and the support vector machine," *Current Bioinformatics*, vol. 8, no. 1, pp. 3–8, 2013.

[43] L. Hua and P. Zhou, "Combining protein-protein interactions information with support vector machine to identify chronic obstructive pulmonary disease related genes," *Molecular Biology*, vol. 48, no. 2, pp. 287–296, 2014.

[44] S. Dohkan, A. Koike, and T. Takagi, "Prediction of protein-protein interactions using support vector machines," in *Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering*, pp. 165–173, Taichung, Taiwan, 2004.

[45] 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, 2018.

[46] J.-Y. An, F.-R. Meng, Z.-H. You, Y.-H. Fang, Y.-J. Zhao, and M. Zhang, "Using the relevance vector machine model combined with local phase quantization to predict protein-protein interactions from protein sequences," *BioMed Research International*, vol. 2016, Article ID 4783801, 9 pages, 2016.

[47] 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.

[48] X. Y. Pan, Y. N. Zhang, and H. B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[49] P. Smialowski, P. Pagel, P. Wong et al., "The Negatome database: a reference set of non-interacting protein pairs," *Nucleic Acids Research*, vol. 38, Supplement 1, pp. D540–D544, 2010.

[50] S. Martin, D. Roe, and J. L. Faulon, "Predicting protein–protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2004.

[51] L. Li, Y. Liang, and R. L. Bass, "GAPWM: a genetic algorithm method for optimizing a position weight matrix," *Bioinformatics*, vol. 23, no. 10, pp. 1188–1194, 2007.

[52] J. Korhonen, P. Martinmäki, C. Pizzi, P. Rastas, and E. Ukkonen, "MOODS: fast search for position weight matrix matches in DNA sequences," *Bioinformatics*, vol. 25, no. 23, pp. 3181-3182, 2009.

[53] J. Yang and S. A. Ramsey, "A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites," *Bioinformatics*, vol. 31, no. 21, pp. 3445–3450, 2015.

[54] P.-T. Yap and R. Paramesran, "An efficient method for the computation of Legendre moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1996–2002, 2005.

[55] K. M. Hosny, "Exact Legendre moment computation for gray level images," *Pattern Recognition*, vol. 40, no. 12, pp. 3597–3605, 2007.

[56] J. D. Zhou, H. Z. Shu, L. M. Luo, and W. X. Yu, "Two new algorithms for efficient computation of Legendre moments," *Pattern Recognition*, vol. 35, no. 5, pp. 1143–1152, 2002.

[57] B. Fu, J. Zhou, Y. Li, G. Zhang, and C. Wang, "Image analysis by modified Legendre moments," *Pattern Recognition*, vol. 40, no. 2, pp. 691–704, 2007.

[58] G. A. Papakostas, E. G. Karakasis, and D. E. Koulouriotis, "Accurate and speedy computation of image Legendre moments for computer vision applications," *Image and Vision Computing*, vol. 28, no. 3, pp. 414–423, 2010.

[59] J. Xu, L. Xiang, Q. Liu et al., "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 119–130, 2016.

[60] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[61] A. Sankaran, P. Pandey, M. Vatsa, and R. Singh, "On latent fingerprint minutiae extraction using stacked denoising sparse autoencoders," in *IEEE International Joint Conference on Biometrics*, pp. 1–7, Clearwater, FL, USA, 2014.

[62] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral-spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, pp. 2438–2442, 2015.

[63] Y. B. Wang, Z. H. You, X. Li et al., "Predicting protein-protein interactions from protein sequences by a stacked sparse autoencoder deep neural network," *Molecular BioSystems*, vol. 13, no. 7, pp. 1336–1344, 2017.

[64] H. Chen, P. Tino, and X. Yao, "Probabilistic classification vector machines," *IEEE Transactions on Neural Networks*, vol. 20, no. 6, pp. 901–914, 2009.

[65] Z. Xue, X. Yu, Q. Fu, X. Wei, and B. Liu, "Hyperspectral imagery classification based on probabilistic classification vector machines," in *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, C. M. Falco and X. Jiang, Eds., Chengu, China, 2016.

[66] H. Chen, P. Tino, and X. Yao, "Efficient probabilistic classification vector machine with incremental basis function selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 356–369, 2014.

[67] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proceedings. Internation Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, 1994.

[68] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[69] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[70] B. Liu, F. Liu, L. Fang, X. Wang, and K. C. Chou, "repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[71] J. R. Bock and D. A. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–134, 2003.

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

[73] 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.

[74] 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, Supplement 8, article S10, 2013.

[75] 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 *Advanced Intelligent Computing Theories and Applications*, pp. 713–720, Springer, Cham, Switzerland, 2015.