Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4783801, 9 pages
http://dx.doi.org/10.1155/2016/4783801

Research Article

Using the Relevance Vector Machine Model Combined with Local Phase Quantization to Predict Protein-Protein Interactions from Protein Sequences

Ji-Yong An,[1] Fan-Rong Meng,[1] Zhu-Hong You,[1,2] Yu-Hong Fang,[1] Yu-Jun Zhao,[1] and Ming Zhang[1]

[1]*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 21116, China*
[2]*Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*

Correspondence should be addressed to Fan-Rong Meng; mengfr@cumt.edu.cn and Zhu-Hong You; zhuhongyou@cumt.edu.cn

Received 5 March 2016; Accepted 12 April 2016

Academic Editor: Xun Lan

We propose a novel computational method known as RVM-LPQ that combines the Relevance Vector Machine (RVM) model and Local Phase Quantization (LPQ) to predict PPIs from protein sequences. The main improvements are the results of representing protein sequences using the LPQ feature representation on a Position Specific Scoring Matrix (PSSM), reducing the influence of noise using a Principal Component Analysis (PCA), and using a Relevance Vector Machine (RVM) based classifier. We perform 5-fold cross-validation experiments on *Yeast* and *Human* datasets, and we achieve very high accuracies of 92.65% and 97.62%, respectively, which is significantly better than previous works. To further evaluate the proposed method, we compare it with the state-of-the-art support vector machine (SVM) classifier on the *Yeast* dataset. The experimental results demonstrate that our RVM-LPQ method is obviously better than the SVM-based method. The promising experimental results show the efficiency and simplicity of the proposed method, which can be an automatic decision support tool for future proteomics research.

1. Introduction

Proteins are crucial molecules that participate in many cellular functions in an organism. Typically, proteins do not perform their roles individually, so detection of PPIs becomes more and more important. Knowledge of PPIs can provide insight into the molecular mechanisms of biological processes and lead to a better understanding of practical medical applications. In recent years, various high-throughput technologies, such as yeast two-hybrid screening methods [1, 2], immunoprecipitation [3], and protein chips [4], have been developed to detect interactions between proteins. Until now, a large quantity of PPI data for different organisms has been generated, and many databases, such as MINT [5], BIND [6], and DIP [7], have been built to store protein interaction data. However, these experimental methods have some shortcomings, such as being time-intensive and costly. In addition, the aforementioned approaches suffer from high rates of false positives and false negatives. For these reasons, predicting unknown PPIs is considered a difficult task using only biological experimental methods.

As a result, a number of computational methods have been proposed to infer PPIs from different sources of information, including phylogenetic profiles, tertiary structures, protein domains, and secondary structures [8–16]. However, these approaches cannot be employed when prior knowledge about a protein of interest is not available. With the rapid growth of protein sequence data, the protein sequence-based method is becoming the most widely used tool for predicting PPIs. Consequently, a number of protein sequence-based methods have been developed for predicting PPIs. For example, Bock and Gough [17] used a support vector machine (SVM) combined with several structural and physiochemical descriptors to predict PPIs. Shen et al. [18] developed

a conjoint triad method to infer human PPIs. Martin et al. [19] used a descriptor called the signature product of subsequences and an expansion of the signature descriptor based on the available chemical information to predict PPIs. Guo et al. [20] used the SVM model combined with an autocorrelation descriptor to predict *Yeast* PPIs. Nanni and Lumini [21] proposed a method based on an ensemble of K-local hyperplane distances to infer PPIs. Several other methods based on protein amino acid sequences have been proposed in previous work [22, 23]. In spite of this, there is still space to improve the accuracy and efficiency of the existing methods.

In this paper, we propose a novel computational method that can be used to predict PPIs using only protein sequence data. The main improvements are the results of representing protein sequences using the LPQ feature representation on a Position Specific Scoring Matrix (PSSM), reducing the influence of noise by using a Principal Component Analysis (PCA), and using a Relevance Vector Machine (RVM) based classifier. More specifically, we first represent each protein using a PSSM representation. Then, a LPQ descriptor is employed to capture useful information from each protein PSSM and generate a 256-dimensional feature vector. Next, dimensionality reduction method PCA is used to reduce the dimensions of the LPQ vector and the influence of noise. Finally, the RVM model is employed as the machine learning approach to carry out classification. The proposed method was executed using two different PPIs datasets: *Yeast* and *Human*. The experimental results are found to be superior to SVM and other previous methods, which prove that the proposed method performs incredibly well in predicting PPIs.

2. Materials and Methodology

2.1. Dataset. To verify the proposed method, two publicly available datasets are used in our study. The datasets are *Yeast* and *Human* that were obtained from the publicly available Database of Interaction Proteins (DIP) [24]. For better implementation, we selected 5594 positive protein pairs to build the positive dataset and 5594 negative protein pairs to build the negative dataset from the *Yeast* dataset. Similarly, we selected 3899 positive protein pairs to build the positive dataset and 4262 negative protein pairs to build the negative dataset from the *Human* dataset. Consequently, the *Yeast* dataset contains 11188 protein pairs and the *Human* dataset contains 8161 protein pairs.

2.2. Position Specific Scoring Matrix. A Position Specific Scoring Matrix (PSSM) is an $M \times 20$ matrix $X = \{X_{ij}: i = 1 \cdots M,\ j = 1 \cdots 20\}$ for a given protein, where M is the length of the protein sequence and 20 represents the 20 amino acids [28–33]. A score X_{ij} is allocated for the jth amino acid in the ith position of the given protein sequence in the PSSM. The score X_{ij} of the position of a given sequence is expressed as $X_{ij} = \sum_{k=1}^{20} p(i,k) \times q(j,k)$, where $p(i,k)$ is the ratio of the frequency of the kth amino acid appearing at position i of the probe to be the total number of probes and $q(j,k)$ is the value of Dayhoff's mutation matrix [34] between the jth and kth amino acids [35–37]. As a result, a high score represents a largely conserved position and a low score represents a weakly conserved position [38–40].

PSSMs are used to predict protein folding patterns, protein quaternary structural attributes, and disulfide connectivity [41, 42]. Here, we also use PSSMs to predict PPIs. In this paper, we used the Position Specific Iterated BLAST (PSI-BLAST) [43] to create PSSMs for each protein sequence. The e-value parameter was set as 0.001, and three iterations were selected for obtaining broadly and highly homologous sequences in the proposed method. The resulting PSSMs can be represented as 20-dimensional matrices. Each matrix is composed of $L \times 20$ elements, where L is the total number of residues in a protein. The rows of the matrix represent the protein residues, and the columns of the matrix represent the 20 amino acids.

2.3. Local Phase Quantization. Local Phase Quantization (LPQ) has been described in detail in the literature [44]. The LPQ method is based on the blur invariance property of the Fourier phase spectrum [45–47]. It is an operator used to process spatial blur in textural features of images. The spatial invariant blurring of an original image $f(x)$ apparent in an observed image $g(x)$ can be expressed as a convolution, given by

$$g(x) = f(x) * h(x), \tag{1}$$

where $h(x)$ is the function of the spread point of the blur, $*$ represents two-dimensional convolutions, and x is a vector of coordinates $[x, y]^T$. In the Fourier domain, this amounts to

$$G(u) = F(u) \cdot H(u), \tag{2}$$

where $G(u)$, $F(u)$, and $H(u)$ are the discrete Fourier transforms (DFT) of the blurred image $g(x)$, the original image $f(x)$, and $h(x)$, respectively, and u is a vector of coordinates $[u, v]^T$. According to the characteristic of the Fourier transform, the phase relations can be expressed as

$$\angle G(u) = \angle F(u) + \angle H(u). \tag{3}$$

When the spread point function $h(x)$ is the center of symmetry, meaning $h(x) = h(-x)$, the Fourier transform of $h(x)$ always has a real value. As a result, its phase can be expressed as a two-valued function, given by

$$\angle H(u) = \begin{cases} 0 & \text{if } H(u) \geq 0 \\ \pi & \text{if } H(u) < 0. \end{cases} \tag{4}$$

This means that

$$\angle G(u) = \angle F(u). \tag{5}$$

The shape of the point spread function $h(x)$ is similar to the Gaussian or Sin function. This ensures that $H(u) \geq 0$ and $\angle G(u) = \angle F(u)$ at low frequencies, which means that the phase characteristics are due to blur invariance. The local phase information can be extracted using the two-dimensional DFT in LPQ. In other words, a short-term

Fourier transform (STFT) computed over a rectangular $M \times M$ neighborhood N_x at each pixel position x of an image $f(x)$ is represented by

$$F(u,x) = \sum_{y \in N_x} f(x-y)\, e^{-j2\pi y u^T} = w_u^T f_x, \tag{6}$$

where w_u is the basis vector of the two-dimensional DFT at frequency u and f_x is another vector containing all M^2 image samples from N_x. Using LPQ, the Fourier coefficients of four frequencies are calculated: $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, and $u_4 = [a, -a]^T$, where a is a small enough number to satisfy $h(u) \geq 0$. As a result, each pixel point can be expressed as a vector, given by

$$\begin{aligned} F_x^c &= \left[F(u_1,x), F(u_2,x), F(u_3,x), F(u_4,x)\right], \\ F_x &= \left[\mathrm{Re}\left\{F_x^c\right\}, \mathrm{Im}\left\{F_x^c\right\}\right]^T. \end{aligned} \tag{7}$$

Then, using a simple scalar quantizer, the resulting vectors are quantized, given by

$$q_j(x) = \begin{cases} 1, & \text{if } g_j(x) \geq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $g_j(x)$ is the jth component of F_x. After quantization, F_x becomes an eight-bit binary number vector, and each component of F_x is assigned a weight of 2^j. As a result, the quantized coefficients are represented as integer values between 0 and 255 using binary coding

$$f_{\mathrm{LPQ}}(x) = \sum_{0}^{7} q_j(x)\, 2^j. \tag{9}$$

Finally, a histogram of these integer values from all image positions is composed and used as a 256-dimensional feature vector in classification. In this paper, the PSSM matrixes of each protein from the *Yeast* and *Human* datasets were converted to 256-dimensional feature vectors using this LPQ method.

2.4. Principal Component Analysis. Principal Component Analysis (PCA) is widely used to process data and reduce the dimensions of datasets. In this way, high-dimensional information can be projected to a low-dimensional subspace, while retaining the main information. The basic principle of PCA is as follows.

A multivariate dataset can be expressed as the following matrix X:

$$X = \begin{pmatrix} x(1) \\ \vdots \\ x(N) \end{pmatrix}, \qquad x(t) = \left[x_1(t), \ldots, x_s(t)\right], \ (t = 1, \ldots, N), \tag{10}$$

where s is the number of variables and N is the number of samplings of each variable. PCA closely related to singular value decomposition (SVD) of matrix and the singular value decomposition of matrix X as follows:

$$X = \sum_{i=1}^{s} a_i b_i c_i^T, \tag{11}$$

where c_i represent feature vector of $X^T X$ and b_i represent feature vector of XX^T and a_i is singular value. If there are m linear relationships between s variables, then m singular values are zero. Any line of X can be expressed as feature vector $(q_1, q_2, \ldots, q_k)$:

$$X^T(t) = \sum_{i=1}^{k} a_i b_i c_i = \sum_{i=1}^{k} r_i(t)\, q_i, \tag{12}$$

where $r_i(t) = x(t) q_i$ is projection $x(t)$ on q_i, feature vector $(q_1, q_2, \ldots, q_k)$ is load vector, and $r_i(t)$ is score.

When there is a certain degree of linear correlation between the variables of matrix, then the projection of final several load vectors of matrix X will be enough small for resulting from measurement noise. As a result, the principal decomposition of matrix X is represented by

$$X = r_1 q_1^T + r_2 q_2^T + \cdots + r_k q_k^T + E, \tag{13}$$

where E is error matrix and can be ignored. This does not bring about the obvious loss of useful information of data. In this paper, for the sake of reducing the influence of noise and improving the prediction accuracy, we reduce the dimensionality of the *Yeast* dataset from 256 to 180 and dimensionality of the *Human* dataset from 256 to 172 in the proposed method by using Principal Component Analysis.

2.5. Relevance Vector Machine. The characteristics of the Relevance Vector Machine have been described in detail in the literature [48]. For binary classification problems, assume that the training sample sets are $\{x_n, t_n\}_{n=1}^N$, $x_n \in R^d$ is the training sample, $t_n \in \{0, 1\}$ represents the training sample label, t_i represents the testing sample label, and $t_i = y_i + \varepsilon_i$, where $y_i = w^T \varphi(x_i) = \sum_{j=1}^N w_j K(x_i, x_j) + w_0$ is the model of classification prediction; ε_i is additional noise, with a mean value of zero and a variance of σ^2, where $\varepsilon_i \sim N(0, \sigma^2)$, $t_i \sim N(y_i, \sigma^2)$. Assuming that the training sample sets are independent and identically distributed, the observation of vector t obeys the following distribution [49–51]:

$$p\left(t \mid x, w, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\left\|t - \varphi w\right\|^2\right], \tag{14}$$

where φ is defined as follows:

$$\varphi = \begin{pmatrix} 1 & k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & k(x_N, x_1) & \ldots & k(x_N, x_N) \end{pmatrix}. \tag{15}$$

The RVM uses sample label t to predict the testing sample label t_*, given by

$$p(t_* \mid t) = \int p\left(t_* \mid w, \sigma^2\right) p\left(w, \sigma^2 \mid t\right) dw\, d\sigma^2. \tag{16}$$

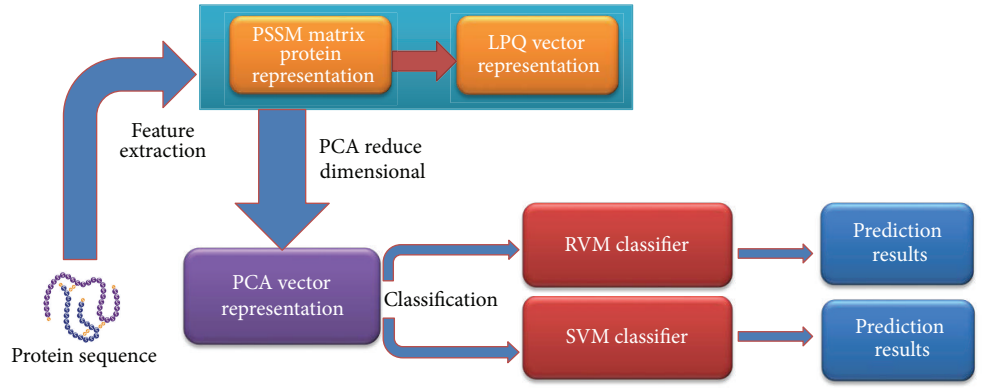

Figure 1: The flow chart of the proposed method.

To make the value of most components of the weight vector w zero and to reduce the computational work of the kernel function, the weight vector w is subjected to additional conditions. Assuming that w_i obeys a distribution with a mean value of zero and a variance of α_i^{-1}, the mean $w_i \sim N(0, \alpha_i^{-1})$, $p(w \mid a) = \prod_{i=0}^{N} p(w_i \mid a_i)$, where a is a hyper-parameters vector of the prior distribution of the weight vector w. Hence,

$$\begin{aligned} &p(t_* \mid t) \\ &\quad = \int p\left(t_* \mid w, a, \sigma^2\right) p\left(w, a, \sigma^2 \mid t\right) dw\, da\, d\sigma^2, \\ &p\left(t_* \mid w, a, \sigma^2\right) = N\left(t_* \mid y\left(x_*; w\right), \sigma^2\right). \end{aligned} \tag{17}$$

Because $p(w, a, \sigma^2 \mid t)$ cannot be obtained by an integral, it must be resolved using a Bayesian formula, given by

$$\begin{aligned} p\left(w, a, \sigma^2 \mid t\right) &= p\left(w \mid a, \sigma^2, t\right) p\left(a, \sigma^2 \mid t\right), \\ p\left(w \mid a, \sigma^2, t\right) &= \frac{p\left(t \mid w, \sigma^2\right) p(w \mid a)}{p\left(t \mid a, \sigma^2\right)}. \end{aligned} \tag{18}$$

The integral of the product of $p(t \mid a, \sigma^2)$ and $p(w \mid a)$ is given by

$$\begin{aligned} &p\left(t \mid a, \sigma^2\right) = (2\pi)^{-N/2} |\Omega|^{-1/2} \exp\left(-\frac{t^T \Omega^{-1} t}{2}\right), \\ &\Omega = \sigma^2 I + \varphi A^{-1} \varphi^T, \quad A = \operatorname{diag}\left(a_0, a_1, \ldots, a_N\right), \\ &p\left(w \mid a, \sigma^2, t\right) \\ &\quad = (2\pi)^{-(N+1)/2} |\Sigma|^{-1/2} \exp\left(-\frac{(w-u)^T (w-u)}{2}\right), \\ &\Sigma = \left(\sigma^{-2} \varphi^T \varphi + A\right)^{-1}, \\ &u = \sigma^{-2} \Sigma \varphi^T t. \end{aligned} \tag{19}$$

Because $p(a, \sigma^2 \mid t) \propto p(t \mid a, \sigma^2) p(a) p(\sigma^2)$ and $p(a, \sigma^2 \mid t)$ cannot be solved by means of integration, the solution is approximated using the maximum likelihood method, represented by

$$\left(a_{MP}, \sigma_{MP}^2\right) = \arg\max_{a, \sigma^2} p\left(t \mid a, \sigma^2\right). \tag{20}$$

The iterative process of a_{MP} and σ_{MP}^2 is as follows:

$$\begin{aligned} a_i^{\text{new}} &= \frac{\gamma_i}{\mu_i^2}, \\ \left(\sigma^2\right)^{\text{new}} &= \frac{\|t - \varphi\mu\|^2}{N - \sum_{i=0}^{N} \mu_i}, \\ \gamma_i &= 1 - a_i \sum i, i, \end{aligned} \tag{21}$$

where $\sum i, i$ is ith element on the diagonal of Σ and the initial value of a and σ^2 can be determined via the approximation of a_{MP} and σ_{MP}^2 by continuously updating using formula (21). After enough iterations, most of a_i will be close to infinity, the value of the corresponding parameters in w_i will be zero, and other a_i values will be close to finite. The resulting corresponding parameters x_i of a_i are now referred to as the relevance vector.

2.6. Procedure of the Proposed Method. In the paper, our proposed method contains three steps: feature extraction, dimensionality reduction using PCA, and sample classification. The feature extraction step contains two steps: (1) each protein from the datasets is represented as a PSSM matrix and (2) the PSSM matrix of each protein is expressed as a 256-dimensional vector using the LPQ method. Dimensional reduction of the original feature vector is achieved using the PCA method. Finally, sample classification occurs in two steps: (1) the RVM model is used to carry out classification based on the datasets from *Yeast* and *Human* whose features have been extracted and (2) the SVM model is employed to execute classification on the dataset of *Yeast*. The flow chart of the proposed method is displayed in Figure 1.

2.7. Performance Evaluation. To evaluate the feasibility and efficiency of the proposed method, five parameters, the

accuracy of prediction (Ac), sensitivity (Sn), specificity (Sp), precision (Pe), and Matthews's correlation coefficient (MCC), were computed. They are represented as follows:

$$\begin{aligned} \text{Ac} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{Sn} &= \frac{\text{TP}}{\text{TP} + \text{TN}}, \\ \text{Sp} &= \frac{\text{TN}}{\text{FP} + \text{TN}} \\ \text{Pe} &= \frac{\text{TP}}{\text{FP} + \text{TP}}, \\ \text{MCC} &= \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}}, \end{aligned} \tag{22}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. True positives stand for the number of true interacting pairs correctly predicted. True negatives are the number of true noninteracting pairs predicted correctly. False positives stand for the number of true noninteracting pairs falsely predicted, and false negatives are the number of true interacting pairs falsely predicted to be noninteracting pairs. Moreover, a Receiver Operating Curve (ROC) was created to evaluate the performance of our proposed method.

3. Results and Discussion

3.1. Performance of the Proposed Method. To avoid the overfitting in the prediction model and to test the reliability of our proposed method, we used 5-fold cross-validation in our experiment. More specifically, the whole dataset was divided into five parts; four parts were employed for training model, and one part was used for testing. Five models were gained from the *Yeast* and *Human* datasets using this method, and each model was executed alone in the experiment. For the sake of ensuring fairness, the related parameters of the RVM model were set up the same for the two different datasets, *Yeast* and *Human*. Here, the Gaussian function was selected as the kernel function with the following parameters: width = 0.6, initapla = $1/N^2$, and beta = 0, where width represents the width of the kernel function, N is the number of training samples, and the value of beta was defined as zero, which represents classification. The experimental results of the prediction models of the RVM classifier combined with Local Phase Quantization and the Position Specific Scoring Matrix and Principal Component Analysis based on the protein sequence information from the two datasets are listed in Tables 1 and 2.

Using the proposed method on the *Yeast* dataset, we achieved the results of average accuracy, sensitivity, precision, and MCC of 96.25%, 92.63%, 92.67%, and 87.27%. The standard deviations of these criteria values were 0.95%, 0.55%, 1.40%, and 1.61%, respectively. Similarly, we also obtained good results of average accuracy, sensitivity, precision, and MCC of 97.92%, 99.187%, 96.77%, and 95.95% on the *Human* dataset. The standard deviations of these criteria values were 0.81%, 0.21%, 1.57%, and 1.58%, respectively.

Table 1: 5-fold cross-validation results shown by using our proposed method on the Yeast dataset.

Testing set	Ac (%)	Sn (%)	Pe (%)	MCC (%)
1	92.76	92.73	92.79	86.56
2	93.79	93.27	93.41	88.34
3	91.28	92.12	90.43	84.08
4	92.27	92.02	92.50	85.72
5	93.17	93.02	93.32	87.27
Average	92.65 ± 0.95	92.63 ± 0.55	92.67 ± 1.40	86.40 ± 1.61

Table 2: 5-fold cross-validation results shown by using our proposed method on the Human dataset.

Testing set	Ac (%)	Sn (%)	Pe (%)	MCC (%)
1	98.10	98.99	97.25	96.27
2	97.67	99.49	96.02	95.45
3	97.37	99.25	95.55	94.87
4	97.24	98.96	95.72	94.63
5	99.26	99.22	99.31	98.54
Average	97.92 ± 0.81	99.18 ± 0.21	96.77 ± 1.57	95.95 ± 1.58

It can be seen from Tables 1 and 2 that the proposed method is accurate, robust, and effective for predicting PPIs. The better performance for predicting PPIs may be attributed to the feature extraction of the proposed method. This approach is novel and effective, and the choice of the classifier is accurate. The proposed feature extraction method contains three data processing steps. First, the PSSM matrix not only describes the order information for the protein sequence but also retains sufficient prior information; thus, it is widely used in other proteomics research. As a result, we converted each protein sequence to a PSSM matrix that contains all the useful information from each protein sequence. Second, because Local Phase Quantization has the advantage of blur invariance in the domain of image feature extraction, information can be effectively captured from the PSSMs using the LPQ method. Finally, while meeting the condition of maintaining the integrity of the information in the PSSM, we reduced the dimensions of each LPQ vector and reduced the influence of noise using Principal Component Analysis. Consequently, the sample information that was extracted using the proposed feature extraction method is very suitable for predicting PPIs.

3.2. Comparison with the SVM-Based Method. Although our proposed method achieved reasonably good results on the *Yeast* and *Human* datasets, its performance must be further validated against the state-of-the-art support vector machine (SVM) classifier. More specifically, we compared the classification performances between SVM and RVM model on the *Yeast* dataset using the same feature extraction method. The LIBSVM tool (available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/) was employed to carry out classification in SVM. Two corresponding parameters of SVM,

Table 3: 5-fold cross-validation results shown by using our proposed method on the Yeast dataset.

Testing set	Ac (%)	Sn (%)	Sp (%)	MCC (%)
SVM + PSSM + LPQ				
1	85.96	84.77	87.13	75.86
2	84.18	82.86	85.43	73.33
3	85.52	84.10	86.97	75.22
4	85.29	84.12	86.47	74.91
5	85.76	86.16	88.45	75.55
Average	85.34 ± 0.69	84.40 ± 1.20	86.89 ± 1.09	74.97 ± 0.98
RVM + PSSM + LPQ				
1	92.76	92.73	92.79	86.56
2	93.79	93.27	93.41	88.34
3	91.28	92.12	90.43	84.08
4	92.27	92.02	92.50	85.72
5	93.17	93.02	93.32	87.27
Average	*92.65 ± 0.95*	*92.63 ± 0.55*	*92.67 ± 1.40*	*86.40 ± 1.61*

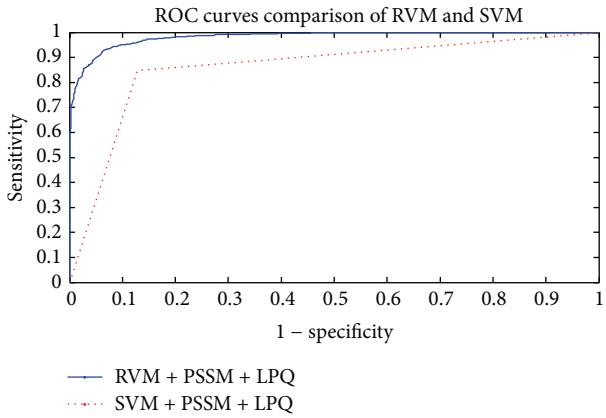

Figure 2: Comparison of ROC curves performed between RVM and SVM on the *Yeast* dataset.

c and g, are optimized using a grid search method. In the experiment, we set $c = 0.7$ and $g = 0.6$ and used a radial basis function as the kernel function.

The prediction results of the SVM and RVM methods on *Yeast* dataset are shown in Table 3, and the ROC curves are displayed in Figure 2. From Table 3, the prediction results of the SVM method achieved 85.34% average accuracy, 84.40% average sensitivity, 86.89% average specificity, and 74.97% average MCC, while the prediction results of the RVM method achieved 92.65% average accuracy, 92.63% average sensitivity, 92.67%, average specificity, and 86.40% average MCC. From these results, we can see that the RVM classifier is significantly better than the SVM classifier. In addition, the ROC curves were analyzed in Figure 2, showing that the ROC curve of the RVM classifier is significantly better than that of the SVM classifier. This clearly proves that the RVM classifier of the proposed method is an accurate and robust classifier for predicting PPIs. The increased classification performance of the RVM classifier compared with the SVM classifier can be explained by two reasons: (1) the obvious advantage of RVM is that the computational work of the kernel function is greatly reduced and (2) RVM overcomes the shortcoming of the kernel function being required to satisfy the condition of Mercer. Due to these reasons, the RVM classifier of our proposed method is significantly better than the SVM classifier. At the same time, it has been proven that the proposed method can yield highly accurate PPI predictions.

3.3. Comparison with Other Methods. In addition, a number of PPI prediction methods based on protein sequences have been proposed. To prove the effectiveness of our proposed method, we compared the prediction ability of our proposed method, which uses an RVM model combined with a Position Specific Scoring Matrix, Local Phase Quantization, and Principal Component Analysis, with existing methods on *Yeast* and *Human* datasets. It can be seen from Table 4 that the average prediction accuracy of the five different methods is between 75.08% and 89.33% for *Yeast* dataset. The prediction accuracies of these methods are lower than that of the proposed method, which is 92.65%. Similarly, the precision and sensitivity of our proposed method are also superior to those of the other methods. At the same time, Table 5 shows the average prediction accuracy between the six different methods and the proposed method on the *Human* dataset. From Table 5, the prediction accuracies yielded by the other methods are between 89.3% and 96.4%. None of these methods obtains higher prediction accuracy than our proposed method. From Tables 4 and 5, it can be observed that the proposed method yielded obviously better prediction results compared to other existing methods based on ensemble classifiers. All these results prove that the RVM classifier combined with Local Phase Quantization and the Position Specific Scoring Matrix and Principal Component Analysis can improve the prediction accuracy relative to current state-of-the-art methods. Our method improves predictions by using a correct classifier and a novel extraction method that captures the useful evolutionary information.

4. Conclusion

Knowledge of PPIs is becoming increasingly more important, which has prompted the development of computational

Table 4: Predicting ability of different methods on the Yeast dataset.

Model	Testing set	Ac (%)	Sn (%)	Pe (%)	MCC (%)
Guo et al.'s work [20]	ACC	89.33 ± 2.67	89.93 ± 3.60	88.77 ± 6.16	N/A
	AC	87.36 ± 1.38	87.30 ± 4.68	87.82 ± 4.33	N/A
Zhou et al.'s work [25]	SVM + LD	88.56 ± 0.33	87.37 ± 0.22	89.50 ± 0.60	77.15 ± 0.68
Yang et al.'s work [26]	Cod1	75.08 ± 1.13	75.81 ± 1.20	74.75 ± 1.23	N/A
	Cod2	80.04 ± 1.06	76.77 ± 0.69	82.17 ± 1.35	N/A
	Cod3	80.41 ± 0.47	78.14 ± 0.90	81.66 ± 0.99	N/A
	Cod4	86.15 ± 1.17	81.03 ± 1.74	90.24 ± 1.34	N/A
You et al.'s work [27]	PCA-EELM	87.00 ± 0.29	86.15 ± 0.43	87.59 ± 0.32	77.36 ± 0.44
The proposed method	*RVM*	*92.65 ± 0.95*	*92.63 ± 0.55*	*92.67 ± 1.40*	*86.40 ± 1.61*

Table 5: Predicting ability of different methods on the Human dataset.

Model	Ac (%)	Sn (%)	Pe (%)	MCC (%)
LDA + RF [28]	96.4	94.2	N/A	92.8
LDA + RoF [28]	95.7	97.6	N/A	91.8
LDA + SVM [28]	90.7	89.7	N/A	81.3
AC + RF [28]	95.5	94.0	N/A	91.4
AC + RoF [28]	95.1	93.3	N/A	91.0
AC + SVM [28]	89.3	94.0	N/A	79.2
The proposed method	*97.92*	*99.18*	*96.77*	*95.95*

methods. Though many approaches have been developed to solve this problem, the effectiveness and robustness of previous prediction models can still be improved. In this study, we explore a novel method using an RVM classifier combined with Local Phase Quantization and a Position Specific Scoring Matrix. From the experimental results, it can be seen that the prediction accuracy of the proposed method is obviously higher than those of previous methods. It is a very promising and useful support tool for future proteomics research. The main improvements of the proposed method come from adopting an effective feature extraction method that can capture useful evolutionary information. Moreover, the results showed that PCA significantly improves the prediction accuracy by integrating the useful information and reducing the influence of noise. In addition, the experimental results show that the RVM model is suitable for predicting PPIs. In conclusion, the proposed method is an efficient, reliable, and powerful prediction model and can be a useful tool for future proteomics research.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Authors' Contributions

The authors wish it to be known that, in their opinion, Ji-Yong An and Zhu-Hong You should be regarded as joint first authors.

Acknowledgments

This work is supported by the National Science Foundation of China, under Grants 61373086 and 61572506, in part by the Shenzhen Foundational Research Funding under Grant JCYJ20150626110425228.

References

[1] A.-C. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[3] H. Yuen, G. Albrecht, H. Adrian et al., "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

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

[5] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT: a molecular INTeraction database," *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

[6] G. D. Bader, B. Doron, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 242–245, 2003.

[7] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "DIP, the database of interacting proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, 2002.

[8] X. Luo, Z. Ming, Z. You, S. Li, Y. Xia, and H. Leung, "Improving network topology-based protein interactome mapping via collaborative filtering," *Knowledge-Based Systems*, vol. 90, pp. 23–32, 2015.

[9] S. Li, Z.-H. You, H. Guo, X. Luo, and Z.-Q. Zhao, "Inverse-free extreme learning machine with optimal information updating," *IEEE Transactions on Cybernetics*, vol. 46, no. 5, pp. 1229–1241, 2015.

[10] 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.

[11] Z.-H. You, S. Li, X. Gao, X. Luo, and Z. Ji, "Large-scale protein-protein interactions detection by integrating big biosensing data

with computational model," *BioMed Research International*, vol. 2014, Article ID 598129, 9 pages, 2014.

[12] 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.

[13] Y.-K. Lei, Z.-H. You, Z. Ji, L. Zhu, and D.-S. Huang, "Assessing and predicting protein interactions by combining manifold embedding with multiple information integration," *BMC Bioinformatics*, vol. 13, supplement 7, article S3, 2012.

[14] 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, article 343, 2010.

[15] X. Lan, R. Bonneville, J. Apostolos, W. Wu, and V. X. Jin, "W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data," *Bioinformatics*, vol. 27, no. 3, Article ID btq669, pp. 428–430, 2011.

[16] X. Lan, H. Witt, K. Katsumura et al., "Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages," *Nucleic Acids Research*, vol. 40, no. 16, pp. 7690–7704, 2012.

[17] J. R. Bock and D. A. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–135, 2003.

[18] 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.

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

[20] 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.

[21] 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.

[22] L. Nanni, "Fusion of classifiers for predicting protein-protein interactions," *Neurocomputing*, vol. 68, pp. 289–296, 2005.

[23] L. Nanni, "Hyperplanes for predicting protein-protein interactions," *Neurocomputing*, vol. 69, no. 1–3, pp. 257–263, 2005.

[24] I. Xenarios, Ł. Salwínski, 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.

[25] Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, *Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence*, Springer, Berlin, Germany, 2011.

[26] 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.

[27] Z. H. You, Y. K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, no. 8, pp. 69–75, 2013.

[28] 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.

[29] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[30] B. Liu, S. Wang, and X. Wang, "DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation," *Scientific Reports*, vol. 5, Article ID 15479, 2015.

[31] X. Chen, C. C. Yan, X. Zhang et al., "WBSMDA: within and between score for MiRNA-disease association prediction," *Scientific Reports*, vol. 6, article 21106, 2016.

[32] 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, Article ID 867516, 9 pages, 2015.

[33] L. Wong, Z.-H. You, Z. Ming, J. Li, X. Chen, and Y.-A. Huang, "Detection of interactions between proteins through rotation forest and local phase quantization descriptors," *International Journal of Molecular Sciences*, vol. 17, no. 1, p. 21, 2015.

[34] M. Dayhoff, "A model of evolutionary change in proteins," *Atlas of Protein Sequence & Structure*, vol. 5, pp. 345–352, 1978.

[35] Z.-H. You, K. C. C. Chan, and P. Hu, "Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest," *PLoS ONE*, vol. 10, no. 5, Article ID e0125811, 2015.

[36] Y. Huang, Z. You, X. Gao, L. Wong, and L. Wang, "Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence," *BioMed Research International*, vol. 2015, Article ID 902198, 10 pages, 2015.

[37] Q. Huang, Z. You, X. Zhang, and Y. Zhou, "Prediction of protein-protein interactions with clustered amino acids and weighted sparse representation," *International Journal of Molecular Sciences*, vol. 16, no. 5, pp. 10855–10869, 2015.

[38] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 13, pp. 4355–4358, 1987.

[39] B. Liu, L. Fang, R. Long, X. Lan, and K. C. Chou, "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 32, no. 3, pp. 362–369, 2016.

[40] L. Bin, C. Junjie, and X. Xiaolong, "Application of learning to rank to protein remote homology detection," *Bioinformatics*, vol. 31, no. 21, pp. 3492–3498, 2015.

[41] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. 1, pp. W65–W71, 2015.

[42] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[43] S. F. Altschul and E. V. Koonin, "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases," *Trends in Biochemical Sciences*, vol. 23, no. 11, pp. 444–447, 1998.

[44] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds., vol. 5099 of *Lecture Notes in Computer Science*, pp. 236–243, 2008.

[45] H. Wang, A. Song, B. Li, B. Xu, and Y. Li, "Psychophysiological classification and experiment study for spontaneous EEG based on two novel mental tasks," *Technology and Health Care*, vol. 23, supplement 2, pp. S249–S262, 2015.

[46] Y. Li and E. B. Olson, "Structure tensors for general purpose LIDAR feature extraction," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '11)*, pp. 1869–1874, Shanghai, China, May 2011.

[47] Y. Li and E. B. Olson, "A general purpose feature extractor for light detection and ranging data," *Sensors*, vol. 10, no. 11, pp. 10356–10375, 2010.

[48] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[49] Y. Li, S. Li, Q. Song, H. Liu, and M. Q.-H. Meng, "Fast and robust data association using posterior based approximate joint compatibility test," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 331–339, 2014.

[50] S. Li and Y. Li, "Nonlinearly activated neural network for solving time-varying complex sylvester equation," *IEEE Transactions on Cybernetics*, vol. 44, no. 8, pp. 1397–1407, 2014.

[51] Y. Li, S. Li, and Y. Ge, "A biologically inspired solution to simultaneous localization and consistent mapping in dynamic environments," *Neurocomputing*, vol. 104, pp. 170–179, 2013.