THE PROTEIN SOCIETY

Improving protein–protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model

Ji-Yong An,[1] Fan-Rong Meng,[1]* Zhu-Hong You,[1]* Xing Chen,[2] Gui-Ying Yan,[3] and Ji-Pu Hu[1]

[1]School of Computer Science Technology, China University of Mining and Technology, Xuzhou, Jiangsu 21116, China
[2]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 21116, China
[3]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China

Received 18 June 2016; Accepted 22 July 2016
DOI: 10.1002/pro.2991
Published online 25 July 2016 proteinscience.org

Abstract: Predicting protein–protein interactions (PPIs) is a challenging task and essential to construct the protein interaction networks, which is important for facilitating our understanding of the mechanisms of biological systems. Although a number of high-throughput technologies have been proposed to predict PPIs, there are unavoidable shortcomings, including high cost, time intensity, and inherently high false positive rates. For these reasons, many computational methods have been proposed for predicting PPIs. However, the problem is still far from being solved. In this article, we propose a novel computational method called RVM-BiGP that combines the relevance vector machine (RVM) model and Bi-gram Probabilities (BiGP) for PPIs detection from protein sequences. The major improvement includes (1) Protein sequences are represented using the Bi-gram probabilities (BiGP) feature representation on a Position Specific Scoring Matrix (PSSM), in which the protein evolutionary information is contained; (2) For reducing the influence of noise, the Principal Component Analysis (PCA) method is used to reduce the dimension of BiGP vector; (3) The powerful and robust Relevance Vector Machine (RVM) algorithm is used for classification. Five-fold cross-validation experiments executed on *yeast* and *Helicobacter pylori* datasets, which achieved very high accuracies of 94.57 and 90.57%, respectively. Experimental results are significantly better than previous methods. 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-BiGP method is significantly better than the SVM-based method. In addition, we achieved 97.15% accuracy on imbalance *yeast* dataset, which is higher than that of balance *yeast* dataset. The promising experimental results show the efficiency and robust of the proposed method, which can be an automatic decision support tool for future proteomics research. For facilitating extensive studies for future proteomics research, we developed a freely available web server called RVM-BiGP-PPIs in Hypertext Preprocessor (PHP) for predicting PPIs. The web server including source code and the datasets are available at http://219.219.62.123:8888/BiGP/.

Keywords: evolutionary information; position specific scoring matrix; proteomics

*Correspondence to: Zhu-Hong You, School of Computer Science and Technology, China University of Mining and Technology, Xuzhou Jiangsu 21116, China, E-mail: zhuhongyou@hotmail.com

Introduction

In an organism, proteins are fundamental molecules, which participate in many cellular functions. Especially, proteins seldom perform their roles alone, so it is much important for PPIs detection. Thus, PPIs detection is an essential step for basic research

Table I. *Description of* Yeast *and* Helicobacter pylori *Protein Sequence Dataset*

Organisms	Number of positive pairs	Number of negative pairs
Yeast	5594	5594
Helicobacter pylori	1458	1458

and practical application, which can provide insight into molecular functions biological processes, and bring about a deep understanding of disease mechanisms, and suggest novel methods for practical medical applications. Until now, many high-throughput approaches, such as yeast two-hybrid (Y2H) screening methods,[1,2] immunoprecipitation,[3] and protein chips,[4] have been used to detect PPIs. However, these experimental methods have some disadvantages, such as time-intensiveness and high cost. Besides, the aforementioned methods suffer from high rates of false positives and false negatives. For the sake of these reasons, it is difficult for predicting unknown PPIs only using biological experimental approaches. Therefore, developing computational methods for PPIs becomes more and more important.

Up to now, variously computational methods[5,6] for PPIs detection have been proposed from different sources of information, including tertiary structures, phylogenetic profiles, protein domains, and secondary structures. However, if it is not available for prior-knowledge about a protein of interest, these methods cannot be used. With the rapid growth of protein sequence data, protein sequence-based method[7,8] is becoming the most widely used tool for predicting PPIs. Consequently, large quantities of protein sequence-based methods for predicting PPIs have been exploited.[9–14] Such as, Sylvain *et al*.[15] proposed a novel protein–protein interaction prediction engine called PIPE, which can detect PPIs for any target pair of the yeast *Saccharomyces cerevisiae* proteins. Xia *et al*.[16] proposed a sequence-based method that selected rotation forest as classifier and employed autocorrelation descriptor as feature extraction method for predicting PPIs. Hamed *et al*.[17] proposed a novel approach to predict interaction of two proteins solely by analyzing their coding sequences. Ming *et al*.[12] proposed a sequence-based method that using support vector machine (SVM) combined with correlation coefficient (CC) transformation. In spite of this, there has still space to improve the accuracy and efficiency of the existing methods.

In this article, we proposed a novel computational method that can be used to predict PPIs only using protein sequence data. Improving the accuracy of PPIs detection is the main purpose in the study. The major novelty of our proposed method includes (1) protein sequences are represented using the Bi-gram probabilities (BiGP) feature representation on a position specific scoring matrix (PSSM), (2) For reducing the influence of noise, principal component analysis (PCA) method is used to reduce the dimension of BiGP vector, (3) using the relevance vector machine (RVM) based classifier. First, each protein sequence is represented using a PSSM. Second, the Bi-gram probabilities (BiGP) descriptor is used to capture useful information from each protein sequence PSSM and generate a 400-dimensional feature vector. Third, The PCA method is employed to reduce the dimensions of the BiGP vector. Finally, the RVM model is employed as the machine learning approach to perform classification. The proposed method was carried out by using two different PPIs datasets (*yeast* and *Helicobacter pylori*). The experimental results are found to be superior to SVM and other previous methods. In addition, to further evaluate the feasibility and efficiency of the proposed method, imbalance *yeast* dataset consists of 5594 positive protein pairs and 16,782 negative protein pairs used to execute using the proposed method. Thus, we compared the prediction accuracy between balance *yeast* and imbalance *yeast*, the prediction accuracy of imbalance *yeast* is higher that of balance *yeast*. Experimental results demonstrate that the proposed method is suitable for predicting PPIs. It proved that the proposed method performs incredibly well for PPIs detection.

Materials and Methodology

Dataset

In this article, *yeast* and *Helicobacter pylori* protein sequence dataset have been used. The two datasets can be obtained from the publicly available database of interaction proteins (DIP).[18] The *yeast* contains 5594 positive protein pairs and 5594 negative protein pairs. Similarly, the *Helicobacter pylori* consist of 1458 positive protein pairs and 1458 negative protein pairs. The description of *Yeast* and *Helicobacter pylori* protein sequence dataset were shown in Table I.

In addition, for further evaluating the proposed method, we created imbalance *yeast* dataset. First, we count the number of without repetition protein sequences on *yeast* dataset, where contain 2530 without repetition protein sequences. A total of 6,400,900 protein pairs were created from 2530 protein. Here, we removed 5594 positive protein pairs from 6,400,900 protein pairs. As a result, we obtained 6,395,306 negative protein pairs. Finally, 5594 positive protein pairs were selected to build the positive pairs and 16,782 negative protein pairs random selected from 6395306 negative protein pairs to build the negative pairs. As a result, the balance *yeast* dataset contains 11,188 protein pairs, the imbalance *yeast* dataset consist of 22,376 protein pairs, and the *Helicobacter pylori* dataset contains 2916 protein pairs.

Position-specific scoring matrix

Position specific scoring matrix (PSSM) was originally employed to detect distantly related proteins, which can be generated from a set of protein sequences.[19] For a given protein sequence, PSSM can be defined as an $M \times 20$ matrix$P=\{P_{ij}i : 1=1 \ldots L, j=1 \ldots 20\}$, where L is a protein sequence length, and 20 represents 20 amino acids. A score P_{ij} for the j_{th} amino acid in the i_{th} position of the query protein sequence is assigned by PSSM. The score P_{ij} can be expressed as $P_{ij}=\sum_{k=1}^{20} m(i,k)\times n(j,k)$, where $m(i,k)$ represents the k_{th} amino acid appearing frequency ratio at position i of the probe, and $n(i,k)$ is the value of Dayhoff's mutation matrix between j_{th} and k_{th} amino acids. As a result, a high score represents a largely conserved position and a small score represents a weakly conserved position.

PSSM is very useful to predict protein quaternary structural attributes, disulfide connectivity, and folding patterns.[20,21] Thus, it is used to predict PPIs in this work. The Position Specific Iterated BLAST (PSI-BLAST)[22] has been employed to build each protein sequence PSSM. To obtain broadly and highly homologous sequences, the e-value parameter of PSI_BLAST was selected as 0.001 and three iterations were chosen. The resulting PSSM can be represented as 20-dimensional matrices. Each matrix contains $L\times$ 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.

Bi-gram probabilities

In this section, the Bi-gram probabilities (BiGP) feature extraction method using PSSM linear probabilities is expressed. The characteristics of the Bi-gram probabilities was originally described in the literature.[23] The Bi-gram probabilities (BiGP) represents the given protein sequence by using its PSSM and the Bi-gram features is calculated using the PSSM probability information. Let P represent the PSSM of a given protein. PSSM has been mentioned in the Position-specific Scoring Matrix section of the article. Thus, the matrix P contains L rows and 20 columns, where L is a protein sequence length. A PSSM element P_{ij} for the j_{th} amino acid in the i_{th} position for a protein sequence can be interpreted as the relative probability of j_{th} amino acid at the i_{th} location of the primary protein sequence, which can be expressed as $P_{ij}=\sum_{j=1}^{20} i : 1=1 \ldots L, j=1 \ldots 20$. The frequency of occurrence of transition from m_{th} amino acid to n_{th} amino acid can be defined as follows:

$$BGP_{mn}=\sum_{i=1}^{L-1} P_{i,m}P_{i+1,n} 1 \le m \le 20, 1 \le n \le 20 \quad (1)$$

The Eq. (1) gives 400 frequencies of occurrences BGP_{mn} for 400 bi-gram transitions, the matrix BiGP is called as the bi-gram occurrence matrix and its 400 elements define our bi-gram feature vector[23] as follows:

$$BF=[BGP_{1,1}, BGP_{1,2} \ldots BGP_{1,20}, BGP_{2,1}, \ldots BGP_{2,20}, \ldots\ldots BGP_{20,1}, \ldots BGP_{20,20}] \quad (2)$$

These bi-gram features can also be represented as follows:

$$BF=[\varphi_{1,}, \varphi_2, \varphi_{3,} \ldots \varphi_{u,}, \ldots \varphi_\theta] \quad (3)$$

Where θ=mn=400 is the dimensionality of the feature vectorBF, the φ_u can be represented as follows:

$$\varphi_u=\begin{cases} BGP_{1,u} & (1 \le u \le 20) \\ BGP_{2,u-20} & (21 \le u \le 40) \\ \ldots\ldots & \\ BGP_{20,u-380} & (381 \le u \le 400) \end{cases} \quad (4)$$

Finally, each protein sequence of *yeast* and *Helicobacter pylori* was converted into a 400-dimensional vector using the Bi-gram Probabilities feature extraction method.

Principal component analysis

Principal component analysis (PCA) is widely used to reduce the dimensional of sample data. In such a way, high-dimensional sample data can be projected to a low-dimensional subspace. At the same time, the useful information can be retained. Suppose a multivariate sample data can be represented as

$$P=\begin{pmatrix} p(1) \\ \cdots \\ p(N) \end{pmatrix}, p(t)=[p_1(t) \ldots p_s(t)],) t=1, \ldots N) \quad (5)$$

Wheres represents the number of variables, and N represents the number of each variable sampling. PCA closely related to singular value decomposition (SVD) of matrix and the singular value decomposition of matrix P can be expressed as

$$P=\sum_{i=1}^{s} a_i b_i c_i^T \quad (6)$$

Where c_i represents feature vector of P^TP and b_i represents feature vector of PP^T, and a_i is singular value. If there are m linear relationships between s variables, then m singular values is zero. Any line of P can be defined as feature vector $(q_1, q_2, \ldots, q_k)$

$$P^T(t)=\sum_{i=1}^{k} a_i b_i c_i=\sum_{i=1}^{k} r_i(t) q_i \quad (7)$$

Where $r_i(t)=x(t)q_i$ represents the projection of $p(t)$ onq_i, feature vector $(q_1, q_2, \ldots, q_k)$ is load vector, and $r_i(t)$ is score.

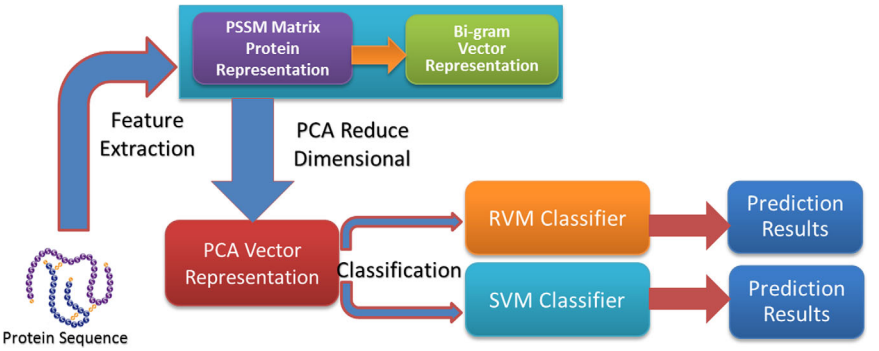

Figure 1. The flow chart of the proposed method.

If there are certain degrees of linear correlation between the variables of matrix, then the projection of final several load vectors of matrix P will become enough small, which resulting from measurement noise. Consequently, the principal decomposition of matrix P represented as

$$P=r_1q_1^T+r_2q_2^T+\cdots+r_kq_k^T+E \tag{8}$$

Where E is error matrix and can be ignored. This not brings about the obvious loss of useful information of data.

In the study, in order to reduce the influence of noise and improve the prediction accuracy, the dimensional of balance *yeast* and imbalance *yeast* and *Helicobacter pylori* have been reduced from 400 to 350 using PCA.

Relevance vector machine

The characteristics of the Relevance Vector Machine described in the literature.[24] We can assume that $\{x_n,t_n\}_{n=1}^N$, $x_n\in R^d$ is the training sample for binary classification problems, where $t_n\in\{0,1\}$ represents the training sample label, t_i represents the label of testing sample, and $t_i=y_i+\varepsilon_i$, where $y_i=w^T\varphi(x_i)=\sum_{j=1}^N w_jK(x_i,x_j)+w_0$ is the classification model; ϵ_i represents additional noise, with a mean value of zero and a variance of σ^2, where $\varepsilon_i\sim N(0,\sigma^2)$, $t_i\sim N(y_i,\sigma^2)$. It is assumed that the training datasets are independent identically distributed; the vector t obeys the following distribution:

$$p(t|x,w,\sigma^2)=(2\pi\sigma^2)^{-N/2}\exp\left[-\frac{1}{2\sigma^2}||t-\varphi w||^2\right] \tag{9}$$

Where φ is defined as follows:

$$\varphi=\begin{pmatrix}1 & k(x_1,x_1)\cdots & k(x_1,x_N))\\ \cdots & \cdots & \cdots\\ 1 & k(x_N,x_1)\ldots & k(x_N,x_N)\end{pmatrix} \tag{10}$$

The sample label t is used to predict the testing sample label t_*, given by

$$p(t_*|t)=\int p(t_*|w,\sigma^2)p(w,\sigma^2|t)dwd\sigma^2 \tag{11}$$

For making the value of most components of the weight vector w zero and reducing the amount of calculation of the kernel function, additional conditions is attached to the weight vector w. Assuming that w_i obeys a distribution with a mean value of zero and a variance of α_i^{-1}, the mean $w_i\sim N(0,\ \alpha_i^{-1}),p(w|a)=\prod_{i=0}^N p(w_i|a_i)$, where a is a hyper-parameters vector of the prior distribution of the weight vector w.

$$p(t_*|t)=\int p(t_*|w,a,\sigma^2)p(w,a,\sigma^2|t)dwdad\sigma^2 \tag{12}$$

$$p(t_*|w,a,\sigma^2)=N(t_*\ |y(x_*;w),\sigma^2). \tag{13}$$

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

$$p(w,a,\sigma^2|t)=p(w|a,\sigma^2,t)p(a,\sigma^2|t) \tag{14}$$

$$p(w|a,\sigma^2,t)=p(t|w,\sigma^2)p(w|a)/p(t|a,\sigma^2) \tag{15}$$

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

$$p(t|a,\sigma^2)=(2\pi)^{-N/2}|\Omega|^{-1/2}\exp\left(-\frac{t^T\Omega^{-1}t}{2}\right) \tag{16}$$

$$\Omega=\sigma^2I+\varphi A^{-1}\varphi^T,\ A=diag(a_0,a_1,\ldots,a_N), \tag{17}$$

$$p(w|a,\sigma^2,t)=(2\pi)^{-(N+1)/2}|\Sigma|^{-1/2}\exp\left(-\frac{(w-u)^T(w-u)}{2}\right) \tag{18}$$

Table II. *Fivefold Cross Validation Results Shown Using Our Proposed Method on* Yeast

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	94.32	94.88	93.79	89.29
2	95.22	95.04	95.46	90.89
3	94.81	93.62	96.00	90.17
4	94.73	94.59	94.59	90.00
5	93.79	93.20	94.45	88.36
Average	94.57 ± 0.005	94.27 ± 0.0.008	94.86 ± 0.009	89.74 ± 0.010

$$\Sigma=\left(\sigma^{-2}\varphi^{T}\varphi+A\right)^{-1} \tag{19}$$

$$u=\sigma^{-2}\Sigma\varphi^{T}t \tag{20}$$

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

$$(a_{\mathrm{MP}},\sigma^2_{\mathrm{MP}})=\arg\max_{a,\sigma^2} p(t|a,\sigma^2) \tag{21}$$

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

$$\begin{cases} a_i^{new}=\dfrac{\gamma_i}{\mu_i^2} \\ (\sigma^2)^{new}=\dfrac{||t-\varphi\mu||^2}{N-\sum_{i=0}^{N}\mu_i} \\ \gamma_i=1-a_i\sum i,i \end{cases} \tag{22}$$

Here $\sum i,i$ is ith element on the diagonal of Σ, and the initial value of a and σ^2 can be decided via the approximation of a_{MP} and σ^2_{MP} using formula (22) continuously renewal. After enough iterations, most of a_i will be close to infinity, 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.

Procedure of the proposed method

In the work, the proposed method consists of three steps: feature extraction, dimensionality reduction using PCA, and sample classification. The feature extraction method includes two steps: (1) each protein sequence is represented as a PSSM matrix; (2) representing each protein sequence PSSM as a 400-dimensional vector by taking advantage of Bi-gram probabilities. Each 400-dimensional vector was converted into 350 dimensional using the PCA method. Finally, sample classification occurs in two steps: (1) the RVM model is employed to execute classification on balance *yeast* and imbalance *yeast* and *Helicobacter pylori* datasets; (2) the SVM model is used to perform classification on *yeast* dataset. The flow chart of the proposed method is shown in Figure 1.

Performance evaluation

In this article, for evaluating the feasibility and effectiveness of the proposed method, four parameters include (1) Accuracy (Ac), (2) Sensitivity (Sn), (3) Precision (Pe), (4) Matthews's correlation coefficient (Mcc), which were calculated. They are represented as follows:

Table III. *Fivefold Cross Validation Results Shown Using Our Proposed Method on* Helicobacter pylori

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	91.77	93.15	90.67	84.88
2	89.37	92.12	87.34	80.96
3	91.42	92.74	90.94	84.28
4	91.08	90.22	90.88	83.70
5	89.21	91.19	87.90	80.73
Average	90.57 ± 0.012	91.88 ± 0.012	89.55 ± 0.018	82.91 ± 0.020

Table IV. *Fivefold Cross Validation Results Shown Using Our Proposed Method on Imbalance Yeast*

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	97.79	95.06	95.85	94.12
2	97.05	93.45	94.79	92.38
3	97.25	94.40	94.97	93.03
4	97.09	93.02	95.23	92.41
5	96.56	93.15	92.81	91.02
Average	97.15 ± 0.004	93.82 ± 0.009	94.73 ± 0.011	92.59 ± 0.011

Table V. *Five-fold Cross Validation Results Shown by Using Our Proposed Method on* Yeast

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
SVM+PSSM+BiGP				
1	85.96	83.60	87.55	75.83
2	86.90	83.68	88.81	77.15
3	87.04	85.34	88.62	77.42
4	86.81	84.21	88.89	77.07
5	86.29	84.82	87.83	76.34
Average	86.60 ± 0.005	84.33 ± 0.007	88.34 ± 0.006	76.76 ± 0.007
RVM+PSSM+BiGP				
1	94.32	94.88	93.79	89.29
2	95.22	95.04	95.46	90.89
3	94.81	93.62	96.00	90.17
4	94.73	94.59	94.59	90.00
5	93.79	93.20	94.45	88.36
Average	94.57 ± 0.005	94.27 ± 0.0.008	94.86 ± 0.009	89.74 ± 0.010

$$\mathrm{Ac}=\frac{TP+TN}{TP+FP+TN+FN}$$

$$\mathrm{Sn}=\frac{TP}{TP+TN}$$

$$\mathrm{Pe}=\frac{TP}{FP+TP}$$

$$\mathrm{Mcc}=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}}$$

Where TN represents true negatives, TP represents true positives, FN represents false negatives and FP represents false positives respectively. True positives represent the count of true interacting pairs correctly predicted. True negatives are the number of true noninteracting pairs predicted correctly. False positives defined as the count of true noninteracting pairs falsely predicted, and false negatives represent true interacting pairs falsely predicted to be noninteracting pairs. Moreover, a receiver operating curve (ROC) was generated to evaluate the performance of our proposed method.

Results and Discussion

Performance of the proposed method

For averting the over-fitting and verifying the efficacy and stability of our proposed method, fivefold cross validation was employed in the experiment. More specifically, we divided into the whole dataset five parts; four parts were selected as training dataset and one part was selected as test dataset. For the sake of ensuring fairness, there are several parameters for RVM model, which were set up the same for the balance *yeast* and imbalance *yeast* and *Helicobacter pylori* datasets. Thus, we chose the Gaussian function as the kernel function. Meanwhile, we set up the three parameters: width = 2.8, initapla = 1/*N* and beta = 0, where width represents the width of Gaussian function, *N* represents the count of training samples, and beta represents classification or regression. Hear, "beta = 0" represents classification. The experimental results of the RVM classifier combined with Bi-gram probabilities and PSSM and PCA, which based on the protein sequence information from balance *yeast* and imbalance *yeast and Helicobacter pylori* datasets are shown in Tables (II–IV).

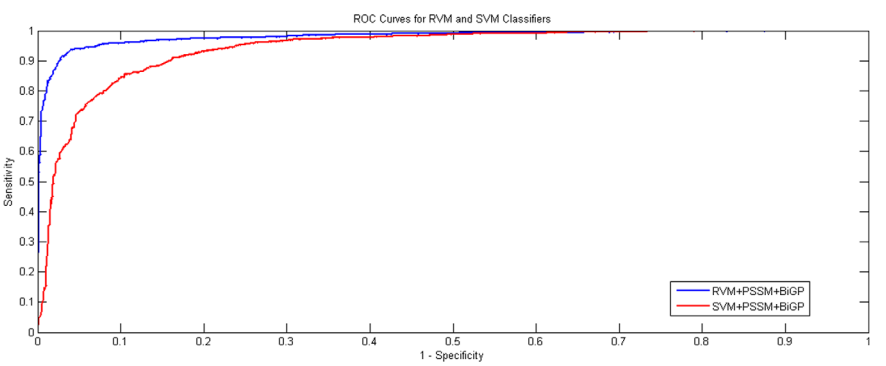

Figure 2. Comparison of ROC curves between RVM and SVM on *yeast* dataset.

Table VI. *Predicting Ability of Different Methods on* Yeast

Model	Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
Guo's work[26]	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's work[27]	SVM + LD	88.56 ± 0.33	87.37 ± 0.22	89.50 ± 0.60	77.15 ± 0.68
Yang's work[28]	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's work[29]	PCA-EELM	87.00 ± 0.29	86.15 ± 0.43	87.59 ± 0.32	77.36 ± 0.44
Proposed method	RVM	94.57 ± 0.005	94.27 ± 0.0.008	94.86 ± 0.009	89.74 ± 0.010

We performed computational experiments on balance y*east* and imbalance y*east and Helicobacter pylori* datasets to verify the efficacy and stability of the propose method. We obtained the results of average accuracy, sensitivity, precision, and Mcc of 94.57, 94.27, 94.86, and 89.74% and the standard deviations of them of 0.005, 0.008, 0.009, and 0.01% on *yeast* dataset, respectively. At the same time, good results of average accuracy, sensitivity, precision, and Mcc of 97.15, 93.82, 94.73, and 92.59% were obtained on imbalance *yeast* dataset and the standard deviations of them of 0.004, 0.009, 0.011, and 0.019% achieved, respectively. Similarly we also achieved good results of average accuracy, sensitivity, precision, and Mcc of 90.57, 91.88, 89.55, and 82.91% on *Helicobacter pylori* dataset and the standard deviations of them of 0.012, 0.012, 0.018, and 0.020%, respectively.

It can be found from Tables II–IV that the proposed method is accurate, robust, and effective for predicting PPIs. The better prediction accuracy achieved may be attributed to feature extraction method and choice of classifier. This feature extraction approach is novel and effective, and the choice of the classifier is accurate. The major improvement of the proposed feature extraction method lies in three reasons: (1) the PSSM matrix is a much useful tool for representing protein sequence, which can not only describes the order information but also retains sufficient prior information for the protein sequence. As a result, each protein sequence represented as a PSSM that contains all the useful information for predicting PPIs. (2) The Bi-gram probabilities represented each protein sequence by its PSSM and calculated the Bi-gram feature using the probability information contained in PSSM. The Bi-grams features from PSSMs can significantly reduce the sparsity level which helps in improving the recognition performance.[23] (3) Under the condition of guaranteeing the integrity of the information of feature vector, for reducing the influence of noise, each Bi-gram vector was reduced dimensional using PCA method. The experiments results demonstrated that the feature vector extracted using the proposed feature extraction method is very fit for predicting PPIs. In addition, it can be observed from Tables II and IV, the prediction accuracy of imbalance y*east* is higher than that of balance y*east* using the proposed method, which further proved that the proposed prediction model is accurate, robust, and effective for PPIs detection.

Table VII. *Predicting Ability of Different Methods on* Helicobacter pylori

Model	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
Nanni[30]	83	86	85.1	N/A
Nanni[31]	84	86	84	N/A
Nanni and Lumini[32]	86.6	86.7	85	N/A
Z-H You[29]	87.5	88.95	86.15	78.13
L Nanni[31]	84	84	86	N/A
Proposed method	90.57	91.88	89.55	82.91

Comparison with the SVM-based method

To further validate the effectiveness of the proposed approach, we compared the prediction accuracies with that of the state-of-the-art support vector machine (SVM) classifier. More specifically, the classification performance was compared between SVM and RVM model on the y*east* dataset using Bi-gram probabilities feature extraction method. The LIBSVM tool[25] was used to carry out classification in SVM. For the classifier, SVM is used with a radial basis function (RBF).The RBF kernel parameters are $c = 0.5$ and $g = 0.6$ optimized by using a grid search method.

The obtained prediction results of the RVM are compared with that of the SVM and shown in Table V on *yeast* dataset. At the same time, the ROC curves are compared between RVM and SVM and displayed in Figure 2. It can be observed from Table V, the SVM classifier achieved 86.60% average Accuracy, 84.33% average sensitivity, 88.34% average precision, and 76.76% average Mcc. However, the RVM classifier of the proposed method achieved 94.57% average accuracy, 94.27% average sensitivity, 94.86%, average precision, and 89.74% average Mcc. It can be seen from these prediction results that the RVM classifier is significantly better than the SVM classifier. Similarly, it is shown in Figure 2 that the ROC curves of RVM classifier is also significantly better than that of SVM classifier. This clearly

Table VIII. *Fivefold Cross Validation Results Shown Using Our Proposed Method on* HPRD

Testing set	Ac (%)	Sn (%)	Pe (%)	Mcc (%)
1	98.92	99.12	98.78	97.87
2	98.50	98.91	98.14	97.04
3	98.70	98.88	98.48	97.43
4	99.00	99.45	98.56	98.02
5	98.82	99.29	98.35	97.68
Average	98.79 ± 0.002	99.13 ± 0.002	98.46 ± 0.019	97.61 ± 0.004

demonstrated that the RVM classifier employed the proposed method is an accurate and robust classifier. The better classification performance of RVM classifier may be attributed to two reasons: (1) The obvious advantage of RVM classifier is that the amount of calculation of the kernel function is greatly reduced; (2) The RVM classifier overcomes the disadvantage that the kernel function required to meet the condition of Mercer. For the sake of these reasons, the RVM classifier used our proposed method is obviously better than the SVM classifier. At the same time, it is proved that the proposed prediction model can obtain higher accuracy for detecting PPIs.

Comparison with other methods

To demonstrate the effectiveness of the propose method, some state-of-the-art methods for PPIs detection were selected to compare with the proposed method that uses a RVM model combined with PSSM, Bi-gram probabilities, and PCA on y*east and Helicobacter pylori* datasets. Experimental results of various methods on y*east* and *Helicobacter pylori* datasets were shown in Tables VI and VII. As we can see from Table VI that the average prediction accuracy of the proposed method is variously higher that of the other five methods on y*east* dataset. Similarly, the precision and sensitivity of our proposed method are also superior to those of the other five methods. At the same time, it can be seen from Table VII that the average prediction accuracy of the proposed method is also significantly higher than that of the five different methods on *Helicobacter pylori* dataset. From Tables VI and VII, it can be observed that the proposed method obtained obviously better prediction results compared to other existing methods. All experiment results proved that the RVM classifier combined with BiGP and the PSSM and PCA can improve the prediction accuracy relative to current state-of-the-art methods. The improvement of prediction accuracy of the proposed method lies in using a correct classifier and a novel feature extraction method.

Performance of the proposed method on HPRD *dataset*

To further illustrate the effectiveness of the propose method, we carry out the experiment on *HPRD* dataset. The *HPRD* dataset contains 36,480 positive protein pairs and 36,630 negative protein pairs. As a result, HPRD dataset contains 73,110 protein pairs. Experimental results on *HPRD* dataset were shown in Table (I and VIII). It can be seen from Table VIII that the proposed method achieved good prediction accuracy. The experiment results further demonstrated that the improvement of prediction accuracy of the proposed method lies in using a correct classifier and a novel feature extraction method.

Conclusion

In this article, we explore a novel computational method for predicting PPIs, called RVM-BiGP. It was constructed by combining an RVM classifier with Bi-gram probabilities and Position Specific Scoring Matrix. Experimental results on two widely used y*east and Helicobacter pylori* datasets showed that the prediction accuracy of the proposed method is significantly higher than that of the previous methods. Compared with some other state-of-the-art methods, the proposed method achieved the best performance. Furthermore, by carrying out the proposed approach on imbalance y*east* dataset, the prediction accuracy is better than that of balance y*east* dataset, which further proved that the proposed prediction model is very powerful for imbalanced data classification. The major improvements of the proposed method lie in employing an effective feature extraction method that can capture useful evolutionary information and significantly reduce the sparsity level which helps in improving the recognition performance. Moreover, PCA can integrate the useful information and reduce the influence of noise, which helps in improving the prediction accuracy. In addition, the experimental results demonstrated that the RVM classifier model is very suitable for PPIs detection. In conclusion, the proposed method is an efficient, reliable, and powerful prediction model and can be a useful tool for future proteomics research. For the future study, more effective feature extraction methods and machine learning techniques will be explored for prediction PPIs.

References

1. Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM

(2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574.
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415:180–183.
4. Snyder M, Zhu H, Bertone P, Bidlingmaier SM, Bilgin M, Casamayor AJ, Gerstein M, Jansen R, Lan N (2004) Global analysis of protein activities using proteome chips. In.: CN; 293:2101–2105.
5. Huang DS, Du JX (2009) A constructive hybrid structure optimization methodology for radia basis probabilistic neural networks. IEEE Trans Neural Netw 19: 2099–2115.
6. Huang DS, Zheng CH (2006) Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics 22:1855–1862.
7. Deng SP, Huang DS (2013) SFAPS: an R package for structure/function analysis of protein sequences based on informational spectrum method. IEEE Intl Conf Bioinform Biomed 69:207–212.
8. Zhu L, You ZH, Huang DS (2013) Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding. Neurocomputing 121:99–107.
9. Huang DS, Yu HJ (2013) Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids. IEEE/ACM Trans Comput Biol Bioinform 10:457–467.
10. Huang DS, Zhang L, Han K, Deng S, Yang K, Zhang H (2014) Prediction of protein–protein interactions based on protein–protein correlation using least squares regression. Curr Prot Pept Sci 15:553–560.
11. Keskin O, Tuncbag N, Gursoy A (2016) Predicting protein–protein interactions from the molecular to the proteome level. Chem Rev 116:4884–4909.
12. Shi MG, Xia JF, Li XL, Huang DS (2010) Predicting protein–protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids 38:891–899.
13. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) MINT: a molecular interaction database. FEBS Lett 513:135–140.
14. Zhu L, Deng SP, Huang DS (2015) A two-stage geometric method for pruning unreliable links in protein–protein networks. IEEE Trans Nanobiosci 14:528–534.
15. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N (2006) PIPE: a protein–protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC Bioinform 7:763–769.
16. Xia JF, Han K, Huang DS (2010) Sequence-based prediction of protein–protein interactions by means of rotation forest and autocorrelation descriptor. Prot Pept Lett 17:137–145.
17. Najafabadi HS, Salavati R (2008) Sequence-based prediction of protein–protein interactions by means of codon usage. Genome Biol 9:1–9.
18. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30:303–305.
19. Gribskov M, Mclachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84:4355–4358.
20. Georgiou DN, Karakasidis TE, Megaritis AC (2013) A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. Maternal Child Health Care China 7:41–48.
21. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2010) A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets. J Theoret Biol 267: 95–105.
22. Altschul SF, Koonin EV (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 23:444–447.
23. Sharma A, Lyons J, Dehzangi A, Paliwal KK (2013) A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Nanobiosci IEEE Trans 320:41–46.
24. Tipping ME (2001) Sparse bayesian learning and the relevance vector machine. J Machine Learn Res 1:211–244.
25. Chang CC, Lin CJ (2011) LIBSVM. A library for support vector machines. ACM Trans Intellig Sys Tech 2: 389–396.
26. Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36:3025–3030.
27. Zhou YZ, Gao Y, Zheng YY (2011) Prediction of protein–protein interactions using local description of amino acid sequence. Berlin, Heidelberg: Springer.
28. Lei Y, Jun-Feng X, Jie G (2010) Prediction of protein–protein interactions from protein sequence using local descriptors. Prot Pept Lett 17:1085–1090.
29. You ZH, Lei YK, Zhu L, Xia J, Wang B (2013) Prediction of protein–protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinform 14: 69–75.
30. Nanni L (2005) Fusion of classifiers for predicting protein–protein interactions. Neurocomputing 68:289–296.
31. Nanni L (2005) Letters: hyperplanes for predicting protein–protein interactions. Neurocomputing 69:257–263.
32. Nanni L (2006) An ensemble of K-local hyperplanes for predicting protein–protein interactions. Neurocomputing 22:1207–1210.