www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 14), pp: 23638-23649

Research Paper

Accurate prediction of protein-protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier

Zheng-Wei Li[1,*], Zhu-Hong You[2,*], Xing Chen[3], Li-Ping Li[2], De-Shuang Huang[4], Gui-Ying Yan[5], Ru Nie[1], Yu-An Huang[6]

[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[2]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China

[3]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

[4]School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

[5]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[6]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China

[*]These authors have contributed equally to this work and are joint First Authors

Correspondence to: *Zhu-Hong You,* ***email:*** *zhuhongyou@ms.xjb.ac.cn*
Xing Chen, ***email:*** *xingchen@amss.ac.cn*

Keywords: *disease, position-specific scoring matrix, Weber Local Descriptor, cancer, protein-protein interactions*

Received: *November 30, 2016* ***Accepted:*** *January 11, 2017* ***Published:*** *February 21, 2017*

ABSTRACT

Identification of protein-protein interactions (PPIs) is of critical importance for deciphering the underlying mechanisms of almost all biological processes of cell and providing great insight into the study of human disease. Although much effort has been devoted to identifying PPIs from various organisms, existing high-throughput biological techniques are time-consuming, expensive, and have high false positive and negative results. Thus it is highly urgent to develop *in silico* methods to predict PPIs efficiently and accurately in this post genomic era. In this article, we report a novel computational model combining our newly developed discriminative vector machine classifier (DVM) and an improved Weber local descriptor (IWLD) for the prediction of PPIs. Two components, differential excitation and orientation, are exploited to build evolutionary features for each protein sequence. The main characteristics of the proposed method lies in introducing an effective feature descriptor IWLD which can capture highly discriminative evolutionary information from position-specific scoring matrixes (PSSM) of protein data, and employing the powerful and robust DVM classifier. When applying the proposed method to *Yeast* and *H. pylori* data sets, we obtained excellent prediction accuracies as high as 96.52% and 91.80%, respectively, which are significantly better than the previous methods. Extensive experiments were then performed for predicting cross-species PPIs and the predictive results were also pretty promising. To further validate the performance of the proposed method, we compared it with the state-of-the-art support vector machine (SVM) classifier on *Human* data set. The experimental results obtained indicate that our method is highly effective for PPIs prediction and can be taken as a supplementary tool for future proteomics research.

INTRODUCTION

In this post-genomic era, protein-protein interactions (PPIs) can provide great insights into the intrinsic mechanisms of biological processes within a cell and so the PPI networks have been drawing increasing attention. Recently, a number of high-throughput biological techniques, such as yeast two hybrid screens [1], mass

spectrometric protein complex identification (MS-PCI) [2] and protein chips [3], have been proposed to identify interactions between proteins. Therefore, a large amount of PPI data from various kinds of organisms has been collected, and a number of databases, like DIP [4], BIND [5] and MINT [6], have also been constructed. However, such experimental methods for identifying PPIs are usually labor-intensive and time-consuming. The PPI pairs identified by these traditional techniques only account for a small part of the entire PPIs network [7, 8]. What's worse, those high-throughput techniques suffer from high rates of false positive and false negative results. All these limitations require robust and effective in silico methods as a complement to biological experimental techniques for protein-protein interactions prediction.

As a beneficial supplement to biological methods, a number of computational methods have been developed to predict protein interactions through different source of information, such as protein domains, phylogenetic profiles, gene co-expression and secondary structures [9–12]. However, such methods need specific domain knowledge which prevents their further applications. Evolutionary information embedded in proteins sequence has good capability for predicting PPIs [13]. Zahiri *et al*. [14] proposed a novel algorithm named PPIevo for detecting PPIs, which extracted the evolutionary feature from position-specific scoring matrixes (PSSM) of protein sequence. Hamp *et al.* [15] combined evolutionary profiles from protein sequence with profile-kernel support vector machines (SVM) to predict PPIs and obtained good results. An *et al.* [16] reported RVM-BiGP prediction model to predict PPIs from protein sequences and the results are very promising. Nevertheless, there is still room to improve the performance of the state-of-the-art prediction methods.

This paper is an extension of our previous work [17]. In this study, we report a novel computational model to predict PPIs using the evolutionary information of protein. The main improvements of the proposed method lie in introducing an effective feature extraction method, namely improved Weber local descriptor (IWLD) and using our newly developed discriminative vector machine (DVM) classifier. Specifically, given a protein sequence of length L, it would first be converted to an L-by-20 position-specific scoring matrix (PSSM). Then, an IWLD descriptor is used to extract discriminative evolutionary information from PSSM and a 256-dimensional histogram feature vector for each protein is constructed accordingly. Next, we combined two histogram vectors from corresponding protein pair into a 512-dimensional feature vector. Furthermore, the dimensionality reduction tool PCA (principal component analysis) is employed to extract the highly discriminatory information and reduce noise information. At last, the DVM classifier is used to carry out classification prediction. In this work, we first evaluated the proposed method on two PPIs data sets, *Yeast* and *H. pylori* and obtained good predictive accuracies of 96.52% and 91.80% respectively. Then, extensive experiments were performed to compare the proposed method with the state-of-the-art SVM classifier based on *Human* data set. Besides, comparisons between our method and other previous methods were also carried out. All the experimental results obtained indicate that the proposed method is impressively effective for PPIs prediction.

RESULTS AND DISCUSSION

Evaluation of predictive ability

To decrease data dependence and avoid over-fitting of prediction model, five-fold cross validation strategy was used in our study. Namely, the whole data set was evenly divided into five subsets, four of which were randomly chosen for training, and the rest for testing. To validate the validity of the proposed method, the random selection was repeated for five times, and five training sets and five validation sets were generated respectively. To be fair, parameters of DVM in different experiments were set to the same values. The predictive results of the proposed method on *Yeast* and *H. pylori* PPIs data sets are shown in Table 1 and Table 2.

It can be observed from Table 1 that when applied to *Yeast* data set, the average accuracy, sensitivity, precision and MCC of the proposed method are 96.52%, 94.86%, 98.11%, and 93.08%, respectively. Similarly, Table 2 shows the results on *H. pylori* data set, it can be observed that the average accuracy obtained using our method is 91.80%, with an average sensitivity of 92.15%, an average precision of 91.47%, and an average MCC of 83.60%. In addition, it can be noticed that the standard deviations of them are also relatively low. For *Yeast* data set, the average standard deviations of accuracy, sensitivity, precision and MCC are 0.46%, 0.59%, 0.48% and 0.92%, respectively. The average standard deviations of accuracy, sensitivity, precision and MCC on *H. pylori* data set are 0.85%, 1.54%, 0.91% and 1.69%, respectively. The ROC curves using five-fold cross-validation on *Yeast* and *H. pylori* data sets are illustrated in Figure 1 and Figure 2, respectively.

From Table 1 and Table 2, it can be drawn that the proposed predictive model combing DVM and IWLD descriptor is accurate and effective for the prediction of PPIs from the two data sets. In our predictive model, PSSM not only provides the order information of protein sequence but also retains sufficient evolutionary information. Next, by using differential excitation and orientation component, the IWLD descriptor has strong ability to maintain local highly discriminative information for PPIs prediction. Besides, the application of PCA reduces the dimensions of IWLD vector, decreases the impact of noise and accelerates the predictive process.

Table 1: Performance of the proposed method using five-fold cross validation on Yeast data set

Test set	Acc (%)	Sen (%)	Pre (%)	MCC (%)
1	95.89	94.06	97.82	91.85
2	96.16	94.41	97.63	92.35
3	96.87	95.23	98.65	93.80
4	96.92	95.14	98.60	93.89
5	96.74	95.44	97.85	93.50
Average	**96.52±0.46**	**94.86±0.59**	**98.11±0.48**	**93.08±0.92**

Table 2: Performance of the proposed method using five-fold cross validation on *H. Pylori* data set

Test set	Acc (%)	Sen (%)	Pre (%)	MCC (%)
1	92.62	93.25	92.95	85.18
2	91.08	89.96	91.27	82.13
3	92.11	93.15	91.28	84.24
4	90.74	91.07	90.44	81.48
5	92.47	93.33	91.41	84.95
Average	**91.80±0.85**	**92.15±1.54**	**91.47±0.91**	**83.60±1.69**

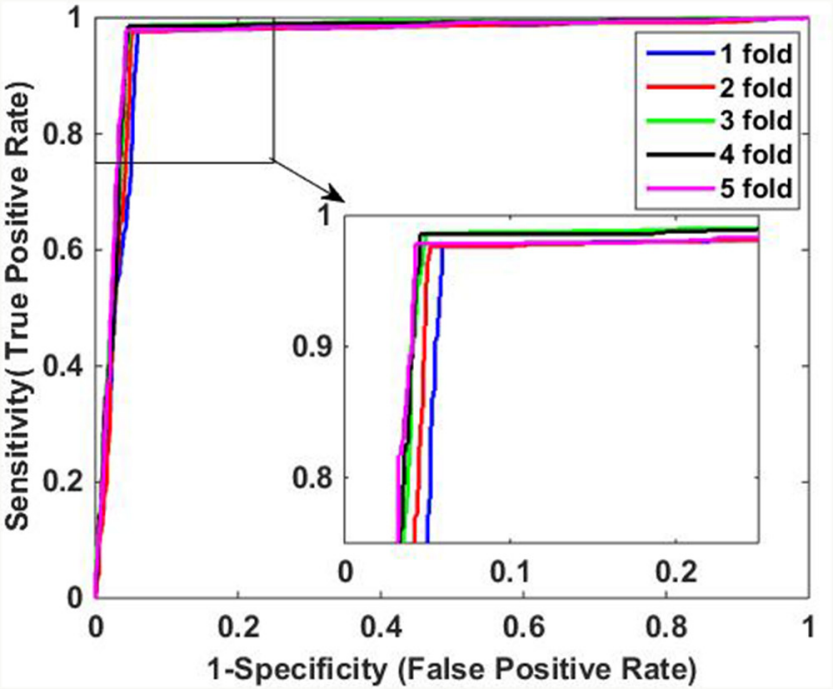

Figure 1: ROC curves of proposed method on *Yeast* data set.

Table 3: Five-fold cross validation results performed on Human data set

Model	Test set	Acc (%)	Sen (%)	Pre (%)	MCC (%)
DVM	**1**	97.18	95.61	98.40	94.37
	2	97.30	95.05	99.62	94.71
	3	96.38	94.73	97.43	92.75
	4	97.73	96.29	98.95	95.48
	5	97.92	96.83	98.65	95.82
	Average	**97.30±0.60**	**95.70±0.87**	**98.61±0.80**	**94.63±1.20**
SVM	**1**	89.89	90.83	88.21	79.79
	2	91.54	91.79	91.57	83.08
	3	89.40	90.78	86.99	78.82
	4	90.93	92.96	88.64	81.95
	5	91.24	91.68	89.66	82.44
	Average	**90.60±0.95**	**91.61±0.89**	**89.01±1.72**	**81.22±1.82**

Consequently, our proposed method is suitable for predicting PPIs from the two data sets.

Comparison with SVM classification model

Support vector machine (SVM) is one of the most widely used classification models for PPIs prediction. In this study, we used LIBSVM toolbox to carry out the prediction of PPIs (available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/). To further verify the performance of the proposed method, we applied SVM to predict PPIs of *Human* data set and compared its performance with DVM. To be fair, the two predictive models adopted same feature extraction method. Here, Gaussian function was chosen by SVM as the kernel function. A general grid search method was employed to optimize SVM's two parameters (kernel width parameter γ, regularization parameter C) and they were tuned to γ=0.01 and C=0.6 respectively.

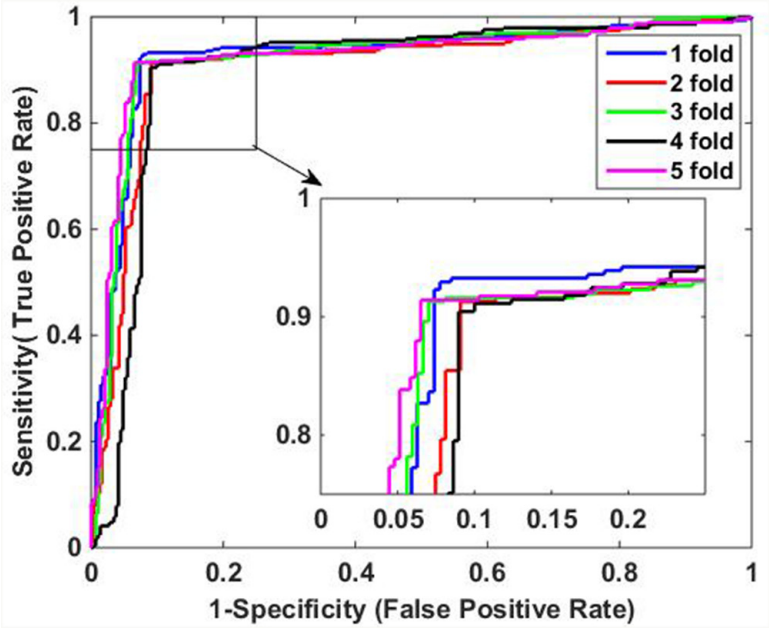

Figure 2: ROC curves of proposed method on *H. Pylori* data set.

The predictive results of the two methods are illustrated in Table 3. When using DVM classifier to identify the PPIs on *Human* data set, we got promising results with average accuracy, sensitivity, precision and MCC of 97.30%, 95.70%, 98.61% and 94.63%, respectively. Meanwhile, SVM-based method had relatively poor performance with lower average accuracy, sensitivity, precision and MCC of 90.60%, 91.61%, 89.01% and 81.22%, which indicate that DVM has better performance than SVM for predicting PPIs. In addition, it can be observed that DVM is more stable than SVM because the former has lower standard deviations of evaluation criteria than the latter. Specifically, DVM-based method yielded standard deviations of accuracy,

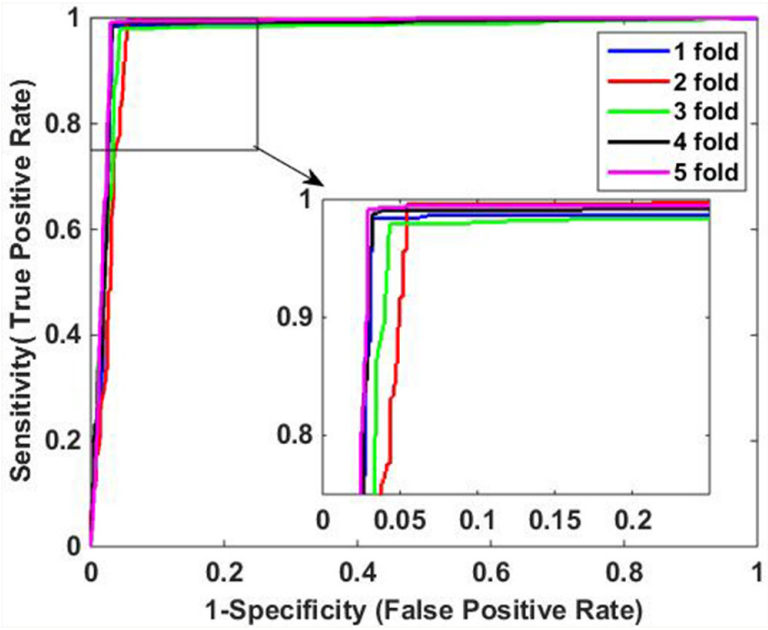

Figure 3: ROC curves of proposed DVM-based method on Human data set.

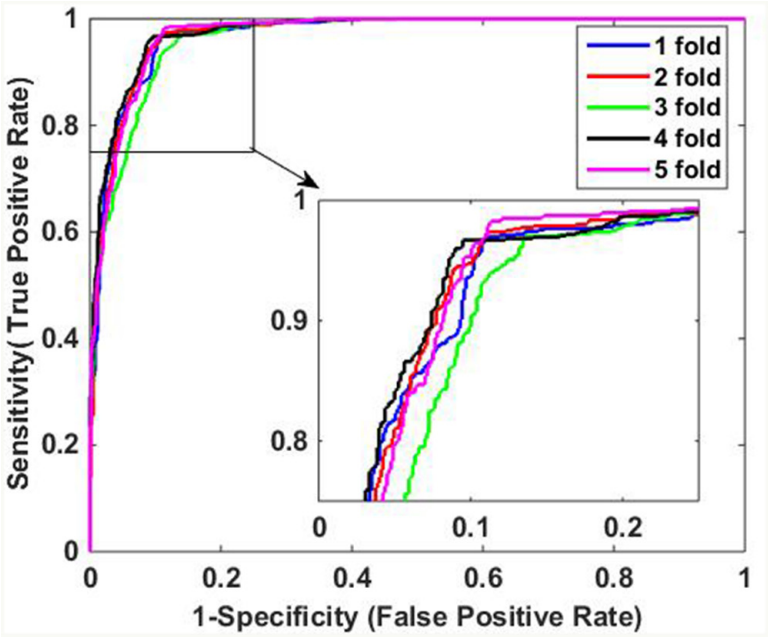

Figure 4: ROC curves of SVM-based method on Human data set.

Table 4: Predictive results of proposed method on five other species

Species	Test pairs	Accuracy
E. coli	6954	76.23%
C.elegans	4013	92.72%
H.sapien	1406	89.40%
H. pylori	1420	86.37%
M.musculus	312	87.69%

Table 5: Predictive results of different methods on Yeast data set

Model	Test set	Acc (%)	Sen (%)	Pre (%)	MCC (%)
Guo [20]	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 [21]	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 [22]	EELM	87.00±0.29	86.15±0.43	87.59±0.32	77.36±0.44
Wong [23]	RF+PR-LPQ	93.92±0.36	91.10±0.31	96.45±0.45	88.56±0.63
Our method	DVM	**96.52±0.46**	**94.86±0.59**	**98.11±0.48**	**93.08±0.92**

Table 6: Predictive results of different methods on *H. Pylori* data set

Model	Acc (%)	Sen (%)	Pre (%)	MCC (%)
Nanni *et al.* [24]	83.00	86.00	85.10	N/A
Nanni *et al.* [25]	84.00	86.00	84.00	N/A
Nanni *et al.* [26]	86.60	86.70	85.00	N/A
You *et al.* [22]	87.50	88.95	86.15	78.13
Martin *et al.* [27]	83.40	79.90	85.70	N/A
Wong *et al.* [23]	89.47	89.18	89.63	81.00
Our method	**91.80**	**92.15**	**91.47**	**83.60**

sensitivity, precision and MCC as low as 0.60%, 0.87%, 0.80% and 1.20%, which is less than the corresponding values of 0.95%, 0.89%, 1.72% and 1.82% of SVM-based method. Furthermore, Figure 3 and Figure 4 show the ROC curves performed by DVM and SVM, respectively. It can be observed that DVM yielded higher average AUC (area under an ROC curve) value than that of SVM classifier.

By analyzing the experimental results, we can conclude that DVM is more effective and robust than SVM in predicting PPIs. There are two possible explanations for the results. (1) Based on k nearest neighbors (kNNs), the robust M-estimator and manifold regularization, DVM decreases the influence of outliers and overcomes the shortcoming of the kernel function required to satisfy the Mercer condition. (2) Although there are three parameters (β, γ, and θ) to be tuned in DVM, those parameters slightly affect the performance of DVM if they are adjusted in suitable ranges. Therefore, DVM is more suitable for predicting PPIs than SVM.

Performance on independent data set

Although our proposed method had achieved good performance for PPIs prediction on *Yeast*, *H. pylori* and *Human* data sets, we still carried out extensive analyses to verify its ability for predicting PPIs from other species (*E. coli, C. elegans*, *H. sapien*, *H. pylori* and *M. musculus*). In the following experiments, we used 11188 samples of *Yeast* data set for training and samples from other five species for testing. The corresponding feature extraction method is same to the previous experiments. The predictive results are listed in Table 4. The basis of this hypothesis is that homologs tend to be similar functional behavior and so they preserve the same PPI [18]. When applying the proposed method to the prediction of PPIs from these five species, the average accuracies of them vary from 76.23 to 92.72. On the one hand, these promising results obtained indicate that *Yeast* protein may have a similar interacting mechanism with other five species and its sequence data is sufficient for the prediction of PPIs from other species; on the other hand, it demonstrates the proposed method has good generalization ability. In addition, the prediction results fully demonstrate that it is possible that PPIs in one species can be employed to identify PPIs in other species.

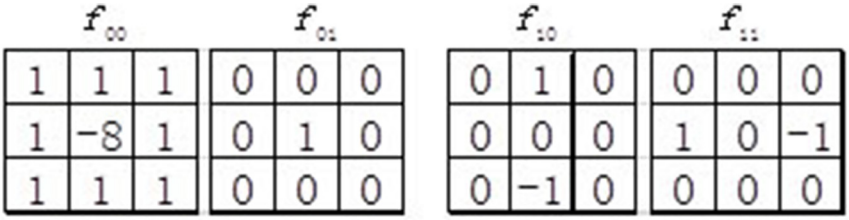

Figure 5: Four filters used in the original WLD.

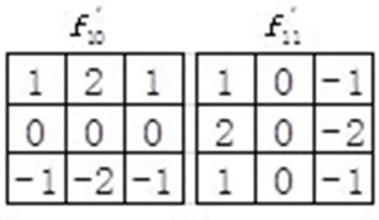

Figure 6: Sobel operators used in the improved WLD (IWLD).

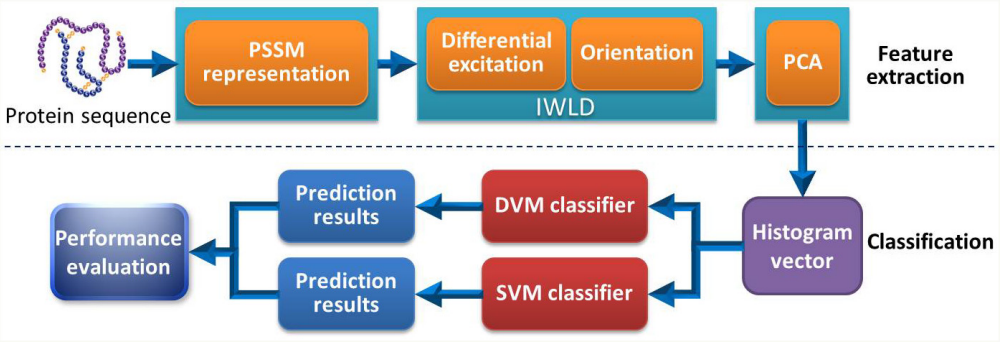

Figure 7: Flow chart of our proposed method for the prediction of PPIs.

Comparison with other methods

So far, a variety of machine-learning based computational methods have been proposed for PPIs prediction. To further validate the effectiveness of our method, we also compared our DVM-based predictive model using IWLD descriptor with several other previous methods (see Table 5 and Table 6) on *Yeast* and *H. pylori* data sets. In Table 5, the prediction accuracy of other previous methods on *Yeast* data set varies from 75.08% to 93.92%, while our proposed method achieves higher value of 96.52%. Similarly, for sensitivity and precision, our predictive model yields better performance than the others. Moreover, the corresponding standard deviations indicate the proposed method is stable and robust. Considering ensemble classifier usually has better performance than single classifier, although RF + PR-LPQ method has lower standard deviations, our method can also be viewed as one of the most competitive computational methods for predicting PPIs.

The similar results of different methods on *H. pylori* data set can also be found in Table 6. The accuracies of other methods vary from 83.00% to 89.47% while our proposed method attains relatively higher value of 91.80%. The same is true for precision, sensitivity and MCC. The predictive results in Table 5 and Table 6 indicate that the DVM-based classifier incorporating IWLD descriptor can improve the performance of PPIs compared with the state-of-the-art methods. The promising prediction results of our method may contribute to the novel feature extraction method which can provide highly discriminative information, and the selection of DVM classifier which has been demonstrated to be robust and powerful [19].

CONCLUSIONS

In this work, we put forward a novel evolutionary information based computational model for predicting PPIs, which combines our newly developed discriminative vector machine classifier (DVM) and an improved Weber local descriptor (IWLD) to capture highly discriminative information. To minimize data dependence and avoid the over-fitting, five-fold cross-validation was adopted accordingly. When applied to *Yeast* and *H.Pylori* data sets, the model achieves promising prediction accuracies of 96.52% and 91.80%, respectively. Additionally, to evaluate the generalization capability of the proposed method, extensive experiments are performed to predict the PPIs on five other species data sets. Besides, it is compared with SVM-based model and other previous works. The achieved results show that the proposed method is very competitive for predicting PPIs and can be taken as a useful supplementary tool to the traditional experimental methods for future proteomics research.

MATERIALS AND METHODS

Golden standard data sets

In this study, we verified the proposed method on a high-confidence PPIs data set *Yeast*, gathered from the publicly available database of interaction proteins (DIP), version DIP_20070219 [4]. All protein pairs were aligned by a multiple sequence alignment tool, CD-HIT [28]. To reduce fragments and similarity, those protein pairs with ≤50 residues or ≥40% sequence identity were all removed. Then the remaining 5594 interacting protein pairs form the positive data set and 5594 additional protein pairs from different subcellular localizations were chosen to construct the negative data set. Therefore, the data set of *Yeast* finally contains 11188 protein pairs of which half are positive samples and half negative samples.

To further test the generality of the proposed method, we also evaluate it on two other PPIs data sets: *Human* and *H. pylori*. The first data set *Human* comes from the human protein references database (HPRD). By using the aforementioned steps, we selected 3899 protein pairs as the positive data set and 4262 additional protein pairs from different subcellular localizations as negative data set. As a result, the *Human* data set finally consists of 8161 protein pairs. Similarly, the second data set *H. pylori* consists of 2916 protein pairs, of which half are interacting pairs and half non-interacting pairs, as described by Martin *et al*.

Improved Weber local descriptor

Inspired by Weber's Law, Chen *et al.* [29] proposed the original Weber local descriptor (WLD) for image recognition, which contains two components, namely differential excitation and orientation. Differential excitation component $\xi(x_i)$ of WLD is the ratio between two terms: One is the relative intensity differences of an interest point x_i against its neighbors; the other is the intensity of x_i itself. We first calculate the intensity differences between x_i and its neighbors with the filter f_{00} (see Figure 5):

$$v_i^{00} = \sum_{j=0}^{p-1}(x_j - x_i) \quad (1)$$

where $x_j (j = 0,1,..,p-1)$ represents the jth neighbor of x_i and p is the number of its neighbors. We then calculate the ratio of the intensity differences v_i^{00} and v_i^{01}:

$$G_r(x_i) = \frac{v_i^{00}}{v_i^{01}} \quad (2)$$

where v_i^{01} is the output of the filter f_{01} (see Figure 5). As described before, v_i^{01} is just the original intensity of x_i. Next, the arctangent function is employed to construct the differential excitation $\xi(x_i)(\in[-\pi/2, \pi/2])$:

$$\xi(x_i) = \arctan\left(G_r(x_i)\right) = \arctan\left(\frac{v_i^{00}}{v_i^{01}}\right) \quad (3)$$

$$= \arctan\left(\sum_{j=0}^{p-1}\left(\frac{x_j - x_i}{x_i}\right)\right)$$

In addition, orientation component of WLD describes the gradient orientation of interest point. In the original WLD, only 4 neighbors of x_i are utilized which may lose some important discriminating information and are sensitive to noise. In our study, we adopted an improved WLD (IWLD) descriptor by introducing Sobel operators (see Figure 6). By taking into account all 8 neighbors of x_i, it can not only preserve sufficient orientation information but also effectively suppress the noise. Thus, the orientation component of IWLD $\gamma(x_i)$ is computed as:

$$\gamma(x_i) = \arctan 2\left(\frac{v_i^{11}}{v_i^{10}}\right) + \pi \quad (4)$$

where v_i^{10} and v_i^{11} denote the outputs of the filters $f_{10}^{'}$ and $f_{11}^{'}$ (see Figure 6).

To perform histogram statistics, the differential excitation $\xi(x_i)$ is quantized into M intervals $l_m (l_m = [\eta_m^l, \eta_m^u], m = 0,2,\ldots,M-1)$, where $\eta_m^l = \left(\frac{m}{M} - 1/2\right)\pi$ is the lower bound and $\eta_m^u = \left(\frac{m+1}{M} - 1/2\right)\pi$ is the upper bound. So, the value of m is calculated as follow:

$$m = mod\left(\left[\frac{\xi(x_i) + \frac{\pi}{2}}{\frac{\pi}{M}}\right], M\right) \quad (5)$$

Similarly, $\gamma(x_i)(\in(0,2\pi))$ is also quantized into T dominant orientations as follow:

$$\Phi_t = f_q(\gamma) = \frac{2t}{T}\pi, \text{ and } t = mod\left(\left[\frac{\gamma(x_i)}{2\pi/T} + \frac{1}{2}\right], T\right) \quad (6)$$

By calculating m, t value of each point in an image, a 1D histogram vector $S = \{s_{m,t}\}$ $(m = 0,1,\ldots,M-1, t = 0,1,\ldots,T-1)$ can be obtained accordingly. To fully mine the local discriminative information, we first divide the image into $V \times H$ sub blocks. Here, V represents the number of sub blocks in vertical direction and H represents the number of sub blocks in horizontal direction, and the histogram vector of each block is obtained accordingly. Then all the histogram vectors of the image are concatenated into the final one-dimensional IWLD feature vector.

In this work, there are four free parameters (M, T, V, H) to be tuned. Through grid search on *Yeast* and *H. pylori* data sets, we chose M=8, T=8, V=H=2 in our experiments and each protein sequence sample is transformed into a 256 dimensional IWLD vector. Next, every two IWLD vectors from corresponding protein pairs are concatenated into a 512 dimensional vector. Then, the dimensionality reduction algorithm PCA is employed to reduce the impact of noises and accelerate the predictive process, and the final 200 dimensional reduced vector is constructed for the subsequent classification.

Discriminative vector machine

Classification is a fundamental issue in pattern recognition field and there exist numerous classification algorithms for different recognition tasks. In this work, our newly developed discriminative vector machine (DVM) classifier [19] was adopted in classification. DVM is a probably approximately correct (PAC) learning classifier which can reduce the error caused by generalization and is very robust. For a given test sample y, the first step of DVM is to find its k nearest neighbors (kNNs) to suppress the effect of outliers. The kNNs of y can be expressed as $X_k = [x_1, x_2, \ldots, x_k]$, where x_i denotes the ith nearest neighbor. Equally, X_k can also be represented as $X_k = [x_{k,1}, x_{k,2}, \ldots, x_{k,c}]$, where $x_{k,j}$ comes from the jth class. So the objective of DVM is to solve the following minimization problem:

$$\min_{\beta_k} \delta \| \beta \|_k + \sum_{i=1}^{d} \varnothing\left(\left(y - X_k\beta_k\right)_i\right) + \gamma \sum_{p=1}^{k}\sum_{q=1}^{k} w_{pq}(\beta_k^p - \beta_k^q)^2 \quad (7)$$

where β_k can be denoted as $\left[\beta_k^1, \beta_k^2, \ldots, \beta_k^k\right]$ or $[\beta_{k,1}, \beta_{k,2}, \ldots, \beta_{k,c}]$, where $\beta_{k,i}$ is the coefficient from the ith class, $\| \beta_k \|$ is a norm of β_k and the corresponding L_2 norm is employed in our calculation, $\left(y - X_k\beta_k\right)_i$ is the ith element of $y - X_k\beta_k$ and $\varnothing$ is a robust M-estimator to improve the robustness of DVM. M-estimator is a generalized maximum likelihood operator proposed by Huber to estimate parameters under the cost function [30]. In this work, a robust Welsch M-estimator ($\varnothing(x) = (1/2)(1 - exp(-x^2))$) is adopted to attenuate error so that outliers would have a less impact on classification. The last section of Eq. (7) is a manifold regularization where w_{pq} is the similarity between the pth and the qth nearest neighbors of y. In this work, w_{pq} is defined as the cosine distance between the pth and the qth NN of y. Then the corresponding Laplacian matrix L can be expressed as

$$L = D - W \quad (8)$$

where W is the similarity matrix whose element is $w_{pq}(p = 1,2,\ldots,k; q = 1,2,\ldots,k)$, D is a diagonal matrix

whose ith element d_i is the sum of $w_{iq}(q=1,2,\ldots,k)$. According to Eq. (8), the last section of Eq. (7) can be rewritten as $\gamma\beta_k^T L\beta_k$. Furthermore, a diagonal matrix $P=\mathrm{diag}(p_i)$ is constructed and its element $p_i\left(i=1,2,\ldots,d\right)$ is denoted as:

$$p_i = e^{\frac{\left((y-X_k\beta_k)\right)_i^2}{\sigma^2}} \tag{9}$$

where σ is the kernel size which can be calculated in the following form:

$$\sigma = \sqrt{(\theta * \left(y - X_k\beta_k\right)^T * (y - X_k\beta_k)/d} \tag{10}$$

where d is the dimension of y and θ is a constant to curb outliers. In this work, it is assigned to 1.0 as in the literature [31]. By merging Eq. (8), (9) and (10), the minimization of Eq. (7) can be converted to the following problem:

$$\arg \min_{\beta_k} \left(y - X_k\beta_k\right)^T P\left(y - X_k\beta_k\right) + \delta \| \beta_k \|_2^2 + \gamma\beta_k^T L\beta_k \tag{11}$$

According to the theory of half-quadratic minimization, the global solution β_k can be described as:

$$\beta_k = (X_k^T P X_k + \delta I + \gamma L)^{-1} X_k^T P y \tag{12}$$

After the related coefficients are calculated, the test sample y can be identified as the ith class if the residual $y - X_{ki}\beta_{ki}$ is the minimum value.

$$R_i = \min_i \| y - X_{ki}\beta_{ki} \|, \quad i=1,2,\ldots,c \tag{13}$$

By means of robust M-estimator and manifold regularization to suppress the effect of outliers and strengthen its discriminatory ability, DVM classifier has better robustness and higher generalization ability than kNNs. In this work, there are two classes in total to be identified: non-interacting protein pair (class 1) and interacting pair (class 1). If the residual R_1 is the minimum distance, the test sample y would be classified as non-interacting protein pair, or it would be identified as interacting protein pair. For three free parameters (δ, γ, θ) of DVM model, it is time-consuming to directly search for their optimal values. It is gratifying that DVM algorithm is so stable that all these parameters only affect the performance slightly if they are set in feasible ranges. Based on above knowledge and through grid search, the parameters δ and γ are set as 1E-3 and 1E-4 respectively. Just as described before, θ is a constant and is always set to 1 throughout the entire process. For large data set, DVM classifier needs to spend relatively more time in finding the representative vector, so multi-dimensional indexing techniques can be adopted to speed up the search process to a certain extent.

Procedure of proposed model

The procedure of our proposed model mainly contains two steps: feature extraction and classification. The feature extraction is also divided into three steps: (1) the PSI-BLAST tool is used to represent each protein sequence and PSSM is obtained accordingly; (2) The PSSM from each protein is transformed into the corresponding histogram vector via IWLD descriptor; (3) Dimensional reduction of the histogram vector is performed by PCA algorithm. In the same way, sample classification also consists of two steps. (1) Based on the data sets of *Yeast*, *H. pylori* and *Human*, DVM model is trained and used to carry out classification; (2) The trained DVM model is then employed to predict the PPIs and its performance is evaluated accordingly. Furthermore, SVM model is also constructed for predicting PPIs on *Human* data set and the corresponding evaluation is also performed. The overall flow chart of our method is shown in Figure 7.

Evaluation criteria

To evaluate the performance of related predictive methods, four criteria, including the accuracy (*Acc*), sensitivity (*Sen*), precision (*Pre*), and Matthews's correlation coefficient (*MCC*), were introduced, which can be calculated as follows:

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$Pre = \frac{TP}{TP+FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \tag{4}$$

where *TP* (true positive) represents the number of interacting protein pairs predicted correctly while *FP* (false positive) denotes the number of non-interacting protein pairs predicted falsely. Similarly, *TN* (true negative) stands for the number of non-interacting protein pairs predicted correctly, and *FN* (false negative) denotes the number of interacting protein pairs predicted falsely. Receiver-operating characteristics (ROC) curve is a standard technique for summarizing classifier performance over a range of trade-offs between TP and FP error rates. In our study, ROC curves were also calculated to evaluate the validity of prediction models.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation of China, under Grants 61373086, 11301517, 61572506, 11301517 and 11631014, in part by Guangdong Natural Science Foundation, under Grant

2014A030313555, and in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences. The authors would like to thank all anonymous reviewers for their constructive advices.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

Author contributions

ZL, ZY conceived the algorithm, prepared the data sets, carried out experiments, carried out the analyses, and wrote the manuscript. XC, LL, DH, GY, YH and RN designed, performed and analyzed experiments and wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A. 2001; 98:4569-74. doi: 10.1073/pnas.061034498.
2. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 2002; 415:180-3. doi: 10.1038/415180a.
3. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A. Global analysis of protein activities using proteome chips. Biophysical Journal. 2001; 293:2101-5. doi: 10.1126/science.1062191.
4. Xenarios I, Salwínski L, Duan X, Higney P, Kim S. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research. 2002; 30:303-5. doi: 10.1093/nar/30.1.303.
5. Bader GD, Betel D, Hogue CWV. BIND: the biomolecular interaction network database. Nucleic acids research. 2003; 31:248-50. doi: 10.1093/nar/gkg056.
6. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a molecular INTeraction database. FEBS Letters. 2002; 513:135-40. doi: 10.1016/S0014-5793(01)03293-8.
7. Kotlyar M, Pastrello C, Pivetta F, Sardo AL, Cumbaa C, Li H, Naranian T, Niu Y, Ding Z, Vafaee F. In silico prediction of physical protein interactions and characterization of interactome orphans. Nature Methods. 2015; 12:308-10. doi: 10.1038/nmeth.3178.
8. Morris JH, Knudsen GM, Verschueren E, Johnson JR, Cimermancic P, Greninger AL, Pico AR. Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. Nat Protoc. 2014; 9:2539-54. doi: 10.1038/nprot.2014.164.
9. Li S, You Z, Guo H, Luo X, Zhao Z. Inverse-free Extreme Learning Machine with Optimal Information Updating. IEEE Transactions on Cybernetics. 2016; 46:1229-41. doi: 10.1109/TCYB.2015.2434841.
10. You ZH, Yin Z, Han K, Huang DS, Zhou X. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. BMC Bioinformatics. 2010; 11:343. doi: 10.1186/1471-2105-11-343.
11. Lan X, Bonneville R, Apostolos J, Wu W, Jin VX. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Bioinformatics. 2011; 27:428-30. doi: 10.1093/bioinformatics/btq669.
12. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002; 18:S233-S40. doi: 10.1093/bioinformatics/18.suppl_1.S233.
13. Yu J, Guo M, Needham CJ, Huang Y, Cai L, R D, Westhead. Simple sequence-based kernels do not predict protein-protein interactions. Bioinformatics. 2010; 26:2610-4. doi: 10.1093/bioinformatics/btq483.
14. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPIevo: protein-protein interaction prediction from PSSM based evolutionary information. Genomics. 2013; 102:237-42. doi: 10.1016/j.ygeno.2013.05.006.
15. Hamp T, Rost B. Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics. 2015; 31:1945-50. doi: 10.1093/bioinformatics/btv077.
16. An JY, Meng FR, You ZH, Chen X, Yan GY, Hu JP. Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. Protein Sci. 2016; 25:1825-33. doi: 10.1002/pro.2991.
17. Li ZW, You ZH, Chen X, Gui J, Nie R. Highly Accurate Prediction of Protein-Protein Interactions via Incorporating Evolutionary Information and Physicochemical Characteristics. International Journal of Molecular Sciences. 2016; 17. doi: 10.3390/ijms17091396.
18. You ZH, Chan KC, Hu P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS One. 2015; 10:e0125811. doi: 10.1371/journal.pone.0125811.
19. Gui J, Liu T, Tao D, Sun Z, Tan T. Representative Vector Machines: A unified framework for classical classifiers. IEEE Transactions on Cybernetics 2015; 46:1877 - 88. doi: 10.1109/TCYB.2015.2457234.
20. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein

interactions from protein sequences. Nucleic Acids Res. 2008; 36:3025-30. doi: 10.1093/nar/gkn159.

21. Yang L, Xia J, Gui J. Prediction of protein-protein interactions from protein sequence using local descriptors. Protein and Peptide Letters. 2010; 17:1085-90. doi: 10.2174/092986610791760306.

22. You Z, Lei Y, Zhu L, Xia J, Wang B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinformatics. 2013; 14:69-75. doi: 10.1186/1471-2105-14-S8-S10.

23. Wong L, You Z, Ming Z, Li J, Chen X, Huang Y. Detection of interactions between proteins through rotation forest and local phase quantization descriptors. Int J Mol Sci. 2016; 17. doi: 10.3390/ijms17010021.

24. Nanni L. Fusion of classifiers for predicting protein–protein interactions. Neurocomputing. 2005; 68:289-96. doi: 10.1016/j.neucom.2005.03.004.

25. Nanni L. Hyperplanes for predicting protein–protein interactions. Neurocomputing. 2005; 69:257-63. doi: 10.1016/j.neucom.2005.05.007.

26. Nanni L, Lumini A. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics. 2006; 22:1207-10. doi: 10.1093/bioinformatics/btl055.

27. Martin S, Roe D, Faulon JL. Predicting protein-protein interactions using signature products. Bioinformatics. 2005; 21:218-26. doi: 10.1093/bioinformatics/bth483.

28. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 2001; 17:282-3. doi: 10.1093/bioinformatics/17.3.282.

29. Chen J, Shan S, He C, Zhao G, ¨inen MP, Chen X, Gao W. WLD: a robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010; 32:1705-20. doi: 10.1109/TPAMI.2009.155.

30. Liu W, Pokharel PP, Principe JC. Correntropy: Properties and Applications in Non-Gaussian Signal Processing. IEEE Transactions on Signal Processing. 2007; 55:5286-98. doi: 10.1109/TSP.2007.896065.

31. He R, Zheng W, Hu B. Maximum correntropy criterion for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011; 33:1561-76. doi: 10.1109/TPAMI.2010.220.