Journal of Theoretical Biology 418 (2017) 105–110

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier

Lei Wang[a,b,1], Zhu-Hong You[c,*,1], Shi-Xiong Xia[a,**], Feng Liu[d], Xing Chen[e], Xin Yan[f], Yong Zhou[a]

[a] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
[b] College of Information Science and Engineering, Zaozhuang University, Zaozhuang, Shandong 277100, China
[c] The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China
[d] China National Coal Association, Beijing 100713, China
[e] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
[f] School of Foreign Languages, Zaozhuang University, Zaozhuang, Shandong 277100, China

ARTICLE INFO

Keywords:
Position-specific scoring matrix
Multiple sequences alignments
Rotation forest
Cancer

ABSTRACT

Protein-Protein Interactions (PPIs) are essential to most biological processes and play a critical role in most cellular functions. With the development of high-throughput biological techniques and *in silico* methods, a large number of PPI data have been generated for various organisms, but many problems remain unsolved. These factors promoted the development of the *in silico* methods based on machine learning to predict PPIs. In this study, we propose a novel method by combining ensemble Rotation Forest (RF) classifier and Discrete Cosine Transform (DCT) algorithm to predict the interactions among proteins. Specifically, the protein amino acids sequence is transformed into Position-Specific Scoring Matrix (PSSM) containing biological evolution information, and then the feature vector is extracted to present protein evolutionary information using DCT algorithm; finally, the ensemble rotation forest model is used to predict whether a given protein pair is interacting or not. When performed on *Yeast* and *H. pylori* data sets, the proposed method achieved excellent results with an average accuracy of 98.54% and 88.27%. In addition, we achieved good prediction accuracy of 98.08%, 92.75%, 98.87% and 98.72% on independent data sets (*C.elegans, E.coli, H.sapiens* and *M.musculus*). In order to further evaluate the performance of our method, we compare it with the state-of-the-art Support Vector Machine (SVM) classifier and get good results. As a web server, the source code and *Yeast* data sets used in this article are freely available at http://202.119.201.126:8888/DCTRF/.

1. Introduction

Proteins are involved in various biological processes, and carry out nearly all cellular functions by interacting with other proteins or DNA to function properly. Knowledge of Protein-Protein Interactions (PPIs) can provide valuable insights into the underlying mechanisms of biological processes and gene functions within a living cell. Hence, detection of PPIs has become an important topic in systems biology and functional genomics. In recent years, various experimental approaches have opened new prospects to detect PPIs. Thus, small-scale biochemical and biophysical experimental techniques, including yeast two-hybrid systems (Ito et al., 2001; Uetz et al., 2000), mass spectrometry (Gavin et al., 2002), and protein chips technology (Zhu et al., 2001) have been widely used to identify interactions between proteins. However, there are some disadvantages of existing experimental methods. For example, they have high false positive rates, high cost and are time-consuming. In addition, current PPI pairs obtained from biological experiments only cover a small fraction of the complete PPI networks (Han et al., 2005). Therefore, computational analysis of PPIs is considered one of the most important and efficient auxiliary methods for inferring PPIs from various kinds of biological datasets.

In fact, much work has been done for predicting PPIs from various data types, including the coevolution method of interacting proteins (Pazos et al., 1997), the phylogenetic profile method (Pazos and Valencia, 2001), the literature mining method (Marcotte et al., 2001), and the fusion/Rosetta Stone method (Marcotte et al., 1999). These

* Corresponding author.
** Corresponding author.
E-mail addresses: zhuhongyou@ms.xjb.ac.cn (Z.-H. You), xiasx@cumt.edu.cn (S.-X. Xia).
[1] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

http://dx.doi.org/10.1016/j.jtbi.2017.01.003
Received 28 May 2016; Received in revised form 24 September 2016; Accepted 4 January 2017
Available online 11 January 2017
0022-5193/

computational methods can be roughly divided into those that are based on function annotation (Ben-Hur and Noble, 2005; Saha et al., 2014; Souiai et al., 2014; Yang and Tang, 2014), structure (Aloy and Russell, 2003; Bock and Gough, 2001; Planas-Iglesias et al., 2013; Priya et al., 2013), and sequence (Ben-Hur and Noble, 2005; Guo et al., 2008; Liu et al., 2013b; Shen et al., 2007; Xia et al., 2010; Yang et al., 2010; You et al., 2013; Zhang et al., 2011). Among them, protein sequence-based methods have the advantage of being implementable even if information about the proteins is not available. Hence, it is a broadly applicable approach, as it only needs information on the protein sequence to distinguish absence and presence of interaction among proteins. For example, Guo et al. (2008) predicted PPIs using support vector machine combined with auto-covariance features extracted from the protein sequences, which led to very accurate prediction. Shen et al. (2007) developed an alternative method that considered the local environments of residues through a conjoint triad method. When applied to predicting *human* PPIs, this method yielded a high prediction accuracy of 83.9%. Martin et al., (2005) proposed a method to predict PPIs using a novel descriptor called signature product, which was an expansion of the signature descriptor and was derived from the chemical information of subsequences. Singhal and Resat (2007) tried to solve this problem by using a method based on a quantitative score measuring domain-domain interactions derived from an available PPI database; and then used the obtained score to predict the probability of the interaction between two proteins. This method has been proven successful for predicting PPIs, but a variety of reasons make its widespread use limited, such as a requirement of a prior experimental knowledge about the query proteins, the high dimensionality of feature representation, etc.

In this study, we present a sequence-based approach for predicting PPIs with Position-Specific Scoring Matrix (PSSM) and Discrete Cosine Transform (DCT) via the Rotation Forest (RF) classifier (Gribskov et al., 1987; Nanni and Lumini, 2009; Rodriguez and Kuncheva, 2006). For a protein's amino acids sequence, the PSSM gives the log-odds score of a particular residue substitution at a specific position based on evolutionary information (Chen and Jeong, 2009; Jones, 1999; Jones and Ward, 2003; Wass and Sternberg, 2008). DCT has the advantages of packing the most information into the fewest coefficients and minimizing reconstruction errors, and only a small amount of information on the details is lost during the processing procedures. Finally, the ensemble RF model is built using the PSSM-derived features as input. RF is a newly proposed multiple classifier system, which is more robust because it can enhance the diversity in the ensemble and the accuracy for an individual classifier at the same time (Ratsch et al., 2002; Braga et al., 2007; Lee and Yang, 2006; Breiman, 2001; Cutler et al., 2007). The proposed method was evaluated using PPIs data for *Yeast* and *H. pylori*, and yielded a high accuracy of 98.54% and 88.27%, respectively. In addition, the proposed method was further tested using an independent PPIs dataset of four other organisms, *C.elegans, E.coli, H.sapiens,* and *M.musculus*, with a high prediction accuracy of 98.08%, 92.75%, 98.87%, and 98.72%, respectively. The experimental results demonstrated that our proposed method helped improving the prediction.

2. Materials and methodology

2.1. Data sources

For a fair comparison of the proposed method, we employed the benchmark data sets that had been used in previous studies. The first data set was the *S.cerevisiae* PPIs data set by Guo *et al.*, which was gathered from the publicly available Database of Interacting Proteins (DIP) (Xenarios et al., 2002; Guo et al., 2008). The reliability of this core subset had been tested by two methods, paralogous verification method (PVM) and expression profile reliability (EPR) (Deane et al., 2002). In our experiment, the core subset that contained 5966 interaction pairs was employed. The protein pairs which contained a protein with fewer than 50 residues were removed because of the possibility that they might be fragments, and the remaining 5943 protein pairs constituted the positive data set. The CD-Hit (Li and Godzik, 2006; Li et al., 2001) algorithm was further used with $< 40\%$ identity to reduce pair wise sequence redundancy. By doing this, the remaining 5594 protein pairs made the final positive data set. For constructing the negative dataset, we selected 5594 additional protein pairs of different subcellular localizations. The final data set consisted of 11188 protein pairs, where half of the protein pairs were from the positive data set and half were from the negative data set. The second data set was the *H. pylori* dataset (available at http://www.cs.sandia.gov/~smartin/software.html), which consisted of 2916 protein pairs (1458 interacting pair and 1458 non-interacting pairs) as described by Martin et al. (2005).

2.2. Position specific scoring matrix (PSSM)

The PSSM was generated from a group of sequences previously aligned by structural or sequence similarity (Gribskov et al., 1987). It has been successfully applied to protein binding site prediction, protein secondary structure prediction, prediction of disordered regions, and protein function prediction. The PSSM was first introduced by Gribskov et al. (1987) for detecting distantly related proteins. A PSSM for a query protein is an $N\times 20$ matrix $P=\{\lambda_{i,j}: i = 1\cdots N \text{ and } j = 1\cdots 20\}$, which contains sequence evolution information and is defined as Eq. (1).

$$P=\begin{bmatrix} \lambda_{1,1} & \lambda_{1,2}\cdots & \lambda_{1,20} \\ \lambda_{2,1} & \lambda_{2,2}\cdots & \lambda_{2,20} \\ \vdots & \vdots\ \vdots & \vdots \\ \lambda_{N,1} & \lambda_{N,2}\cdots & \lambda_{N,20} \end{bmatrix} \tag{1}$$

where N is the length of the protein sequence and $\lambda_{i,j}$ in the i row of PSSM is the probability of the ith residue being mutated into type j of 20 native amino acids during the evolutionary process of in the protein from multiple sequence alignments.

In our experiment, we used Position-Specific Iterated BLAST (PSI-BLAST) (Altschul and Koonin, 1998; Altschul et al., 1997) to create PSSM for each protein sequence. PSI-BLAST is currently the most used application that compares PSSM profiles to detect remotely related homologous proteins or DNA. The default parameters in PSI-BLAST were used except for the e-value and number of iterations. To obtain broad and high homologous sequences, an e-value of 0.001 and three iterations were selected for PSI-BLAST to search against the *SwissProt* database. Applications of PSI-BLAST and *SwissProt* database can be download at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

2.3. Discrete Cosine Transform

Feature representation is critical in machine learning based applications, and in this paper we used Discrete Cosine Transform (DCT) combined with PSSM-based features to predict PPIs. DCT has the advantages of packing the most information into the fewest coefficients and minimizing reconstruction errors, with only a small amount of information lost during the processing procedures. Such advantages make DCT suitable and favorable for our experiment of predicting PPIs. Discrete cosine transform can be described as follows:

$$DCT(u, v)=\rho(u)\rho(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x, y)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N}$$
$$0 \le u \le M-1, \quad 0 \le v \le N-1 \tag{2}$$

where

$$\rho(u)=\begin{cases}\sqrt{\frac{1}{M}}, & u=0\\ \sqrt{\frac{2}{M}}, & 1\le u\le M-1\end{cases} \tag{3}$$

$$\rho(v)=\begin{cases}\sqrt{\frac{1}{N}}, & v=0\\ \sqrt{\frac{2}{N}}, & 1\le v\le N-1\end{cases} \tag{4}$$

$f(x, y)\in P^{N\times M}$ is the input signal matrix and here denotes the $N\times 20$ PSSM. In our experiment, the final DCT feature descriptor that represents a protein sequence is obtained by choosing the first 400 coefficients.

2.4. Rotation Forest classifier

Rotation Forest (RF) is a newly proposed ensemble classifier, which is built from a set of decision trees. For each tree, the concentrated extract of the Bootstrap sample from the original training sets up a new training set. Then, the feature set of the new training set is individually and randomly divided into several subsets with a linear transformation method individually. Therefore, by converting the set of all the features of each tree, a full feature set can be reconstructed. Because a small rotary axis can create a completely different tree, the diversity of the ensemble system can be ensured by transformation. Finally, the major vote rule can be used to output all the trees.

Assuming $\{x_i, y_i\}$ contains N training samples, wherein $x_i=(x_{i1}, x_{i2}, \ldots, x_{iD})$ be a D-dimensional feature vector, $y_i$is the corresponding labels, and F is the feature set. Let X be an $n\times D$ matrix, which is composed of n observation feature vector compositions, and then $X=(x_1, x_2, \ldots, x_n)^T$, $Y=(y_1, y_2, \ldots, y_n)^T$. Suppose that the feature set is split randomly into K subsets with the approximate size; there are L decision trees in a rotation forest, denoted by $C_1, C_2, \ldots, C_L$. The training set for an individual classifier C_i is preprocessed with the following steps:

(1) Split F into K disjointed subsets randomly. As a result, each feature subset contains $M=n/K$ features (Here suppose that K is a factor of n).
(2) Select the features corresponding to a subset of $F_{i,j}$ contained in the corresponding column from the training data X, to form a new matrix $X_{i,j}$. Then, a bootstrap subset of objects is drawn from $X_{i,j}$ with the size of 75 percent of the data set to form a new training set, which is denoted by $X'_{i,j}$.
(3) The Principal Component Analysis (PCA) technique is applied to $X'_{i,j}$ for producing the coefficients in a matrix $D_{i,j}$ in which jth column coefficient as the characteristic component jth.
(4) Construct a sparse rotation matrix R_j with the obtained coefficients in matrix $D_{i,j}$ as follows:

$$R_j=\begin{bmatrix} a_{i,1}^{(1)},\ldots,a_{i,1}^{(M_1)} & 0 & \cdots 0\\ 0 & a_{i,2}^{(1)},\cdots,a_{i,2}^{(M_2)} & \cdots 0\\ \vdots\ \vdots & \ddots & \vdots\\ 0\ 0 & \cdots a_{i,k}^{(1)},\ldots,a_{i,k}^{(M_k)} & \end{bmatrix} \tag{5}$$

In the prediction phase, given a test sample x, let $d_{i,j}(XR_i^a)$ be the probability produced by the classifier C_i for the hypothesis that x belongs to class y_i. Then, the confidence for a class can be calculated by the average combination method as follows:

$$\mu_j(x)=\frac{1}{L}\sum_{i=1}^{L} d_{i,j}(XR_i^a) \tag{6}$$

Then, the test sample x is easily assigned to the class with the largest confidence.

3. Results and discussions

3.1. Evaluation measures

In this paper, 5-fold cross-validation test is adopted to evaluate the prediction models by means of the overall prediction accuracy (Accu.), sensitivity (Sen.), precision (Prec.), and Matthews Correlation Coefficient (MCC) as defined, respectively, by:

$$Accu.=\frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$Sen.=\frac{TP}{TP+FN} \tag{8}$$

$$Prec.=\frac{TP}{TP+FP} \tag{9}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{10}$$

where true positive (TP) denotes the number of positive samples that are predicted as positive, false positive (FP) denotes the number of negative samples that are predicted as positive, true negative (TN) denotes the number of negative samples that are predicted as negative, and false negative (FN) denotes the number of positive samples that are predicted as negative. In addition, we also produce the Receiver Operating Characteristic (ROC) curve (Zweig and Campbell, 1993) to assess the prediction performance. An ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR), which is known as sensitivity, against the false positive rate (FPR), which is known as the fall-out and can be calculated as (1-specificity) at various threshold settings.

3.2. Assessment of prediction ability

In order to obtain good experimental results, the corresponding parameters for the Rotation Forest were first optimized. Based on previous studies (Rodriguez and Kuncheva, 2006), we selected PCA as a transformation method for the rotation forest. In addition, the J48 decision tree from WEKA library (Chatfield, 2004) was chosen here as the base classifier. In this study, the two parameters were optimized via a grid search within a possible interval of K (the number of feature subsets), and L (the number of decision trees), where the results are illustrated in Fig. 1. It can be found from Fig. 1 that the average prediction accuracy of the rotation forest performed well and kept improving with ensemble size. However, it can be found in the experiments that the improvement became diminished when the value of the parameters K and L were both greater than 5. Meanwhile, its time consumption also became larger. As a result, both K and L were

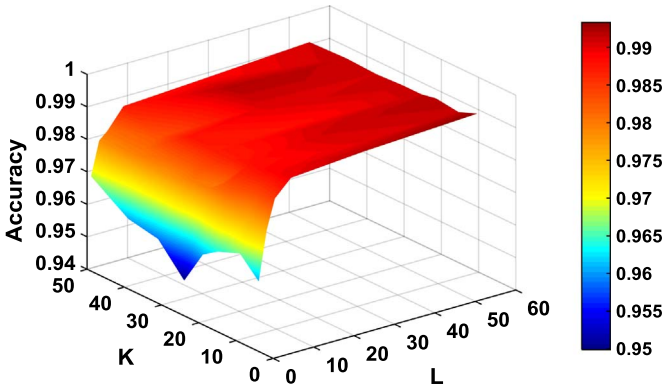

Fig. 1. Accuracy surface obtained of rotation forest for optimizing regularization parameters K and L.

Table 1
5-fold cross-validation results obtained by using the proposed method on the *Yeast* PPIs data set.

Testing set	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
1	98.48	100.00	96.94	97.00
2	98.61	99.72	97.43	97.26
3	98.70	99.91	97.53	97.44
4	98.61	100.00	97.23	97.27
5	98.30	99.73	96.93	96.66
Average	**98.54 ± 0.16**	**99.87 ± 0.14**	**97.21 ± 0.27**	**97.13 ± 0.30**

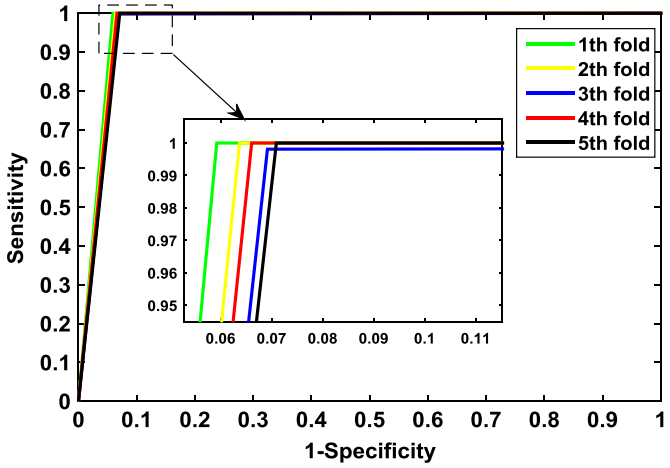

Fig. 2. ROC curves performed by using the proposed method on the *Yeast* PPIs data set.

set to 5 in this experiment.

Considering the numerous samples used in this work, 5-fold cross-validation was used to investigate the training set, which could avoid over-fitting the prediction model and test the stability of the proposed model. Specifically, the whole data set was divided into five parts where four parts were used for training and the rest was used for testing. Thus five models were generated for the five sets of data. The prediction results of position-specific scoring matrix combined with the rotation forest are shown in Table 1. The ROC curves performed on *Yeast* data set is shown in Fig. 2. In this figure, the x-axis depicts the false positive rate (FPR) while the y-axis depicts the true positive rate (TPR) (Fig. 3).

3.3. Comparison with the Proposed Method on the H. pylori Data Set

To further assess the performance of the proposed method, we also tested the ability of trained classifiers on the *H. pylori* data set, which was described by Martin et al. (2005). When using the proposed method to predict PPIs of the *H. pylori* data set, we obtained results for the average accuracy, precision, sensitivity, and MCC of 88.27%, 91.37%, 84.62%, and 79.29%, respectively. The prediction results are shown in Table 2.

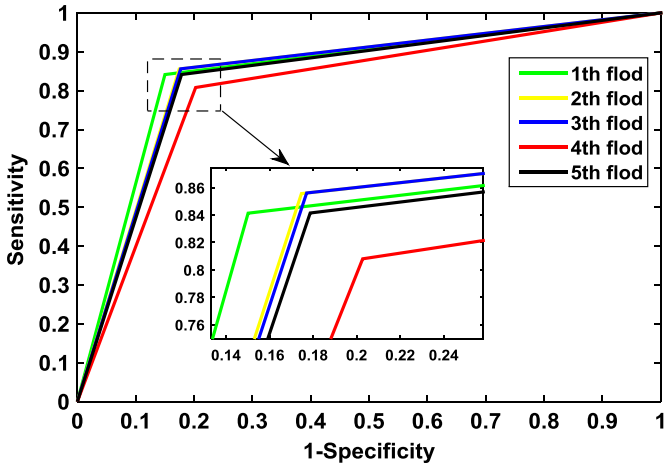

Fig. 3. ROC curves performed by using the proposed method on the *H. pylori* PPIs data set.

Table 2
5-fold cross-validation results obtained by using the proposed method on the *H. pylori* PPIs dataset.

Testing set	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
1	88.16	89.25	86.46	79.11
2	86.79	92.40	81.00	76.95
3	89.54	91.70	86.99	81.24
4	85.93	89.71	81.88	75.77
5	90.92	93.82	86.79	83.38
Average	**88.27 ± 2.02**	**91.37 ± 1.90**	**84.62 ± 2.93**	**79.29 ± 3.10**

The high accuracies show that PSSM combined with the rotation forest is feasible and effective for predicting PPIs. In addition, the low standard deviation of these criteria values indicates that the proposed method is stable and robust. This good performance lies in the fact that the feature extraction method not only depicts the order information of protein sequences but also retains sufficient prior information from the PSSM matrix, which was generated from a group of sequences previously aligned by structural or sequence similarity. Because the process of forming a protein interaction network contains numerous amino acid changes, the substitution rates will help to reveal whether the two proteins interact with each other. Actually, protein pairs with high similarity are more likely to interact with each other, and the similarity between protein sequences depends on their divergence time and substitution rates. Thus, the proposed method we propose here uses evolutionary information and accurately predicts protein-protein interactions.

3.4. Comparison with previous method

Predicting PPIs has been attempted by many machine learning models, most of which are based on the traditional classification. In order to further evaluate the proposed method, we compared it with the state-of-the-art Support Vector Machine (SVM) classifier. Particularly, we used the same feature extraction method to compare the classification performance between the rotation forest and SVM classifier. We downloaded the LIBSVM tools at www.csie.ntu.edu.tw/~cjlin/libsvm. Using the grid search method of SVM with optimized parameters, we set $c=0.5$ and $g=0.6$. From Table 3, it can be observed that, when using SVM to predict PPIs for the *Yeast* data set, we obtained good results with the average accuracy, precision, sensitivity, and MCC of 90.05%, 90.57%, 89.45%, and 82.09%, respectively. For the *Yeast* data set, most of the SVM-based methods performed with average standard values lower than those of the proposed method. In addition, the higher standard deviation indicated that the SVM-based model was less stable than ours.

Many computational methods used to predict PPIs have been proposed. Then, we focused on the *Yeast* and *H. pylori* data sets to compare the rotation forest prediction model using the PSSM algorithm with the existing methods. Table 3 shows the average prediction results of the other six different methods on *Yeast* data set. We can see that the accuracy obtained by these methods were between 75.08% and 89.33%. The average accuracy, precision and sensitivity of these methods were lower than those of the proposed method, which were 98.54%, 99.87%, 97.21%, and 97.13%, respectively. Table 4 shows the prediction performance on the *H. pylori* data set using six different methods. We can see that the accuracies obtained by these methods were between 75.80% and 87.50%, which were also lower than the proposed method.

Table 3
Performance comparison of different methods on the *Yeast* PPIs data set.

Model	Test set	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
Guos' work (Guo et al., 2008)	ACC	89.33 ± 2.67	88.87 ± 6.16	89.93 ± 3.68	N/A
	AC	87.36 ± 1.38	87.82 ± 4.33	87.30 ± 4.68	N/A
Zhous' work (Zhou et al., 2011)	SVM + LD	88.56 ± 0.33	89.50 ± 0.60	87.37 ± 0.22	77.15 ± 0.68
Yangs' work (Yang et al., 2010)	Cod1	75.08 ± 1.13	74.75 ± 1.23	75.81 ± 1.20	N/A
	Cod2	80.04 ± 1.06	82.17 ± 1.35	76.77 ± 0.69	N/A
	Cod3	80.41 ± 0.47	81.86 ± 0.99	78.14 ± 0.90	N/A
	Cod4	86.15 ± 1.17	90.24 ± 0.45	81.03 ± 1.74	N/A
Yous' work (You et al., 2013)	PCA-EELM	87.00 ± 0.29	87.59 ± 0.32	86.15 ± 0.43	77.36 ± 0.44
Our method	**SVM+PSSM**	**90.05 ± 0.86**	**90.57 ± 0.95**	**89.45 ± 1.46**	**82.09 ± 1.38**
	RF+PSSM	**98.54 ± 0.16**	**99.87 ± 0.14**	**97.21 ± 0.27**	**97.13 ± 0.30**

Table 4
Performance comparison of different methods on the *H. pylori* PPIs data set.

Model	Accu. (%)	Prec. (%)	Sen. (%)	MCC (%)
Phylogentic bootstrap (Bock and Gough, 2003)	75.80	80.20	69.80	N/A
HKNN (Nanni, 2005)	84.00	84.00	86.00	N/A
Signature products (Martin et al., 2005)	83.40	85.70	79.90	N/A
Ensemble of HKNN (Nanni and Lumini, 2006)	86.60	85.00	86.70	N/A
Boosting (Liu et al., 2013a)	79.52	81.69	80.37	70.64
Ensemble ELM (You et al., 2013)	87.50	86.15	88.95	78.13
Our method	**88.27**	**91.37**	**84.62**	**79.29**

Table 5
Prediction results on four species based on our model.

Species	Test pairs	Accu. (%)
C.elegans	4013	98.08
E.coli	6954	92.75
H.sapiens	1412	98.87
M.musculus	313	98.72

3.5. Performance on independent data set

Because our method performed well on both the *Yeast* and *H. pylori* data sets, we chose an independent data set to evaluate the practical prediction capability of our final model. It is worth noting that there is a biological hypothesis that in a given species, a large number of physically interacting proteins have coevolved, so these proteins might also interact with proteins from other organisms. In these experiments, we constructed our final prediction model using all 11,188 protein pairs of the *Yeast* data set as the training set with the optimal parameters. The same feature extraction method based on the PSSM method was applied to the feature vector of protein pairs from the other four data sets as RF test input. The four datasets including *C.elegans, E.coli, H.sapiens* and *M.musculus* were collected from the DIP database as our independent test data set. All of the test samples were positive. The performance of our method is summarized in Table 5. The prediction performance accuracies on *C.elegans, E.coli, H.sapiens* and *M.musculus* achieved by our method are 98.08%, 92.75%, 98.87% and 98.72%, respectively. When predicting the PPIs of the *C.elegans, H.sapiens* and *M.musculus* datasets, the accuracy of our method is higher than 98%. This result shows that the proposed method performance well in cross-species PPIs prediction. Interestingly, these results indicate that the yeast protein sequence information is sufficient for predicting the PPIs of other species. In addition, this means that the proposed method has a strong ability for predicting cross-species protein-protein interactions.

4. Conclusions and discussion

In this article, a novel computational method that combined protein evolutionary information embedded in PSSM and Rotation Forest classifier was developed for PPI prediction. This method, 5-fold cross-validation on six different PPI data set including *Yeast*, *H. pylori*, *C.elegans*, *E.coli*, *H.sapiens,* and *M.musculus* data sets, achieves a high prediction accuracy of 98.54%, 88.27%, 98.08%, 92.75%, 98.87%, and 98.72% respectively, which is significantly better than previous methods. The main improvements come from the use of the informative features extracted from the PSSM matrix using DCT algorithm, from the powerful and robust RF classifier. All of these experiment results demonstrate that the PSSM combined with the rotation forest can improve the accuracy of the prediction. And our method could become a useful supplementary tool to address many other bioinformatics problems. In conclusion, the proposed method is an efficient, reliable, powerful prediction model and can be a useful tool for future proteomics research. For the future study, more effective features extraction method and machine learning techniques will be explored for PPIs prediction.

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported in part by the National Science Foundation of China, under Grants 61373086, 61572506, 11301517, 11631014, in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, in part by the National Key Research and Development Plan, under Grant 2016YFC0600908, and in part by Graduate Education Innovation project of Jiangsu Province, under Grants KYLX16_0535. The authors would like to thank all anonymous reviewers for their constructive advices.

References

Aloy, P., Russell, R.B., 2003. InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 19, 161–162. http://dx.doi.org/10.1093/bioinformatics/19.1.161.

Altschul, S.F., Koonin, E.V., 1998. Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases. Trends Biochem. Sci. 23, 444–447. http://dx.doi.org/10.1016/s0968-0004(98)01298-5.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.

Ben-Hur, A., Noble, W.S., 2005. Kernel methods for predicting protein-protein interactions. Bioinformatics 21, I38–I46. http://dx.doi.org/10.1093/bioinformatics/bti1016.

Bock, J.R., Gough, D.A., 2001. Predicting protein-protein interactions from primary

structure. Bioinformatics 17, 455–460. http://dx.doi.org/10.1093/bioinformatics/17.5.455.
Bock, J.R., Gough, D.A., 2003. Whole-proteome interaction mining. Bioinformatics 19, 125–134. http://dx.doi.org/10.1093/bioinformatics/19.1.125.
Braga, P.L., Oliveira, A.L.I., Ribeiro, G.H.T., Meira, S.R.L., 2007. Ieee. Bagging predictors for estimation of software project effort. International Joint Conference on Neural Networks, Orlando, FL, pp. 1595–1600.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. http://dx.doi.org/10.1023/a:1010933404324.
Chatfield, C., 2004. Statistical data mining and knowledge discovery. J. R. Stat. Soc. Ser. a-Stat. Soc. 167, 567–568. http://dx.doi.org/10.1111/j.1467-985X.2004.02057_4.x.
Chen, X.-W., Jeong, J.C., 2009. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 25, 585–591. http://dx.doi.org/10.1093/bioinformatics/btp039.
Cutler, D.R., Edwards, T.C., Jr., Beard, K.H., Cutler, A., Hess, K.T., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792. http://dx.doi.org/10.1890/07-0539.1.
Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D., 2002. Protein interactions - two methods for assessment of the reliability of high throughput observations. Mol. Cell. Proteom. 1, 349–356. http://dx.doi.org/10.1074/mcp.M100037-MCP200.
Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147. http://dx.doi.org/10.1038/415141a.
Gribskov, M., McLanchlan, A.D., Eisenberg, D., 1987. Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA 84, 4355–4358. http://dx.doi.org/10.1073/pnas.84.13.4355.
Guo, Y., Yu, L., Wen, Z., Li, M., 2008. Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. Nucleic Acids Res. 36, 3025–3030. http://dx.doi.org/10.1093/nar/gkn159.
Han, J.D.J., Dupuy, D., Bertin, N., Cusick, M.E., Vidal, M., 2005. Effect of sampling on topology predictions of protein-protein interaction networks. Nat. Biotechnol. 23, 839–844. http://dx.doi.org/10.1038/nbt1116.
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. http://dx.doi.org/10.1073/pnas.061034498.
Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202. http://dx.doi.org/10.1006/jmbi.1999.3091.
Jones, D.T., Ward, J.J., 2003. Prediction of disordered regions in proteins from position specific score matrices. Proteins-Struct. Funct. Bioinforma. 53, 573–578. http://dx.doi.org/10.1002/prot.10528.
Lee, T.-H., Yang, Y., 2006. Bagging binary and quantile predictors for time series. J. Econ. 135, 465–497. http://dx.doi.org/10.1016/j.jeconom.2005.07.017.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. http://dx.doi.org/10.1093/bioinformatics/btl158.
Li, W.Z., Jaroszewski, L., Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283. http://dx.doi.org/10.1093/bioinformatics/17.3.282.
Liu, B., Yi, J., Aishwarya, S.V., Lan, X., Ma, Y., Huang, T.H.M., Leone, G., Jin, V.X., 2013a. QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. Bmc Genom., 14. http://dx.doi.org/10.1186/1471-2164-14-s8-s3.
Liu, C.H., Li, K.-C., Yuan, S., 2013b. Human protein-protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence. Bioinformatics 29, 92–98. http://dx.doi.org/10.1093/bioinformatics/bts620.
Marcotte, E.M., Xenarios, I., Eisenberg, D., 2001. Mining literature for protein-protein interactions. Bioinformatics 17, 359–363. http://dx.doi.org/10.1093/bioinformatics/17.4.359.
Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein-protein interactions from genome sequences. Science (New York, N.Y.)285, 751-753, doi:http://dx.doi.org/10.1126/science.285.5428.751.
Martin, S., Roe, D., Faulon, J.L., 2005. Predicting protein-protein interactions using signature products. Bioinformatics 21, 218–226. http://dx.doi.org/10.1093/bioinformatics/bth483.
Nanni, L., 2005. Hyperplanes for predicting protein-protein interactions. Neurocomputing 69, 257–263. http://dx.doi.org/10.1016/j.neucom.2005.05.007.
Nanni, L., Lumini, A., 2006. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics 22, 1207–1210. http://dx.doi.org/10.1093/bioinformatics/btl055.
Nanni, L., Lumini, A., 2009. Ensemble generation and feature selection for the identification of students with learning disabilities. Expert Syst. Appl. 36, 3896–3900. http://dx.doi.org/10.1016/j.eswa.2008.02.065.
Pazos, F., Valencia, A., 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14, 609–614. http://dx.doi.org/10.1093/protein/14.9.609.
Pazos, F., Helmer-Citterich, M., Ausiello, G., Valencia, A., 1997. Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 271, 511–523. http://dx.doi.org/10.1006/jmbi.1997.1198.
Planas-Iglesias, J., Bonet, J., Garcia-Garcia, J., Marin-Lopez, M.A., Feliu, E., Oliva, B., 2013. Understanding Protein-Protein Interactions Using Local Structural Features. J. Mol. Biol. 425, 1210–1224. http://dx.doi.org/10.1016/j.jmb.2013.01.014.
Priya, S.B., Saha, S., Anishetty, R., Anishetty, S., 2013. A matrix based algorithm for protein-protein interaction prediction using domain-domain associations. J. Theor. Biol. 326, 36–42. http://dx.doi.org/10.1016/j.jtbi.2013.02.016.
Ratsch, G., Mika, S., Scholkopf, B., Muller, K.R., 2002. Constructing boosting algorithms from SVMs: an application to one-class classification. Ieee Trans. Pattern Anal. Mach. Intell. 24, 1184–1199. http://dx.doi.org/10.1109/tpami.2002.1033211.
Rodriguez, J.J., Kuncheva, L.I., 2006. Rotation forest: a new classifier ensemble method. Ieee Trans. Pattern Anal. Mach. Intell. 28, 1619–1630. http://dx.doi.org/10.1109/tpami.2006.211.
Saha, I., Zubek, J., Klingstrom, T., Forsberg, S., Wikander, J., Kierczak, M., Maulik, U., Plewczynski, D., 2014. Ensemble learning prediction of protein-protein interactions using proteins functional annotations. Mol. Biosyst. 10, 820–830. http://dx.doi.org/10.1039/c3mb70486f.
Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K, Li, Y., Jiang, H., 2007. Predictina protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337–4341. http://dx.doi.org/10.1073/pnas.0607879104.
Singhal, M., Resat, H., 2007. A domain-based approach to predict protein-protein interactions. Bmc Bioinforma., 8. http://dx.doi.org/10.1186/1471-2105-8-199.
Souiai, O., Guerfali, F., Ben Miled, S., Brun, C., Benkahla, A., 2014. In silico prediction of protein-protein interactions in human macrophages. BMC Res. Notes 7. http://dx.doi.org/10.1186/1756-0500-7-157.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M.J., Johnston, M., Fields, S., Rothberg, J.M., 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627.
Wass, M.N., Sternberg, M.J.E., 2008. ConFunc - functional annotation in the twilight zone. Bioinformatics 24, 798–806. http://dx.doi.org/10.1093/bioinformatics/btn037.
Xenarios, I., Salwinski, L., Duan, X.Q.J., Higney, P., Kim, S.M., 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. http://dx.doi.org/10.1093/nar/30.1.303.
Xia, J.-F., Han, K., Huang, D.-S., 2010. Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor. Protein Pept. Lett. 17, 137–145.
Yang, L., Tang, X., 2014. Protein-protein interactions prediction based on iterative clique extension with gene ontology filtering. Sci. World J. doi:http://dx.doi.org/10.1155/2014/523634.
Yang, L., Xia, J.-F., Gui, J., 2010. Prediction of protein-protein interactions from protein sequence using local descriptors. Protein Pept. Lett. 17, 1085–1090.
You, Z.-H., Lei, Y.-K., 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 Bioinforma., 14. http://dx.doi.org/10.1186/1471-2105-14-s8-s10.
Zhang, Y.-N., Pan, X.-Y., Huang, Y., Shen, H.-B., 2011. Adaptive compressive learning for prediction of protein-protein interactions from primary sequence. J. Theor. Biol. 283, 44–52. http://dx.doi.org/10.1016/j.jtbi.2011.05.023.
Zhou, Y.Z., Gao, Y., Zheng, Y.Y., 2011. Prediction of protein-protein interactions using local description of amino acid sequence. Adv. Comput. Sci. Educ. Appl., Pt Ii 202, 254–262.
Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M., Snyder, M., 2001. Global analysis of protein activities using proteome chips. Science 293, 2101–2105. http://dx.doi.org/10.1126/science.1062191.
Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.