Expert Systems With Applications 176 (2021) 114876

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Prediction of protein–protein interactions based on elastic net and deep forest

Bin Yu [a,b,c,d,1,*], Cheng Chen [a,c,1], Xiaolin Wang [a,c], Zhaomin Yu [a,c], Anjun Ma [e], Bingqiang Liu [f]

[a] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China
[b] School of Life Sciences, University of Science and Technology of China, Hefei 230027, China
[c] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China
[d] Key Laboratory of Computational Science and Application of Hainan Province, Haikou 571158, China
[e] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA
[f] School of Mathematics, Shandong University, Jinan 250100, China

ARTICLE INFO

Keywords:
Protein-protein interactions
Multi-information fusion
Elastic net
Deep forest

ABSTRACT

Prediction of protein–protein interactions (PPIs) helps to grasp molecular roots of disease. However, web-lab experiments to predict PPIs are limited and costly. Using machine-learning-based frameworks can not only automatically identify PPIs, but also provide new ideas for drug research and development from a promising alternative. We present a novel deep-forest-based method for PPIs prediction. Firstly, pseudo amino acid composition (PAAC), autocorrelation descriptor (Auto), multivariate mutual information (MMI), composition-transition-distribution (CTD), amino acid composition position-specific scoring matrix (AAC-PSSM), and dipeptide composition PSSM (DPC-PSSM) are adopted to extract and construct the pattern of PPIs. Secondly, elastic net is utilized to optimize the initial feature vectors and boost the predictive performance. Finally, we ensemble XGBoost, random forest, and extremely randomized trees to construct deep forest model via cascade architecture for PPIs prediction (GcForest-PPI). Benchmark experiments reveal that the proposed approach outperforms other state-of-the-art predictors on *Saccharomyces cerevisiae* and *Helicobacter pylori*. We also apply GcForest-PPI on independent test sets, CD9-core network, crossover network, and cancer-specific network. The evaluation shows that GcForest-PPI can boost the prediction accuracy, complement experiments and improve drug discovery.

1. Introduction

The study of the protein–protein interactions (PPIs) of molecular mechanisms is essential (Alberts, 1998; Amar, Hait, Izraeli, & Shamir, 2015; Schadt, 2009). The disorder of the PPI network structure can cause abnormalities in cell life activities. Because of the progress of high-throughput technologies, lots of PPIs via web-lab experimental verification have emerged. Multiple PPIs sources lead to the generation of PPIs databases, containing the DIP (Database of Interacting Protein) (Xenarios et al., 2002), and HPRD (Human Protein Reference Database) (Peri, Navarro, Amanchy, & Kristiansen, 2003). The detection of PPIs relied on computational methods could reduce the web-lab limitations and an effective, accurate, useful machine learning (ML) tool can predict large scale PPIs.

Several genomic features have been leveraged in ML-based technologies, including but not limited to, protein structure information, gene neighbors, sequence composition information, gene expression, physicochemical information, position information, and evolutionary information (Deng et al., 2014; Guo, Yu, Wen, & Li, 2008; Yu et al., 2016). Zhang et al. (Zhang & Tang, 2016) designed a PPI tool based on gene ontology. However, when structure information cannot be in hand, the domain-based method does not work. Kovács et al. (2019) used network paths of length three to perform link prediction. This approach can offer structural and evolutionary reference to detect PPIs. Zahiri et al. (Zahiri, Yaghoubi, Mohammad-Noori, Ebrahimpour, & Masoudi-Nejad, 2013) extracted evolutionary information via PPIevo from the position-specific scoring matrix (PSSM) and received better performance and robustness on HPRD dataset.

* Corresponding author at: College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China.
E-mail addresses: yubin@qust.edu.cn (B. Yu), chenchengqust@163.com (C. Chen), wangxiaol_qust@163.com (X. Wang), yzm_qust@163.com (Z. Yu), anjun.ma@osumc.edu (A. Ma), bingqiang@sdu.edu.cn (B. Liu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2021.114876
Received 1 August 2019; Received in revised form 22 February 2021; Accepted 3 March 2021
Available online 10 March 2021

With the collection of various omics data and development of popular feature extraction methods, integrating multiple features mentioned above is necessary. Zhang et al. (Zhang, Yu, Xia, & Wang, 2019) integrated different descriptors to obtain complimentary information. The constructed ensemble predictor was valid for interactions prediction. Lian et al. (Lian, Yang, Li, Fu, & Zhang, 2019) proposed an ML-based predictor for human-bacteria PPIs. They fused traditional sequence-based and two network-property-related descriptors introduced by them. Then, individual random forest (RF) model was constructed for each feature encoding scheme. Finally, the noisy-OR algorithm was employed to predict human-bacteria PPIs. The designed pipeline revealed that the introduced network-based methods could represent important network topology properties. However, considering the effect of high-dimension condition after feature fusion, feature selection is a good choice by retaining effective feature subset, which could easily process noisy and redundant data. For example, You et al. (You et al., 2014) utilized multi-scale continuous and discrete and minimum redundancy maximum relevance (mRMR), and evaluation of this pipeline indicates mRMR did enhance the success of PPIs prediction.

Recently, more and more deep learning (DL) based algorithms emerged for PPIs prediction. Hashemifar et al. (Hashemifar, Neyshabur, Khan, & Xu, 2018) proposed a sequence-based convolutional neural network learning to infer PPIs called DPPI, and DL could obtain the high-level evolutional representations from PSSM. Yadav et al. (Yadav, Ekbal, Saha, Kumar, & Bhattacharyya, 2019) constructed Bi-LSTM model based on stacked algorithm for the identification of PPIs, which combined multiple levels features using shortest dependency path. Then the information via embedding layer was input into the stacked Bi-LSTM model. Lei et al. (Lei et al., 2019) presented a multimodal deep polynomial network called MDPN. For the first stage, high-level features were produced using deep polynomial network based on BLOSUM62 and hydrophobic property. For the second stage, extreme leaning machine was taken as classifier to identify PPIs. Chen et al (Chen et al., 2019) presented a PPIs predictive framework PIPR using siamese residual convolution neural network (RCNN). This architecture can extract local and contextualized information. However, DL also has the following limitations: (*i*) the number of layers and the number of nodes for the neural network need to be determined before training the DL model (Krizhevsky, Sutskever, & Hinton, 2012); (*ii*) the to-be-optimized parameters of DL are diverse on different data, requiring substantial efforts in adjusting the parameters (Krizhevsky, Sutskever, & Hinton, 2012; Lecun, Bottou, Bengio, & Haffner, 1998; Simonyan, & Zisserman, 2015); and (*iii*) DL requires a lot of data for training (Silver et al., 2018).

Compared with DL, tree ensemble methods have good properties and achieve excellent performance. For example, Feng et al. (Feng & Zhou, 2017) proposed a tree ensemble AutoEncoder (eforest), which can do backward reconstruction using tree-based approach (maximal compatible rule). They utilized forest to perform the process of encoding and decoding for the first time. The experimental results showed that eforest can effectively eliminate noisy information compared with the auto-encoder network. Feng et al. (Feng, Yu, & Zhou, 2018) proposed a multi-layered GBDT, which can effectively learn hierarchical features through stacking multiple layers. It is known that deep forest (DF) model had fewer hyper-parameters setting and higher flexibility than neural-based DL models (Zhou & Feng, 2017, 2019). DF can deal with non-differential issues without requiring backpropagation algorithms and learn high-level feature information through cascade structure to avoid overfitting. The cascade structure of DF can extract high-level feature information from raw PPIs feature space, and it is very helpful to deal with classification task.

According to the above discussion, we can conclude this: (*i*) PPIs prediction method can fuse various features, including network-based, secondary structure-based, PSSM-based, amino acid-based features. However, the high-dimension data can generate redundant and irrelevant features. Using appropriate feature coding strategy and feature selection algorithm can obtain good initial feature information. (*ii*) Compared with DL, multi-grained cascade forest is great and robust, hence, can be effectively used to handle classification problems. The raw GcForest framework proposed by Zhou & Feng is composed of RF and extremely randomized trees (Extra-Trees). In essence, good and diverse base classifier can generate a better ensemble classifier. Considering XGBoost is a great boosting algorithm with strong generation ability, which is different from the tree-based RF and Extra-Trees, a new DF architecture composed of XGBoost, RF and Extra-Trees is constructed. The constructed tree-based DF contains good and diverse base classifier, which could obtain complementary advantages and essential features. In this way, the improved GcForest can achieve good prediction performance and generalization ability.

Considering the validity of the feature fusion, selection, and the superiority of DF, we propose a new PPI prediction method based on DF, so-called GcForest-PPI, where GcForest represents multi-Grained Cascade Forest. The physicochemical information, sequence information, and evolutionary information are retrieved by PAAC, Auto, MMI, CTD, AAC-PSSM, and DPC-PSSM. Then, elastic net is used to select variables highly relevant to the category labels. Next, the optimal features are input into GcForest to identify PPIs based on the known PPIs. Finally, the five-fold cross-validation shows that GcForest-PPI achieves higher accuracy than the state-of-the-art predictors. Cross-species prediction is performed using *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens,* and *Mus musculus* as independent datasets with the accuracy of 98.58%, 99.04%, 96.01%, and 96.30%, respectively. We also found that (*i*) the PPIs of a CD9-core network are all predicted successfully; (*ii*) GcForest-PPI can predict PPIs in a crossover network and can reveal the biological functions for the Wnt-related pathway; and (*iii*) the PPIs of the cancer-specific network are also all predicted successfully, providing new ideas for studying the associations of drug-disease and drug-target for developing new drugs of cancer treatment.

The organization of this paper is as follows. In Section 1, we introduce the research background and situation of PPIs prediction. Section 2 describes the algorithms of feature encoding and selection, GcForest construction and evaluation, and datasets. In Section 3, we list the experimental results, including parameter optimization, independent testing and network prediction. In Section 4, we summarize the conclusion and the future research direction.

2. Materials and methods

In this section, materials and methods consist of five subsections, which are consistent with the GcForest-PPI pipeline using the sequence as input. (2.1) Feature extraction. Encoding sequence using six descriptors. (2.2) Elastic net. Selecting the effective feature subset without losing information. (2.3) Deep forest. Building up the GcForest-PPI model through integrating Extra-Trees, RF and XGBoost. (2.4) Proposed methodology. Introducing layers and structure of deep forest and the pipeline of GcForest-PPI. (2.5) Performance evaluation. Listing the evaluation indicators. (2.6) Datasets. Listing the nine PPIs datasets for evaluating the GcForest-PPI.

2.1. Feature extraction

We use six feature coding schemes to obtain the physicochemical information, sequence information and evolutionary information, including Pseudo Amino Acid Composition (PAAC), Autocorrelation descriptor (Auto), Multivariate Mutual Information (MMI), Composition-Transition-Distribution (CTD), Amino Acid Composition PSSM (AAC-PSSM) and Dipeptide Composition PSSM (DPC-PSSM).

2.1.1. Physicochemical information

PAAC and Auto are utilized to extract the physicochemical and composition information. At present, PAAC has shown good properties in proteomics field (Cui et al., 2019; Qiu et al., 2018; Yu et al., 2021, 2017, 2018; Zhang et al., 2021). Auto includes Morean-Broto, Moran,

and Geary (Chen, Zhang, Ma, & Yu, 2019; Chen et al., 2018). The seven physicochemical properties in Auto can be obtained in Supplementary Table S1. The PAAC encoding feature vector x_u can be defined as:

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (1 \leqslant u \leqslant 20) \\ \dfrac{\omega \theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (20+1 \leqslant u \leqslant 20+\lambda) \end{cases} \tag{1}$$

where f_i represents amino acid composition information, θ_j represents layer sequence correlation factor calculated using hydrophobicity, hydrophilicity, and side-chain mass, $\omega = 0.05$ (Chou, 2001). The shortest length of protein in benchmark PPIs dataset is 12. So the λ must satisfy $\lambda < 12$ and the dimension of PAAC is $20 + \lambda$.

We use A_i to characterize the i - th amino acids and $P(A_i)$ represents the normalized physicochemical values. The $\overline{P}$ can be employed as the mean value for specific physicochemical property in whole protein sequence. The Eqs. (2), (3), (4) represent Moreau-Broto, Moran, Geary, respectively.

$$NMBA(l) = \frac{MBA(l)}{N-l}, \qquad l = 1, 2, \cdots, lag \tag{2}$$

$$MA(l) = \frac{\frac{1}{N-l}\sum_{i=1}^{N-l}(P(A_i) - \overline{P})(P(A_{i+l}) - \overline{P})}{\frac{1}{N}\sum_{i=1}^{N}(P(A_i) - \overline{P})^2}, \qquad l = 1, 2, \cdots, lag \tag{3}$$

$$GA(l) = \frac{\frac{1}{2(N-l)}\sum_{i=1}^{N-l}(P(A_i) - P(A_{i+l}))^2}{\frac{1}{N}\sum_{i=1}^{N}(P(A_i) - \overline{P})^2}, \quad l = 1, 2, \cdots, lag \tag{4}$$

where $MBA(l) = \sum_{i=1}^{N-l} P(A_i)P(A_{i+l})$, l represents the l - th value in the obtained Auto-based vector. N is the number of residues of one sequence, and lag is the parameter that needs to be adjusted manually ($lag < 12$). The dimension of Auto is $3 \times 7 \times lag$.

2.1.2. Sequence information

MMI (Ding, Tang, & Guo, 2016, 2017) and CTD are utilized to obtain sequence information (Zhang, Yu, Xia, & Wang, 2019). MMI can represent the information entropy and group-based features. CTD can obtain the distribution pattern and effective sequence information.

For MMI, the amino acid residues can be classified into seven classes (Supplementary Table S2). The algorithm flowchart of MMI is shown in the Supplementary Fig. S1. For a given protein sequence, we can define various 2-gram $I(a,b)$ and 3-gram $I(a,b,c)$ features. The information entropy can be expressed as:

$$I(a,b) = f(a,b) ln(\frac{f(a,b)}{f(a)f(b)}) \tag{5}$$

where $f(a,b)$ represents frequency 2-gram (a, b) for given sequence. $f(a)$ represents frequency of a.

$$I(a,b,c) = I(a,b) - I(a,b|c) \tag{6}$$

where a, b, c are types of amino acid in triplet, and $I(a,b|c) = H(a|c) - H(a|b,c)$ which could be described as:

$$H(a|c) = -\frac{f(a,c)}{f(c)} ln(\frac{f(a,c)}{f(c)}) \tag{7}$$

$$H(a|b,c) = -\frac{f(a,b,c)}{f(b,c)} ln(\frac{f(a,b,c)}{f(b,c)}) \tag{8}$$

Finally, each protein sequence yields 84-dimensional 3-gram features and 28-dimensional 2-gram features. The dimension of MMI is 119.

In CTD (Chen et al., 2018), amino acids are grouped into three groups based on hydrophobicity: polar (P), neutral (N), and hydrophobic (H) (Supplementary Table S3). $N(r)$ represents the character type r in the replaced sequence, and N is sequence length. Given sequence MTTTVPKVFAFHEF. It can be represented as '32223213323213' according to Hydrophobicity_PRAM900101. '1' represents polar, '2' represents neutral, '3' represents hydrophobicity.

$$Composition(r) = \frac{N(r)}{N}, \qquad r \in \{P, \quad N, \quad H\} \tag{9}$$

The composition generates grouped information, the frequency of '1' is $2/14 = 0.1429$, the frequency of '2' is $6/14 = 0.4286$, the frequency of '3' is $6/14 = 0.4286$.

The T descriptor is defined as follows:

$$T(r,s) = \frac{N(r,s) + N(s,r)}{N-1}, \qquad r,s \in \{(P,N), \quad (N,H), (H,P)\} \tag{10}$$

where $N(r,s)$ represents dipeptide frequency, the value of (P,N) is $2/13 = 0.1538$, the value of (N,H) is $6/13 = 0.4615$, the value of (H,P) is $2/13 = 0.1538$.

For each group (P, N and H), we obtain the pattern information of the first, 25%, 50%, 75% and 100% of the encoded grouped sequence. Take '3' for example, there are 6 residues encoded '3'. The first '3' is 1. The second '3' is $25\% \times 6 = 1$. The third '3' is $50\% \times 6 = 3$. The fourth '3' is $75\% \times 6 = 4$. The fifth '3 is $100\% \times 6 = 6$. The position in the first, the second, the third, the fourth, the fifth '3' of whole sequence are 1, 1, 8, 9, 14, respectively. So the distribution descriptor for '3' are (1/14),(1/14)(8/14)(9/14)(14/14).

2.1.3. Evolutionary information

Evolutionary information in the position-specific scoring matrix (PSSM) is essential in proteomics (Supplementary File S1). The amino acid composition PSSM (AAC-PSSM) and dipeptide composition PSSM (DPC-PSSM) are utilized to generate evolutionary information. Some researchers have used PSSM to leverage encoding information, including the identification of drug-target interaction (Shi et al., 2019), detecting protein–protein interaction site (Wang et al., 2019; Wei, Han, Yang, Shen, & Yu, 2016; Zhang, Li, Quan, Chen, & Lü, 2019).

PSSM are converted to feature vector by AAC-PSSM via equation (11)

$$P_{AAC} = (P_1, P_2, \cdots, P_j, \cdots P_{20})^T \qquad (j = 1, 2, \cdots, 20) \tag{11}$$

where $P_j = \frac{1}{L}\sum_{i=1}^{L} p_{ij} \qquad (j = 1, 2, \cdots 20)$, P_j represents the composition evolutionary information of the j amino acid residue. And the dimension of AAC-PSSM is 20.

ACC-PSSM only represents the composition information from PSSM, and loses the position information, which is insufficient to fully represent the evolutionary information. DPC-PSSM can reflect the sequence-order information of PSSM, the encoding feature vector can be expressed as

$$P_{DPC} = (D_{1,1}, D_{1,2}, \cdots, D_{1,20}, D_{2,1}, D_{2,2}, \cdots, D_{2,20}, \cdots, D_{20,20}) \tag{12}$$

where $D_{i,j} = \frac{1}{L-1}\sum_{k=1}^{L-1} p_{k,i} \times p_{k+1,j}$, the dimension of DPC-PSSM is 400.

2.2. Elastic net

Elastic Net (EN) (Zou & Hastie, 2005) is a regularization-based feature selection approach. EN not only keeps the sparse model of least absolute shrinkage and selection operator (LASSO) but also maintains the regularization properties of the ridge regression. α, β are penalty terms, which represent a compromise strategy between LASSO and ridge regression. The objective function of EN is

$$\min_w \frac{1}{2*n}\|y - Xw\|_2^2 + \alpha*\beta\|w\|_1 + \frac{1}{2}\alpha*(1-\beta)\|w\|_2^2 \tag{13}$$

where X is the sample matrix, y is the label, n represents sample number, and w indicates the regression coefficient.

2.3. Deep forest

Deep forest (DF) is a forest-based ensemble learning method for trees (Zhou & Feng, 2017, 2019), which can represent high-level feature information by cascade structure. Zhou et al. used two random forest (RF) (Breiman, 2001), and two extremely randomized trees (Extra-Trees) (Geurts, Ernst, & Wehenkel, 2006) to construct the deep forest. Considering the boosting algorithm achieves higher computation accuracy and better model generation ability. XGBoost (Chen & Guestrin, 2016), which combines the linear model, regularized objective and second-order approximation via boosting algorithm to avoid overfitting, can reduce computational costs and enhance predictive performance. Meanwhile, sub-samples speed up the parallel computing in the process of tree learning. So we develop a new deep forest architecture to implement GcForest, which is composed of four XGBoost, four RF and four Extra-Trees at each level of the cascade (except the last one). XGBoost is a variant of gradient boosting decision tree whose base classifier is regression tree. The base classifiers of the RF and Extra-Trees are decision tree. In this way, an outstanding deep forest contains good and diverse base classifier. Then the deep layer architecture GcForest can obtain complementary advantages and essential features.

XGBoost is an ensemble algorithm. Given dataset $D = \{(x_i, y_i)(|D| = n, x_i \in R^m, y_i \in R)\}$, the loss function of XGBoost is shown as

$$Loss(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{14}$$

where *Loss* is the convex objective function, Ω penalizes the complexity of XGBoost, f_k is k - th regression tree.

$$Loss^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{15}$$

where $g_i = \partial_{y^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$,$h_i = \partial^2_{y^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$represents the first order and second order gradient statistics.

RF is a bagging ensemble classifier using random bootstrap. Gini coefficient (Gastwirt, 1972) is employed as the evaluation to split the node for tree learning. There are two main differences between Extra-Trees and RF. (*i*) Extra-Trees uses all training set to generate decision tree. (*ii*) Each tree is segmented and grows at each node by randomly selecting a feature.

2.4. Proposed methodology

In this paper, we develop a new deep forest architecture to implement GcForest, which is composed of four XGBoost, four RF and four Extra-Trees at each level of the cascade. XGBoost is an excellent variant of gradient boosting decision tree whose base classifier is regression tree. The base classifiers of the RF and Extra-Trees are decision tree. The constructed tree-based DF contains good and diverse base classifier, which could obtain complementary advantages and essential features. In this way, the improved GcForest can achieve good prediction performance and generalization ability. The cascade structure is shown in Fig. 1A. Suppose there are two classes to predict, each forest will output a two-dimensional class vector, and each layer will generate a 24-dimensional new class vector. In Fig. 1A, the features in white color (four XGBoost, four RF and four Extra-Trees) represent $4 \times 2 \times 3 = 24$ new high-level features. Concatenated with the raw feature vector (the red color), the dimension is $24 + 476 = 500$ per layer. For the last layer, the dimension is 24 without raw feature vector. The dimension of output feature vector is 2. And the Supplementary Fig. S2 is the more detailed representation of Fig. 1A. The newly generated class vectors are concatenated with the raw protein feature vectors to produce multi-level features. The output class probability score of the last layer is shown in Fig. 1B.

As illustrated in Fig. 1, given an instance, each forest can produce an estimate of class distribution by calculating the percentage of different types of training samples at the leaf node. The number of iterations on XGBoost is set to 500. The RF includes 500 decision trees and randomly

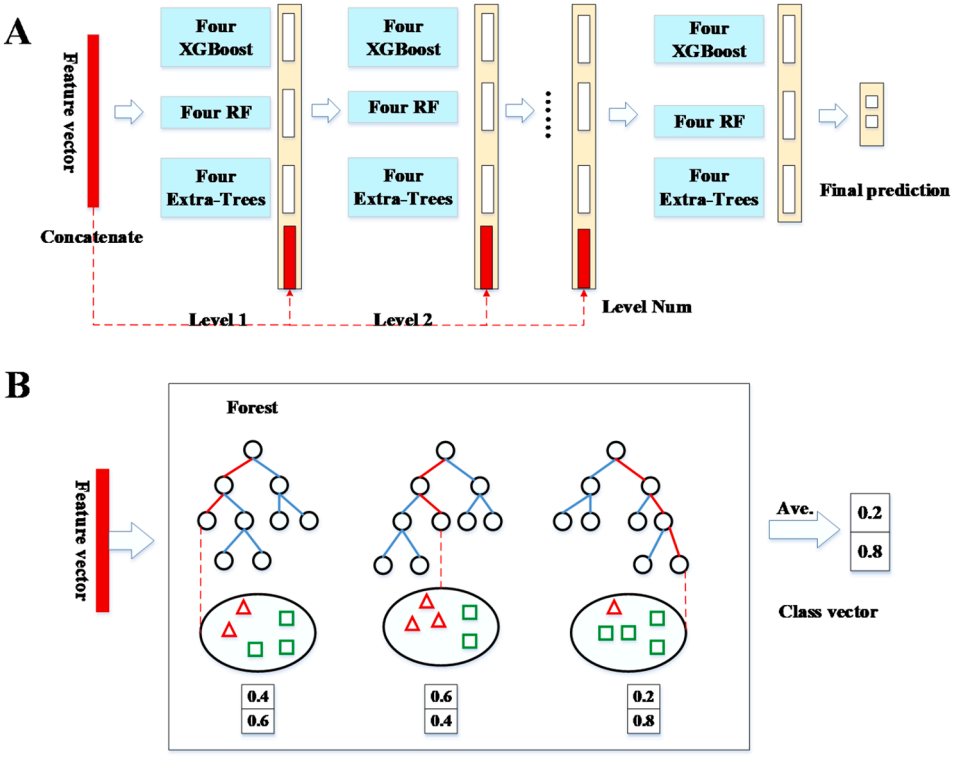

Fig. 1. The GcForest structure and the generation of a class vector. (A) Illustration of GcForest structure. (B) The generation of class vector. Different marks in leaf nodes represent different classes.

selects $\sqrt{d}$ features as candidate subsets (d is the dimension of dataset). The Extra-Trees consist of 500 trees.

To reduce overfitting of GcForest, the class vector is generated by each forest using five-fold cross-validation. Specifically, each sample will be employed as training set four times. Then, the class vectors are concatenated to produce augmented class vectors. The feature information is obtained from known sequences in the previous study, but they may generate noisy data inputs. It is reasonable to extract high-level feature information for prediction, and the probability output is employed as the next level of the forest. So, DF has good generalization ability, and the deep structure can exploit potential information from PPIs. The proposed methodology of GcForest-PPI is shown in Fig. 2 and the detailed process for PPIs network prediction can be described as follows.

Step 1: Protein pairs. The input are interacting pairs and non-interacting pairs.

Step 2: Feature extraction. The effective initial coding information of PPIs could be obtained by PAAC, Auto, MMI, CTD, AAC-PSSM and DPC-PSSM. These descriptors can produce complimentary information by integrating physicochemical, position, sequence, composition and evolutionary information.

Step 3: Feature selection. The elastic net based on L1 and L2 regularization can eliminate redundancy and retain essential variables. Adjusting the parameters α and β via five-fold cross-validation to generate effective subset for identifying PPIs.

Step 4: Deep forest and model construction. The important feature representations can be obtained for binary PPIs prediction task via Step 2 and Step 3. Then XGBoost, RF and Extra-Trees are integrated via cascade architecture to implement the task, and the predictive tool GcForest-PPI for PPIs based on deep forest is built up on *S. cerevisiae* and *H. pylori*.

Step 5: Model evaluation. We apply GcForest-PPI on four cross-species datasets, CD9-core network, crossover network and cancer-specific network. Then list the comparison of GcForest-PPI with the state-of-the-art predictors and plot the three types of PPIs networks.

2.5. Performance evaluation

In order to evaluate GcForest-PPI, the evaluation indicators included Recall, Precision, Accuracy (ACC) and Matthews correlation coefficient (MCC) (Cui et al., 2019; Du et al., 2017; Tian et al., 2019; Yu et al., 2018).

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{18}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{19}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Receiver operating characteristic (ROC) curve (Shi et al., 2019; Wang et al., 2019) and AUC, Precision recall curve (PR) (Davis & Goadrich, 2006) and AUPR are also indicators to assess the GcForest-PPI (Chen et al., 2020).

2.6. Datasets

Nine PPIs benchmark datasets are utilized to test GcForest-PPI model. We select the *S. cerevisiae* and *H. pylori* as training set. *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens* and *Mus musculus* were employed as PPIs independent datasets (Zhou, Gao, & Zheng, 2011). These datasets have been widely used to evaluate the performance of PPIs prediction, such as ACC + SVM (Guo, Yu, Wen, & Li, 2008), Code4 + KNN (Yang, Xia, & Gui, 2010), LD + SVM (Zhou, Gao, & Zheng, 2011), MCD + SVM (You et al., 2014), LRA + RF (You, Li, & Chan, 2017), DeepPPI (Du et al., 2017), DPPI (Hashemifar, Neyshabur, Khan, & Xu, 2018), SVM (Martin, Roe, & Faulon, 2005), Ensemble of HKNN (Nanni & Lumini, 2006), DCT + WSRC (Huang, You, Gao, Wong, & Wang, 2015), MIMI + NMBAC + RF (Ding, Tang, & Guo, 2016), PCA-EELM (You, Lei, Zhu, Xia, & Wang, 2013).

The first set was *S. cerevisiae* from DIP core database (Xenarios et al., 2002). And all protein pairs were identified by the tool CD-HIT (Li, Jaroszewski, & Godzik, 2001). The protein sequences with ⩽50 residues were removed, and sequence similarity ⩾40% were filtered. So golden standard positive (GSP) set includes 5,594 protein pairs, which have been tested for reliability by the expression profile reliability (EPR) and paralogous verification method (PVM) (Deane, Salwinski, Xenarios, & Eisenberg, 2002). A total of 5,594 protein pairs with different subcellular location were selected as golden standard negative (GSN).

The *H. pylori* dataset was validated using the yeast two-hybrid technique (Rain et al., 2001) and built up by Martin et al. (Martin, Roe, & Faulon, 2005), where 1,458 interacting pairs were set as GSP, and 1,458 non-interacting pairs were set as GSN. A one-core network (16 interacting protein pairs) (Yang et al., 2006), a Wnt-related pathway crossover network (96 interacting protein pairs) (Stelzl et al., 2005), and cancer-specific network dataset (108 interacting protein pairs) (Amar, Hait, Izraeli, & Shamir, 2015) were adopted to predict PPIs networks based on GcForest-PPI. The one-core network and Wnt-related pathway cross network have been used in references (Ding, Tang, & Guo, 2016; Lei et al., 2019; Shen et al., 2007), and achieved good performance,

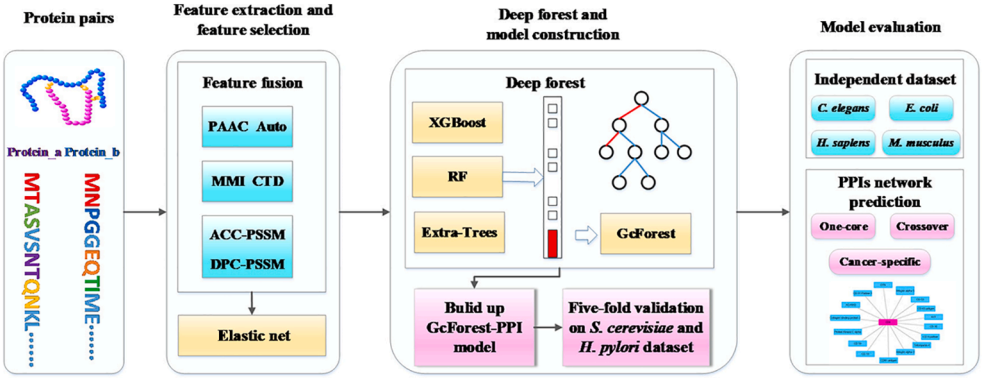

Fig. 2. The overall framework of GcForest-PPI. First, input the protein pairs and utilize PAAC, Auto, MMI, CTD, AAC-PSSM and DPC-PSSM to encode feature values. Then using elastic net to find effective, significant, and valuable feature subset. Finally, the GcForest-PPI model is constructed based on deep forest. The output of GcForest-PPI should decide whether protein pairs are PPIs or non-PPIs.

which can be employed to study the signal pathway. In order to mine the important disease genes in the PPIs network, the cancer-specific network is utilized to perform this study. The information about species, number of proteins and interactions can be obtained from Table S4.

3. Results and discussion

In this section, we illustrate the process of parameter optimization, the superiority of GcForest-PPI and biological significance. (3.1) Parameter selection of the feature extraction. The optimization λ and *lag* of PAAC and Auto. (3.2) Elastic net performs better than other dimensionality reduction method. EN is utilized to eliminate redundant features. (3.3) GcForest performs better than other classifiers. Performing binary PPIs prediction task based on the tree-based deep architecture. (3.4) Comparison with other state-of-the-art PPIs prediction methods. (3.5) Prediction results on four independent species. (3.6) PPIs network prediction. CD9-core, Wnt-related pathway and cancer-specific network are used to study the insight of signaling pathway and pathogenesis of disease.

3.1. Parameter selection of the feature extraction

The parameter λ of PAAC indicates the order information in the coding process. The parameter *lag* represents the interval of two residues in the computational process of Auto. For different λ and *lag* values, deep forest is adopted to construct the predictor. The prediction results are listed in Supplementary Table S5 and S6. The intuitive parameter changes of accuracy are shown in Fig. 3.

As shown in Fig. 3, we can notice that the peaks of the *S. cerevisiae* and *H. pylori* datasets are same for the PAAC. Hence, we determine $\lambda = 11$ in PAAC. For the Auto algorithm, the peak point of *S. cerevisiae* is 11, and the peak point of *H. pylori* is 5. Considering that we use the *S. cerevisiae* dataset as the train set to predict the independent test set, we set $lag = 11$ to unify the parameter *lag*. In order to objectively select the optimal parameter in feature extraction, the mean prediction accuracy (MACC) of two datasets is employed to perform parameter optimization. If we use this indicator, the optimal parameter in Auto descriptor is still $lag = 11$, which is shown Table S7. PAAC and Auto can mine the sequence physicochemical information. MMI and CTD obtain sequence and composition pattern through grouping amino acids. The PSSM can be converted to important evolutionary representation related to PPIs through AAC-PSSM and DPC-PSSM. The protein pair vectors are concatenated to fully characterize pairwise relations whose dimension is 2148. In addition, we compare the feature encoding method of GcForest-PPI with other excellent protein descriptors (Nanni, Lumini, & Brahnam, 2014). The prediction results on two datasets are shown in Supplementary Table S8 and Table S9.

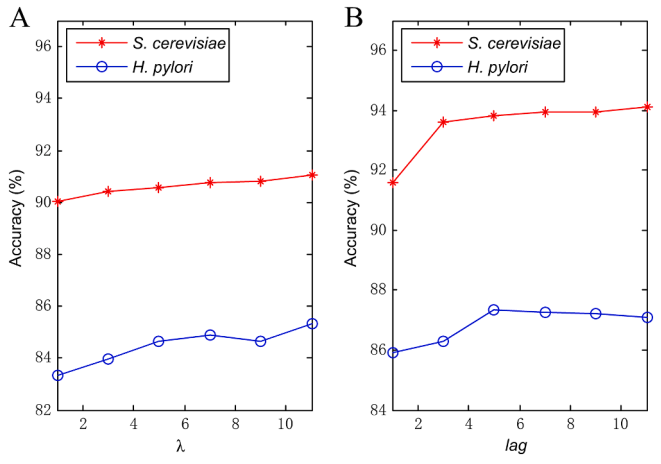

Fig. 3. The parameter optimization of PAAC and Auto for *S. cerevisiae* and *H. pylori*. (A) The λ represents the parameter need to be adjusted in PAAC. (B) The *lag* represents the parameter need to be adjusted in AD.

3.2. Elastic net performs better than other dimension reduction method

The elastic net feature selection was employed to optimize the feature set. The parameters of the elastic net are $\alpha = 0.03$ and $\beta = 0.1$ (Supplementary File S2, Table S10 and S11). The numbers of optimal features are 476 and 516 on *S. cerevisiae* and *H. pylori*, respectively. We also use principal component analysis (PCA) (Wold, Esbensen, & Geladi, 1987), kernel principal component analysis (KPCA) (Schölkopf, Smola, & Müller, 1998), local linear embedding (LLE) (Roweis & Saul, 2000), spectral embedding (SE) (Ng, Jordan, & Weiss, 2002), singular value decomposition (SVD) (Wall, Rechtsteiner, & Rocha, 2002), and semi-supervised dimensionality reduction (SSDR) (Zhang, Zhou, & Chen, 2007) to eliminate redundant information. Then the GcForest-PPI framework is constructed based on deep forest via five-fold cross-validation (Supplementary Table S12). The ROC curves, PR curves and AUC values, AUPR values are shown in Fig. 4, and Supplementary Table S13, respectively.

It is noticed that the accuracy of EN exceeds the PCA, KPCA, LLE, SE, SVD, SSDR (0.9864 vs. 0.9603, 0.9497, 0.9302, 0.9243, 0.9664, 0.8425) for *S. cerevisiae*. EN is 2.61% higher than PCA (0.9864 vs. 0.9603) and 14.39% higher than SSDR (0.9864 vs. 0.8425) from Fig. 4A. From Fig. 4B, we can know that compared with other methods, the robustness of the elastic net is optimal. The AUC value of EN outperforms PCA, KPCA, LLE, SE, SVD, SSDR (0.9816 vs. 0.9545, 0.9402, 0.9088, 0.9172, 0.9605, 0.7999). From the PR curve of Fig. 4C, EN achieves relatively high accuracy compared with PCA, KPCA, LLE, SE, SVD, SSDR (0.9485 vs. 0.9019, 0.9230, 0.8888, 0.8509, 0.9129, 0.8560) in terms of AUPR. Fig. 4D plots the PR curve and the AUPR of EN is 2.4%-11.95% higher than PCA, KPCA, LLE, SE, SVD, SSDR (0.9449 vs. 0.8873, 0.9209, 0.8845, 0.8356, 0.9007, 0.8254) on the *H. pylori*. And the AUPR of EN is 5.76% higher than PCA (0.9449 vs. 0.8873). Therefore, EN can effectively eliminate redundant information that has little correlation with PPIs, and retain important feature subset information, which provides effective feature fusion information for deep forest. Without EN, the feature representation vectors of PPIs contain redundancy and noise information. In this way, the GcForest-PPI model can not perform good prediction accuracy. EN also achieves better prediction performance in other bioinformatics filed, such as the prediction of RNA-binding proteins and protein crotonylation sites (Liu, Yu, Chen, Han, & Yu, 2020; Sun et al., 2020). In order to evaluate the importance of different information, we list the raw features and optimal features (After feature selection) from different feature information (Fig. 5). The comparison between raw features and optimal feature subset indicates the sequence information is more effective than physicochemical and evolutionary information when detecting PPIs.

3.3. GcForest performs better than other classifiers

In order to verify the validity of the improved DF architecture, we list the comparison results between raw DF (Zhou & Feng, 2017, 2019) and the improved DF on *S. cerevisiae*, *H. pylori*, and four independent test sets (Table 1 and 2).

From Table 1, GcForest-PPI (This paper) is 0.54% and 0.82% higher than Zhou & Feng on *S. cerevisiae* and *H. pylori*, respectively. For MCC, this paper achieves better performance than Zhou & Feng on the two training sets. From Table2, we can see that GcForest-PPI (This paper) is 0.07%, 0.64%, 1.09% and 2.94% than GcForest-PPI (Zhou & Feng) on *H. sapiens*, *M. musculus*, *C. elegans*, and *E. coli*, respectively. For *H. Pylori*, *H. Sapiens* and *H. Musculus*, the accuracy difference between this paper and Zhou & Feng is small. Maybe because the sample numbers of these three datasets are much smaller than other datasets. This means the improved GcForest can both obtain good prediction performance on big and small datasets. Especially, for *E. coli*, this paper is about 3% higher

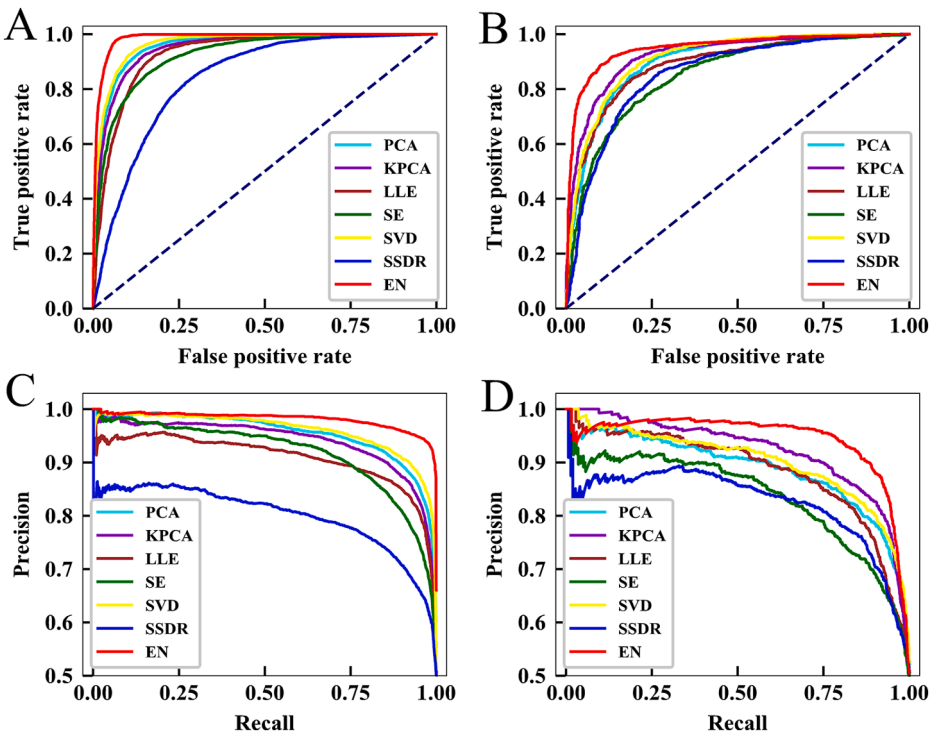

Fig. 4. Predictive performance of PCA, KPCA, LLE, SE, SVD, SSDR and EN via five-fold cross-validation. (A-B) The ROC curves of *S. cerevisiae* and *H. pylori*. (C-D) The PR curves of *S. cerevisiae* and *H. pylori*.

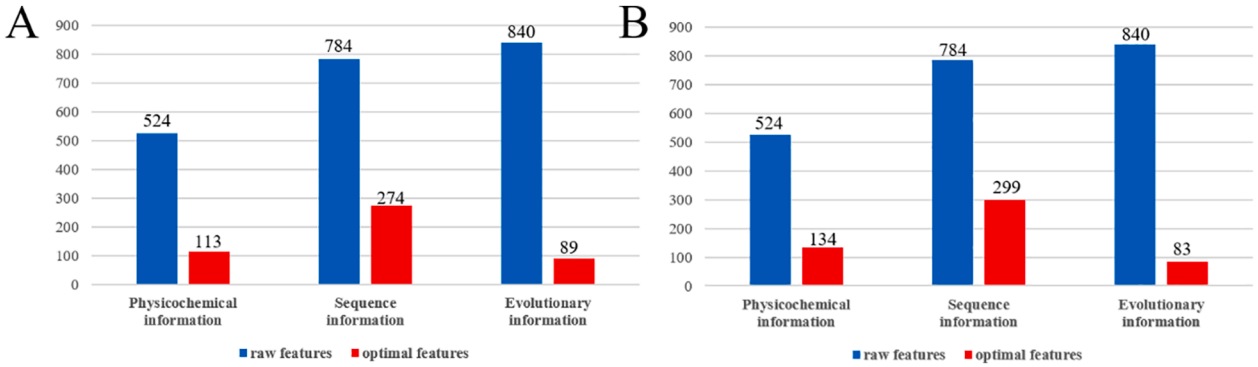

Fig. 5. The raw features and optimal features from different feature information. (A) The raw features and optimal features of *S. cerevisiae* dataset. (B) The raw features and optimal features on *H. pylori* data set.

Table 1
Prediction results between the reference (Zhou & Feng) and this paper on *S. cerevisiae*, *H. pylori* dataset.

Dataset	Methods	ACC (%)	Recall (%)	Precision (%)	MCC
S. cerevisiae	Zhou & Feng	94.90	91.94	97.72	0.8995
	This paper	**95.44**	**92.72**	**98.05**	**0.9102**
H. pylori	Zhou & Feng	88.44	88.67	88.20	0.7694
	This paper	**89.26**	**89.71**	**88.95**	**0.7857**

Table 2
The comparison result of this paper and the reference (Zhou & Feng) using GcForest on independent datasets.

Species	ACC (%)	
	GcForest-PPI (This paper)	GcForest-PPI (Zhou & Feng)
H. sapiens	98.58	98.51
M. musculus	99.04	98.40
C. elegans	96.01	94.92
E. coli	96.30	93.36

than GcForest-PPI (Zhou & Feng). From the above discussion of Table 1 and Table 2, we can conclude that the improved GcForest model is efficient and effective for PPIs prediction through combining XGBoost, RF and Extra-Trees.

For the parameter setting of DF, we choose four XGBoost, four RF and four Extra-Trees models to construct DF architecture. Zhou & Feng used two RF and two Extra-Trees models. But two is small, which could not mine more potential non-linear relationship between raw sequence information and PPIs label. The larger number of base classifiers are needed. However, this change may increase model complexity. We take a trade-off between computational complexity and model accuracy, and we set the number of base classifier as 4. We also set this value as 2, 4, 6 to construct the PPIs predictor. Experimental results indicate the 4 is the best. And the running time is also considerable. We also use logistic regression (LR) (Yu, Huang, & Lin, 2011), Naïve Bayes (NB) (Friedman,

Geiger, & Pazzanzi, 1997), K nearest neighbors (KNN) (Nigsch et al., 2006), AdaBoost (Freund & Schapire, 1997), random forest (RF) (Breiman, 2001) and SVM (Cortes & Vapnik, 1995) six classifiers to predict PPIs (Supplementary Table S14). The results of the ROC curves and the PR curves, AUC and AUPR are shown in Fig. 6, Table S15, respectively.

From Fig. 6A, the ROC curve of DF performs the best on *S. cerevisiae* compared with LR, NB, KNN, AdaBoost, RF and SVM classifiers (0.9864 vs. 0.9298, 0.7914, 0.9503, 0.9750, 0.9762, 0.9653). The AUC of GcForest is increased by 2.11% over SVM (0.9864 vs. 0.9653). From Fig. 6B, the AUC value of GcForest is higher than LR, NB, KNN, AdaBoost, RF, SVM (0.9816 vs. 0.9189, 0.7721, 0.9270, 0.9693, 0.9704, 0.9616). GcForest is 1.12%-20.95% higher than the other six machine-learning-based algorithms. From Fig. 6C, the PR curve indicates that GcForest is superior to LR, NB, KNN, AdaBoost, RF, SVM for predicting PPIs (0.9485 vs. 0.8996, 0.8022, 0.8722, 0.9225, 0.9427, 0.9264) in terms of AUPR on *S. cerevisiae*. GcForest is 0.58%-14.63% higher than the other six classifiers. Fig. 6D indicates the AUPR of GcForest is higher than LR, NB, KNN, AdaBoost, RF, SVM (0.9449 vs. 0.9091, 0.7653, 0.8764, 0.9229, 0.9447, 0.9246). GcForest is 3.58%, 2.03% higher than LR, SVM (0.9449 vs. 0.9091), (0.9449 vs. 0.9246).

We use DF to predict PPIs using XGBoost, random forest and extremely randomized trees to construct cascade forest for the first time. The high-level feature information can be extracted, and probability output from the previous layer is input into the next level. The experimental results show that GcForest is superior to LR, NB, KNN, AdaBoost, RF and SVM classifiers. Tree-ensemble methods can mine the potential feature information of protein interaction pairs through layer-by-layer learning, and thus it can fit the non-linear relationship to determine whether a pair is interacting or non-interacting. DF can have flexible hyperparameter adjustment, high efficiency, and good scalability.

3.4. Comparison with other state-of-the-art PPIs prediction methods

To verify the validity of the GcForest-PPI model, we listed the results of ACC + SVM (Guo, Yu, Wen, & Li, 2008), Code4 + KNN (Yang, Xia, & Gui, 2010), LD + SVM (Zhou, Gao, & Zheng, 2011), MCD + SVM (You et al., 2014), LRA + RF (You, Li, & Chan, 2017), DeepPPI (Du et al., 2017), and DPPI (Hashemifar, Neyshabur, Khan, & Xu, 2018) on *S. cerevisiae* in Table 3. The results of SVM (Martin, Roe, & Faulon, 2005), Ensemble of HKNN (Nanni & Lumini, 2006), DCT + WSRC (Huang, You, Gao, Wong, & Wang, 2015), MCD + SVM (You et al., 2014), MIMI + NMBAC + RF (Ding, Tang, & Guo, 2016), PCA-EELM

Table 3
Comparison of different PPIs prediction methods on *S. cerevisiae* dataset.

Method	ACC (%)	Recall (%)	Precision (%)	MCC
ACC + SVM[a]	89.33 ± 2.67	89.93 ± 3.68	88.87 ± 6.16	N/A
Code4 + KNN[b]	86.15 ± 1.17	81.03 ± 1.74	90.24 ± 1.34	N/A
LD + SVM[c]	88.56 ± 0.33	87.37 ± 0.22	89.50 ± 0.60	0.7715 ± 0.0068
MCD + SVM[d]	91.36 ± 0.36	90.67 ± 0.69	91.94 ± 0.62	0.8421 ± 0.0059
LRA + RF[e]	94.14 ± 1.8	91.22 ± 1.6	97.10 ± 2.1	0.8896 ± 0.026
DeepPPI[f]	94.43 ± 0.30	92.06 ± 0.36	96.65 ± 0.59	0.8897 ± 0.0062
DPPI[g]	94.55	92.24	96.68	N/A
GcForest-PPI	**95.44 ± 0.18**	**92.72 ± 0.44**	**98.05 ± 0.25**	**0.9102 ± 0.0035**

Note: N/A means not available. The values behind ± represent the standard deviation.

[a] Results reported by (Guo, Yu, Wen, & Li, 2008) and the paper uses the holdout validation.

[b] Results reported by (Yang, Xia, & Gui, 2010).

[c] Results reported by (Zhou, Gao, & Zheng, 2011).

[d] Results reported by (You et al., 2014).

[e] Results reported by (You, Li, & Chan, 2017).

[f] Results reported by (Du et al., 2017).

[g] Results reported by (Hashemifar, Neyshabur, Khan, & Xu, 2018). The GcForest-PPI achieves better prediction performance than other state-of-the-art methods. The deep forest can also effectively predict PPIs using tree-based deep architecture.

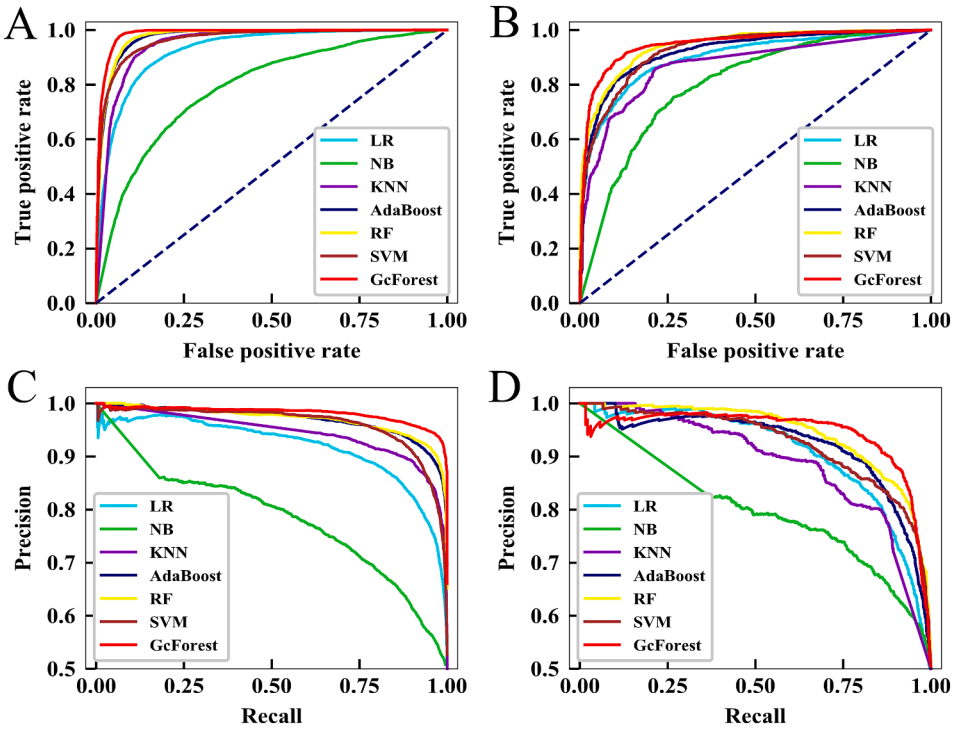

Fig. 6. Predictive performance of LR, NB, KNN, AdaBoost, RF, SVM, and GcForest via five-fold cross-validation. (A-B) The ROC curves illustrating the prediction of *S. cerevisiae* and *H. pylori*. (C-D) The PR curves representing the performance on the *S. cerevisiae* and *H. pylori*.

(You, Lei, Zhu, Xia, & Wang, 2013), DeepPPI (Du et al., 2017) on *H. pylori* are shown in Table 4.

From Table 3, we can see that the GcForest-PPI achieves the best prediction performance. The ACC and Recall based on GcForest-PPI are higher than DPPI, DeepPPI and ACC + SVM (0.89%, 1.01% and 6.11% on ACC) and (0.48%, 0.66% and 2.79% on Reall). DPPI used convolutional neural network, and DeepPPI leveraged deep neural network. The two DL-based framework obtained lower performance than DF-based GcForest-PPI. For Recall, GcForest-PPI is 1.5% higher than LRA + RF whose classifier is traditional tree-based algorithm. On Precision, GcForest-PPI is 9.18% higher than the ACC + SVM, which leverage the popular SVM-based and physicochemical property-based feature encoding method. We can know, GcForest is superior to DL-based, tree-based classifier and SVM, and that means the cascade deep forest is effective.

From Table 4, we can see that the ACC and Recall of GcForest-PPI are superior to PCA-EELM, MIMI + NMBAC + RF, MCD + SVM on the *H. pylori*. Since the three pipelines can not perform the effective feature fusion, and they just used one or two feature encoding algorithms. The ACC of GcForest-PPI is 3.03% and 5.86% higher than DeepPPI and SVM respectively. DeepPPI and SVM tools do not contain the process of dimensional reduction, and the noisy information of the raw feature space may affect the performance. The MCC of GcForest-PPI is superior to the DCT + WSRC and MCD + SVM. The superiority indicate that deep structure can obtain better results than shallow structure. In conclusion, deep forest combined with elastic net can effectively predict PPIs.

3.5. Prediction results on four independent species

The pros and cons of GcForest-PPI are further evaluated on *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus*, and the numbers of interacting pairs can be obtained in Supplementary Table S4. Then the whole samples of the *S. cerevisiae* are regarded as training set. The results of GcForest-PPI and DPPI (Hashemifar, Neyshabur, Khan, & Xu, 2018), DeepPPI (Du et al., 2017), MLD + RF (You, Chan, & Hu, 2015), DCT + WSRC (Huang et al., 2015) are shown in Table 5.

From Table 5, we can know that GcForest-PPI performs better than DeepPPI, MLD + RF, and DCT + WSRC with the ACC values of 98.58%, 99.04% and 96.01%, especially on *H. sapiens* and *M. musculus*. This also demonstrates that the combination of cascade structure and elastic net can characterize PPIs information using *S. cerevisiae* dataset. In other words, it is possible that PPIs of one species can predict cross-species and the co-evolved relationship can be mined via cascade structure based on XGBoost, RF and Extra-Trees.

Table 4
Comparison of different PPIs prediction methods on *H. pylori* dataset.

Method	ACC (%)	Recall (%)	Precision (%)	MCC
SVM[a]	83.40	79.90	85.70	N/A
Ensemble of HKNN[b]	86.60	86.70	85.00	N/A
DCT + WSRC[c]	86.74	86.43	87.01	0.7699
MCD + SVM[d]	84.91	83.24	86.12	0.7440
MIMI + NMBAC + RF[e]	87.59	86.81	88.23	0.7524
PCA-EELM[f]	87.50	88.95	86.15	0.7813
DeepPPI[g]	86.23	89.44	84.32	0.7263
GcForest-PPI	**89.26 ± 1.07**	**89.71 ± 2.26**	**88.95 ± 1.36**	**0.7857 ± 0.0212**

Note: N/A means not available. The values behind ± represent the standard deviation.

[a] Results reported by (Martin, Roe, & Faulon, 2005) and this paper uses ten-fold cross-validation.

[b] Results reported by (Nanni & Lumini, 2006) and this paper uses ten-fold cross-validation.

[c] Results reported by (Huang, You, Gao, Wong, & Wang, 2015), and this paper used ten-fold cross-validation.

[d] Results reported by (You et al., 2014).

[e] Results reported by (Ding, Tang, & Guo, 2016).

[f] Results reported by (You, Lei, Zhu, Xia, & Wang, 2013).

[g] Results reported by (Du et al., 2017). The GcForest-PPI achieves better prediction performance than other state-of-the-art methods. The deep forest combined with elastic net can also effectively predict PPIs.

Table 5
Comparison of performance of the proposed method with other state-of-the-art predictors on the independent dataset.

Species	ACC (%)				
	GcForest-PPI	DPPI[a]	DeepPPI[b]	MLD+RF[c]	DCT+WSR[d]
H. sapiens	**98.58**	96.24	94.84	94.19	82.22
M. musculus	**99.04**	95.84	92.19	91.96	79.87
C. elegans	**96.01**	95.51	93.77	87.71	81.19
E. coli	96.30	**96.66**	91.37	89.30	66.08

Note:

[a] Results reported by (Hashemifar, Neyshabur, Khan, & Xu, 2018).

[b] Results reported by (Du et al., 2017).

[c] Results reported by (You, Chan, & Hu, 2015).

[d] Results reported by (Huang, You, Gao, Wong, & Wang, 2015). GcForest-PPI achieves the highest accuracy on *H. sapiens*, *M. musculus* and *C. elegans*.

3.6. PPIs network prediction

We use the one-core network, Wnt-related pathway network and cancer-specific network to evaluate the pipeline of GcForest-PPI. It provides some reference for identifying PPIs from unknown PPIs networks. The one-core network is a simple CD9-core network including 17 genes. The second is a crossover network for Wnt-related pathway. This network has 78 genes consisting of 96 PPI pairs. The cancer-specific network (Amar, Hait, Izraeli, & Shamir, 2015) consists of 78 genes, which are of importance in DNA replication and cancer pathways. The interaction pairs in the cancer-specific network are derived from the IntAct database (Kerrien et al., 2007).

For the task of PPIs network prediction, the *S. cerevisiae* is used to build GcForest-PPI model with the optimal parameters. The input of task should be the real protein–protein interaction pairs that have been experimentally verified, and the output values are prediction scores to determine whether the given pairs interact or not. In other words, GcForest-PPI model can output whether the given PPIs are successfully predicted or not. The detained process for PPIs network prediction can be described as: (*i*) The protein pairs are converted to 2,148-dimensional feature vector by PAAC, Auto, MMI, CTD, AAC-PSSM, and DPC-PSSM (where λ is 11 in PAAC and *lag* is 11 in Auto). (*ii*) We select 476 important features via elastic net ($\alpha = 0.03$ and $\beta = 0.1$). (*iii*) Deep forest-based model GcForest-PPI using random forest, Extra-trees and XGBoost is constructed. The results of three types PPIs networks are shown in Figs. 7, 8, and 9.

As shown in Fig. 7, when using the GcForest-PPI model to predict a one-core network, all PPIs of the network are predicted successfully (16/16). The accuracy of GcForest-PPI is superior to Shen et al. (Shen et al., 2007) and Ding et al. (Ding, Tang, & Guo, 2016) (100% vs. 81.25%, 87.50%). CD9 plays a crucial role in sperm egg fusion, and myoblast fusion (Yang et al., 2006). The palmitoylation of CD9 contributed to the interaction between CD81 and CD53 (Charrin et al., 2002).

From Fig. 8, we successfully predicted interacting protein pairs with accuracy of 97.92% on crossover network (94/96). GcForest-PPI is 21.88% higher than Shen et al. (Shen et al., 2007) (97.92% vs. 76.04%). GcForest-PPI is 3.13% higher than that of Ding et al. (Ding, Tang, & Guo, 2016) (97.92% vs. 94.79%). However, the relationship ROCK1-CRMP1 is not identified successfully. Maybe because ROCK1-CRMP1 is the part of non-canonical Wnt pathway, and GcForest-PPI is not suitable in this case. AXIN1 plays an important role in multiple pathways (Luo & Lin, 2004). GcForest-PPI can truly identify the interactions between AXIN1 and its neighboring proteins.

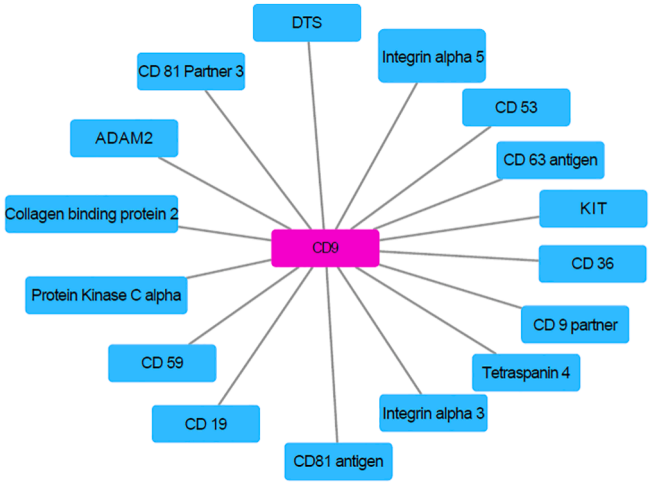

Fig. 7. Predicted results on the three types PPIs networks. Predicted results of PPIs networks of a one-core network for CD9. All 16 PPIs are truly predicted.

All PPIs in the cancer-specific network are successfully predicted (108/108). The cancer-specific network is composed of two sub-networks (Fig. 9). The first sub-network is composed of 14 genes, where TP53 is the main hub. At the molecular level, TP53 is a gene associated with breast cancer (Andrysik et al., 2017). The second sub-network is a PPIs network consisting of 64 genes, where two down-regulated genes NDEL1 and GABARAPL1 link to two sub-modules (Wynne & Vallee, 2018). NDEL1 and LIS1 are essential for development, and they are thought to relate with cytoplasmic dynein (Hebbar et al., 2008). NDEL1 contains phosphorylation sites for CDK1, CDK5 (Mori et al., 2007). CDK5 phosphorylation of NDEL1 affects lysosome motility in axons, indicating CDK5 is important in cell growth and development (Klinman & Holzbaur, 2015; Pandey & Smith, 2011). NDEL1 is also closely related to the development of some diseases (Doobin, Kemal, Dantas, & Vallee, 2016). All PPIs are predicted successfully on the cancer-specific network, indicating that the GcForest-PPI can provide new ideas to elucidate disease mechanisms, and design of new drugs.

4. Conclusion

With the rapid development of big data mining technology, the study of well-established computational predictive framework based on proteomics data is necessary. Using machine learning to automatically predict PPIs can provide reference for grasping disease pathogenesis, drug discovery and repositioning. We present a novel approach Gcforest-PPI for identifying PPIs, which uses PAAC, Auto, MMI, CTD, AAC-PSSM and DPC-PSSM to extract physicochemical features, sequence features and evolutionary features of PPIs. Then we use the elastic net to eliminate noise from extracted vectors, which could combine the advantages of L1 and L2 regularization and generate a sparse model and group effects. The comparison between raw features and optimal feature subset indicates the sequence information is more effective than

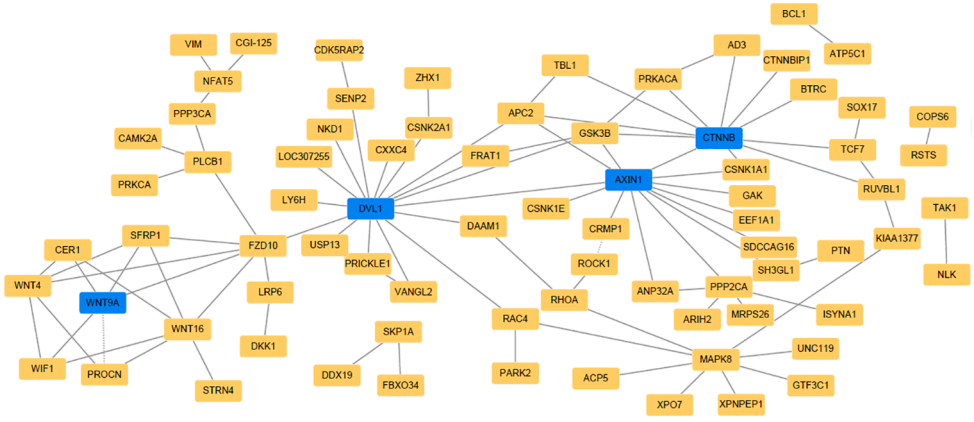

Fig. 8. Pedicted results of a crossover network, where WNT9A, CXXC4, AXIN1 and ANP32A are linked in the Wnt-related pathway. The solid lines and dotted lines represent the interactions of true prediction and false prediction, respectively.

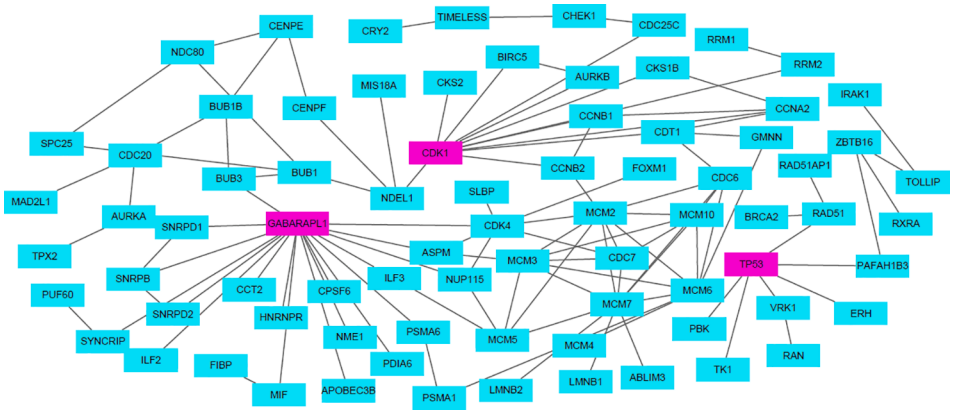

Fig. 9. Predicted results of PPIs networks of the cancer-specific differential genes. The network is composed of two components. The first component is marked in red and the second component is marked in blue. NDEL1 and GABARAPL1 connect the first component. TP53 is the main hub in the second component of this network. All 108 PPIs are truly predicted.

physicochemical and evolutionary information when detecting PPIs. In this paper, we develop a new deep forest architecture to implement GcForest, in which all levels (except the last one) of the cascade are composed of four XGBoost, four RF and four Extra-Trees. XGBoost is an excellent variant of gradient boosting decision tree whose base classifier is regression tree. The base classifiers of the RF and Extra-Trees are decision tree. The constructed tree-based DF contains good and diverse base classifier, which could obtain complementary advantages and essential features. In this way, the improved GcForest can achieve good prediction performance and generalization ability. The results of *S. cerevisiae* and *H. pylori* indicate that GcForest-PPI can effectively identify PPIs. The prediction results of *C. elegans*, *E. coli*, *H. sapiens*, and *M. musculus* show that GcForest-PPI is capable of cross-species prediction and PPIs in *S. cerevisiae* include representation information of other species. Finally, the satisfactory scalability of the model is demonstrated by the one-core network, crossover network and cancer-specific network dataset, which can provide new ideas for exploring disease pathogenesis. In summary, GcForest-PPI can be a useful predictive tool for bioinformatics and proteomics.

Feature extraction from protein sequences is a key step based on machine learning. Although we combine the physicochemical and position information, sequence and composition information, and evolutionary information from primary interacting protein pairs, the comprehensive important features related to PPIs is still not elucidated. We are developing a python tool for effective multi-information fusion to provide an online platform for further studying the biological function of PPIs. DL has powerful representation learning ability and can mine more abstract essential features. Capsule neural network is a new deep learning framework. How to use capsule neural network is the next research direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank anonymous reviewers for valuable suggestions and comments. This work was supported by the National Natural Science Foundation of China (No. 61863010), the Key Research and Development Program of Shandong Province of China (No. 2019GGX101001), and the Key Laboratory Open Foundation of Hainan Province (No. JSKX202001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2021.114876.

References

Alberts, B. (1998). The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell, 92*(3), 291–294.
Amar, D., Hait, T., Izraeli, S., & Shamir, R. (2015). Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Research, 43*(16), 7779–7789.
Andrysik, Z., Galbraith, M. D., Guarnieri, A. L., Zaccara, S., Sullivan, K. D., Pandey, A., et al. (2017). Identification of a core tp53 transcriptional program with highly distributed tumor suppressive activity. *Genome Research, 27*(10), 1645–1657.
Breiman, L. (2001). Random forest. *Machine Learning, 45*, 5–32.
Charrin, S., Manié, S., Oualid, M., Billard, M., Boucheix, C., & Rubinstein, E. (2002). Differential stability of tetraspanin/tetraspanin interactions: Role of palmitoylation. *FEBS Letters, 516*, 139–144.
Chen, C., Zhang, Q., Ma, Q., & Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemometrics and Intelligent Laboratory Systems, 191*, 54–64.
Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., et al. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Computers in Biology and Medicine, 123*, 103899. https://doi.org/10.1016/j.compbiomed.2020.103899
Chen, M., Ju, C. J. T., Zhou, G., Chen, X., Zhang, T., Chang, K. W., et al. (2019). Multifaceted protein-protein interaction prediction based on siamese residual RCNN. *Bioinformatics, 35*, i305–i314.
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics, 34*, 2499–2502.
Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *PROTEINS: Structure Function, and Genetics, 43*(3), 246–255.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
Cui, X., Yu, Z., Yu, B., Wang, M., Tian, B., & Ma, Q. (2019). UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. *Chemometrics and Intelligent Laboratory Systems, 184*, 28–43.
Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240).
Deane, C. M., Salwiński, Ł., Xenarios, I., & Eisenberg, D. (2002). Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics, 1*(5), 349–356.
Deng, L., Zhang, Q. C., Chen, Z., Meng, Y., Guan, J., & Zhou, S. (2014). Predhs: a web server for predicting protein-protein interaction hot spots by using structural neighborhood properties. *Nucleic Acids Research, 42*, W290–W295.
Ding, Y., Tang, J., & Guo, F. (2016). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinformatics, 17*, 398.
Ding, Y., Tang, J., & Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Information Science, 418-419*, 546–560.
Doobin, D. J., Kemal, S., Dantas, T. J., & Vallee, R. B. (2016). Severe nde1-mediated microcephaly results from neural progenitor cell cycle arrests at multiple specific stages. *Nature Communications, 7*, 12551.
Du, X., Sun, S., Hu, C., Yao, Y.u., Yan, Y., & Zhang, Y. (2017). DeepPPI: Boosting prediction of protein-protein interactions with deep neural networks. *Journal of Chemical Information and Modeling, 57*(6), 1499–1510.
Feng, J., Yu, Y., & Zhou, Z. H. (2018). Multi-layered gradient boosting decision trees. In *International Conference on Neural Information Processing Systems* (pp. 3555–3565).
Feng, J., & Zhou, Z. H. (2017). Autoencoder by forest. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (pp. 2967–2973).
Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.
Friedman, N., Geiger, D., & Pazzanzi, M. (1997). Bayesian network classifiers. *Machine Learning, 2*, 131–163.
Gastwirt, J. (1972). The estimation of lorenz curve and gini index. *The Review of Economics and Statistics, 54*, 306–316.
Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.
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 Research, 36*, 3025–3030.
Hashemifar, S., Neyshabur, B., Khan, A. A., & Xu, J. B. (2018). Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics, 34*, i802–i810.
Hebbar, S., Mesngon, M. T., Guillotte, A. M., Desai, B., Ayala, R., & Smith, D. S. (2008). Lis1 and Ndel1 influence the timing of nuclear envelope breakdown in neural stem cells. *Journal of Cell Biology, 182*, 1063–1071.
Huang, Y. A., You, Z. H., Gao, X., Wong, L., & Wang, L. (2015). Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. *Biomed Research International, 2015*, Article 902198.
Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., et al. (2007). IntAct-open source resource for molecular interaction data. *Nucleic Acids Research, 35*(Database), D561–D565.
Klinman, E., & Holzbaur, E. F. (2015). Stress-induced cdk5 activation disrupts axonal transport via lis1/ndel1/dynein. *Cell Reports, 12*(3), 462–473.
Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., et al. (2019). Network-based prediction of protein interactions. *Nature Communications, 10*, Article 1240. https://doi.org/10.1038/s41467-019-09177-y
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems* (pp. 1097–1105).
Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
Lei, H., Wen, Y., You, Z., Elazab, A., Tan, E.-L., Zhao, Y., et al. (2019). Protein-protein interactions prediction via multimodal deep polynomial network and regularized extreme learning machine. *IEEE Journal of Biomedical and Health Informatics, 23*(3), 1290–1303.
Li, W., Jaroszewski, L., & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics, 17*(3), 282–283.
Lian, X., Yang, S., Li, H., Fu, C., & Zhang, Z. (2019). Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. *Journal of Proteome Research, 18*(5), 2195–2205.

Liu, Y., Yu, Z., Chen, C., Han, Y.u., & Yu, B. (2020). Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net. *Analytical Biochemistry, 609*, Article 113903. https://doi.org/10.1016/j.ab.2020.113903

Luo, W., & Lin, S.-C. (2004). Axin: A master scaffold for multiple signaling pathways. *Neurosignals, 13*(3), 99–113.

Martin, S., Roe, D., & Faulon, J.-L. (2005). Predicting protein-protein interactions using signature products. *Bioinformatics, 21*(2), 218–226.

Mori, D., Yano, Y., Toyo-oka, K., Yoshida, N., Yamada, M., Muramatsu, M., et al. (2007). NDEL1 phosphorylation by Aurora-A kinase is essential for centrosomal maturation, separation, and TACC3 recruitment. *Molecular and Cellular Biology, 27*(1), 352–367.

Nanni, L., & Lumini, A. (2006). An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics, 22*(10), 1207–1210.

Nanni, L., Lumini, A., & Brahnam, S. (2014). An Empirical study of different approaches for protein classification. *Scientific World Journal, 2014*, 1–17.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *International Conference on Neural Information Processing Systems* (pp. 849–856).

Nigsch, F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E., & Mitchell, J. B. O. (2006). Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *Journal of Chemical Information Modeling, 46*(6), 2412–2422.

Pandey, J. P., & Smith, D. S. (2011). A Cdk5-dependent switch regulates Lis1/ Ndel1/ dynein-driven organelle transport in adult axons. *Journal of Neuroscience, 31*(47), 17207–17219.

Peri, S. J., Navarro, D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., et al. (2003). Development of human protein reference database as aninitial platform for approaching systems biology in humans. *Genome Research, 13*, 2363–2371.

Qiu, W., Li, S., Cui, X., Yu, Z., Wang, M., Du, J., et al. (2018). Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. *Journal of Theoretical Biology, 450*, 86–103.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., et al. (2001). The protein-protein interaction map of helicobacter pylori. *Nature, 409*(6817), 211–215.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*, 2323–2326.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature, 461*(7261), 218–223.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America, 104*(11), 4337–4341.

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., & Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics, 111*(6), 1839–1852.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science, 362*(6419), 1140–1144.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, arXiv: 1409.1556v6.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell, 122*(6), 957–968.

Sun, X., Jin, T., Chen, C., Cui, X., Ma, Q., & Yu, B. (2020). RBPro-RF: Use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net. *Chemometrics and Intelligent Laboratory Systems, 197*, 103919. https://doi.org/10.1016/j.chemolab.2019.103919

Tian, B., Wu, X., Chen, C., Qiu, W., Ma, Q., & Yu, B. (2019). Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. *Journal of Theoretical Biology, 462*, 329–346.

Wall, M. E, Rechtsteiner, A., & Rocha, L. M. (2002). Singular value decomposition and principal component analysis. In: A Practical Approach to Microarray Data Analysis, pp. 91–109.

Wang, X. Y., Yu, B., Ma, A., Chen, C., Liu, B. Q., & Ma, Q. (2019). Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics, 35*, 2395–2402.

Wei, Z. S., Han, K.e., Yang, J. Y., Shen, H. B., & Yu, D. J. (2016). Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing, 193*, 201–212.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1-3), 37–52.

Wynne, C. L., & Vallee, R. B. (2018). Cdk1 phosphorylation of the dynein adapter nde1 controls cargo binding from g2 to anaphase. *The Journal of Cell Biology, 217*, 3019–3029.

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). The Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research, 30*, 303–305.

Yadav, S., Ekbal, A., Saha, S., Kumar, A., & Bhattacharyya, P. (2019). Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein-protein interaction. *Knowledge-Based Systems, 166*, 18–29.

Yang, L., Xia, J. F., & Gui, J. (2010). Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein and Peptide Letters, 17*, 1085–1090.

Yang, X. H., Kovalenko, O. V., Kolesnikova, T. V., Andzelm, M. M., Rubinstein, E., Strominger, J. L., et al. (2006). Contrasting effects of EWI proteins, integrins, and protein palmitoylation on cell surface CD9 organization. *The Journal of Biological Chemistry, 281*(18), 12976–12985.

You, Z. H., Chan, K. C. C., & Hu, P. (2015). Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS One, 10*.

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 Bioinformatics, 14*, S10.

You, Z.-H., Li, X., & Chan, K. CC. (2017). An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers. *Neurocomputing, 228*, 277–282.

You, Z. H., Zhu, L., Zheng, C. H., Yu, H. J., Deng, S. P., & Ji, Z. (2014). Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics, 15*(Suppl 15), S9. https://doi.org/10.1186/1471-2105-15-S15-S9

Yu, B., Chen, C., Zhou, H., Liu, B., & Ma, Q. (2021). GTB-PPI: Predict protein-protein interactions based on L1-regularized logistic regression and gradient tree boosting. *Genomics, Proteomics & Bioinformatics.*. https://doi.org/10.1016/j.gpb.2021.01.001

Yu, B., Li, S., Chen, C., Xu, J., Qiu, W., Wu, X., et al. (2017). Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition. *Chemometrics and Intelligent Laboratory Systems, 167*, 102–112.

Yu, B., Li, S., Qiu, W., Wang, M., Du, J., Zhang, Y., et al. (2018). Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. *BMC Genomics, 19*(1). https://doi.org/10.1186/s12864-018-4849-9

Yu, H. F., Huang, F. L., & Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning, 85*(1–2), 41–75.

Yu, J., Vavrusa, M., Andreani, J., Rey, J., Tufféry, P., & Guerois, R. (2016). Interevdock: A docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Research, 44*(W1), W542–W549.

Zahiri, J., Yaghoubi, O., Mohammad-Noori, M., Ebrahimpour, R., & Masoudi-Nejad, A. (2013). PPIevo: Protein-protein interaction prediction from PSSM based evolutionary information. *Genomics, 102*(4), 237–242.

Zhang, B., Li, J., Quan, L., Chen, Y.u., & Lü, Q. (2019). Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing, 357*, 86–100.

Zhang, D., Zhou, Z. H., & Chen, S. (2007). Semi-supervised dimensionality reduction. In *SIAM Conference on Data Mining* (pp. 629–634).

Zhang, L., Yu, G., Xia, D., & Wang, J. (2019). Protein-protein interactions prediction based on ensemble deep neural networks. *Neurocomputing, 324*, 10–19.

Zhang, S.-B., & Tang, Q.-R. (2016). Protein-protein interaction inference based on semantic similarity of gene ontology terms. *Journal of Theoretical Biology, 401*, 30–37.

Zhang, Q., Zhang, Y., Li, S., Han, Y., Jin, S., Gu, H., et al. (2021). Accurate prediction of multi-label protein subcellular localization through multi-view feature learning with RBRL classifier. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbab012

Zhou, Y. Z., Gao, Y., & Zheng, Y. Y. (2011). Prediction of protein-protein interactions using local description of amino acid sequence. *Communications in Computer and Information Science, 202*, 254–262.

Zhou, Z. H., & Feng, J. (2017). Deep forest: Towards an alternative to deep neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 3553–3559).

Zhou, Z. H., & Feng, J. (2019). Deep forest. *National Science Review, 6*, 74–86.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, 67*(2), 301–320.