Chemometrics and Intelligent Laboratory Systems 191 (2019) 54–64

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion

Cheng Chen [a,b,1], Qingmei Zhang [a,b,1], Qin Ma [c], Bin Yu [a,b,d,e,*]

[a] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China
[b] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao, 266061, China
[c] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA
[d] School of Life Sciences, University of Science and Technology of China, Hefei, 230027, China
[e] School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha, 410114, China

ARTICLE INFO

Keywords:
Protein-protein interactions
Multi-information fusion
Elastic net
LightGBM

ABSTRACT

Protein-protein interactions (PPIs) play an important role in cell life activities such as transcriptional regulation, signal transduction and drug signal transduction. The study of PPIs has become a research hotspot in bioinformatics. However, the identification of PPIs using experimental methods is time-consuming and costly. PPIs prediction based on machine learning is very important. This paper proposes a new protein-protein interactions prediction method called LightGBM-PPI. First, pseudo amino acid composition, autocorrelation descriptor, local descriptor, conjoint triad are employed to extract feature information. Secondly, we use the elastic net to select the optimal feature subset and eliminate redundant features. Finally, the LightGBM is employed as the classifier to predict PPIs and the LightGBM-PPI model is built up. Five-fold cross-validation shows that the prediction accuracy of the *Helicobacter pylori* and *Saccharomyces cerevisiae* datasets are 89.03% and 95.07%, respectively. The prediction accuracy of *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens* and *Mus musculus* are 90.16%, 92.16%, 94.83% and 94.57%, respectively, which are superior to the state-of-the-art prediction methods. To further evaluate the advantages and disadvantages of the model, we use one-core network and the crossover network for the Wnt-related pathway to predict PPIs, which can provide new ideas for drug design and disease prevention. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/LightGBM-PPI/.

1. Introduction

Protein-protein interactions (PPIs) play an important role in cell life activities, including DNA synthesis, gene transcription and translation, immune response, enzyme catalysis, and signal transduction [1]. The study of protein-protein interactions can explore and elucidate the mechanisms of related proteins in cells. In addition, it also can help to provide new ideas in the functional module identification [2], disease gene ranking [3], understanding disease mechanisms [4], new drug development and human disease prevention and treatment. However, the traditional experimental methods are time-consuming and costly, and there is a problem that the generalization ability is weak and the false positive rate is high. The machine learning methods are urgently required to predict the PPIs.

As we can know, the prediction of protein 3D structures and their ligand compounds is critical for rational drug design. Experiments for determining the 3D structure are time-consuming and expensive [5,6]. With the development of high-throughput sequencing technology and the rapid development of computational biology, many bioinformatics tools have been developed [7,8], including protein post-translational modification sites [9], drug-target interactions [10], sigma-54 promoter predictions [11] and DNA-methylation sites [12]. Inspired by this, it is particularly important to predict PPIs based on computational methods.

The primary work of predicting PPIs using machine learning is to extract the feature information of protein sequences, which converts inconsistent protein sequence strings into numerical vectors [13]. Because machine learning methods only handle numerical vectors of uniform length. Chou's PseAAC [14] can effectively mine the sequence information of protein sequences, and has been widely used in the field of

* Corresponding author. College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.chemolab.2019.06.003
Received 13 April 2019; Received in revised form 20 May 2019; Accepted 7 June 2019
Available online 8 June 2019

bioinformatics [15–17]. At the same time, four powerful softwares 'PseAAC' [18], 'PseAAC-Builder' [19], 'propy' [20] and 'PseAAC-General' [21] can effectively convert protein character signals into numerical signals. These software packages can not only obtain the PseAAC information, but also generate the domain, gene ontology, and evolutionary information of protein sequences [22]. Recently, a more powerful web server Pse-in-One [23] and its updated version Pse-in-One 2.0 [24] have been developed to extract useful feature information from proteins, DNA and RNA sequences and provide a friendly visual interface.

Du et al. [25] proposed DeepPPI, i.e. a PPIs prediction method based on deep neural networks. They used amino acid composition, dipeptide composition, composition, transition and distribution, quasi-sequence-order descriptors, and amphiphilic pseudo amino acids composition for feature extraction. The *Helicobacter pylori* (*H. pylori*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) datasets were employed to predict PPIs with the prediction accuracy of 86.23% and 94.43%, respectively. Tian et al. [26] used pseudo amino acid composition, auto covariance descriptor and group weight coding to obtain protein sequence information. And the wavelet denoising was used to remove the redundant information after feature fusion. The support vector machine was employed as a classifier on the *H. pylori* and *S. cerevisiae* data sets. Li et al. [27] generated feature vectors using auto covariance descriptors and conjoint triad method to predict PPIs based on deep learning. Göktepe et al. [28] extracted features from protein sequences using Bi-Gram representations, pseudo-amino acid composition and weighted skip-sequential conjoint triad. The principal component analysis is employed to eliminate redundant and noise information.

In addition to the feature extraction method, the selection of the classifier is critical to predict PPIs. Common classifiers include support vector machine [29–33], decision tree [34], logistic regression [35] and random forest [36–38]. Shen et al. [39] proposed a new PPIs prediction method using conjoint triad and support vector machine to predict PPIs. Pan et al. [40] proposed a new LDA-RF model, which used conjoint triad to represent the local sequential features, and random forest was employed as a classifier. Hashemifar et al. [41] proposed a sequence-based deep learning prediction method DPPI, which predicted PPIs by convolutional neural networks and random projection module.

Although a number of PPIs prediction methods based on machine learning are proposed, there is still much room for improvement based on machine learning. First, the feature information of protein sequences is not clearly elucidated for predicting the effects of protein-protein interactions. Secondly, the multi-information fusion will bring redundancy and noise information. How to choose the appropriate dimensionality reduction method for feature selection is urgently needed. Finally, the development of experimental techniques produces a large amount of PPIs data. How to make full use of experimental data to design effective prediction algorithms is very necessary.

Inspired by the above discussion this paper proposes a protein-protein interactions prediction method LightGBM-PPI. The sequence information and physicochemical information of the PPIs are characterized by pseudo-amino acid composition, autocorrelation descriptor, conjoint triad and local descriptor. In order to perform effective feature fusion and eliminate redundant and noise information in the initial feature vector, we use the elastic net to determine the important subset of features. In terms of PPIs prediction, we first use LightGBM to construct a PPIs prediction model. Five-fold cross-validation shows that the prediction performance of LightGBM is better than support vector machine, random forest, Naïve Bayes and k-nearest neighbors on *H. pylori* and *S. cerevisiae* datasets. Cross-species predictions are performed using *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens*, and *Mus musculus* with the prediction accuracy of 90.16%, 92.16%, 94.83%, and 94.57%, respectively. To further evaluate the merits of the LightGBM-PPI model, we verify the validity of the method on the one-core network and the Wnt-related pathway cross-network dataset. The experimental results show that our method can effectively predict the PPIs of single-core network and signaling pathway network.

The recent publications indicate [42,43], Chou's 5-step rule is of great importance, and the details are as follows: (i) construct a benchmark dataset to train and test the predictor; (ii) using effective mathematical formulas to characterize the protein sequence to reflect the correlation between the sample data and label class; (iii) introduce or develop an effective algorithm to complete the prediction; (iv) use cross-validation methods to objectively assess the prediction accuracy; (v) establish the user-friendly web server. Below, we are to describe how to deal with these steps one by one.

2. Materials and methods

2.1. Dataset

We used eight benchmark datasets to evaluate the pros and cons of the predictive model. The first dataset was the *S. cerevisiae* constructed by Guo et al. [44], and had 11,888 protein pairs, including 5594 interacting pairs and 5594 non-interacting pairs. In the dataset, the length of protein sequence is greater than 50 or equal to 50, and the protein pairs have less than 40% sequence identity [45]. The second was the *H. pylori* constructed by Martin et al. [46], containing 2,916 protein pairs (1,458 interacting pairs and 1,458 non-interacting pairs). In addition, the four datasets constructed by Zhou et al. [47] were selected as independent test datasets to test the validation of the prediction model. They were *Caenorhabditis elegans* (4,013 PPIs pairs), *Escherichia coli* (6,954 PPIs pairs), *Homo sapiens* (1,412 PPIs pairs) and *Mus musculus* datasets (313 PPIs pairs). Finally, we also used two PPIs network datasets to verify the validity of the model, the first was one-core network [48], including 16 pairs of PPIs, and the second dataset was the crossover network for the Wnt-related pathway [49], including 96 pairs of PPIs.

2.2. Feature extraction

2.2.1. Pseudo amino acid composition

Pseudo amino acid composition (PseAAC) can represent amino acid composition information and amino acid order information. It has been widely used in the field of bioinformatics, including protein site prediction [50], protein subcellular localization prediction [51–53], protein structure class prediction [54], protein submitochondrial prediction [55] and so on.

Let the feature vector of the PseAAC be:

$$X = [x_1, x_2, \cdots, x_{19}, x_{20}, x_{20+1}, \cdots, x_{20+\lambda}]^T (\lambda < L) \tag{1}$$

where L represents the length of the protein sequence, and the calculation equation of the first 20 elements and the latter λ element in the vector is as shown in equation (2):

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (1 \le u \le 20) \\ \dfrac{\omega \theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \theta_j}, & (20+1 \le u \le 20+\lambda) \end{cases} \tag{2}$$

In equation (2), f_i is the frequency of occurrence of the number of amino acids in the protein after normalization, θ_j is $j-$ th layer sequence correlation factor, ω is a weight factor. PseAAC can generate $20+\lambda$ dimensional feature vectors to characterize PPIs.

2.2.2. Autocorrelation descriptor

When using autocorrelation descriptor (AD) [56] for feature extraction, the length of the numerical sequence should be unified, and the influence of long range and short range interactions can also be considered. The autocorrelation descriptor can obtain the physicochemical

information contained in the protein sequence, which improves the accuracy of PPIs to some extent. The seven physicochemical properties for the 20 native amino acids are shown in Supplementary materials Table S1. Three autocorrelation descriptors are used in this paper.

(a) Moreau-Broto autocorrelation descriptor is defined as:

$$NMBA(l) = \frac{NBA(l)}{N-l}, \quad l = 1, 2, \cdots, lag \tag{3}$$

where $MBA(l) = \sum_{i=1}^{N-l} P(AA_i)P(AA_{i+l})$, AA_i and AA_{i+l} are $i-th$ and $i+l-th$ amid acid. $P(AA_i)$ and $P(AA_{i+l})$ represent $i-th$ and $i+l-th$ physicochemical property, and lag represents the interval between two amid acid.

(b) Moran autocorrelation descriptor is defined as:

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

where $\overline{p}$ represents the average value of the physicochemical property of each amino acid.

(c) Geary autocorrelation descriptor is defined as:

$$GA(l) = \frac{\frac{1}{2(N-l)}\sum_{i=1}^{N-l}(P(AA_i)-P(AA_{i+l}))^2}{\frac{1}{N}\sum_{i=1}^{N}(P(AA_i)-\overline{P})^2}, \quad l = 1, 2, \cdots, lag \tag{5}$$

It can be known from the definitions of the above three autocorrelation descriptors that we can obtain $3 \times n \times lag$ dimensional feature vector (n represents the number of kinds of the amino acid).

2.2.3. Conjoint triad

Electrostatic interactions (including interactions between hydrogen bonds) and hydrophobic interactions can have an effect on protein-protein interactions, but electrostatic and hydrophobic interactions are affected by the dipoles and volumes of the amino acid side chains. Shen et al. [39] proposed conjoint triad (CT) based on dipoles and volumes physicochemical properties to extract effective information from protein sequences.

First, 20 amino acids are clustered into seven classes based on dipoles and volumes of side chains. Considering the interaction between the amino acid and its vicinal amino acids, the three continuous amino acids are regarded as a unit, so that we can obtain $7 \times 7 \times 7 = 343$ triad types. We calculate the frequency of occurrence of each triad called f_i ($i = 1, 2, \cdots, 343$). Then the 343 dimensional feature vectors are obtained according to equation (6).

$$d_i = \frac{f_i - \min\{f_1, f_2, \cdots, f_{343}\}}{\max\{f_1, f_2, \cdots, f_{343}\}}, \quad i = 1, 2, \cdots, 343 \tag{6}$$

2.2.4. Local descriptor

Local descriptor (LD) takes the discontinuity of the sequence into account [57,58]. First, the 20 amino acids are divided into seven classes through physicochemical property, the amino acid groupings are shown in Supplementary materials Table S2. Then, the protein sequence is divided into multiple sequence segments. And the discontinuous regions can lead to spatial proximity by the folding of the protein, which characterize the effective feature information of protein. So the LD feature extraction method could provide important sequence information and physicochemical information for PPIs prediction.

The interactions between continuous sequences and discontinuous sequences of protein are also possible. In order to obtain the effective feature information, each protein sequence is divided into 10 segments. For each segment sequence, we calculate three local descriptors in order:

(1) Composition (C): the frequency of each amino acid in a local segment sequence.
(2) Transition (T): the frequency of the dipeptide which is the two adjacent amino acids from different groups.
(3) Distribution (D): The distribution pattern of the first, 25%, 50%, 75%, and last position of amino acid for each segment sequence.

In this way, 63 features can be obtained on each segment of the protein. Among them, 10 local fragments can generate 630 dimensional feature vector for each sample.

2.3. Elastic net

Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, consider the linear regression model with squared error as the loss function, and the optimization goal is:

$$\min_{\omega} \sum_{i=1}^{m} (y_i - \omega^T x_i)^2 \tag{7}$$

The shortcoming of above loss function is that when there are many features, it is easy to tend to be overfitting, so equation (7) can be written as equation (8) using a convex linear combination of L_1 norm and L_2 norm.

$$\min_{w} \frac{1}{2 \times n}\|y - Xw\|_2^2 - \alpha \times \beta\|w\|_1 + \frac{1}{2}\alpha \times (1-\beta)\|w\|_2^2 \tag{8}$$

where n is the number of samples. w is the weight coefficient. α, β are the penalty parameters. Equation (8) is called elastic net (EN) [59]. EN has good performance when dealing with the correlation between feature factors.

2.4. LightGBM

Gradient boosting decision tree (GBDT) is a popular classifier algorithm [60]. Given a training set $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where x are the data samples and y are the class labels. We use $F(x)$ to represent the estimated function and the optimization goal of GBDT is to minimize the loss function $L(y, F(x))$:

$$\widehat{F} = \arg\min_{F} E_{x,y}[L(y, F(x))] \tag{9}$$

Then, we can obtain the iterative criterion of the GBDT using a line search to minimize the loss function.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{10}$$

where $\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$, m is the iteration number, $h_m(x)$ represents the base decision tree.

But when the number of samples is large or the feature dimension is high, the efficiency and accuracy of GBDT still cannot obtain satisfactory results. GBDT is an ensemble algorithm whose base classifier is decision tree, so the main cost is to find the best split points when learning decision trees. Ke et al. [61] proposed a highly efficient gradient boosting decision tree using gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) called LightGBM.

For GBDT, the information gain is usually employed to split each node. LightGBM uses GOSS to determine the split point via calculating variance gain. First, sorting the absolute values of the gradients of the training examples in descending order and the top $a \times 100\%$ data samples of gradient values are selected called A. Then the subset B whose size is $b \times |A^c|$ is randomly selected from the retained samples A^c. Finally, the instances are split through the estimated variance $\widetilde{V}_j(d)$ on $A \cup B$.

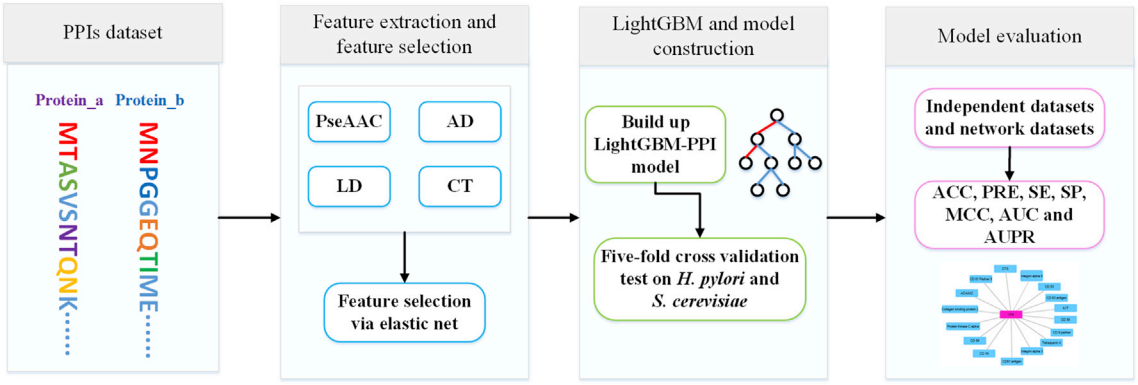

Fig. 1. The overall framework of LightGBM-PPI for protein-protein interactions prediction.

$$\widetilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i\in A_l} g_i + \frac{1-a}{b}\sum_{x_i\in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i\in A_r} g_i + \frac{1-a}{b}\sum_{x_i\in B_r} g_i\right)^2}{n_r^j(d)}\right) \tag{11}$$

where $A_l = \{x_i \in A : x_{ij} \le d\}$, $A_r = \{x_i \in A : x_{ij} > d\}$, $B_l = \{x_i \in B : x_{ij} \le d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$, g_i represents the negative gradient of the loss function, $\frac{1-a}{b}$ is employed to normalize the sum of gradients.

High-dimensional features often tend to be sparse, and many sparse features are exclusive. In this way the EFB could be employed to speed up the training of GBDT. In other words, LightGBM can bundle exclusive features into a single feature and the feature scanning algorithm could be designed to build the same feature histograms from the feature bundles. So, the computation complexity of LightGBM is reduced to $O(data * bundle)$ from $O(data * feature)$, where $bundle \ll feature$.

In summary, LightGBM is an effective GBDT implementation with GOSS and EFB to improve computational efficiency without hurting the accuracy. GOSS is employed to split the optimal node through calculating variance gain. EFB can speed up the training process of GBDT via bundling many exclusive features to fewer dense features. LightGBM is an ensemble model whose base classifier is decision tree, which could be trained in sequence by fitting the negative gradients of loss function. According to equation (10), the LightGBM model $F_M(x)$ can be obtained through the weighted combination scheme.

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{12}$$

where M is the maximum number of iteration and $h_m(x)$ is the base decision tree. The LightGBM code is available at https://github.com/Microsoft/LightGBM.

2.5. Model construction and performance evaluation

For convenience, the protein-protein interactions prediction method proposed in this study is called LightGBM-PPI. As the important biological topics show [62,63], using flowchart to study the intrinsic mechanisms of biomedical systems can provide more intuitive and useful biology information. Graphic approaches could strengthen the illustration of the prediction results. So we plot Fig. 1 to show the general framework of LightGBM-PPI. The framework is implemented using MATLAB 2014a and Python 3.6. The experimental experiment is carried out on a computer with 2.30 GHz, 32.0 GB of RAM and Windows operating system.

The specific steps of LightGBM-PPI for protein-protein interactions prediction method are described as:

1) PPIs dataset. Input the protein-protein interactions datasets the *S. cerevisiae*, *H. pylori*, four independent protein-protein interaction dataset (*C. elegans*, *E. coli*, *H. sapiens* and *M. musculus*) and one-core network, the crossover network for the Wnt-related pathway.
2) Feature extraction and feature selection. PseAAC, AD, CT, and LD are used to transform the protein sequence signal into numerical signal, and the four feature extraction methods are fused. Then the protein pairs are concatenated to construct initial feature vectors. The redundant and irrelevant features are eliminated via elastic net, and the optimal feature subset can be retained for PPIs prediction.
3) LightGBM and model construction. Input the optimal feature vector into the LightGBM classifier, which could mine non-linear relationship between the sequence features and class label to predict PPIs. First, the GOSS is employed to split the optimal node using variance gain and EFB is used to bundle exclusive features into dense features. Then, the LightGBM is trained in the direction of the negative gradient of the loss function in each iteration. Finally, the LightGBM-PPI is built up through the weighted combination scheme via five-fold cross-validation on *H.pylori* and *S. cerevisiae*.
4) Model evaluation. The optimal feature vectors of four independent test sets are extracted according to step 2), and the *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens* and *Mus musculus* are used to evaluate the pros and cons of the model. In order to further evaluate the validation of LightGBM-PPI, one-core network and the crossover network for the Wnt-related pathway network are employed to predict PPIs.

The current conventional indicators are valid only for single-label systems. Recently, multi-label systems are more popular [64,65], so a completely different indicators are needed. To more intuitively evaluate the predictive performance of the model, we use the symbols introduced by Chou in the field of signal peptide prediction [66,67]. The Accuracy (ACC), Precision (PRE), Sensitivity (SE), Specificity (SP) and Matthew's correlation coefficient (MCC) are formulated as:

$$ACC = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{13}$$

$$PRE = 1 - \frac{N_+^-}{N^+ - N_-^+ + N_+^-} \tag{14}$$

$$SE = 1 - \frac{N_-^+}{N^+} \tag{15}$$

$$SP = 1 - \frac{N_+^-}{N^-} \tag{16}$$

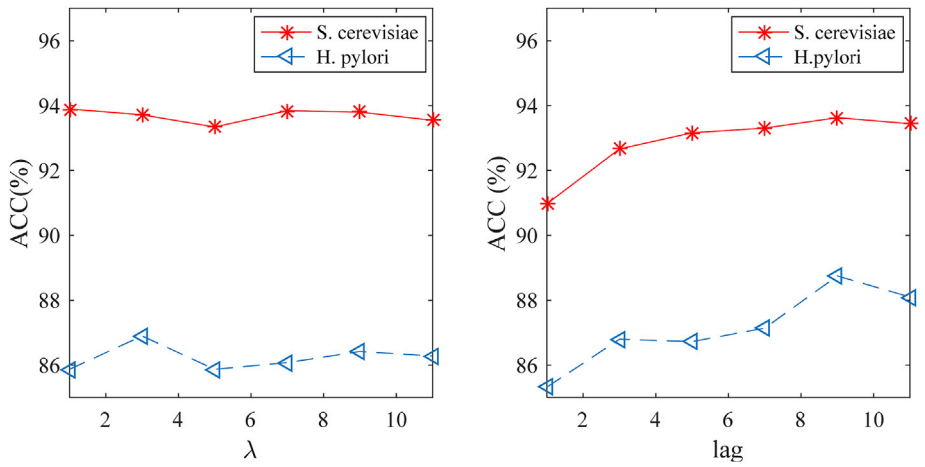

Fig. 2. ACC values corresponding to different λ values and *lag* values.

$$MCC = \frac{1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \tag{17}$$

where N^+ represents the number of samples protein-protein interactions, N^- represents the number of protein-protein interaction samples that do not interact. N_-^+ indicates the number of samples where the interacting protein pairs are incorrectly predicted as non-interacting protein pairs. N_+^- indacates the number of samples where the non-interacting protein pairs are incorrectly predicted as interacting protein pairs. The ROC curve and PR curve [68,69] are also the important indicators to evaluate the generalization ability.

3. Results and discussion

3.1. Selection of optimal parameters λ and lag

The selection of parameters plays an important role in the construction of model. When using pseudo amino acid composition and autocorrelation descriptor to extract the features of the protein sequence, the selection of the λ and *lag* values has a great effect on the result of the feature extraction. In this paper, the *H. pylori* and the *S. cerevisiae* are used as the research objects. The parameters are set to 1, 3, 5, 7, 9, and 11, respectively, and the LightGBM is employed as classifier on *H. pylori* and *S. cerevisiae*. The accuracy, precision, sensitivity, specificity and Matthew's correlation coefficient of the model are shown in Supplementary materials Table S3 and Table S4. Fig. 2 plots the effect of different parameters on *H. pylori* and *S. cerevisiae*.

It can be seen from Fig. 2, when using PseAAC to extract features, the ACC values of the two datasets change with the change of the λ value. In order to unify the parameters of the model, we determine $\lambda = 3$. Therefore, each protein sequence can generate $20 + \lambda = 23$ dimensional vector. In the process of AD feature extraction, when the parameter *lag* is set as 9, the ACC values on the *H. pylori* and the *S. cerevisiae* datasets are the highest. So this paper chooses $lag = 9$. Therefore, when feature extraction is performed using the autocorrelation descriptor, a $3 \times n \times lag = 3 \times 7 \times 9 = 189$ dimensional feature vector can be obtained for each protein sequence ($n = 7$ is the number of physicochemical properties).

PseAAC and AD characterize the physicochemical information of protein sequences, CT and LD characterize the sequence information of PPIs. The four feature extraction methods can more fully characterize the PPIs, but it may bring more redundant information. We use elastic net to eliminate the redundancy and noise information, and the effective information of the PPIs is obtained. All (EN) represents the fusion of four feature extraction methods and dimensional reduction through elastic net. We compare All (EN) with other different feature extraction methods. The prediction results are shown in Table 1.

Table 1
Comparison of different feature extraction methods on *H. pylori* and *S. cerevisiae*.

Dataset	Algorithm	ACC (%)	PRE (%)	SE (%)	SP (%)	MCC
H. pylori	PseAAC	86.90	85.59	88.75	85.05	0.7386
	AD	88.24	88.23	88.27	88.20	0.7649
	CT	87.55	87.14	88.13	86.97	0.7512
	LD	87.99	86.50	90.05	85.94	0.7606
	All (EN)	**89.03**	88.36	89.99	88.06	0.7814
S. cerevisiae	PseAAC	93.72	96.52	90.70	96.73	0.8759
	AD	93.63	96.59	90.45	96.80	0.8744
	CT	93.94	96.96	90.72	97.16	0.8807
	LD	94.57	96.95	92.04	97.10	0.8927
	All (EN)	**95.07**	97.82	92.21	97.94	0.9030

It can be seen from Table 1. The prediction performance of different feature extraction methods is different. For *H. pylori*, the prediction accuracy using PseAAC, AD, CT, LD, and All (EN) are 86.90%, 88.24%, 87.55%, 87.99%, and 89.03%, respectively. The ACC of AD is 1.34% higher than that of PseAAC. And the ACC of LD is only 0.44% higher than that of CT. We can see when the four feature extraction methods are fused and the elastic net is used to select optimal feature subset, the prediction accuracy is significantly improved. The accuracy of All (EN) is 89.03%, which is 0.79% higher than that of AD, 2.13% higher than that of PseAAC. Meanwhile, the MCC of All (EN) is 4.28%, 1.65%, 3.02%, and 2.08% higher than those of PseAAC, AD, CT, and LD, respectively. For *S. cerevisiae*, the accuracy of PseAAC, AD, CT, LD and ALL (EN) are 93.72%, 93.63%, 93.94%, 94.57% and 95.07%, respectively. And when elastic net is employed to eliminate redundancy and noise information, the accuracy of All (EN) is 0.5% higher than LD. The MCC values of All (EN) are 2.71%, 2.86%, 2.23%, and 1.03% higher than those of PseAAC, AD, CT, and LD, respectively. In conclusion, fusing four feature extraction methods and using elastic net to select optimal feature subset can improve the prediction accuracy of PPIs on *H. pylori* and *S. cerevisiae*. Therefore, we combine the four feature extraction methods of PseAAC, AD, CT and LD to characterize sequence information and physicochemical information of PPIs.

3.2. The effect of feature selection algorithms

Feature selection is of great importance when using machine learning to analyze the bioinformatics data [70,71]. In order to evaluate the advantages and disadvantages of the elastic net feature selection method, we also use LASSO [72], L1-regularization linear support vector machine

Table 2
Effect of different dimensionality reduction methods on *H. pylori* and *S. cerevisiae*.

Dataset	Algorithm	ACC (%)	PRE (%)	SE (%)	SP (%)	MCC
H. pylori	LASSO	84.29	82.91	86.49	82.10	0.6870
	LinearSVC	87.65	86.13	89.78	85.53	0.7538
	ET	88.27	86.89	90.19	86.35	0.7663
	EN	**89.03**	88.36	89.99	88.06	0.7814
S. cerevisiae	LASSO	93.25	95.58	90.70	95.80	0.8662
	LinearSVC	94.00	96.19	91.63	96.37	0.8811
	ET	93.31	95.32	91.08	95.53	0.8670
	EN	**95.07**	97.82	92.21	97.94	0.9030

(LinearSVC) [73] and ExtraTrees (ET) [74] to determine optimal feature subset. The optimal feature vectors are input into the LightGBM classifier to predict PPIs via five-fold cross-validation. We obtain the ACC, PRE, SE, SP and MCC values of the two datasets of *H. pylori* and *S. cerevisiae*. The prediction results are shown in Table 2. The ROC and PR curves of LASSO, LinearSVC, ET and EN are shown in Fig. 3.

It can be seen from Table 2 that for the two datasets, the selection of different dimensionality reduction methods has a great influence on the accuracy of PPIs prediction. The prediction performance of EN is the best. And the ACC, PRE, SE, SP, MCC of EN are 89.03%, 88.36%, 89.99%, 88.06% and 0.7814, respectively. The ACC of EN is 4.74%, 1.38%, and 0.76% higher than the LASSO, LinearSVC, and ET, respectively. The MCC value of EN is 9.44%, 2.76%, and 1.51% higher than those of LASSO, LinearSVC, and ET. For the *S. cerevisiae* dataset, the ACC, PRE, SE, SP, and MCC are 95.07%, 97.82%, 92.21%, 97.94% and 0.9030, respectively. The ACC is 1.82%, 1.07%, and 1.76% higher than those of LASSO, LinearSVC, and ET, respectively. The MCC value of EN is 3.68%, 2.19%, and 3.6% higher than those of LASSO, LinearSVC, and ET, respectively.

As we can see from Fig. 3 (A-B), when EN is used to predict the dataset *H. pylori*, the AUC value of EN is 0.9534, which is 3.86%, 0.09%, and 1.42% higher than those of LASSO, LinearSVC and ET, respectively. For the dataset *S. cerevisiae*, the AUC value of EN is 0.9875, which is higher than other feature selection methods. The AUC values of LASSO, LinearSVC and ET are 1.13%, 0.51% and 1.28% lower than EN, respectively. From Fig. 3 (C-D), for the two datasets *H. pylori* and *S. cerevisiae*, compared with other feature selection methods, EN achieves the better precision almost at every recall value. When EN is used to predict the dataset *H. pylori*, its AUPR value is 0.9370, which is 1.81% and 0.3% higher than LASSO and LinearSVC, respectively. For the dataset *S. cerevisiae*, the AUPR value of EN is 0.9897, which is 0.65%, 0.41% and 1.45% higher than those of LASSO, LinearSVC and ET, respectively. In summary, the predictive effect of elastic net dimension reduction is better than LASSO, LinearSVC and ET. So, we use elastic net to select the optimal feature subset.

3.3. Selection of classification algorithms

In order to construct an effective PPIs prediction model, the choice of classifier is crucial. According to the discussion in section 3.1, the parameter of the pseudo amino acid composition is set as $\lambda = 3$ and the parameter of the autocorrelation descriptor is set as $lag = 9$. We can obtain 1185-dimensional feature vectors by fusing four feature extraction algorithms. Then, the two sequence feature vectors of the protein pairs are concatenated as the final feature vector whose dimension is 2370. According to the discussion in section 3.2, we set the elastic net as the best dimensional reduction method. The 227-dimensional feature vectors on *H. pylori* and 179-dimensional feature vectors on *S. cerevisiae* are obtained by elastic net feature selection. In order to evaluate the performance of LightGBM, we compare LightGBM with Naïve Bayes (NB) [75], k-nearest neighbor (k-NN) [76], support vector machine (SVM) [77], and random forest (RF) [78] on *H. pylori* and *S. cerevisiae* to predict PPIs. The number of neighbor in k-NN is 5, and the support vector machine selects the radial basis kernel function. The number of base decision tree in random forest is 500, the maximum number of iteration of LightGBM is

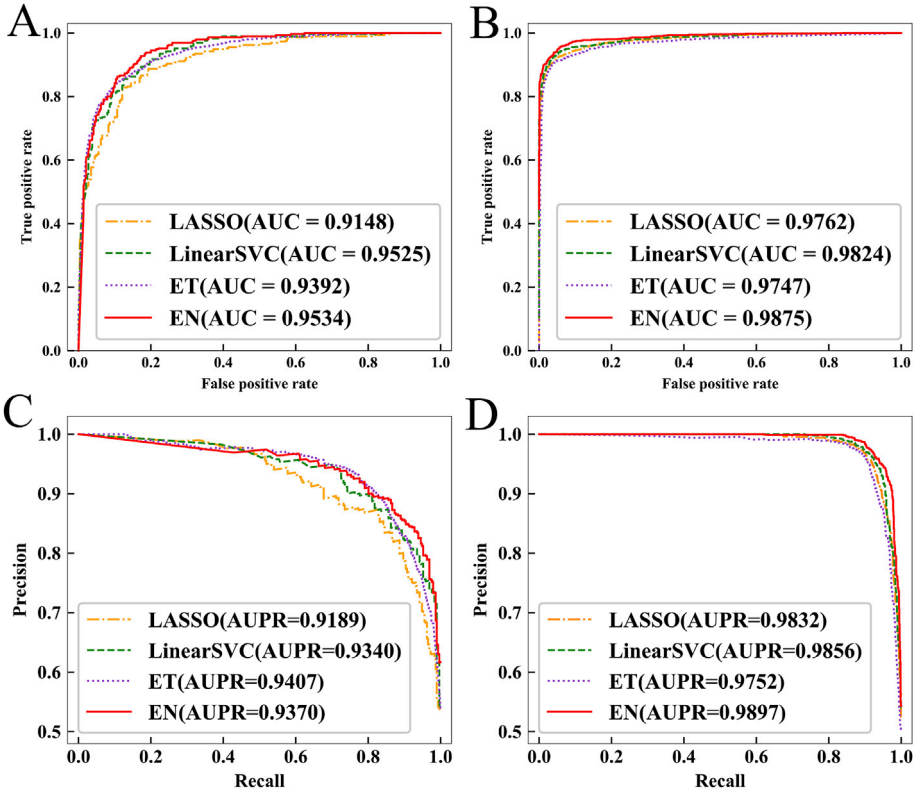

Fig. 3. (A-D) Comparison of ROC and PR curves for different dimensional methods. (A-B) The ROC curves of different dimensional methods on *H. pylori* and *S. cerevisiae*. (C-D) The PR curves of different dimensional methods on *H. pylori* and *S. cerevisiae*.

Table 3
Prediction performance of *H.pylori* and *S. cerevisiae* based on five-fold cross-validation.

Dataset	Classifier	ACC (%)	PRE (%)	SE (%)	SP (%)	MCC
H .pylori	NB	71.60	67.36	83.88	59.32	0.4462
	k-NN	76.85	70.72	91.84	61.86	0.5631
	SVM	78.87	79.05	78.81	78.94	0.5781
	RF	81.38	81.90	80.80	81.96	0.6285
	LightGBM	**89.03**	88.36	89.99	88.06	0.7814
S. cerevisiae	NB	74.70	75.03	74.03	75.37	0.4940
	k-NN	84.82	87.02	81.87	87.77	0.6978
	SVM	83.21	83.25	83.16	83.27	0.6643
	RF	91.39	94.37	88.04	94.74	0.8297
	LightGBM	**95.07**	97.82	92.21	97.94	0.9030

set to 500. The prediction results on *H. pylori* and *S. cerevisiae* under different classifiers are shown in Table 3. In order to evaluate the robustness of classifiers, we plot ROC and PR curves, as shown in Fig. 4.

As we can see from Table 3, for *H. pylori* dataset, the ACC of LightGBM reaches 89.03%, which is 7.65% higher than RF, and 17.43% higher than NB. The LightGBM is also higher than other classifiers with a PRE of 88.36%, SE of 89.99%, SP of 88.06%, and MCC of 0.7814. For *S. cerevisiae*, the LightGBM achieves the best prediction performance with an ACC of 95.07%, PRE of 97.82%, SE of 92.21%, SP of 97.94% and MCC of 0.9030. LightGBM achieves the best performance on *S. cerevisiae* and *H. pylori*.

As we can see from Fig. 4 (A-B), when LightGBM is used to predict the dataset *H. pylori*, the AUC value of LightGBM is 0.9534, which is 13.95%, 6.95%, 7.57%, and 6.37% higher than those of NB, KNN, SVM, and RF, respectively. For the dataset *S. cerevisiae*, the AUC value of LightGBM is 0.9875, which is significantly higher than other classifiers. The AUC values of NB, KNN, SVM, and RF are 17.05%, 6.2%, 7.32%, and 2.26% lower than that of LightGBM, respectively. From Fig. 4 (C-D), for the two datasets *H. pylori* and *S. cerevisiae*, compared with other classifiers, LightGBM achieves the better precision almost at every recall value. When LightGBM is used to predict the dataset *H. pylori*, its AUPR value is 0.9370, which is 23.17%, 9.99%, 7.66%, and 4.4% higher than NB, KNN, SVM, and RF, respectively. For the dataset *S. cerevisiae*, the AUPR value of LightGBM is 0.9897, which is higher than other four classifiers. The AUPR values of NB, KNN, SVM, and RF are 15.89%, 6.97%, 7.25%, and 3.49% lower than the AUPR values of LightGBM, respectively.

By analyzing the ACC, PRE, SE, SP, MCC and ROC curves and PR curves of the *H. pylori* and *S. cerevisiae*, the LightGBM achieves the highest accuracy, robustness and generalization ability. LightGBM can accelerate the training process and improve the operation efficiency when determining the split point. At the same time, LightGBM can preserve the excellent property of gradient boosting decision tree, which can effectively predict PPIs.

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

Methods	ACC (%)	SE (%)	PRE (%)	MCC
PCA-EELM [a]	87.50	88.95	86.15	0.7813
MCD [b]	84.91	83.24	86.12	0.7440
MMI+NMBAC [c]	87.59	86.81	88.23	0.7524
DCT+WSRC [d]	86.74	86.43	87.01	0.7699
Signature products [e]	83.40	79.90	85.70	N/A
HKNN [f]	84.00	86.00	84.00	N/A
Ensemble of HKNN [g]	86.60	86.70	85.00	N/A
LightGBM-PPI	89.03	89.99	88.36	0.7814

Note: N/A means not available.
[a] Results reported by Ref. [57].
[b] Results reported by Ref. [58].
[c] Results reported by Ref. [79].
[d] Results reported by Ref. [80].
[e] Results reported by Ref. [46].
[f] Results reported by Ref. [81].
[g] Results reported by Ref. [82].

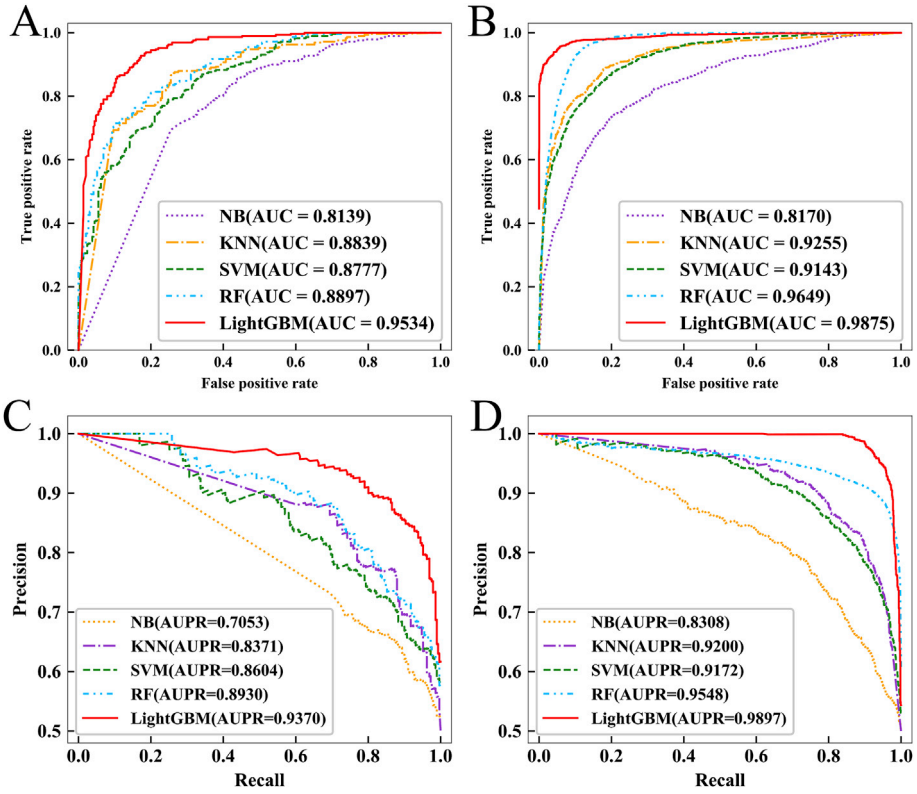

Fig. 4. (A-D) Comparison of ROC and PR curves for different classifiers. (A-B) The ROC curves of different classifiers on *H. pylori* and *S. cerevisiae*. (C-D) The PR curves of different classifiers on *H. pylori* and *S. cerevisiae*.

Table 5
Comparison results of different PPIs prediction methods on *S. cerevisiae*.

Methods	ACC (%)	SE (%)	PRE (%)	MCC
DeepPPI [a]	94.43 ± 0.30	92.06 ± 0.36	96.65 ± 0.59	0.8897 ± 0.62%
PR-LPQ [b]	93.92 ± 0.36	91.10 ± 0.31	96.45 ± 0.45	0.8856 ± 0.63%
MCD [c]	91.36 ± 0.36	90.67 ± 0.69	91.94 ± 0.62	0.8421 ± 0.59%
LD+KNN [d]	86.15 ± 1.17	81.03 ± 1.74	90.24 ± 1.34	N/A
SVM [e]	88.56 ± 0.33	87.37 ± 0.22	89.50 ± 0.60	0.7715 ± 0.68%
ACC+SVM [f]	89.33 ± 2.67	89.93 ± 3.68	88.87 ± 6.16	N/A
LightGBM-PPI	95.07 ± 0.51	92.21 ± 0.61	97.82 ± 0.48	0.9030 ± 1.01%

Note: N/A means not available. The values behind ± represent the standard deviation.
[a] Results reported by Ref. [25].
[b] Results reported by Ref. [83].
[c] Results reported by Ref. [58].
[d] Results reported by Ref. [84].
[e] Results reported by Ref. [47].
[f] Results reported by Ref. [44].

Table 6
Comparison of different PPIs prediction methods on independent test sets.

Species	LightGBM-PPI	DeepPPI[a]	DCT+WSRC [b]	SVM [c]
H. sapiens	94.83	93.77	82.22	76.27
M. musculus	94.57	91.37	79.87	76.68
C. elegans	90.16	94.84	81.19	75.73
E. coli	92.16	92.19	66.08	71.24

Note.
[a] Results reported by Ref. [25].
[b] Results reported by Ref. [80].
[c] Results reported by Ref. [47].

3.4. Comparision with other methods

For the prediction of protein-protein interactions, various prediction methods have been proposed. In order to more objectively evaluate the prediction performance of the constructed model, we compare its prediction results with the other prediction methods on the same dataset. First, PseAAC, AD, CT and LD are used for feature extraction. The redundant and noisy information is eliminated through elastic net and the LightGBM is used as classifier. The comparison results of *H. pylori* and *S. cerevisiae* are shown in Table 4 and Table 5. In order to further verify the predictive performance of LightGBM-PPI, the *S. cerevisiae* is used as the training set, and the *Caenorhabditis elegans*, *Escherichia coli*, *Homo sapiens* and *Mus musculus* are employed as independent test sets. The prediction results are shown in Table 4.

It can be seen from Table 4 that the prediction accuracy of the LightGBM-PPI model on the *H. pylori* dataset is 89.03%, which is 1.44%–5.63% higher than other prediction methods. The overall prediction accuracy of the LightGBM-PPI is 5.03% higher than that of the HKNN. In addition, the SE, PRE and MCC of the model proposed in this paper are 89.99%, 88.36% and 0.7814, respectively, and the SE is 1.04%–10.09% higher than other methods. By comparison, the prediction performance of the proposed method is significantly better than others.

It can be seen from Table 5 that the ACC, SE, PRE and MCC of the proposed model LightGBM-PPI on the *S. cerevisiae* dataset are 95.07%, 92.21%, 97.82% and 0.9030, respectively. Compared with other methods, the ACC value of this paper is 0.64%–8.92% higher than other methods. And the ACC of LightGBM-PPI is 8.92% higher than that of LD+KNN, 5.74% higher than that of ACC+SVM, and 6.51% higher than that of SVM. In addition, the MCC of the proposed model is 1.33%–13.15% higher than other methods. Compared with other methods, the proposed method achieves satisfactory performance.

As can be seen from Table 6, for the dataset *H. sapiens*, our method LightGBM-PPI has an ACC of 94.83%, which is 18.56%, 12.61%, 1.06% higher than those of SVM, DCT+WSRC, DeepPPI. For the dataset *M. musculus*, the ACC value of the prediction model LightGBM-PPI is 94.57%, which is 3.2%, 14.7% and 17.89% higher than those of DeepPPI, DCT+WSRC and SVM, respectively. For the dataset *C. elegans*, the ACC value of LightGBM-PPI is 90.16%, which is 8.97% higher than that of DCT+WSRC and 14.43% higher than that of SVM. For the dataset *E. coli*, the ACC of LightGBM-PPI is 92.16%, which is 26.08% and 20.92% higher than DCT+WSRC and SVM, respectively. The model is trained using *S. cerevisiae* and achieves good prediction performance on four independent test sets, indicating that the proposed LightGBM-PPI model can characterize important PPIs information and perform cross-species prediction. In other words, it is feasible that PPIs produced by one species can predict the other species.

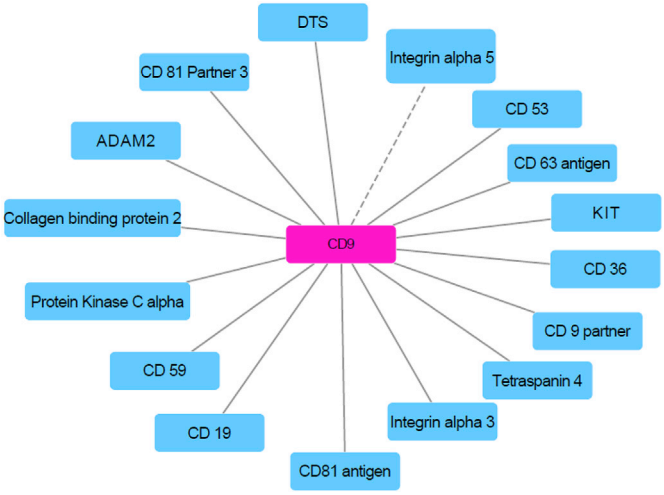

Fig. 5. Predicted results of PPIs networks of a one-core network for CD9. Core and satellite proteins are colored red and blue, respectively. The solid lines are the true prediction, and the dotted line is the false prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.5. PPIs network prediction

The study of protein-protein interactions network and the corresponding signaling pathway is important for understanding the structure and function of proteins. Our method predicts two important PPIs networks, the one-core network and the crossover network for the Wnt-related pathway. The one-core PPIs network is a simple single-core PPIs network which consists of a core protein that interacts with many other proteins [48]. The network consists of 17 genes, of which CD9 is the core protein, and CD9 is a tetrameric protein that plays an important role in cell viability and tumor suppression. The second is the crossover network for Wnt-related pathway constructed from 78 genes, and it is composed with multi-core networks. Stelzl et al. [49] demonstrated the interaction topology using yeast two-hybrid experiments. The pathway network plays a crucial role in tumor growth and tumor formation. The PseAAC, AD, LD and CT are employed to encode the protein pairs and the elastic net is used to select 179 important features to characterize PPIs. Finally, the *S. cerevisiae* dataset is used as the training set, and the single-core network and the cross-network of Wnt-related pathway are used as the test set. The prediction results of the one-core network and the crossover network for the Wnt-related pathway network are shown in Fig. 5 and Fig. 6, respectively. The solid line is true prediction, and the dotted line is false prediction.

For the one-core network, the core protein is CD9. CD9 is an important protein that interacts with a variety of related factors to form a typical single-core network. It can be seen from Fig. 5 that the proposed method can identify 15 of the 16 protein interactions with an accuracy of 93.75%, and Ding et al. [79] has an accuracy of 87.50%. The ACC of LightGBM-PPI is 6.25% higher than Ding et al. [79]. Biologically

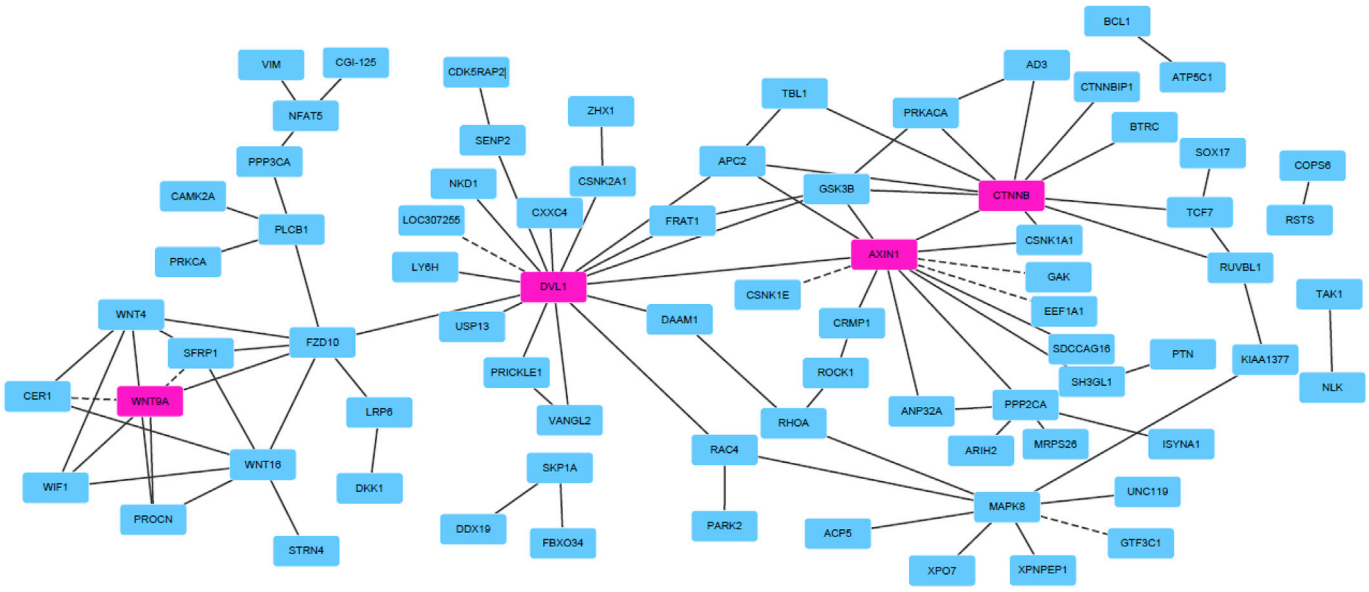

Fig. 6. The predicted results of a crossover network for the Wnt-related pathway. The solid lines are the true prediction, and the dotted lines are the false prediction.

speaking, the PPIs network is generally a crossover network. The crossover network for the Wnt-related pathway is a typical signal pathway crossover network. As we can see from Fig. 6, LightGBM-PPI can predict 89 of the 96 protein-protein interactions with a prediction accuracy of 92.70%. The prediction results of Shen et al. [39] can predict 73 of the 96 protein-protein interactions. Our method is better than the prediction accuracy of Shen et al. [39].

For one-core network, only one protein-protein interaction cannot be predicted successfully. But other protein-protein interactions linked to CD9 are all predicted successfully. Wnt-related pathway network plays an important role in growth and development [85] and Dupuytren's disease [86], and LightGBM-PPI can effectively predict PPIs in Wnt-related pathway network. In summary, our method has excellent performance for PPIs network prediction and can provide some reference for exploring disease pathogenesis, drug design and tumor suppression.

4. Conclusion

The study of protein-protein interactions is critical for understanding the biological processes, which plays an important role in cell life activities, such as transcriptional regulation, signal transduction, and drug signaling. Using machine learning methods to predict PPIs is important for improving cell growth and development and further understanding the life activities of cells. In this paper, a new machine learning-based PPIs prediction method LightGBM-PPI is proposed. For *S. cerevisiae* and *H. pylori*, we use PseAAC, AD, LD, CT four feature extraction methods to extract sequence information and physicochemical information of PPIs. In order to eliminate redundant and noise information, we use the elastic net to select the optimal feature subset and effectively retain the important feature information to characterize PPIs. LightGBM can reduce the running time and improve the computational efficiency, which can effectively predict the protein-protein interactions. The LightGBM-PPI method has also achieved good prediction accuracy on *H. sapiens*, *M. musculus*, *C. elegans*, and *E. coli*, indicating that our method has also achieved good performance for cross-species prediction. The prediction results on the one-core network and the crossover network for the Wnt-related pathway indicate that the LightGBM-PPI method can provide new insights into signaling pathway study, drug target prediction, and understanding disease pathogenesis. Although we can characterize the important feature information of PPIs through multi-information fusion and elastic net, it is not enough to completely elaborate the feature information of PPIs. How to mine the structural information and evolutionary information of protein pairs is the next research direction. Deep learning has powerful feature learning ability, and using deep learning to predict protein-protein interactions is also the next step.

Some recent literatures have shown that publicly accessible user-friendly web servers could promote the development of the computational tools [87–89]. At the same time, useful web servers can increase the influence and even lead to driving an unprecedented revolution [90] in medicinal chemistry. A web server will be provided for the proposed prediction method LightGBM-PPI in future work.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (No. 61863010), the Natural Science Foundation of Shandong Province of China (Nos. ZR2017MA014 and ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), and the Scientific Research Fund of Hunan Provincial Key Laboratory of Mathematical Modeling and Analysis in Engineering (No. 2018MMAEZD10). This work used the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation (No. ACI-1548562).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.06.003.

References

[1] H.X. Zhou, Y.B. Shan, Prediction of protein interaction sites from sequence profile and residue neighbor list, Proteins 44 (2001) 336–343.
[2] S. Navlakha, A. Gitter, Z. Barjoseph, A network-based approach for predicting missing pathway interactions, PLoS Comput. Biol. 8 (2012), e1002640.
[3] D.H. Le, Y.K. Kwon, Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization, Comput. Biol. Chem. 44 (2013) 1–8.
[4] J.F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, M. Vidal, Towards a proteome-scale map of the human protein-protein interaction network, Nature 437 (2005) 1173–1178.

[5] K. Oxenoid, Y.S. Dong, C. Cao, T. Cui, Y. Sancak, A.L. Markhard, Z. Grabarek, L. Kong, Z. Liu, B. Ouyang, Y. Cong, V.K. Mootha, J.J. Chou, Architecture of the mitochondrial calcium uniporter, Nature 533 (2016) 269–273.
[6] M.J. Berardi, W.M. Shih, S.C. Harrison, J.J. Chou, Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching, Nature 476 (2011) 109–113.
[7] K.C. Chou, A.G. Tomasselli, R.L. Heinrikson, Prediction of the tertiary structure of a caspase-9/inhibitor complex, FEBS Lett. 470 (2000) 249–256.
[8] Y. Ma, S.Q. Wang, W.R. Xu, R.L. Wang, K.C. Chou, Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach, PLoS One 7 (2012), e38546.
[9] Y. Xu, J. Ding, L.Y. Wu, K.C. Chou, iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PLoS One 8 (2013), e55844.
[10] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, K.C. Chou, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. 33 (2015) 2221–2233.
[11] H. Lin, E.Z. Deng, H. Ding, W. Chen, K.C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.
[12] Z. Liu, X. Xiao, W.R. Qiu, K.C. Chou, iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77.
[13] B. Yu, Y. Zhang, A simple method for predicting transmembrane proteins based on wavelet transform, Int. J. Biol. Sci. 9 (2013) 22–33.
[14] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.
[15] M. Behbahani, H. Mohabatkar, M. Nosrati, Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition, J. Theor. Biol. 411 (2016) 1–5.
[16] S. Akbar, M. Hayat, iMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences, J. Theor. Biol. 455 (2018) 205–211.
[17] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.
[18] H.B. Shen, K.C. Chou, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[19] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-Builder, A cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions, Anal. Biochem. 425 (2012) 117–119.
[20] D.S. Cao, Q.S. Xu, Y.Z. Liang, propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.
[21] P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (2014) 3495–3506.
[22] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[23] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, K.C. Chou, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.
[24] B. Liu, H. Wu, K.C. Chou, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nat. Sci. 9 (2017) 67–91.
[25] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, Y. Zhang, DeepPPI: boosting prediction of protein-protein interactions with deep neural networks, J. Chem. Inf. Model. 57 (2017) 1499–1510.
[26] B.G. Tian, X. Wu, C. Chen, W.Y. Qiu, Q. Ma, B. Yu, Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach, J. Theor. Biol. 462 (2019) 329–346.
[27] H. Li, X.J. Gong, H. Yu, C. Zhou, Deep neural network based predictions of protein interactions using primary sequences, Molecules 23 (2018) 1923.
[28] Y.E. Göktepe, H. Kodaz, Prediction of protein-protein interactions using an effective sequence based combined method, Neurocomputing 303 (2018) 68–74.
[29] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, D. Plewczynski, PPI_SVM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables, Cell. Mol. Biol. Lett. 16 (2011) 264–278.
[30] M. Rashid, S. Ramasamy, P.S.R. Gajendra, A simple approach for predicting protein-protein interactions, Curr. Protein Pept. Sci. 11 (2010) 589–600.
[31] S. Dohkan, A. Koike, T. Takagi, Improving the performance of an SVM-based method for predicting protein-protein interactions, Silico Biol. 6 (2006) 515–529.
[32] Z. Ju, J.J. He, Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC, J. Mol. Graph. Model. 76 (2017) 356–363.
[33] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.
[34] G.T. Valente, M.L. Acencio, C. Martins, N. Lemke, The development of a universal in silico predictor of protein-protein interactions, PLoS One 8 (2013), e65587.
[35] Y. Qi, Z. Bar-Joseph, J. Klein-Seetharaman, Evaluation of different biological data and computational classification methods for use in protein interaction prediction, Proteins 63 (2006) 490–500.
[36] X.W. Chen, M. Liu, Prediction of protein-protein interactions using random decision forest framework, Bioinformatics 21 (2005) 4394–4400.
[37] N. Lin, B. Wu, R. Jansen, M. Gerstein, H. Zhao, Information assessment on predicting protein-protein interactions, BMC Bioinf. 5 (2004) 154.
[38] I. Saha, J. Zubek, T. Klingström, S. Forsberg, J. Wikander, M. Kierczak, U. Maulik, D. Plewczynski, Ensemble learning prediction of protein-protein interactions using proteins functional annotations, Mol. Biosyst. 10 (2014) 820–830.
[39] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Natl. Acad. Sci. Unit. States Am. 104 (2007) 4337–4341.
[40] X.Y. Pan, Y.N. Zhang, H.B. Shen, Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features, J. Proteome Res. 9 (2010) 4992–5001.
[41] S. Hashemifar, B. Neyshabur, A.A. Khan, J. Xu, Predicting protein-protein interactions through sequence-based deep learning, Bioinformatics 34 (2018) i802–i810.
[42] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, K.C. Chou, iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics 111 (2019) 96–102.
[43] X. Cheng, S.G. Zhao, X. Xiao, K.C. Chou, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (2017) 341–346.
[44] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, Nucleic Acids Res. 36 (2008) 3025–3030.
[45] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (2001) 282–283.
[46] S. Martin, D. Roe, J.L. Faulon, Predicting protein-protein interactions using signature products, Bioinformatics 21 (2005) 218–226.
[47] Y.Z. Zhou, Y. Gao, Y.Y. Zheng, Prediction of protein-protein interactions using local description of amino acid sequence, Commun. Comput.Inf. Sci. 202 (2011) 254–262.
[48] X.H. Yang, O.V. Kovalenko, T.V. Kolesnikova, M.M. Andzelm, E. Rubinstein, J.L. Strominger, M.E. Hemler, Contrasting effects of EWI proteins, integrins, and protein palmitoylation on cell surface CD9 organization, J. Biol. Chem. 281 (2006) 12976–12985.
[49] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E.E. Wanker, A human protein-protein interaction network: a resource for annotating the proteome, Cell 122 (2005) 957–968.
[50] X.W. Cui, Z.M. Yu, B. Yu, M.H. Wang, B.G. Tian, Q. Ma, UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components, Chemomet. Intell. Lab. 184 (2019) 28–43.
[51] B. Yu, L. Shan, W.Y. Qiu, M.H. Wang, J.W. Du, Y. Zhang, X. Chen, Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction, BMC Genomics 19 (2018) 478.
[52] B. Yu, S. Li, C. Chen, J.M. Xu, W.Y. Qiu, X. Wu, R.X. Chen, Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition, Chemomet. Intell. Lab. 167 (2017) 102–112.
[53] B. Yu, S. Li, W.Y. Qiu, C. Chen, R.X. Chen, L. Wang, M.H. Wang, Y. Zhang, Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising, Oncotarget 8 (2017) 107640–107665.
[54] B. Yu, L.F. Lou, S. Li, Y. Zhang, W.Y. Qiu, X. Wu, M.H. Wang, B.G. Tian, Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising, J. Mol. Graph. Model. 76 (2017) 260–273.
[55] W.Y. Qiu, S. Li, X.W. Cui, Z.M. Yu, M.H. Wang, J.W. Du, Y.J. Peng, B. Yu, Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition, J. Theor. Biol. 45 (2018) 86–103.
[56] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C. Chou, J. Song, iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.
[57] Z.H. You, Y.K. Lei, L. Zhu, J.F. Xia, B. Wang, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, BMC Bioinf. 14 (2013) S10.
[58] Z.H. You, L. Zhu, C.H. Zheng, H.J. Yu, S.P. Deng, Z. Ji, Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set, BMC Bioinf. 15 (2014) S9.
[59] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B. 67 (2005) 301–320.
[60] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001) 1189–1232.
[61] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: 31st Conference Neural Information Processing Systems NIPS, 2017, pp. 3146–3154.
[62] K.C. Chou, Graphic rules in steady and non-steady enzyme kinetics, J. Biol. Chem. 264 (1989) 12074–12079.
[63] K.C. Chou, Graphic rule for drug metabolism systems, Curr. Drug Metabol. 11 (2010) 369–378.
[64] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, K.C. Chou, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524–3531.
[65] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, K.C. Chou, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.

[66] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.
[67] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.
[68] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, Q. Ma, Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique, Bioinformatics (2019) bty995, https://doi.org/10.1093/bioinformatics/bty995.
[69] H. Shi, S.M. Liu, J.Q. Chen, X. Li, Q. Ma, B. Yu, Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure, Genomics (2019). https://doi.org/10.1016/j.ygeno.2018.12.007.
[70] B. Yu, Y. Zhang, The analysis of colon cancer gene expression profiles and the extraction of informative genes, J. Comput. Theor. Nanosci. 10 (2013) 1097–1103.
[71] C.Z. Kang, Y.H. Huo, L.H. Xin, B.G. Tian, B. Yu, Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine, J. Theor. Biol. 463 (2019) 77–91.
[72] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Stat. Soc. B. 58 (1996) 267–288.
[73] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[74] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[75] N. Friedman, D. Geiger, M. Pazzanzi, Bayesian network classifiers, Mach. Learn. 2 (1997) 131–163.
[76] F. Nigsch, A. Bender, B.V. Buuren, J. Tissen, E. Nigsch, J.B. Mitchell, Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization, J. Chem. Inf. Model. 46 (2006) 2412–2422.
[77] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[78] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32.
[79] Y. Ding, J. Tang, F. Guo, Predicting protein-protein interactions via multivariate mutual information of protein sequences, BMC Bioinf. 17 (2016) 398.
[80] Y.A. Huang, Z.H. You, X. Gao, L. Wong, L. Wang, Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence, BioMed Res. Int. 2015 (2015) 902198.
[81] L. Nanni, Hyperplanes for predicting protein-protein interactions, Neurocomputing 69 (2005) 257–263.
[82] L. Nanni, A. Lumini, An ensemble of K-local hyperplanes for predicting protein-protein interactions, Bioinformatics 22 (2006) 1207–1210.
[83] L. Wong, Z.H. You, S. Li, Y.A. Huang, G. Liu, Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor, in: International Conference on Intelligent Computing, ICIC, 2015, pp. 713–720.
[84] L. Yang, J.F. Xia, J. Gui, Prediction of protein-protein interactions from protein sequence using local descriptors, Protein Pept. Lett. 17 (2010) 1085–1090.
[85] E.P.M. Ten Dam, M.M. Van Beuge, R.A. Bank, R.A. Bank, P.M.N. Werker, Further evidence of the involvement of the Wnt signaling pathway in Dupuytren's disease, J. Cell Commun. Signal. 10 (2016) 33–40.
[86] G. Ruishuang, N. Teppei, J.F. Mulvaney, V.Y.W. Lin, A.S.B. Edge, A. Dabdoub, Comprehensive expression of Wnt signaling pathway genes during development and maturation of the mouse cochlea, PLoS One 11 (2016), e0148339.
[87] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.
[88] X. Cheng, X. Xiao, K.C. Chou, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics 34 (2018) 1448–1456.
[89] K.C. Chou, X. Cheng, X. Xiao, pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset, Genomics (2019), https://doi.org/10.1016/j.ygeno.2018.08.007.
[90] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.