Chemometrics and Intelligent Laboratory Systems 197 (2020) 103919

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

RBPro-RF: Use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net

Xiaomeng Sun [a,b,1], Tingyu Jin [a,b,1], Cheng Chen [a,b,1], Xiaowen Cui [a,b], Qin Ma [c], Bin Yu [a,b,d,*]

[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

ARTICLE INFO

Keywords:
RNA-Binding proteins
Feature extraction
SMOTE
Elastic net
Random forest

ABSTRACT

RNA-proteins interaction is essential for the regulation of gene expression, cell defense and developmental regulation and other life activities, so applying machine learning to predict RNA-binding proteins (RBPs) has become a research hotspot in bioinformatics. We propose a new method to predict RNA-binding proteins called RBPro-RF. First, the feature vectors of the protein sequence are extracted by fusing composition-transition-distribution (C-T-D), pseudo-amino acid composition (PseAAC) and position-specific scoring matrix-400 (PSSM-400). Secondly, the synthetic minority oversampling technique (SMOTE) and the edited nearest neighbor (ENN) are employed to balance samples. Then, elastic net (EN) is used to eliminate redundant features and retain the important features to represent RBPs. Finally, the optimal feature vectors are input into random forest classifier to predict RBPs. Ten-fold cross-validation indicates the ACC and MCC of the training set are 97.43% and 0.933, respectively. In addition, the accuracies of three independent test sets *Human, S. cerevisiae and A. thaliana* are 95.63%, 88.82%, and 92.35%, respectively, which are superior to the state-of-the-art prediction methods. In summary, experimental results show that our method can significantly improve the accuracy of RNA-binding proteins prediction. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/RBPro-RF/.

1. Introduction

RNA-proteins interaction plays an essential role in many biological activities, such as regulation of gene expression [1], mRNA localization [2], cell defense and developmental regulation [3,4], etc. In addition, RNA-binding proteins (RBPs) are significantly important in many diseases [5]. For example, they can regulate the expression of specific proteins in neurons through post-transcriptional regulation [6], which affect neurodevelopmental and neurological diseases. And its role in diabetes [7], kidney disease [8], cancer diagnosis [9] has also received extensive attention. Therefore, the interaction between protein and RNA are of great significance in cell function. Currently, researchers can obtain RNA-protein complexes experimentally, which allows RNA-proteins interaction to be analyzed at the molecular level. Since the redundant sequences are removed with a certain degree of similarity and only a very small amount of protein sequences are remained. It is difficult to obtain a RNA-protein complex [10,11], and this work is expensive and time-consuming. Therefore, it is essential to study how to use machine learning methods to predict RNA-binding proteins in the field of bioinformatics.

The 3D structure of proteins or their complexes with ligands are essential for proper drug design. Although X-ray crystallography can be used to determine these structures, not all proteins can be successfully crystallized. NMR [12,13] is indeed a powerful tool for determining the 3D structure of membrane proteins, but it is also time-consuming and costly. Therefore, researchers predict 3D protein structure through the way of structural bioinformatics [14–17]. With the explosive growth of biological sequences in the post-genomic era, sequential bioinformatics and structural bioinformatics develop rapidly [18], such as identifying S-prenylation sites in proteins [19], identifying protein molecule functions [20] and detecting formylation sites from protein sequences [21] of which computational biology is playing an increasingly important role in

* 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.2019.103919
Received 11 September 2019; Received in revised form 25 December 2019; Accepted 27 December 2019
Available online 29 December 2019

stimulating the development of new drugs. In view of this, the research of RNA binding protein prediction in this paper also uses the calculation method.

In the field of RBPs prediction, researchers extract sequence and structure information of proteins and further use computational methods to make predictions. For example, Paz et al. [22] proposed a structure-based approach called BindUP and developed a web server (http://www.cbi.seu.edu.cn/PRBP/) to predict RNA-binding proteins, which mainly extracted the electrostatic and other properties of the protein surface information. The support vector machine (SVM) was chosen as a classifier, and the value of AUC was 0.94. Zhao et al. [23] proposed a RBPs prediction method based on the structural similarity score. The Z-score was obtained through calculating the structural similarity score using the TM-align program. Then, according to known complex structures, the statistical energy function DFIRE was employed to predict RNA-binding proteins. Furthermore, Zhao et al. [24] proposed a sequence template-based method SPOT_seq, which combined protein sequence and structure information to achieve an accuracy of 98% on test set. However, this method was biased to the prediction of RNA-binding proteins because of the imbalance of positive and negative sample numbers. In order to solve this problem, it was necessary to resample the imbalanced samples. For example, Wang et al. [25] used the bagging method [26] and SVM classifier to construct the prediction model. First, bagging was adopted to resample instances, and then the sequence information, structure information and evolutionary information of the protein sequences were merged to input into the SVM classifier. The experimental results indicated that bagging could effectively improve the accuracy and generalization ability of the model.

Since the structural information of most proteins is unknown, there are significant limitations when adopting the structure-based approach to construct the prediction models of RNA-binding proteins. Currently, the most widely used method is to extract sequence information of proteins, including evolutionary information, physicochemical information, and amino acid composition information, etc. Kumar et al. [27] extracted the evolutionary information of protein sequences via PSSM-400, and SVM was employed as a classifier to construct an RNA-binding proteins prediction model. The Matthews Correlation Coefficient (MCC) and Accuracy (ACC) were 0.66, 82.89%, respectively. Meanwhile, they also developed a PSSM-SVM model which could distinguish tRNA, mRNA, and RNA-binding proteins, and the values of ACC were 76.7%, 79.28%, and 90.14%, respectively. Han et al. [28] used C-T-D to extract the physicochemical information of protein sequences, where C, T, and D represented composition, transformation, and distribution descriptors, respectively, and SVM was applied to establish the prediction model. PseAAC [29] was a method which could extract the sequence order information of proteins, which used hydrophobicity, hydrophilicity and side chain mass to change the character signal into the numerical signal. Recently, researchers have fused multiple feature information to predict RBPs. A multi-feature fusion method called PRBP was proposed by Ma et al. [30], which predicted the RNA-binding residues of a given protein sequence and then determined whether it was a RNA-binding protein. This method merged evolutionary information, physicochemical information and amino acid composition feature to construct feature vectors, which were subsequently input into the random forest classifier. Prediction results concluded that multi-information fusion could significantly improve the overall performance of the model.

With the development of machine learning and data mining, amounts of new methods are proposed. Zheng et al. [31] proposed a Deep-RBPPred method based on deep learning by improving RBPPred model. They used less physicochemical properties and deep convolutional neural networks instead of SVM to overcome the shortcomings of the original model. The ACC and MCC on three independent test sets were all significantly improved. Bressin et al. [32] developed a model called TriPepSVM that directly used the string kernels to process protein sequences using tri-peptide frequencies and combined SVM to predict human and bacterial RNA-binding proteins. Li et al. [33] proposed a method called DeBooster which extracted sequence information of proteins and predicted RNA-binding proteins using multiple decision trees with boosting method. They applied the basic bag-of-words model to encode the features of protein sequences for the first time. Then the RNA-binding proteins in CLIP-seq data were predicted based on deep boosting. To a certain extent, the impact of experimental noise generated by using CLIP-seq technology could be weakened. However, the above methods still have some shortcomings, which mainly include three aspects: (1) There is a data imbalance problem in the process of RBPs prediction, which is caused by the imbalance of positive and negative samples. And the existing machine learning methods can not deal with such problems very well. (2) It is not effective to predict RBPs based on the single information of protein sequence [34], and there is still much room for prediction of RBPs using machine learning. (3) Multi-information fusion [35–37] can effectively represent the biological information of RBPs, but it inevitably brings redundancy and noise information. Therefore, how to eliminate irrelevant features and retain important feature subset is also a problem that needs to be solved urgently.

Enlightened by this, we propose a new method for predicting RNA-binding proteins called RBPro-RF. First, we use composition-transition-distribution (C-T-D), pseudo-amino acid composition (PseAAC) and position-specific scoring matrix-400 (PSSM-400) to extract feature vectors of protein sequences. So the physicochemical information, pseudo-amino acid information and evolutionary information of protein sequences are obtained. Secondly, we apply the SMOTE and the ENN algorithm to solve the imbalance problem of positive and negative samples. The redundancy and noise information can be reduced by the elastic net. Then, the optimal feature vectors are input into random forest classifier for RBPs prediction. Finally, we evaluate the predictive performance of RBPro-RF model via ten-fold cross-validation, and the ACC, MCC, and AUC of the training set are 97.43%, 0.933, and 0.995, respectively. Furthermore, the ACC of the independent test sets *Human, S. cerevisiae and A. thaliana* achieve 95.63%, 88.82% and 92.35%, respectively. The experimental results show that the RBPro-RF model can significantly improve the prediction effect of RNA-binding proteins prediction.

To develop a really useful predictor, we need to follow Chou's 5-steps rule which is demonstrated and summarized by a series of publications [38–41]: (1) select or construct a valid benchmark dataset to train and test the predictor; (2) use effective formulations to represent samples to truly reflect the internal connection between the dataset and target; (3) introduce or develop powerful algorithms to make predictions; (4) use cross-validation to objectively evaluate the expected prediction accuracy; (5) establish a user-friendly web-server accessible for the public.

2. Materials and methods

2.1. Datasets

This study selects four different datasets constructed by Zhang et al. [42], including one training set and three independent test sets. Zhang et al. first selected RNA-binding proteins in the UniProt database [43], using PISCES [44] to select non-RNA-binding proteins with a sequence identity of 25%, X-ray resolution higher than 3.0 Å, and amino acid number between 50 and 10000 in the PDB database [45]. Then, CD-HIT [46] was employed to remove redundant sequences of all proteins with sequence identity greater than or equal to 25%. Afterward, the proteins whose structural and evolutionary information could not be obtained using PSI-BLAST [47,48] were removed. Finally, 2780 RNA-binding proteins and 7093 non-RNA-binding proteins were remained, which are chosen as the training set in this paper. Furthermore, the independent test sets from *Human*, *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Arabidopsis thaliana* (*A. thaliana*) are constructed using the same method, including 967 *Human* RNA-binding proteins and 597 non-RNA-binding proteins, 354 *S. cerevisiae* RNA-binding proteins and 135 non-RNA-binding proteins, and 456 *A. thaliana* RNA-binding proteins

and 37 non-RNA-binding proteins. The training set is used to select the optimal dimensionality reduction method, classifier and parameters, and three independent test sets are employed to evaluate the performance and generalization ability of the model.

2.1.1. Composition-transition-distribution

The different physicochemical properties of amino acids play an essential role in affecting the structure and function of proteins. We select twelve physicochemical properties, including hydrophobicity, standardized vdW volumes, polarity, polarizability, charge, secondary structure, solvent accessibility, surface tension, molecular weight, solubility in water, number of hydrogen bond donor in side chain, number of hydrogen bond acceptor in side chain. Composition-Transition-Distribution (C-T-D) [49,50] is employed to encode feature information using the above twelve properties, where C represents composition descriptor for calculating the proportion of each type of amino acid in the entire sequence, T represents a transition descriptor for indicating the frequency of conversion between two types of amino acids, and D represents a distribution descriptor used to calculate the overall distribution of various properties of amino acids throughout the sequence. We apply PROFEAT [51] to get the feature encoding data, the specific URL is http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi. In the website, we can obtain $(3+3+15) \times 12 = 252$ dimensional vectors representing each protein sequence.

2.1.2. Pseudo-amino acid composition

Chou et al. [29] proposed the concept of PseAAC which was a vital feature extraction method, and it has been widely used in many areas of computational proteomics, including structure prediction [52], protein-protein interaction prediction [53], protein subcellular localization prediction [54,55], and prediction of protein mitochondrial position [56], etc. Therefore, a powerful open access software has recently been established, called "PseAAC-General" [57]. Encouraged by the successful use of PseAAC to process protein sequences, the concept of PseKNC [58] was developed to generate various feature vectors for DNA/RNA sequences [59]. Particularly, Liu et al. [60] developed a more comprehensive Pse-in-One web server, which has been updated to Pse-in-One 2.0 [61], where users can extract feature information of DNA, RNA or protein sequences as needed. We use PseAAC to encode protein information via Pse-in-One 2.0 [62]. The obtained sequence feature vectors can be described as:

$$P = [p_1, p_2 \cdots p_{20}, p_{20+1} \cdots p_{20+\lambda}] \tag{1}$$

Each component of the protein is given by the following formula:

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k} & 1 \le u \le 20 \\ \dfrac{\omega \tau_{u-20}}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k} & 20+1 \le u \le 20+\lambda \end{cases} \tag{2}$$

where ω represents the weighting factor, Chou set it as 0.05 in the literature, and we also adopt the same setting [29]. τ_k is the sequence order information, f_u is the frequency of the u-th amino acid occurred in the protein sequences [63]. Besides, the first 20-dimensional components of the feature extracted by PseAAC represent the amino acid composition information, the latter λ dimensional feature vectors represent sequence order information and physicochemical information, which is related to the hydrophilicity, hydrophobicity, and side chain molecular.

Different parameters λ have a certain effect on the final prediction result, and λ needs to be smaller than the minimum length of the protein sequence in all data sets, so we set $\lambda \le 29$. Based on this principle, we extract feature information using PseAAC on the online server constructed by Shen et al. [29], and then the optimal λ is determined according to the prediction accuracy.

2.1.3. Position-specific scoring matrix-400

Kumar et al. [27] first applied a position-specific scoring matrix to characterize protein sequence evolutionary information for RBPs prediction, meanwhile, PSSM is also widely used in DNA binding site prediction [64], protein-protein interaction prediction [65,66], protein subcellular localization prediction [67,68], drug-target interaction prediction [69] and single amino acid polymorphism prediction [70], etc. In our study, we would obtain 20×20 dimensional feature vectors through PSSM-400. The encoding process of the feature vectors are as follows: First, the protein sequences are compared with the non-redundant database Swiss-Prot via the PSI-BLAST program [71], where the number of iterations is set as 3, the E-Value threshold is set as 0.001, and the remaining parameters are set by default. Finally, we obtain the $L \times 20$ position-specific scoring matrix (PSSM) for each protein sequence, where L represents the length of each protein sequence. Each row of the PSSM matrix represents the log likelihood score for amino acid substitutions at the corresponding positions in the query sequence.

$$PSSM = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,20} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{pmatrix}_{L \times 20} \tag{3}$$

where $P_{i,j}$ represents the score of the amino acid mutating from the $i-th$ position of the query sequence to the $j-th$ type amino acid when the element of the $i-th$ row and $j-th$ column of the matrix are in the process of evolution, and the value ranges from -9 to 11 integers. Since each protein sequence varies in length, causing the number of feature descriptors for each sample different. Nevertheless, the random forest algorithm requires the same number of features for each protein sequence. Hence, we apply the following method to convert each $L \times 20$ PSSM into a 20×20 dimensional matrix.

First, all values in the specific scoring matrix (PSSM) of each protein sequence are normalized to be between 0 and 1:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

where x^* represents the standardized data and x represents the pre-standardized data, $x_{\min}$ represents the minimum value in the specific scoring matrix (PSSM), and $x_{\max}$ represents the maximum value. Then we select the rows belonging to the same amino acid from PSSM to obtain 20 submatrices, where each of which has N (the frequency of each amino acids in the protein sequence) rows and 20 columns. Finally, all the rows of each submatrix are added and merged into the 400-dimensional feature vectors.

2.2. SMOTE method

In this study, the SMOTE oversampling algorithm is applied to handle the imbalance problem, which was first proposed by Chawla et al. [72]. The basic idea is to analyze the minority samples, and synthesize new samples artificially based on them, then add them to the dataset. We apply the edited nearest neighbors (ENN) [73] to process the oversampled positive and negative samples data in order to eliminate redundant samples.

SMOTE algorithm could synthesize $N \times T$ new samples using T minority samples of the training set, where N is a positive integer. The specific steps are as follows:

Suppose the feature vectors of a certain minority sample to be x_i, $x_i \in \{1, 2 \cdots T\}$.

1. For the samples x_i in the minority samples, we calculate the Euclidean distance from this sample to other samples in the minority samples, and find k nearest neighbor samples, denoted as $x_{i(n)}, n \in \{1, 2 \cdots k\}$.
2. A sample $x_{i(nn)}$ is randomly selected from the k nearest neighbors, and then generate a random value from 0 to 1. The new sample x_{i1} could be synthesized using linear interpolation between $x_{i(nn)}$ and x_i.
3. Repeat step 2 for N iterations to synthesize N new samples $x_{inew}, i \in 1, 2 \cdots N$.

The principle of ENN is that, for a sample from majorities, if more than half of its K neighbor points do not belong to the majorities, the sample will be rejected. In order words, if all K neighbor points do not belong to the majority samples, and the sample will also be rejected.

In this paper, C-T-D, PSSM-400, and PseAAC are combined to extract the features of each protein sequence, and the SMOTE and ENN are applied to balance samples. Therefore, balanced and more effective samples could be obtained.

2.3. Elastic net

Zou et al. [74] proposed elastic net (EN) based on LASSO, which could solve the multicollinearity problem between variables, so that it could process high-dimensional samples more effectively. EN has been widely applied in the field of disease diagnosis, such as relapsing-remitting multiple sclerosis [75]. The naive elastic net can be expressed as follows.

$$L(\alpha_1, \alpha_2, \beta) = |y - X\beta|^2 + \alpha_2 \sum_{i=1}^{m} \beta_j^2 + \alpha_1 \sum_{i=1}^{m} |\beta_i| \tag{5}$$

where α_1, α_2 represent the penalty parameter, β_i represents the regression coefficient of the i-th independent variable, m represents the number of independent variables, y represents the response variables,$y = (y_1, y_2 \cdots y_n)^T$. Let $\lambda = \alpha_2/(\alpha_1 + \alpha_2)$, then the above model can be expressed as:

$$\begin{aligned} &\widehat{\beta}^{NEN} = \operatorname{argmin}|Y - X\beta|^2 \\ &subject\ to\ \ \lambda \sum_{i=1}^{m} \beta_i^2 + (1-\lambda) \sum_{i=1}^{m} |\beta_i| \le \gamma \end{aligned} \tag{6}$$

where γ represents a non-negative parameter, and $\lambda \sum_{i=1}^{m} \beta_i^2 + (1-\lambda) \sum_{i=1}^{m} |\beta_i|$ represents a penalty function of EN. Based on the naive elastic net model, the elastic net can be obtained through scaling the regression coefficients.

$$\widehat{\beta}^{EN} = (1 + \alpha_2)\widehat{\beta}^{NEN} \tag{7}$$

We apply the elastic net algorithm to remove the redundant information representing RBPs information, so the more essential feature information and better prediction performance could be obtained.

2.4. Random forest

The random forest (RF) [76] is the ensemble method whose basic classifier is the decision tree. The RF has excellent anti-noise ability, which could be suitable for processing large data sets and balancing the errors caused by the imbalance of positive and negative samples. Moreover, it has been widely applied in protein subcellular localization. [77], drug-target interaction prediction [78], and prediction of protein-protein interaction sites [79], etc.

The random forest applies the decision tree as the base predictor, and the prediction result is obtained combined with bagging algorithm via the voting principle. Bagging is a resampling technique, which can improve the generalization ability.

When constructing a single decision tree using the random forest algorithm, it is assumed that the training set contains n samples, and the number of feature dimension of each sample is M. First, n samples are randomly selected from the dataset through resampling to construct the basic decision tree. Afterward, $m(m \le M)$ feature components are randomly selected as nodes from M dimensional feature vectors. Finally, we select the optimal split node according to the *Gini* index. Each decision tree grows to the maximum extent without pruning, and the final prediction result is determined based on the majority voting principle. It is a binary classification problem to predict whether a protein sequence is an RNA-binding protein. RBP is labeled as 1 and non-RBP as 0. The random forest is a classification measurement algorithm based on a tree structure, the number of decision trees is set as 500, and the algorithm is implemented using Python's Scikit-learn library [80].

2.5. Performance evaluation and model construction

For the RNA-binding proteins prediction model constructed in this paper, the performance of the model is evaluated by the four indicators of specificity (SP), sensitivity (SN), accuracy (ACC), and matthews correlation coefficient (MCC), which are defined as follows.

$$SP = \frac{TN}{TN + FP} \tag{8}$$

$$SN = \frac{TP}{TP + FN} \tag{9}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \tag{10}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{11}$$

where *TN* represents true negative, i.e. predicting the correct number of non-RNA-binding proteins. *TP* represents true positive, i.e. predicting the correct number of RNA-binding proteins. *FN* represents false-negative, i.e. the number of RNA-binding proteins that are incorrectly predicted to be non-RNA-binding proteins. *FP* represents false-positive, i.e. the number of non-RNA-binding proteins that are incorrectly predicted to be RNA-binding proteins. The above indicators can be converted into the following simpler and more intuitive equations [81–83].

$$SP = 1 - \frac{N_+^-}{N^-} \tag{12}$$

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

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

$$MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \tag{15}$$

where N^+ represents the total number of RNA-binding proteins, N^- represents the total number of non-RNA-binding proteins, N_-^+ represents the number of RNA-binding proteins which are incorrectly predicted to be non-RNA-binding proteins, and N_+^- represents the number of non-RNA-binding proteins which are incorrectly predicted to be RNA-binding proteins. MCC is a comprehensive indicator which can measure the relationship between prediction results and real results. In addition, the ROC curve [84] and PR curve can reflect the generalization performance of the model, and the area under the ROC curve (AUC) and the area under the PR curve (AUPR) are used as a quantitative indicator of the model's robustness. The larger the AUC and AUPR are, the greater

the generalization performance of the model is.

We use 10-fold cross-validation method to evaluate the model's performance on the training set, and the three independent test sets are used to verify the predictive and generalization ability of the model. For convenience, the RNA-binding proteins prediction model proposed in this paper is called RBPro-RF, and the calculation process is as shown in Fig. 1 below. The experimental environment is Windows Server 2012R2 Intel (R) Xeon (TM) CPU E5-2650 @ 2.30 GHz 2.30 GHz with 32.0 GB of RAM, implemented by MATLAB2014a and Python3.6 programming.

The specific steps of the RBPro-RF prediction model proposed in this paper are as follows:

(1) We obtain four RBPs datasets systematically, including one training set and three independent test sets, and then add them the corresponding labels.
(2) C-T-D is applied to extract the physicochemical information and $(3+3+15)\times 12 = 252$ dimensional vectors are generated. For PseAAC, we extract $20+\lambda$ dimensional feature vectors. PSSM-400 is adapted to represent evolutionary information, and then we can obtain a 20×20 dimensional vector. Therefore, the $252+(20+\lambda)+400$ dimensional feature vectors can be obtained by fusing the three feature extraction methods.
(3) The SMOTE and ENN are employed to balance the positive and negative samples. Then elastic net is applied to eliminate redundant features on the balanced data, which achieves better performance than other dimensional reduction methods.
(4) The optimal feature vectors are input into the random forest classifier for RBPs prediction, and the RBPro-RF model is built up. Meanwhile, the comparison of different classifiers indicates RF achieve better prediction performance.
(5) Calculate the AUC, AUPR, ACC, SN, SP, and MCC via ten-fold cross-validation on the training set. The ROC and PR curves are also plotted to evaluate the predictive performance of the model.
(6) The model is also tested using three independent test sets *Human, S. cerevisiae and A. thaliana.* And the prediction results indicate RBPro-RF yields good accuracy for RBPs prediction.

3. Results and discussion

3.1. The parameter selection of PseAAC

The parameters in the feature extraction algorithm have a vital effect on a prediction model. In this paper, PseAAC reflects the sequence information of protein. In the process of PseAAC feature extraction, the parameter λ represents the sequence correlation factor related to the proteins. If the parameter λ is set too large, the dimension of the feature vectors will be too high, which will cause more redundant information to be generated and adversely affect the prediction of RBP. If the parameter λ is too small, it may cause the loss of protein feature information. The parameter λ should be selected according to the length of the protein sequences in the dataset. Since the minimum sequence length of the dataset is 29, we set the parameter values as 0, 5, 10, 15, 20, 25, in order. For different λ values, we use random forest to predict RBPs. The final prediction results are shown in Table 1.

As we can see from Table 1, there are significant differences in model performance by choosing different parameter values. When the parameter λ is equal to 15, the ACC is 97.29% which is the best. The SP and SN also reach the maximum value, which is significantly higher than that of other parameter values. At the same time, when the parameter λ is equal to 15, the AUC value is 0.996, which is the highest. In other words, the feature selection method under this parameter value is the most robust. Meanwhile, the MCC of the model reaches 0.936. Therefore, we choose the parameter λ to be 15.

In order to more intuitively evaluate the prediction accuracy of the model with different parameters, we plot Supplementary materials Fig. S1 to show the change of the ACC when different λ values are selected on the training data set. From Supplementary materials Fig. S1, we can conclude that the prediction accuracy changes with the change of the value of parameter λ. When the parameter λ value is equal to 15, the ACC is the highest. Therefore, we choose 15 as the optimal parameter value, and $20+\lambda = 35$ dimensional feature vectors are generated using PseAAC.

3.2. The effect of feature extraction algorithm

Effective feature information is vital for RBPs prediction [85]. In order to construct an RBPs prediction model, we select three different

Table 1
The prediction results using different parameters λ on the training set.

Parameters	ACC (%)	SP (%)	SN (%)	MCC	AUC
0	96.82	90.64	99.54	0.925	0.995
5	96.85	90.73	99.52	0.926	0.995
10	97.04	91.14	99.60	0.930	0.995
15	**97.29**	91.93	99.61	**0.936**	**0.996**
20	96.84	90.68	99.54	0.925	0.995
25	96.83	90.75	99.49	0.925	0.995

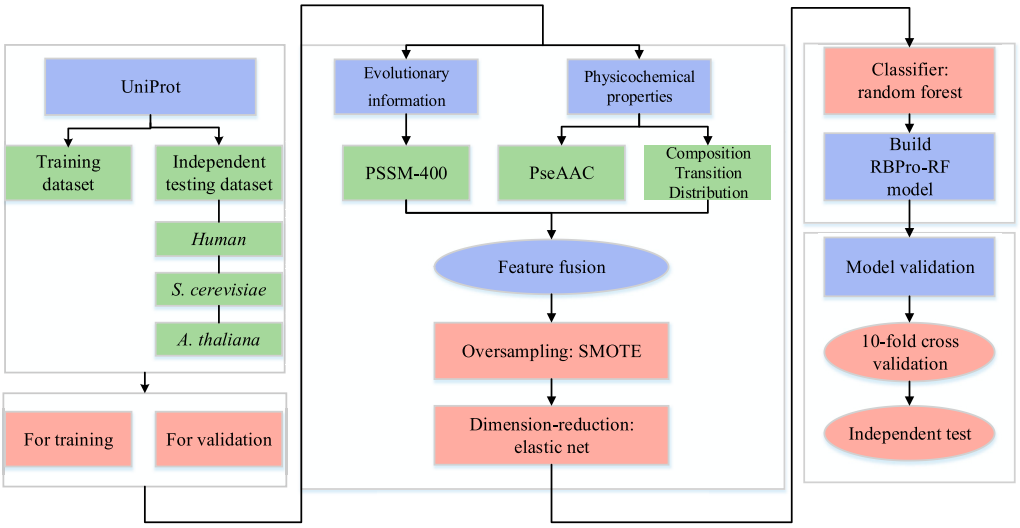

Fig. 1. The pipeline of RBPro-RF model for RNA-binding proteins prediction.

Table 2
The comparison results of different feature extraction methods on the training set.

Parameters	ACC (%)	SP (%)	SN (%)	MCC
PSSM-400	94.80	93.95	95.35	0.891
C-T-D	97.25	85.01	99.91	0.904
PseAAC	94.55	86.26	99.23	0.884
PSSM-CTD-PseAAC	**97.37**	90.97	99.64	**0.931**

feature extraction algorithms. C-T-D reflects the physicochemical information, the PseAAC reflects the sequence information of protein sequences, and the PSSM-400 algorithm reflects protein evolutionary information. We can obtain $(252) + (20 + \lambda) + (400)$ dimensional feature vectors by fusing C-T-D, PseAAC, and PSSM-400. In order to evaluate the validation of feature fusion [86,87], we compare four different feature combinations, including PSSM-400, C-T-D, PseAAC, PSSM-CTD-PseAAC. PSSM-CTD-PseAAC represents the fusion of PSSM-400, C-T-D, and PseAAC. In addition, the SMOTE and ENN are employed to deal with the data imbalance problem. The feature vectors are input into the random forest to predict RBPs via ten-fold cross-validation. The experimental results of the RBPs prediction model are shown in Table 2.

Firstly, it can be seen from Table 2 that for the single feature algorithm, the C-T-D achieves the optimal performance with the accuracy of 97.25%, which is about 2.5% higher than the PSSM-400 and PseAAC, and the SN and MCC values of the C-T-D algorithm also reach the highest. Hence, it indicates that physicochemical information is superior to evolutionary information and sequence information of protein sequences when predicting RBPs, which can further improve the prediction accuracy of the model. Secondly, the PSSM-CTD-PseAAC has the highest ACC of 97.37%, which is 2.57%, 0.12%, and 2.82% higher than those of PSSM-400, C-T-D, PseAAC, respectively. And the MCC of PSSM-CTD-PseAAC is 0.931, which is 4%, 2.7%, and 4.7% higher than PSSM-400, C-T-D and PseAAC, respectively, indicating that the overall performance of the method achieves optimal results. In addition, the ACC of PSSM-CTD-PseAAC is only 0.12% higher than C-T-D, which shows that physicochemical properties play an important role in improving ACC. In summary, the prediction indexes of the three feature fusion algorithms have reached the highest value, indicating the feature fusion improves the model performance.

The RNA-binding proteins prediction model in this paper is a binary classification problem, which is used to distinguish positive and negative samples. It can be evaluated by ROC curve and PR curve, and the robustness and generalization performance of the prediction model with different feature extraction algorithms are compared in Fig. 2.

As can be seen from Fig. 2 (A), for the training set, the AUC and AUPR values of the PseAAC algorithm are higher than the other two algorithms when only one type of feature information is extracted, indicating that the prediction model constructed by this method has better robustness and generalization ability. The area under the ROC curve of the PSSM-CTD-PseAAC combined with the three algorithms reaches the maximum, and the AUC reaches the highest value of 0.995, i.e., the prediction model using the PSSM-CTD-PseAAC feature extraction method has the optimal robustness. PSSM-CTD-PseAAC is 0.6%, 1%, and 0.4% higher than PSSM-400, C-T-D, and PseAAC, respectively. From Fig. 2 (B), the PSSM-CTD-PseAAC achieves better precision than other three feature extraction methods at almost every recall value. And the AUPR of PSSM-CTD-PseAAC reaches the maximum value of 0.990, which is 0.8%, 2.1%, and 0.1% higher than the PSSM-400, C-T-D, and PseAAC algorithms, respectively. It shows that the model performance using PSSM-CTD-PseAAC algorithm is the best.

3.3. The effectiveness of the SMOTE algorithm

The training set contains 2780 RNA-binding proteins and 7093 non-RNA-binding proteins, and the number of positive and negative samples are obviously imbalanced. Since the premise of most classifier algorithms is that the sample data classification ratio is almost equal, the prediction model obtained by training tends to favor non-RNA-binding proteins because of the more negative samples [88], resulting in a higher error rate of RNA-binding proteins with fewer samples. In practical applications, researchers pay great attention to whether the model can predict the minority samples. Hence, in order to weaken the impact of imbalanced data on the final prediction accuracy, the researchers have proposed a variety of effective methods, roughly summarized into two categories [89]: one is from the perspective of machine learning algorithms, which uses the modification of the algorithm to solve the problem [90,91]. The other is to directly process dataset through sampling, such as random over-sampling, under-sampling, multi-classifier system (MCS) [92], and SMOTE, etc.

In this study, the feature vectors extracted by C-T-D, PseAAC, and PSSM-400 are combined. The imbalanced data is processed using SMOTE and ENN. And the elastic net is performed to reduce redundancy and noise information. Then the RNA-binding protein prediction model is constructed by applying random forest as a classifier. Afterward, we compare the results of RBPro-RF model using SMOTE with the results of model without SMOTE, and the specific results are shown in Table 3. Finally, according to the ACC and other indicators, we could determine whether the oversampling process needs to be performed.

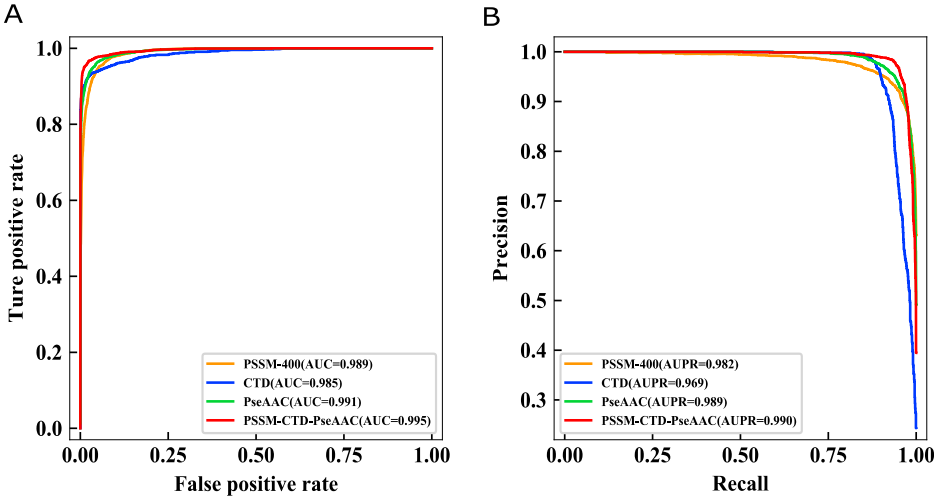

Fig. 2. Performance evaluation of prediction model with four feature extraction methods. (A) ROC curves of four feature extraction methods. (B) PR curves of four feature extraction methods.

Table 3
The RBPs prediction results before and after the SMOTE method on the training set.

Methods	ACC (%)	SP (%)	SN (%)	MCC	AUC
With SMOTE	**97.43**	91.26	99.61	**0.933**	**0.995**
Without SMOTE	91.19	97.01	76.33	0.777	0.969

Table 4
The comparison of different dimensionality reduction methods on RBPs training set.

Methods	ACC (%)	SP (%)	SN (%)	MCC
LLE	92.84	74.40	99.37	0.812
MDS	84.74	41.70	99.99	0.587
LE	93.27	84.20	96.48	0.824
ET	93.10	89.50	96.70	0.865
EN	**97.43**	91.26	99.61	**0.933**

From Table 3, it can be seen that the ACC of the RNA-binding proteins prediction model oversampling by SMOTE algorithm is 97.43%, the MCC is 0.933, and the AUC is 0.995. Without SMOTE, the ACC, MCC, and AUC are 91.19%, 0.777 and 0.969, which achieve worse performance than with SMOTE. Therefore, the feature vectors of the imbalanced data oversampled by SMOTE algorithm can improve the model prediction accuracy, comprehensive performance and robustness when constructing the RNA-binding proteins prediction model.

3.4. The selection of dimensionality reduction methods

When the dimension of the feature vectors are too high, the presence of redundant information may have a bad effect on prediction performance. Therefore, it is necessary to select the optimal feature subset to eliminate noise information. For the prediction of RNA-binding proteins, taking the most effective dimensionality reduction method can effectively improve the prediction accuracy. In this study, we compare the prediction accuracy corresponding to different dimensionality reduction method. Five different dimensionality reduction methods are chosen, including locally linear embedding (LLE) [93], multidimensional scaling (MDS) [94], laplacian eigenmaps (LE) [95], Extra-Trees (ET) [96], elastic Net (EN). The EN algorithm combines the advantages of LASSO and ridge regression. It can not only achieve sparse processing of high-dimensional variables, but also retain important variables from raw variables and handle multiple collinearity problems among partial variables. Hence, the more strongly related variables can be incorporated in the final model. For the protein sequences in the training set, the optimal feature extraction algorithms of C-T-D, PseAAC, and PSSM-400 are adopted, and the SMOTE is applied to oversample in order to obtain the balanced data. Furthermore, the processed feature vectors of different dimensionality reduction methods, and the optimal feature vectors are input into a random forest classifier for prediction. Based on the ten-fold cross--validation, the ACC, SP, SN, and MCC can all be obtained, as shown in Table 4.

As can be seen from Table 4, the accuracy of the RNA-binding protein prediction model constructed through different dimensionality reduction methods has significant differences. Among them, EN has the highest ACC of 97.43%, which is 4.59%, 12.69%, 4.16%, and 4.33% higher than LLE, MDS, LE, and ET. Meanwhile, the MCC of EN reaches the maximum value of 0.933, which is 12.1%, 34.6%, 10.9%, and 6.80% higher than other methods, indicating that the overall performance of the model is optimal. In addition, the ACC and MCC of LLE, MDS, LE, and ET all decrease compared with the situation without dimensionality reduction, while the ACC and MCC values are increased when EN is used. And when evaluating models on independent test sets, the performance of the model using the dimensionality reduction method is significantly better than that of the model without the dimensionality reduction method. Supplementary materials Fig. S2 shows the ACC of the RBPs prediction model with different dimensionality reduction methods, and it can be clearly seen that the model prediction accuracy is the highest when adopting EN dimension reduction.

For different dimensionality reduction methods, we adopt the ROC curve and the PR curve to evaluate the robustness and generalization ability of the prediction model. Fig. 3 is the ROC curve and the PR curves obtained by RNA-binding proteins prediction model with five different dimensionality reduction methods, respectively.

From Fig. 3, for the area under the ROC curve, EN obtains the highest AUC of 0.995, which is 1.0%, 0.6%, 1.7%, and 1.4% higher than ET, MDS, LLE, and LE, respectively, indicating that the robustness of the RNA-binding proteins prediction model constructed by the five dimensionality reduction methods all achieve excellent results, but the EN method works optimal. For the area under the PR curve in Fig. 3, the AUPR of EN method also reaches the highest value of 0.990, indicating that the RNA-binding proteins prediction model constructed by this method is optimal for both prediction accuracy and robustness. Therefore, we choose EN to reduce the dimension so that the redundant information of the feature vectors can be removed and the prediction accuracy would be improved.

3.5. The selection of classification algorithms

With the development of artificial intelligence and big data, machine

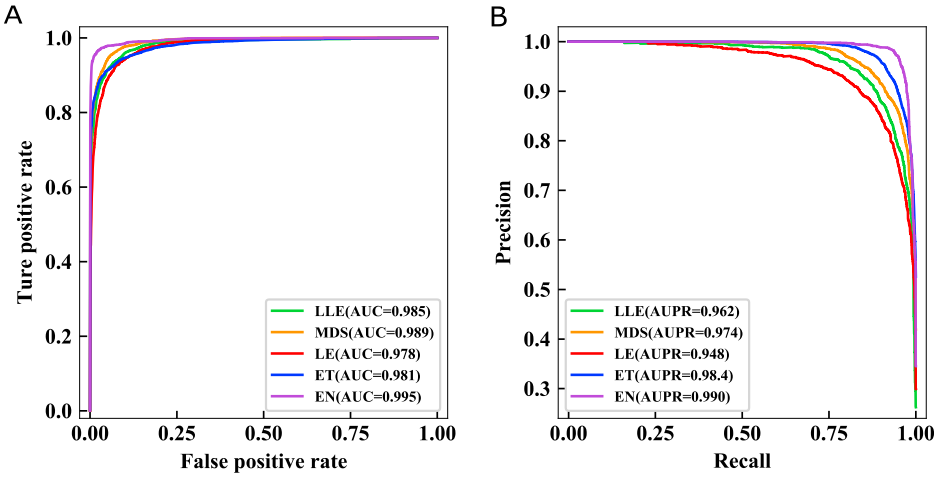

Fig. 3. Performance evaluation of predictive models using different dimensionality reduction methods. (A) ROC curves of five dimensionality reduction methods. (B) PR curves of five dimensionality reduction methods.

Table 5
The RBPs prediction results using different classifiers on the training set.

Classifier	ACC (%)	SP (%)	SN (%)	MCC
KNN	90.40	63.45	99.94	0.748
Logistic regression	94.81	87.12	97.53	0.864
Naïve Bayes	80.63	54.84	89.76	0.492
Decision tree	90.78	89.82	91.74	0.816
MLP	94.84	89.07	96.89	0.872
SVM	89.66	87.71	90.35	0.749
Random forest	**97.43**	**91.26**	**99.61**	**0.933**

learning algorithms such as random forest, decision tree, and Naïve Bayes are increasingly being applied to the research of bioinformatics [97]. In order to select the most effective method to predict RBPs, we select six different classifiers for discussion, including K-nearest neighbor (KNN) [98], logistic regression [99], Naïve Bayes [100], decision tree [101], multilayer perceptron (MLP) [102], support vector machine (SVM) [103], random forest [104]. K-nearest neighbor is a simple common algorithm for classifying the attributes of K sample points with the nearest samples. In this paper, the value of parameter K is chosen to be 10, and the weight of neighbors is set as the reciprocal of distance. Logistic regression is a probabilistic nonlinear regression model for researching the relationship among class labels and independent variables, and the inverse of regularization strength C is set as 10^4, the solver selects sag. Naïve Bayes is a simple probability classifier adopting Bayesian principle, and the MultinomialNB classifier is as a classifier, that is, the prior distribution is selected as a multi-distribution event model. The decision tree model is a tree-based machine learning method. Multilayer perceptron (MLP) is a classifier based on feedforward artificial neural network which adopts a reverse algorithm and a neural network model consisting of multiple node layers. SVM is a classification method that adopts the known effective algorithm to obtain the global optimal solution of the objective function, and the linear kernel function is chosen. The random forest is established using a number of decision trees in a random way, and the final result is determined by the majority of voting. Besides, the number of decision trees is chosen to be 500, and the max-depth is 10. In our study, the retain parameters of the classifier are set as default values. For the protein sequence of the training set, first, we extract feature vectors by fusing C-T-D, PseAAC, and PSSM-400 algorithms. Secondly, the imbalanced data is sampled by SMOTE and the elastic net is adapted to remove the redundant information. Finally, the obtained feature vectors are input into seven different classifiers for prediction, and the prediction results are shown in Table 5.

As can be seen from the data in Table 5, the use of different classifiers has a significant effect on the performance of the model. The accuracy of the RNA-binding proteins prediction model using the Naive Bayes classifier is 80.63%, which is much lower than the other six methods, indicating that this classifier is not suitable for the model classification prediction. The ACC of KNN, logistic regression, Naïve Bayes, decision tree, MLP, and SVM are 90.40%, 94.81%, 80.63%, 90.78%, 94.84%, 89.66% respectively, which also achieve poorer prediction performance than RF. The accuracy of RNA-binding protein prediction model constructed by random forest is the highest, reaching 97.43%, which is 7.03%, 2.62%, 16.80%, 6.65%, 2.59%, 7.77% higher than KNN, logistic regression, Naïve Bayes, decision tree, MLPC, SVM classifiers, respectively. Moreover, MCC also reaches the highest value of 0.933, and SP and SN are 91.26% and 99.61%, respectively. For seven different classifiers, we use ROC curve and PR curve to evaluate the robustness and generalization ability of the prediction model. Fig. 4 shows ROC curve and PR curve obtained by the RNA-binding protein prediction model with different classifiers.

It can be seen from Fig. 4 that the area under the ROC curve of random forest has the highest AUC value of 0.995, which is 1.1%, 1.0%, 15.5%, 3.0%, 0.1%, 3.4% higher than KNN, logistic regression, Naïve Bayes, decision tree, MLP, and SVM classifiers, respectively. Meanwhile, the AUPR also reaches the highest value of 0.990, so it can be concluded that the random forest algorithm is the best from the prediction accuracy, the robustness or the comprehensive performance of the model. Since we mainly seek the optimal classifier based on the prediction accuracy and the comprehensive performance of the model, we choose random forest as the classifier of the prediction model.

3.6. Comparison with state-of-the-art approaches

At present, amounts of methods have been proposed for RNA-binding protein prediction. Zhang et al. [42] constructed an RBPPred prediction model based on SVM classifier and constructed four data sets, including a training set and three independent test sets. In our study, we adopt the datasets to establish RBPro-RF model. In order to evaluate this model more effectively, we compare the final results of RBPPred [42] and RBPro-RF prediction model based on the ten-fold cross-validation method, as shown in Table 6.

For the training set, it can be concluded from Table 6 that the accuracy of the RBPro-RF prediction model after combining the three kinds of feature information is 97.43%, which is 4.79% higher than the RBPPred prediction model adopting the same dataset, indicating the model we establish reaches higher prediction accuracy. In addition, the AUC of the RBPro-RF model is 0.995, which is higher than the AUC of the RBPPred model, indicating that our model has relatively better robustness and generalization ability. Moreover, the MCC of the RBPro-RF model reaches 0.933, which is 17.9% higher than the RBPPred model, indicating that the overall performance of our model is greatly improved. In

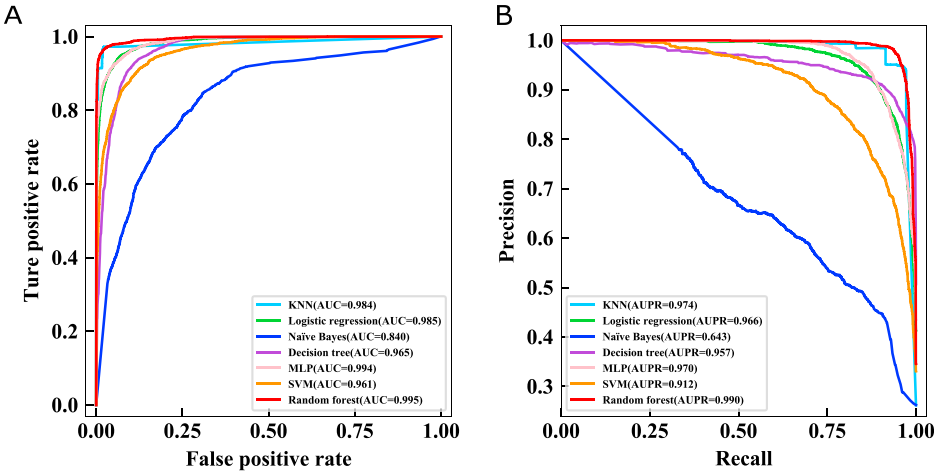

Fig. 4. Performance evaluation of predictive models using different classifiers. (A) ROC curves for seven classifiers. (B) PR curves for seven classifiers.

Table 6
Comparison of RBPPred model and RBPro-RF model on the training set.

Results	ACC (%)	SP (%)	SN (%)	MCC	AUC
RBPPred_All Features	92.64	96.50	82.77	0.754	0.946
RBPro-RF_ All Features	97.43	91.26	99.61	0.933	0.995

Table 7
Comparison of RBPPred and RBPro-RF on the independent test set.

Results	ACC (%)	SP (%)	SN (%)	MCC
RBPPred_*Human*	89.00	96.65	84.28	0.788
RBPro-RF _*Human*	**95.63**	95.62	95.65	**0.902**
RBPPred_*S. cerevisiae*	87.73	91.85	86.16	0.729
RBPro-RF _*S. cerevisiae*	**88.82**	84.40	99.27	**0.779**
RBPPred_*A. thaliana*	87.02	94.59	86.40	0.537
RBPro-RF _*A. thaliana*	**92.35**	91.45	94.31	**0.833**

summary, the performance of the RNA-binding proteins prediction model of this paper is greatly better than the RBPPred model.

To further verify the predictive ability of our model, the RBPro-RF model is employed to predict RBP on three independent data sets of *Human*, *S. cerevisiae* and *A. thaliana*. We list the results of the comparison with RBPPred model, as shown in Table 7.

From Table 7, we can obtain that for the three test sets of *Human, S. cerevisiae and A. thaliana*, the prediction accuracy of RBPro-RF model is 95.63%, 88.82%, and 92.35%, respectively, which is 6.63%, 1.09%, 5.33% higher than the RBPPred model, respectively, and the highest model prediction accuracy reaches 95.63%. The MCC reaches 0.902, 0.779, and 0.833, respectively, which is 11.4%, 5.0%, and 29.6% higher than the RBPPred model, indicating that the overall performance of the model is also greatly enhanced on the test set. The results indicate that the RBPro-RF model can significantly improve the prediction accuracy of RNA-binding proteins, and has better generalization ability. After verification on independent test sets, it further indicates that the model has high prediction accuracy and wide application. In order to more intuitively compare the prediction performance of the three independent test sets under the RBPro-RF model and the RBPPred model, Fig. 5 is a histogram of the prediction accuracy ACC of the three independent test sets under the above two models.

It can be seen from Fig. 5 that the prediction accuracy of the RBPro-RF model constructed in this paper is significantly higher than the RBPPred model. Besides, the ACC of the independent test set *Human* and *A.thaliana* are significantly improved, indicating that RBPro-RF model achieves excellent prediction performance and generalization ability.

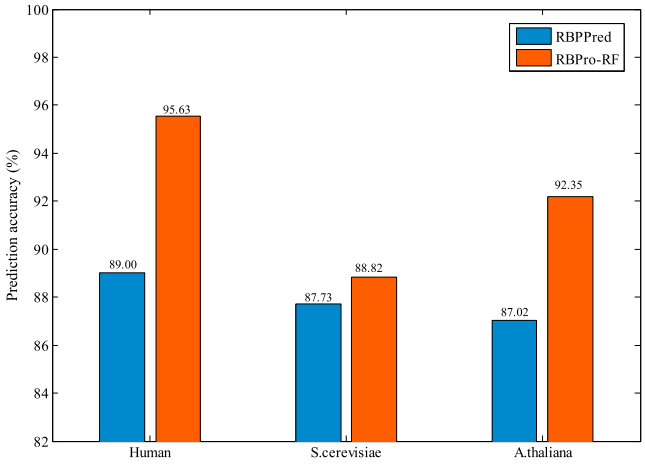

Fig. 5. Comparison of RBPro-RF and RBPPred on the independent test set.

4. Conclusion

RNA-binding proteins play an important role in various life activities of organisms. Because a large number of protein sequences are collected into the database, and traditional experimental methods are time-consuming and costly, which can not meet the needs of life science research. The machine learning-based RNA-binding protein prediction methods become a research hotspot. In this paper, we propose a new method RBPro-RF for RNA-binding proteins prediction. First, we fuse C-T-D, PseAAC and PSSM-400 to extract the feature vectors of protein sequences. Secondly, we apply SMOTE and ENN to balance the samples. Then the elastic net is employed to eliminate the redundant features and the random forest is chosen to predict the RNA-binding proteins. The C-T-D feature extraction method represents different physicochemical information. The PseAAC algorithm can comprehensively illuminate the sequence information of RBPs. And PSSM-400 can characterize the sequence of evolutionary information. We use SMOTE and ENN to solve the problem of imbalance between positive and negative samples. The elastic net can effectively eliminate the redundant information and retain the important features to represent RBPs information, which could improve the prediction accuracy of the model. The random forest algorithm is suitable for processing large data sets, which has strong anti-noise ability and fast training speed. Based on the ten-fold cross-validation method, the prediction accuracy of the RBPro-RF model on training set achieves 97.43%, and the overall average prediction accuracy on the three independent test sets *Human, S. cerevisiae and A. thaliana* achieve 95.63%, 88.82%, and 92.35%, respectively. Compared with the-state-of-the-art methods, our model can effectively improve the prediction accuracy of RNA-binding proteins. Although the RBPro-RF model can achieve good performance for RBPs prediction, there is still much room for improvement. In the next step, we will research the imbalance of RNA-binding protein datasets and construct larger datasets. Deep learning has shown better performance in bioinformatics, which could improve prediction accuracy and model robustness, and using deep learning to predict RBPs is also our next research orientation. As pointed out in Refs. [105,106] and demonstrated in a series of publications [107–109] when presenting a new discovery, a user-friendly and publicly accessible web-server will significantly enhance its impact [18], so we will make efforts in the future work to provide a web-server to satisfy the different needs of users. Besides, some pioneering papers from Chairman of Nobel Prize Committee Sture Forsen and follow-up papers [110–112] can promote in-depth investigation of this topic, and we will use these papers for future research.

Declaration of competing interest

The authors declare no conflict of interest.

CRediT authorship contribution statement

Xiaomeng Sun: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Tingyu Jin:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Cheng Chen:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Xiaowen Cui:** Formal analysis, Investigation, Writing - original draft. **Qin Ma:** Formal analysis, Writing - original draft, Writing - review & editing. **Bin Yu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review & editing.

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 Key Research and

Development Program of Shandong Province of China (2019GGX101001), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), and College Students' Innovative Entrepreneurial Training Program (No. 201810426204). 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.2019.103919.

References

[1] S. Gerstberger, M. Hafner, T. Tuschl, A census of human RNA-binding proteins, Nat. Genet. 15 (2014) 829–845.
[2] J.B. Dictenberg, S.A. Swanger, L.N. Antar, R.H. Singer, G.J. Bassell, A direct role for FMRP in activity-dependent dendritic mRNA transport links filopodial-spine morphogenesis to fragile X syndrome, Dev. Cell 14 (2008) 926–939.
[3] M. Flodstrom-Tullberg, M. Hultcrantz, A. Stotland, A. Maday, D. Tsai, C. Fine, B. Williams, R. Silverman, N. Sarvetnick, RNase L and double-stranded RNA-dependent protein kinase exert complementary roles in islet cell defense during coxsackievirus infection, J. Immunol. 174 (2005) 1171–1177.
[4] B. Tian, P.C. Bevilaequa, A. Diegelman-Parente, M.B. Mathews, The double-stranded RNA-binding motif: interference and much more, Nat. Rev. Mol. Cell Biol. 5 (2004) 1013–1023.
[5] T.A. Cooper, L. Wan, G. Dreyfuss, RNA and disease, Cell 136 (2009) 777–793.
[6] F. Bolognani, M.A. Merhege, J. Twiss, N.I. Perrone-Bizzozero, Dendritic localization of the RNA-binding protein HuD in hippocampal neurons: association with polysomes and upregulation during contextual learning, Neurosci. Lett. 371 (2004) 152–157.
[7] R.G. Fred, N. Welsh, The importance of RNA binding proteins in preproinsulin mRNA stability, Mol. Cell. Endocrinol. 297 (2009) 28–33.
[8] A. Kamijo-Ikemori, T. Sugaya, K. Matsui, T. Yokoyama, K. Kimura, Roles of human liver type fatty acid binding protein in kidney disease clarified using hL-FABP chromosomal transgenic mice, Nephrology 16 (2011) 539–544.
[9] P. Bielli, R. Busa, M.P. Paronetto, C. Sette, The RNA-binding protein sam68 is a multifunctional player in human cancer, Endocr. Relat. Cancer 18 (2011) R91–R102.
[10] A. Ke, J.A. Doudna, Crystallization of RNA and RNA-protein complexes, Methods 34 (2004) 408–414.
[11] L.G. Scott, M. Hennig, RNA structure determination by NMR, Methods Mol. Biol. 452 (2008) 29–61.
[12] 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.
[13] J. Dev, D. Park, Q. Fu, J. Chen, H.J. Ha, F. Ghantous, T. Herrmann, W. Chang, Z. Liu, G. Frey, M.S. Seaman, B. Chen, J.J. Chou, Structural basis for membrane anchoring of HIV-1 envelope spike, Science 353 (2016) 172–175.
[14] K.C. Chou, Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor, Biochem. Biophys. Res. Commun. 319 (2004) 433–438.
[15] K.C. Chou, Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein, J. Proteome Res. 4 (2005) 1681–1686.
[16] X.B. Li, S.Q. Wang, W.R. Xu, R.L. Wang, Novel inhibitor design for hemagglutinin against H1N1 influenza virus by core hopping method, PLoS One 6 (2011), e28111.
[17] K.C. Chou, Modeling the tertiary structure of human cathepsin-E, Biochem. Biophys. Res. Commun. 331 (2005) 56–60.
[18] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.
[19] N.Q.K. Le, E.K.Y. Yapp, Y.Y. Ou, H.Y. Yeh, iMotor-CNN: identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule, Anal. Biochem. 575 (2019) 17–26.
[20] W. Hussain, Y.D. Khan, N. Rasool, S.A. Khan, SPrenylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins, J. Theor. Biol. 468 (2019) 1–11.
[21] Q. Ning, Z. Ma, X. Zhao, dForml(KNN)-PseAAC: detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components, J. Theor. Biol. 470 (2019) 43–49.
[22] I. Paz, E. Kligun, B. Bengad, Y. Mandel-Gutfreund, BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins, Nucleic Acids Res. 44 (2016) W568–W574.
[23] H. Zhao, Y. Yang, Y. Zhou, Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets, Nucleic Acids Res. 39 (2011) 3017–3025.
[24] H.Y. Zhao, Y.D. Yang, Y.Q. Zhou, Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction, RNA Biol. 8 (2011) 988–996.
[25] Q. Wang, Z.H. Luo, J.C. Huang, Y.H. Feng, Z.L. Liu, A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM, Comput. Intell. Neurosci. 2017 (2017) 1827016.
[26] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[27] M. Kumar, M.M. Gromiha, G.P.S. Raghava, M. Kumar, M.M. Gromiha, G.P.S. Raghava, SVM based prediction of RNA-binding proteins using binding residues and evolutionary information, J. Mol. Recognit. 24 (2015) 303–313.
[28] L.Y. Han, Prediction of RNA-binding proteins from primary sequence by a support vector machine approach, RNA 10 (2004) 355–368.
[29] 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.
[30] X. Ma, J. Guo, K. Xiao, X. Sun, PRBP: prediction of RNA-Binding proteins using a random forest algorithm combined with an RNA-Binding residue predictor, IEEE/ACM Trans. Comput. Biol. Bioinform. 12 (2015) 1385–1393.
[31] J.F. Zheng, X.L. Zhang, X.Y. Zhao, X.X. Tong, X. Hong, J. Xie, S.Y. Liu, Deep-RBPPred: predicting RNA binding proteins in the proteome scale based on deep learning, Sci. Rep. 8 (2017) 15264.
[32] A. Bressin, R. Schulte-Sasse, D. Figini, E.C. Urdaneta, B.M. Beckmann, A. Marsico, TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs, Nucleic Acids Res. 47 (2019) 4406–4417.
[33] S. Li, F. Dong, Y. Wu, S. Zhang, C. Zhang, X. Liu, T. Jiang, J. Zeng, A deep boosting based approach for capturing the sequence binding preferences of RNA-binding proteins from high-throughput CLIP-seq data, Nucleic Acids Res. 45 (2017) e129.
[34] B. Yu, Y. Zhang, A simple method for predicting transmembrane proteins based on wavelet transform, Int. J. Biol. Sci. 9 (2013) 22–33.
[35] H. Zhou, C. Chen, M. Wang, Q. Ma, B. Yu, Predicting Golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion, IEEE Access 7 (2019) 144154–144164.
[36] B. Yu, W.Y. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, Q. Ma SubMito-XGBoost, Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics (2019), https://doi.org/10.1093/bioinformatics/btz734.
[37] C. Chen, Q. Zhang, Q. Ma, B. Yu, LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion, Chemometr. Intell. Lab. Syst. 191 (2019) 54–64.
[38] X. Du, Y. Diao, H. Liu, S. Li, MsDBP: exploring DNA-binding proteins by integrating multi-scale sequence information via chou's 5-steps rule, J. Proteome Res. 18 (2019) 3119–3132.
[39] M. Kabir, S. Ahmad, M. Iqbal, M. Hayat, iNR-2L: A two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families, Genomics (2019), https://doi.org/10.1016/j.ygeno.2019.02.006.
[40] M. Tahir, H. Tayara, K.T. Chong, iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule, Chemometr. Intell. Lab. Syst. 189 (2019) 96–101.
[41] I. Nazari, M. Tahir, H. Tayari, K.T. Chong, iN6-Methyl (5-step): identifying RNA N6-methyladenosine sites using deep learning mode via Chou's 5-step rules and Chou's general PseKNC, Chemometr. Intell. Lab. Syst. (2019), https://doi.org/10.1016/j.chemolab.2019.103811.
[42] X.L. Zhang, S.Y. Liu, RBPPred: predicting RNA-binding proteins from sequence using SVM, Bioinformatics 33 (2017) 854–862.
[43] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H.Z. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S.L. Yeh, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 32 (2004) D115–D119.
[44] G. Wang, R.L. Dunbrack, PISCES: a protein sequence culling server, Bioinformatics 19 (2003) 1589–1591.
[45] K.B. Cook, H. Kazan, K. Zuberi, Q. Morris, T.R. Hughes, RBPDB: a database of RNA-binding specificities, Nucleic Acids Res. 39 (2011) D301–D308.
[46] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[47] C.N. Magnan, P. Baldi, SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, Bioinformatics 30 (2014) 2592–2597.
[48] S.F. Altschul, T.L. Madden, A.A. Schffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[49] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 8700–8704.
[50] T. Huang, X.H. Shi, P. Wang, Z. He, K.Y. Feng, L. Hu, Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks, PLoS One 5 (2010), e10972.
[51] P. Zhang, L. Tao, X. Zeng, C. Qin, S.Y. Chen, F. Zhu, S.Y. Yang, Z.R. Li, W.P. Chen, Y.Z. Chen, PROFEAT update: a protein features web server with added facility to compute network descriptors for studying omics-derived networks, J. Mol. Biol. 429 (2016) 416–425.
[52] B. Yu, L.F. Lou, S. Li, Y.S. 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.
[53] 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.
[54] 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, Chemometr. Intell. Lab. Syst. 167 (2017) 102–112.
[55] 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.
[56] 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) 85–103.
[57] 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.
[58] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.
[59] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (2015) 2620–2634.
[60] 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.
[61] B. Liu, F. Liu, X.L. Wang, J. Chen, L.Y. 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.
[62] 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.
[63] Y.D. Cai, S.L. Lin, Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence, Biochim. Biophys. Acta 1648 (2003) 127–133.
[64] S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, BMC Bioinform. 6 (2005) 33.
[65] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, A. Masoudi-Nejad, PPIevo: protein-protein interaction prediction from PSSM based evolutionary information, Genomics 102 (2013) 237–242.
[66] X.Y. Wang, B. Yu, A.J. Ma, C. Chen, B.Q. Liu, Q. Ma, Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique, Bioinformatics 35 (2019) 2395–2402.
[67] B. Yu, S. Li, W.Y. Qiu, M.H. Wang, J.W. Du, Y.S. 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.
[68] 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.
[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 11 (2019) 1839–1852.
[70] T. Huang, C. Wang, G. Zhang, SySAP: a system-level predictor of deleterious single amino acid polymorphisms, Protein Cell 3 (2012) 38–43.
[71] P. Mundra, M. Kumar, K.K. Kumar, K. Valadi Jayaraman, D. Bhaskar Kulkarni, Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM, Pattern Recognit. Lett. 28 (2007) 1610–1615.
[72] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[73] B. Tang, H.B. He, Enn: extended nearest neighbor method for pattern recognition, IEEE Comput. Intell. Mag. 10 (2015) 52–60.
[74] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. B. 67 (2005) 301–320.
[75] C. Zhao, A.G. Deshwar, Q. Morris, Relapsing-remitting multiple sclerosis classification using elastic net logistic regression on gene expression data, Syst. Biomed. 14 (2013) 247–253.
[76] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[77] M. Tahir, A. Khan, A. Majid, A. Lumini, Subcellular localization using fluorescence imagery: utilizing ensemble classification with diverse feature extraction strategies and data balancing, Appl. Soft Comput. 13 (2013) 4231–4243.
[78] R.S. Olayan, H. Ashoor, V.B. Bajic, DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches, Bioinformatics 34 (2017) 1164–1173.
[79] B.Q. Li, K.Y. Feng, L. Chen, T. Huang, Y.D. Cai, Prediction of protein-protein interaction sites by random forest algorithm with mRMR and ifs, PLoS One 7 (2012), e43927.
[80] F. Pedregosa, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, G. Louppe, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[81] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, K.C. Chou, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther. Nucleic Acids 7 (2017) 155–163.
[82] B. Liu, F. Yang, K.C. Chou, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Mol. Ther. Nucleic Acids 7 (2017) 267–277.
[83] 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.
[84] J. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.
[85] J. Rahman, N.I. Mondal, K.B. Islam, A.M. Hasan, Feature fusion based SVM classifier for protein subcellular localization prediction, J. Integr. Bioinform. 13 (2016) 23–33.
[86] G. Huang, Y. Zhang, L. Chen, N. Zhang, T. Huang, Y.D. Cai, Prediction of multi-type membrane proteins in human by an integrated approach, PLoS One 9 (2014), e93553.
[87] T. Huang, S. Niu, Z. Xu, Y. Huang, X. Kong, Y.D. Cai, Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties, PLoS One 6 (2011), e22940.
[88] P. Kang, S. Cho, EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems, in: ICONIP 2006: Neural Information Processing, Springer Berlin Herdelberg, 2006, pp. 837–846.
[89] R. Batuwita, V. Palade, microPred: effective classification of pre-miRNAs for human miRNA gene prediction, Bioinformatics 25 (2009) 989–995.
[90] R. Akbani, S. Kwek, N. Japkowicz, Applying support vector machines to imbalanced datasets, in: ECML 2004: Machine Learning, Springer Berlin Heidelberg, 2004, pp. 39–50.
[91] T. Imam, K.M. Ting, J. Kamruzzaman, z-SVM: An SVM for improved classification of imbalanced data, in: AI 2006: Advances in Artificial Intelligence, Springer Berlin Heidelberg, 2006, pp. 264–273.
[92] R. Batuwita, V. Palade, microPred: effective classification of pre-miRNAs for human miRNA gene prediction, Bioinformatics 25 (2009) 989–995.
[93] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[94] A.M. Bronstein, M.M. Bronstein, R. Kimmel, Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching, Proc. Natl. Acad. Sci. U. S. A 103 (2006) 1168–1172.
[95] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[96] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[97] C. Wang, A modified machine learning method used in protein prediction in bioinformatics, Int. J. Bioautomotion 19 (2015) 25–36.
[98] H.B. Shen, K.C. Chou, Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types, Biochem. Biophys. Res. Commun. 334 (2005) 288–292.
[99] B. Wang, X.F. Wang, P. Howell, X. Qian, K. Huang, A.I. Riker, J. Ju, Y. Xi, A personalized microRNA microarray normalization method using a logistic regression model, Bioinformatics 26 (2010) 228–234.
[100] Y.C. Dou, X.M. Guo, L.L. Yuan, D.R. Holding, C. Zhang, Differential expression analysis in RNA-seq by a naive Bayes classifier with local normalization, BioMed Res. Int. 2015 (2015) 789516.
[101] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Sys. Man Cyber. 21 (1991) 660–674.
[102] S.K. Pal, S. Mitra, Multilayer perceptron, fuzzy sets, and classification, IEEE Trans. Neural Netw. 3 (1992) 683–697.
[103] 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.
[104] 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, Chemometr. Intell. Lab. Syst. 184 (2019) 28–43.
[105] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.
[106] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.
[107] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362–369.
[108] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524–3531.
[109] X. Cheng, W.Z. Lin, X. Xiao, pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics 35 (2019) 398–406.
[110] K.C. Chou, Biological functions of low-frequency vibrations (phonons). III. Helical structures and microenvironment, Biophys. J. 45 (1984) 881–889.
[111] I.W. Althaus, A.J. Gonzales, J.J. Chou, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase, J. Biol. Chem. 268 (1993) 14875–14880.
[112] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56.