frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 25 February 2021
doi: 10.3389/fbioe.2021.647113

Prediction of lncRNA–Protein Interactions via the Multiple Information Integration

Yifan Chen[1,2], *Xiangzheng Fu*[1], *Zejun Li*[2], *Li Peng*[3] *and Linlin Zhuo*[4]*

[1] College of Information Science and Engineering, Hunan University, Changsha, China, [2] School of Computer and Information Science, Hunan Institute of Technology, Hengyang, China, [3] College of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China, [4] Department of Mathematics and Information Engineering, Wenzhou University Oujiang College, Wenzhou, China

The long non-coding RNA (lncRNA)–protein interaction plays an important role in the post-transcriptional gene regulation, such as RNA splicing, translation, signaling, and the development of complex diseases. The related research on the prediction of lncRNA–protein interaction relationship is beneficial in the excavation and the discovery of the mechanism of lncRNA function and action occurrence, which are important. Traditional experimental methods for detecting lncRNA–protein interactions are expensive and time-consuming. Therefore, computational methods provide many effective strategies to deal with this problem. In recent years, most computational methods only use the information of the lncRNA–lncRNA or the protein–protein similarity and cannot fully capture all features to identify their interactions. In this paper, we propose a novel computational model for the lncRNA–protein prediction on the basis of machine learning methods. First, a feature method is proposed for representing the information of the network topological properties of lncRNA and protein interactions. The basic composition feature information and evolutionary information based on protein, the lncRNA sequence feature information, and the lncRNA expression profile information are extracted. Finally, the above feature information is fused, and the optimized feature vector is used with the recursive feature elimination algorithm. The optimized feature vectors are input to the support vector machine (SVM) model. Experimental results show that the proposed method has good effectiveness and accuracy in the lncRNA–protein interaction prediction.

Keywords: feature representation, mutual information, structure analysis, support vector machine, lncRNA protein interactions

OPEN ACCESS

Edited by:
Zhibin Lv,
University of Electronic Science and Technology of China, China

Reviewed by:
Qiu Xiao,
Hunan Normal University, China
Wei Zhang,
Changsha University, China

***Correspondence:**
Linlin Zhuo
zhuoninnin@163.com

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *29 December 2020*
Accepted: *19 January 2021*
Published: *25 February 2021*

Citation:
Chen Y, Fu X, Li Z, Peng L and Zhuo L (2021) Prediction of lncRNA–Protein Interactions via the Multiple Information Integration. Front. Bioeng. Biotechnol. 9:647113. doi: 10.3389/fbioe.2021.647113

INTRODUCTION

Long non-coding RNA (lncRNA)–protein interactions play an important role in the post-transcriptional gene regulation, polyadenylation, splicing, and translation, and predicting lncRNA–protein interactions helps to understand lncRNA-related activities (Mittal et al., 2009; Ray et al., 2013). With the rapid advancement of high-throughput technologies and the rapid increase of lncRNA and protein sequence data, predicting lncRNA–protein interactions by traditional biological experimental approaches, such as RNA-pulldown, RNA immunoprecipitation, and other biological experiments, is expensive and time-consuming. In recent years, computational methods, especially machine learning methods, have been widely used in the field of bioinformatics. For example, Link prediction paradigms have been used to predict drug targets

(Munir et al., 2019; Srivastava et al., 2019; Zeng et al., 2019, 2020; Ru et al., 2020; Wang et al., 2020), enhancer promoter interactions (Hong et al., 2019; Cai et al., 2020a), disease genes (Zeng et al., 2017a; Ji et al., 2019; Kuang et al., 2019; Wang et al., 2019; Peng et al., 2020), link prediction (Xiao et al., 2018, 2019, 2020), circular RNAs (Zeng et al., 2017b; Xiao et al., 2019), microRNAs (miRNAs) (Xiao et al., 2018, 2020; Zeng et al., 2018; Hajieghrari et al., 2019; Jeyaram et al., 2019; Zhang X. et al., 2019), and peptide recognition (Bai et al., 2019; Cai et al., 2020b; Fu et al., 2020; Zhang and Zou, 2020). In addition, computational intelligence such as evolutionary algorithms (Song et al., 2020a,b) and unsupervised learning (Lambrou et al., 2019; Noureen et al., 2019; Zhang L. et al., 2019; Zou et al., 2020) can be applied to the field of bioinformatics. Given the efficient performance of machine learning methods in predicting lncRNA–protein interactions, the number of researchers considering machine learning methods as the first choice for predicting lncRNA–protein interactions have been increasing.

The general process of machine learning methods for predicting lncRNA–protein interactions is as follows. First, raw lncRNA and protein data are mined and analyzed separately to extract the characteristic information of lncRNA and protein. Algorithms are then designed to compute the lncRNA–protein interactions and obtain their relationships. Finally, prediction results are verified and can be used to guide biological experiments in reverse, which can reduce the cost of biological experiments and improve the efficiency of research. Currently, machine learning-based methods for predicting lncRNA–protein interactions can be divided into two main categories.

(1) Construction of prediction models on the basis of lncRNA and protein features. The feature information of lncRNA and protein can be extracted using feature extraction methods based on sequence information, structure, and various physicochemical properties, which are fused to construct feature vectors. Feature vectors are fed into machine learning classification algorithms to construct prediction models for lncRNA–protein interaction relationships. Bellucci et al. (2011) have proposed the catRAPID model for predicting lncRNA–protein interactions, which combines the protein molecular secondary structure and the position information and extracts and inputs more than 100 dimensions of feature information from protein and non-coding RNA into the random forest (RF) and the support vector machine (SVM) to train the prediction model. Muppirala et al. (2011) have developed the RPISeq method, which utilizes only lncRNA and protein sequence information and uses SVM and RF classifiers to construct a model for the prediction of lncRNA–protein association interactions. Wang et al. (2013) have applied the plain Bayesian to construct prediction models for predicting lncRNA–protein interactions on the basis of the study of Lu et al. (2013) have proposed a method called the lncPro, which extracts amino acid and nucleotide sequence information and applies the Fisher's linear discriminant method to construct the prediction model. Subsequently, Suresh et al. (2015) have proposed the RPI–Pred method, which extracts the sequence and the structural feature information of lncRNAs and proteins and the high-order 3D structural features of proteins to construct prediction models. However, the low conserved nature of lncRNA sequences makes the prediction algorithm based on lncRNA and protein feature information perform poorly in terms of accuracy and the prediction efficiency and needs to be enhanced.

(2) Heterogeneous network-based prediction model. Given the development of related experimental techniques and the accumulation of research results in the field of lncRNA, many lncRNA–protein interaction relationships have been experimentally confirmed, and researchers have successively proposed many network-based prediction algorithms to study the interaction relationships between lncRNAs and proteins. Li et al. (2015) have constructed lncRNA and protein similarity networks and combined the existing lncRNA and protein interaction data to predict unknown lncRNA–protein interaction relationships and proposed a heterogeneous network-based method called the LPIHN. The LPIHN method predicts unknown lncRNA–protein interaction relationships by constructing a heterogeneous network with the restart random walk (RWR) implemented on the constructed network to predict novel lncRNA–protein associations. Ge et al. (2016) have introduced a network dichotomy method called the LPBNI. This method performs a resource allocation procedure in the lncRNA–protein dichotomous network to evaluate candidate proteins for each lncRNA for the prediction of interaction deletions. Hu et al. (2017) have proposed a semisupervised method called the LPI–ETSLP, which reveals lncRNA–protein correlations and does not require negative samples. On the one hand, the number of known action–relationship pairs is sparse compared with the huge number of lncRNAs and proteins and directly affects the network construction and the performance of the network link prediction. On the other hand, lncRNAs or proteins with only one action–relationship in which the data behave as isolated nodes in the network and most algorithms based on network link prediction cannot effectively predict isolated nodes.

Based on the above analysis, this paper proposes a multifeature information fusion method based on lncRNA and protein sequence features and heterogeneous network topological features to predict lncRNA and protein interaction relationships. First, a novel feature extraction method based on the topological feature information of lncRNA and protein heterogeneous networks is proposed to extract the topological network features of lncRNA and protein, lncRNA sequence mutual information, the basic statistical information of lncRNA sequence bases and lncRNA expression profile features, and the evolutionary information and the composition–transition–distribution (CTD) feature information of protein sequences. Then, the above features are fused, and the fused feature information are input into the SVM to train and construct the lncRNA–protein prediction model.

MATERIALS AND METHODS

Framework of the Proposed Method

In this paper, we propose a multi-information fusion-based lncRNA–protein association prediction model consisting of three main phases, namely, (1) dataset preparation, (2) feature extraction and optimization, and (3) model training and

prediction. In the dataset preparation, candidate lncRNA and protein sequences and their interaction data are usually collected from validated databases and related literature. Good training and test sets are usually required to build a high-quality prediction model. The training set is used for model training, and the test set is used to verify the transferability and the reliability of the training model. In the feature extraction and optimization, lncRNA and protein topological network features are proposed, and the protein sequence, Position Specific Scoring Matrix (PSSM), lncRNA sequence, and lncRNA expression spectrum features are extracted. Feature vectors are usually optimized by removing some irrelevant features to improve the performance of the feature information. In the model training and prediction, the SVM is used to train the input training set, and the grid search provides SVM training parameters for the construction of the training model. The prediction is performed on the given set of prediction vectors. The overall framework of the entire lncRNA–protein association prediction model is shown in **Figure 1**.

Datasets

With the development of high-throughput sequencing technologies, many public databases are available for scientists to study lncRNA–protein interactions. The NPInter database includes experimentally validated information on interactions between non-coding RNAs and other biomolecules (e.g., proteins, RNAs, and genomic DNA). The NONCODE (Liu et al., 2005) database is a comprehensive annotation database covering all types of non-coding RNAs except tRNAs and rRNAs. The NONCODE4.0 database contains 141,353 lncRNA sequence data, covering the lncRNA sequence data required in this paper. The UniProt database (Consortium, 2018) can provide the protein sequence data required in this paper. Through the abovementioned public databases, the datasets required to study lncRNA–protein interactions can be obtained and may help in the conduct of the study.

The acquisition and the preprocessing of datasets usually consist of two main steps, i.e., candidate data collection and invalid data rejection. (1) Candidate data collection, human lncRNA, and its association term data are extracted from the NPInter V2.0 database (Yuan et al., 2013; Hao et al., 2016), and 4,870 pairs of experimentally identified lncRNA–protein interaction datasets, which include 1,114 lncRNAs and 96 proteins, are obtained. Then, the lncRNA sequence information is obtained from the NONCODE 4.0 database, and the protein sequence information is obtained from the UniProt database. (2) Eliminate invalid data; since a few lncRNA sequence data are not available in some candidate datasets, proteins and lncRNAs with unavailable sequence information should be removed. In addition, some lncRNAs that only interact or are related to one protein or proteins that only interact or are related to one lncRNA have usually low correlation and potentially noisy information. Therefore, such data are excluded.

A dataset containing 4,158 lncRNA–protein interactions (including 990 lncRNAs and 27 proteins) is constructed in this paper through the above data processing steps.

Features Extraction

In this paper, five types of feature information, namely, lncRNA–protein network topology features, protein evolution information (Shao et al., 2020), protein sequence features (Liu et al., 2019), lncRNA sequence features, and lncRNA expression profile feature information, are extracted for the lncRNA–protein association prediction.

lncRNA–Protein Network Topology Features

The lncRNA–protein network can be regarded as a heterogeneous undirected graph. Suppose that the lncRNA–protein network contains N lncRNAs and M proteins and that the sets of lncRNAs and proteins are denoted by L and P, respectively, then $L = \{l_1, l_2, l_3, \ldots, l_N\}$, and $P = \{p_1, p_2, p_3, \ldots, p_M\}$. The set of edges E of this bipartite graph is denoted by $E = \{e_{ij} \mid l_i \in L,\ p_j \in P, e_{ij} = e_{ji}\}$.

If any node l_i and p_j have an interaction, then $e_{ij} = 1$, and vice versa $e_{ij} = 0$. The interaction feature L_{ij} between any lncRNA node l_i and protein node p_j is denoted as the set of edge values of node l_i and all other protein nodes except node p_j, i.e., $e_{ij} \notin L_{ij}$,$L_{ij} = \{e_{i1}, e_{ij-1}, e_{ij+1}, \ldots, e_{iM}\}$. Similarly, the interaction feature P_{ji} between any protein node p_j and protein node l_i is denoted as the set of edge values of node p_j and all other lncRNAs nodes except node l_j. Then, $e_{ji} \notin P_{ji}$,$P_{ji} = \{e_{j1}, e_{j\,i-1}, e_{j\,i+1}, \ldots, e_{jN}\}$.

The lncRNA–protein network topology is characterized as:

$$\mathrm{LPNet}_{ij} = L_{ij} \cup P_{ji},\ \ i = 1, \ldots, N, j = 1, \ldots, M. \tag{1}$$

As a result, we can obtain 1,015-dimensional network features.

Protein Evolutionary Feature Information

The protein evolutionary feature information is extracted using our previously proposed K-PSSM-composition method (Fu et al., 2018). The K-PSSM-composition feature extraction method is derived from the PSSM-composition feature extraction method. The PSSM-composition, which is proposed by Sharma et al. (2015), is used to extract protein sequence features for the prediction of the protein subcellular localization. The PSSM-composition feature extraction method can mine the evolutionary information of protein sequences but loses the mutual information between 20 amino acid residues and the local information of protein sequences. For this reason, we propose the K-PSSM-composition feature method to alleviate the above problems. In this paper, we have applied the K-PSSM-composition method to extract features from the obtained protein sequence data for the collection of the protein evolutionary feature information. The K-PSSM-composition feature is calculated as shown below.

$$\begin{aligned} &K\_PSSM\_composition \\ &= \left[PSSM\_com(1), \ldots, PSSM\_com(\lambda)\right]_{1\times(400*k)} \end{aligned} \tag{2}$$

Here, $\lambda = 1, \ldots K$; $PSSM\_com(\lambda)$ denotes the submatrix features, the calculation of which is shown in Equation (3)

$$PSSM\_\mathrm{com}(\lambda) = \left[F^A, F^R, \ldots, F^\varphi\right]_{1\times 400} \tag{3}$$

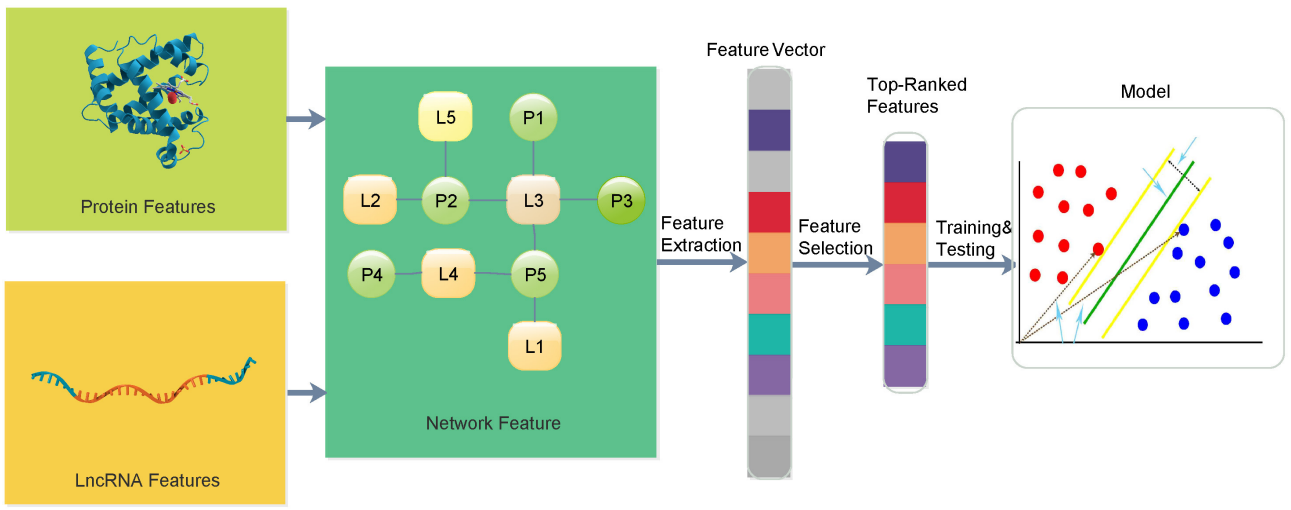

FIGURE 1 | The overall framework of the proposed method for lncRNA–protein interactions.

Here, φ denotes the 20 amino acid residues {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. F^{φ} represents the row sum of amino acid residues in the sub-PSSM matrix. In this study, $k = 1$; thus, we obtain a total of 400 dimensional features.

Protein Sequence Feature Information

In this paper, we have used the CTD (Cai et al., 2003) to extract protein sequence features, which represent the distribution patterns of specific structural or physicochemical properties in a protein or peptide sequence. Twenty amino acids are divided into three groups on the basis of different amino acid properties and represented by three feature descriptors, namely, composition (C), transition (T), and distribution (D). C denotes the percentage frequency of a specific set of amino acid properties in the calculated protein sequence, T depicts the percentage frequency of amino acids characterizing a specific property followed by another property, and D denotes the amino acid fragment describing a specific property of the whole protein sequence. Thirteen physicochemical properties have been used to calculate CTD features. Here, we use the iFeature (Chen et al., 2018) to set default parameters to extract CTD feature information and obtained a total of 504 dimensional features.

lncRNA Sequence Features

The extracted lncRNA sequence feature information contains two categories, namely, the lncRNA sequence mutual and the base compositional feature information. The lncRNA sequence mutual information is extracted using our previously proposed PSFMI feature extraction method (Fu et al., 2019) by using the entropy and the mutual information to calculate the interdependence between two bases on a given lncRNA sequence. Specifically, the 3- and the 2-gram mutual information (MI) are calculated as the characteristic information of a given lncRNA sequence.

In this study, we used entropy and MI to calculate the interdependence between bases on a given lncRNA sequence. Specifically, the 3-gram MI and the 2-gram MI were calculated separately as the characteristic information of the given lncRNA sequences. The procedure of the 3-gram triplet mutual information calculation is shown in Equation (4).

$$MI(x, y, z) = MI(x, y) - MI(x, y|z) \tag{4}$$

Here x, y, and z denote three bases that are consecutively adjacent to each other, and the equations for the calculation of MI(x, y) and conditional mutual information MI(x, $y|z$) are as follows.

$$MI(\mathrm{x}, \mathrm{y}|\mathrm{z}) = H(\mathrm{x}|\mathrm{z}) - \mathrm{H}(\mathrm{x}|\mathrm{y}, \mathrm{z}) \tag{5}$$

$$MI(\mathrm{x}, \mathrm{y}) = p(\mathrm{x}, \mathrm{y})^* \log\left(\frac{\mathrm{p}(\mathrm{x}, \mathrm{y})}{\mathrm{p}(\mathrm{x})*\mathrm{p}(\mathrm{y})}\right) \tag{6}$$

$$MI(\mathrm{x}, \mathrm{y}) = MI(\mathrm{y}, \mathrm{x}) \tag{7}$$

Where $H(x|z)$ and $H(x|y, z)$ are calculated as follows:

$$H(\mathrm{x}) = p(\mathrm{x})^* \log(\mathrm{p}(\mathrm{x})) \tag{8}$$

$$H(\mathrm{x}|\mathrm{z}) = -\frac{p(\mathrm{x}, \mathrm{z})}{p(\mathrm{z})} \log\left(\frac{p(\mathrm{x}, \mathrm{z})}{p(\mathrm{z})}\right) \tag{9}$$

$$H(\mathrm{x}|\mathrm{y}, \mathrm{z}) = -\frac{p(\mathrm{x}, \mathrm{y}, \mathrm{z})}{p(\mathrm{y}, \mathrm{z})} \log\left(\frac{p(\mathrm{x}, \mathrm{y}, \mathrm{z})}{p(\mathrm{y}, \mathrm{z})}\right) \tag{10}$$

Where $p(x)$ denotes the frequency of occurrence of base x in the lncRNA sequence, $p(x, y)$ denotes the frequency of occurrence of 2 grams of bases x and y in the lncRNA sequence, and $p(x, y, z)$ denotes the frequency of occurrence of 3 grams of bases x, y, and z in the lncRNA sequence. The values of $p(x)$, $p(x, y)$, and $p(x, y, z)$ can be calculated by Equations (11)–(13) as follows.

TABLE 1 | Parameter description in the SVM-RFE + CBR method.

Parameter	Value	Describe
kerType	2	Kernel type, see libsvm. linear: 0; rbf:2
rfeC	16	Parameter C in SVM training
rfeG	0.0078	Parameter g in SVM training
useCBR	True	Whether or not use CBR
Rth	0.9	Corrcoef threshold for highly corr features

$$p(\mathrm{x}) = \frac{\mathrm{N_x} + \varepsilon}{\mathrm{L}} \tag{11}$$

$$p(\mathrm{x,y}) = \frac{\mathrm{N_{xy}} + \varepsilon}{\mathrm{L} - 1} \tag{12}$$

$$p(\mathrm{x,y,z}) = \frac{\mathrm{N_{xyz}} + \varepsilon}{\mathrm{L} - 2} \tag{13}$$

Here, N_x denotes the number of bases x that appear in the pre-miRNA sequence and L is the length of the given lncRNA sequence. The ε in Equations (11–13), denoting a very small positive real number, is used to avoid using 0 as the denominator.

For the lncRNA base composition feature information, given any lncRNA sequence, we have calculated the percentage of 4 nucleotide (i.e., A, C, G, and T) and 16 dinucleotide (e.g., AA, AG, and AC) types in each lncRNA sequence separately and obtained 20-dimensional feature vectors. The lncRNA sequence mutual information and the lncRNA base composition feature information have 19 and 16 dimensions, respectively. Thus, the total number of lncRNA sequence feature dimensions is 35; i.e., the dimensionality of the feature vector is 35 dimensions.

lncRNA Expression Profile Features

In this paper, we have obtained the lncRNA expression profile information from the NONCODE4.0 database, which contains 170,601 lncRNA expression profile data. The expression profiles describe the expression of lncRNAs in 24 types of human tissues or cells. Thus, the lncRNA expression profile features contain 24-dimensional feature vectors.

By the above analysis, we can extract a total of 1,978 (1,015 + 400 + 504 + 35 + 24) dimensional features obtained.

Feature Optimization

The feature space of lncRNA–protein interactions consists of five features, namely, lncRNA–protein network topology, lncRNA sequence, lncRNA expression profile, protein CTD information, and protein sequence evolution information features. Compared with individual features, the fusion of multiple features can capture increased sequence information, which leads to improved prediction performance. However, the fusion of multiple features produces a high-dimensional redundant feature and may lead to problems, such as excessive training time and bias in performance. Therefore, in this paper, we have used the SVM Recursive Feature Elimination (SVM-RFE) and Correlation Bias Reduction (CBR) (Yan and Zhang, 2015) to optimize the feature set.

The SVM-RFE algorithm proposed by Tolosi and Lengauer (2011) has been successfully applied to many system biology problems. The CBR algorithm has been used to reduce potential biases in linear and non-linear SVM-RFE. In this study, we use the algorithm SVM-RFE + CBR (Yan and Zhang, 2015), which consists of a combination of SVM-RFE and CBR, to optimize the feature vectors. The specific process is as follows: first, all features are ranked using SVM-RFE + CBR (Yan and Zhang, 2015) to select a set of features with the top score; second, the selected features are reorganized into new, ordered features; and finally, these new features are fed into the predictive classifier to generate a training model. Thus, we can obtain the ranked list of features through the SVM-RFE and CBR and select a set of top-ranked feature information to enable the optimal selection of features.

In the SVM-RFE + CBR method, we used the following parameters: kerType, rfeC, rfeG, useCBR, Rth. The values and descriptions of these parameters are shown in **Table 1**. The rest of the required parameters use the default settings of the SVM-RFE + CBR method.

Classification Algorithm

In this paper, we choose SVM as the classifier to build the prediction model. Specifically, the open source Library of Support Vector Machines (LIBSVM) is used for model training and construction. The LIBSVM toolbox can be downloaded for free at http://www.csie.ntu.edu.tw/~cjlin/libsvm. We integrated the toolbox in the Matrix Laboratory (MATLAB) workspace to build predictive models. The specific form of the kernel function has a large impact on the performance of the SVM. The Gaussian radial basis kernel function (RBF) has good results for non-linear classification and is widely used for bioinformatics classification; therefore, we choose RBF as the kernel function for SVM. A grid search based on five-fold cross-validation was applied to optimize the SVM parameters γ and the penalty parameter C. The grid search yielded the optimal $C = 256$ and $\gamma = 0.002$ set as their values.

Measurements

Several measures were used to evaluate the performance of the lncRNA–protein interaction prediction method comprehensively (Jin et al., 2019; Manavalan et al., 2019; Manayalan et al., 2019; Su et al., 2019a,b, 2020a,b; Qiang et al., 2020). The receiver operating characteristic curve was based on specificity and sensitivity. The area under the receiver operator characteristic curve (AUC) and the area under precision-recall curve (AUPR) were used as evaluation metrics (Wei et al., 2014, 2017a,b; Tang et al., 2020). The AUC provided a measure of classifier performance. A high AUC value indicated improved performance of the classifier. However, for class imbalance problems, the AUPR penalizes false positives in the evaluation and is more suitable than the AUC. In addition, the Matthew correlation coefficient (MCC) was used to assess the prediction performance. The MCC considered true and false and positive and negative and was usually a balanced measure that could be used even if these classes had different sizes. Sensitivity (SE), specificity (SP), precision (PR), accuracy

TABLE 2 | Performance of different feature subsets on the benchmark dataset.

Methods	ACC (%)	SE (%)	SP (%)	MCC	F1 score (%)	AUC (%)	AUPR (%)
LDNet	90.56	77.94	97.14	0.603	64.36	89.32	71.10
Pro	85.87	69.19	98.65	0.290	26.61	57.33	27.92
lRNA	84.47	52.91	**99.77**	0.067	2.83	52.29	20.34
lRNA + Pro	86.17	68.11	98.22	0.323	31.79	79.11	47.94
lRNA + LDNet	**90.81**	**78.69**	97.20	**0.615**	**65.52**	**90.99**	**73.75**
CTD + LDNet	90.62	78.25	97.18	0.606	64.62	89.02	71.32

The best values are shown in boldface.

TABLE 3 | Comparison of performance with different excellent algorithms.

Methods	ACC (%)	F1 score (%)	AUC (%)	AUPR (%)
IRWNRLPI	90.09	65.16	**91.50**	71.38
LPI–ETSLP	88.34	59.78	88.76	64.38
RWR	95.36	36.03	83.32	28.93
LPBNI	**95.81**	38.68	85.86	33.06
RPISeq–RF	46.62	14.81	39.49	6.31
RPISeq–SVM	48.23	14.93	39.87	6.98
Our method	90.82	**65.91**	90.97	**74.39**

The best values are shown in boldface.

(ACC), and MCC are defined as follows.

$$SE = \frac{TP}{TP + FN} \tag{14}$$

$$SP = \frac{TN}{TN + FP} \tag{15}$$

$$PR = \frac{TP}{TP + FP} \tag{16}$$

$$F1 - score = 2 \times \frac{SE \times PR}{SE + PR} \tag{17}$$

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

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

TP, TN, FP, and FN indicate the number of true positives, true negatives, false positives, and false negatives, respectively.

RESULTS AND DISCUSSION

Analysis of the Effect of Different Feature Information Subsets on the Experimental Performance

The effect of different feature subsets on the experimental performance was analyzed to evaluate the effect of different feature information on the lncRNA–protein prediction performance. We compared each feature subset and their two-by-two combinations on the benchmark dataset separately.

The lncRNA sequence and the lncRNA expression profile features had feature vector dimensions of 35 and 24, respectively. These features were combined for the dimensionality of the lncRNA feature information be 59 and named as lRNA features for convenience. The CTD features of protein sequences were 273 dimensions, and the K-PSSM-composition features of protein evolutionary information were 400 dimensions. The CTD and K-PSSM-composition features were combined and named as Pro features. Thus, the Pro features of proteins were 673 dimensions. The lncRNA–protein topological network features were named LDNet features, and their total feature dimension was 1,015 dimensions. Therefore, six subsets of features [i.e., lRNA, Pro, and LDNet and their two-by-two combinations (i.e., lRNA + Pro, lRNA + LDNet, and Pro + LDNet)] were obtained. To evaluate the effect and the importance of each feature subset on the prediction results, this paper uses the SVM classifier to train the prediction model, and the grid search algorithm was employed to adjust the parameters of the SVM so that each feature subset achieves the best accuracy in the same threshold range. Five-fold cross-validation tests were conducted on these six feature subsets. Experimental results are shown in **Table 2**.

The experimental results of the six feature subsets constructed in this paper by five-fold cross-validation tests are shown in **Table 2**. The ACC, SE, MCC, F score, AUC, and AUPR values of LDNet features were 90.56, 77.94, 0.603, 64.36, 89.32, and 71.10%, respectively, and higher than those of lRNA and Pro features. For the F1 score, AUC, and AUPR metrics, the LDNet features were higher by 37.75, 31.99, and 43.18%, respectively, than the Pro features, which ranked second in these three feature subsets. Therefore, the LDNet features performed the best in the separate experiments for the three feature subsets of LDNet, Pro, and lRNA, which indicated that the LDNet was the best for the lncRNA–protein association prediction because the LDNet was the largest and far exceeded the two other feature subsets.

The ACC, SE, MCC, F score, AUC, and AUPR values for lRNA + LDNet features were 90.81, 78.69, 0.615, 65.52, 90.99, and 73.75%, respectively, and were the maximum values in these six feature subsets (**Table 1**). The values of these metrics for Pro + LDNet and lRNA + LDNet feature subsets were close. The F1 score, AUC, and AUPR values for the lRNA + Pro feature subset were 31.79, 79.11, and 47.94%, respectively, which were lower than the first two combined features and even lower than the LDNet feature subset. Therefore, the lRNA + LDNet features performed best in predicting lncRNA–protein

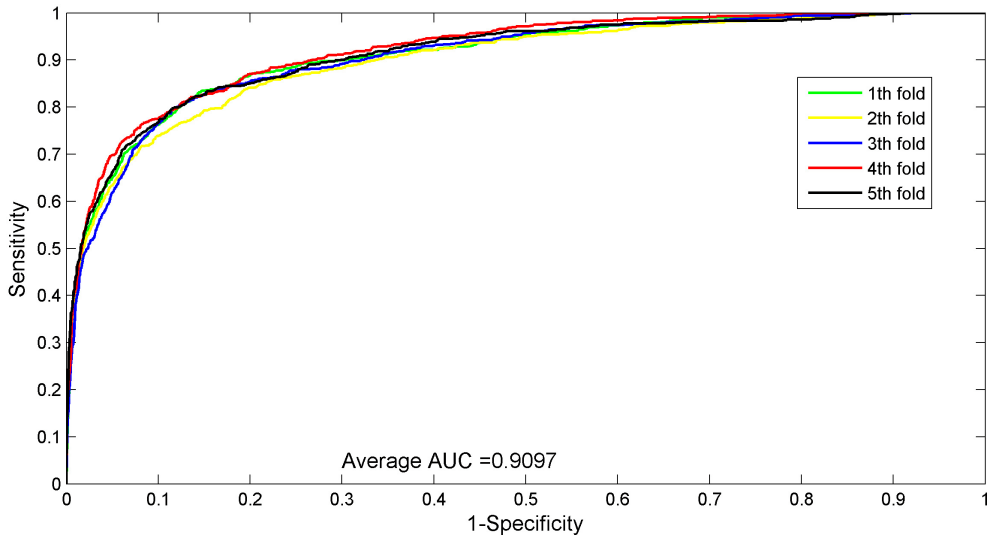

FIGURE 2 | ROC curves for five-fold cross-validation tests of the benchmark dataset.

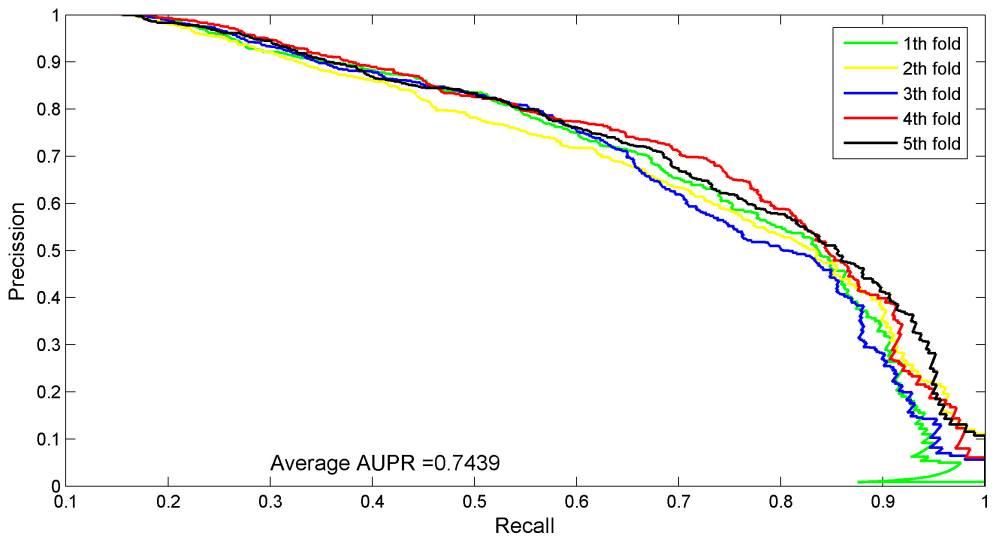

FIGURE 3 | AUPR curves for five-fold cross-validation tests of the benchmark dataset.

interactions. Among lRNA and LDNet features, the LDNet was the main decisive feature subset, which also indicated that the lncRNA and protein network topology-based features proposed in this paper had the greatest effect on the prediction performance. In addition, the performance of each feature subset in the two-by-two combination was better than the feature performance value of each feature subset individually.

Comparison With Existing Approaches

We selected the following six excellent methods for experimental comparison on the benchmark dataset to compare the performance of our proposed method with existing excellent methods. These six methods included IRWNRLPI (Zhao et al., 2018), LPI–ETSLP (Hu et al., 2017), RWR (Kohler et al., 2008), LPBNI (Li et al., 2015), RPISeq–RF (Muppirala et al., 2011), and RPISeq–SVM (Muppirala et al., 2011). The RPISeq–RF and the RPISeq–SVM models are prediction methods that extract and input lncRNA and protein features into RF or SVM predictors, whereas the IRWNRLPI, LPI–ETSLP, RWR, LPBNI, and RPISeq–RF algorithms are prediction methods that are based on heterogeneous networks constructed from lncRNAs and proteins. On the benchmark dataset, a five-fold cross-validation test was performed separately, and four evaluation metrics, namely, ACC, F1 score, AUC, and AUPR, were selected to evaluate the performance of different algorithms. Experimental results are shown in **Table 3**.

The experimental results of each evaluation index for predicting lncRNA–protein interactions are listed in **Table 3**. First, we compared the values of AUPR, which were 64.38% (LPI–ETSLP), 28.93% (RWR), 33.06% (LPBNI), 6.31% (RPISeq–RF), 6.98% (RPISeq–SVM), and 71.38% (IRWNRLPI) lower than 74.39% in our method and indicated that our method predicted reliable results.

The AUC value of our method was 90.97%, which ranked the second among all methods, and was close to the first ranked IRWNRLPI (91.50%) method and 2.21% higher than the third ranked LPI–ETSLP method. These results showed that our method had very good prediction performance. We plotted the curves of AUPR and ROC for the five-fold cross-validation tests to demonstrate the AUPR and the AUC values, respectively (**Figures 2**, **3**).

Next, we further analyzed the ACC and the F1 score values of these prediction models. The ACC of our method was 90.96% smaller than those of RWR (95.36%) and LPBNI (95.81) but better than that of IRWNRLPI (90.09%) because of very few experimentally validated lncRNA–protein interactions, which were far less than the unknown lncRNA–protein association relationships in the benchmark dataset. Therefore, the use of F1 score values to evaluate the performance of different methods than the ACC evaluation was reasonable. The F1 score value of our method was 65.91%, which was the highest among all methods and higher than those of the RWR (36.03%) and the LPBNI (38.68%). Therefore, the combined results of all experiments further demonstrated the good performance of our method in predicting lncRNA–protein associations. Notably, the four evaluation metrics (AUC, AUPR, ACC, and F1 score) of our method, which constructed prediction models on the basis of lncRNA and protein features, were more remarkable than RPISeq–RF and RPISeq–SVM.

CONCLUSIONS

lncRNAs are involved in the regulation of gene expression at the transcriptional level, epigenetics, and other life activity processes by interacting with RNA-binding proteins. Therefore, related research on the prediction of lncRNA–protein interaction relationship is beneficial in the excavation and the discovery of the mechanism of lncRNA function and action occurrence.

In this paper, a computational model for lncRNA–protein interaction relationship prediction based on the multisource information fusion is proposed. A method for representing the topological feature information of the network of lncRNA–protein interactions is proposed. Subsequently, protein evolutionary information, protein CTD sequence information features, lncRNA sequence mutual information features, and lncRNA expression profile information are extracted, and the recursive feature elimination algorithm is used to optimize feature vectors. The obtained optimized feature vectors are fed into SVM to predict lncRNA–protein interactions. Our proposed method is experimentally compared with six excellent lncRNA–protein prediction algorithms by using five-fold cross-validation tests on benchmark datasets, and experimental results show that our proposed method achieves the best performance values in AUPR and F1 score, illustrating the effectiveness and the accuracy of the proposed method in lncRNA–protein association prediction methods.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YC, XF, and LZ conceived the concept of the work. YC, XF, and LP performed the experiments. YC, ZL, and LZ wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

The work was supported in part by the National Natural Science Foundation of China (62002111, 61902125, and 61672223), in part by the China Postdoctoral Science Foundation (2019M662770), and in part by the Hunan Provincial Natural Science Foundation of China (2019JJ50187).

REFERENCES

Bai, Y., Dai, X., Ye, T., Zhang, P., Yan, X., Gong, X., et al. (2019). PlncRNADB: a repository of plant lncRNAs and lncRNA-RBP protein interactions. *Curr. Bioinform.* 14, 621–627. doi: 10.2174/1574893614666190131161002

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nat. Methods* 8:444. doi: 10.1038/nmeth.1611

Cai, C. Z., Han, L. Y., Ji, Z., Chen, X., and Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697. doi: 10.1093/nar/gkg600

Cai, L., Ren, X., Fu, X., Peng, L., Gao, M., and Zeng, X. (2020a). iEnhancer-XG: Interpretable sequence-based enhancers and their strength predictor. *Bioinformatics.* btaa914. doi: 10.1093/bioinformatics/btaa914. [Epub ahead of print].

Cai, L., Wang, L., Fu, X., Xia, C., Zeng, X., and Zou, Q. (2020b). ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. *Brief. Bioinform.* bbaa367. doi: 10.1093/bib/bbaa367. [Epub ahead of print].

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Consortium, T. U. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515.

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131

Fu, X., Zhu, W., Cai, L., Liao, B., Peng, L., Chen, Y., et al. (2019). Improved pre-miRNAs identification through mutual information of pre-miRNA sequences and structures. *Front. Genet.* 10:119. doi: 10.3389/fgene.2019.00119

Fu, X., Zhu, W., Liao, B., Cai, L., Peng, L., and Yang, J. (2018). Improved DNA-binding protein identification by incorporating evolutionary information into the Chou's PseAAC. *IEEE Access* 6, 66545–66556. doi: 10.1109/ACCESS.2018.2876656

Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding RNA–protein interactions. *Genomics Proteomics Bioinform.* 14, 62–71. doi: 10.1016/j.gpb.2016.01.004

Hajieghrari, B., Farrokhi, N., Goliaei, B., and Kavousi, K. (2019). *In silico* identification of conserved MiRNAs from *Physcomitrella patens* ESTs and their target characterization. *Curr. Bioinform.* 14, 33–42. doi: 10.2174/1574893612666170530081523

Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., et al. (2016). NPInter v3. 0: an upgraded database of noncoding RNA-associated interactions. *Database* 2016:baw057. doi: 10.1093/database/baw057

Hong, Z., Zeng, X., Wei, L., and Liu, X. (2019). Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043. doi: 10.1093/bioinformatics/btz694

Hu, H., Zhu, C., Ai, H., Zhang, L., Zhao, J., Zhao, Q., et al. (2017). LPI-ETSLP: lncRNA–protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. *Mol. Biosyst.* 13, 1781–1787. doi: 10.1039/C7MB00290D

Jeyaram, C., Philip, M., Perumal, R. C., Benny, J., Jayakumari, J. M., and Ramasamy, M. S. (2019). A computational approach to identify novel potential precursor miRNAs and their targets from hepatocellular carcinoma cells. *Curr. Bioinform.* 14, 24–32. doi: 10.2174/1574893613666180413150351

Ji, J., Tang, J., Xia, K.-,j., and Jiang, R. (2019). LncRNA in tumorigenesis microenvironment. *Curr. Bioinform.* 14, 640–641. doi: 10.2174/1574893614071909171616 54

Jin, Q., Meng, Z., Tuan, D. P., Chen, Q., Wei, L., and Su, R. (2019). DUNet: a deformable network for retinal vessel segmentation. *Knowl. Based Syst.* 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Kuang, L., Zhao, H., Wang, L., Xuan, Z., and Pei, T. (2019). A novel approach based on point cut set to predict associations of diseases and LncRNAs. *Curr. Bioinform.* 14, 333–343. doi: 10.2174/1574893613666181026122045

Lambrou, G. I., Sdraka, M., and Koutsouris, D. (2019). The "Gene Cube": a novel approach to three-dimensional clustering of gene expression data. *Curr. Bioinform.* 14, 721–727. doi: 10.2174/1574893614666190116170406

Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting long noncoding RNA and protein interactions using heterogeneous network model. *Biomed Res. Int.* 2015:671950. doi: 10.1155/2015/671950

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47:e127. doi: 10.1093/nar/gkz740

Liu, C., Bai, B., Skogerbø, G., Cai, L., Deng, W., Zhang, Y., et al. (2005). NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic acids research* 33(Suppl. 1), D112–D115. doi: 10.1093/nar/gki041

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14:651. doi: 10.1186/1471-2164-14-651

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Manayalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. *Bioinformatics* 35, 2757–2765. doi: 10.1093/bioinformatics/bty1047

Mittal, N., Roy, N., Babu, M. M., and Janga, S. C. (2009). Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20300–20305. doi: 10.1073/pnas.0906940106

Munir, A., Malik, S. I., and Malik, K. A. (2019). Proteome mining for the identification of putative drug targets for human pathogen *Clostridium tetani*. *Curr. Bioinform.* 14, 532–540. doi: 10.2174/1574893613666181114095736

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC Bioinform.* 12:489. doi: 10.1186/1471-2105-12-489

Noureen, N., Fazal, S., Qadir, M. A., and Afzal, M. T. (2019). HCVS: pinpointing chromatin states through hierarchical clustering and visualization scheme. *Curr. Bioinform.* 14, 148–156. doi: 10.2174/1574893613666180402141107

Peng, L., Zhou, D., Liu, W., Zhou, L., Wang, L., Zhao, B., et al. (2020). Prioritizing human microbe-disease associations utilizing a node-information-based link propagation method. *IEEE Access* 8, 31341–31349. doi: 10.1109/ACCESS.2020.2972283

Qiang, X., Zhou, C., Ye, X., Du, P.-F., Su, R., and Wei, L. (2020). CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief. Bioinform.* 21, 11–23. doi: 10.1093/bib/bby091

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499:172. doi: 10.1038/nature12311

Ru, X., Wang, L., Li, L., Ding, H., Ye, X., and Zou, Q. (2020). Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm. *Comput. Biol. Med.* 119:103660. doi: 10.1016/j.compbiomed.2020.103660

Shao, J., Yan, K., and Liu, B. (2020). FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network. *Brief. Bioinform.* bbaa144. doi: 10.1093/bib/bbaa144

Sharma, R., Dehzangi, A., Lyons, J., Paliwal, K., Tsunoda, T., and Sharma, A. (2015). Predict gram-positive and gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into Chou's general PseAAC. *IEEE Trans. Nanobiosci.* 14, 915–926. doi: 10.1109/TNB.2015.2500186

Song, B., Li, K., Orellana-Martín, D., Valencia-Cabrera, L., and Pérez-Jiménez, M. J. (2020a). Cell-like P systems with evolutional symport/antiport rules and membrane creation. *Inform. Comput.* 275:104542. doi: 10.1016/j.ic.2020.104542

Song, B., Zeng, X., and Rodríguez-Patón, A. (2020b). Monodirectional tissue P systems with channel states. *Inf. Sci.* 546, 206–219. doi: 10.1016/j.ins.2020.08.030

Srivastava, N., Mishra, B. N., and Srivastava, P. (2019). *In-silico* identification of drug lead molecule against pesticide exposed-neurodevelopmental disorders through network-based computational model approach. *Curr. Bioinform.* 14, 460–467. doi: 10.2174/1574893613666181112130346

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020a). Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.* 21, 408–420. doi: 10.1093/bib/bby124

Su, R., Liu, X., Wei, L., and Zou, Q. (2019a). Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009

Su, R., Liu, X., Xiao, G., and Wei, L. (2020b). Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. *Brief. Bioinform.* 21, 996–1005. doi: 10.1093/bib/bbz022

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019b). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *Ieee Acm Trans. Comput. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/TCBB.2018.2858756

Suresh, V., Liu, L., Adjeroh, D., and Zhou, X. (2015). RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Res.* 43, 1370–1379. doi: 10.1093/nar/gkv020

Tang, Y.-J., Pang, Y.-H., and Liu, B. (2020). IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformaitcs* 36, 5177–5186. doi: 10.1093/bioinformatics/btaa667

Tolosi, L., and Lengauer, T. (2011). Classification with correlated features. *Bioinformatics* 27, 1986–1994. doi: 10.1093/bioinformatics/btr300

Wang, J., Wang, H., Wang, X., and Chang, H. (2020). Predicting drug-target interactions via FM-DNN learning. *Curr. Bioinform.* 15, 68–76. doi: 10.2174/1574893614666190227160538

Wang, L., Xuan, Z., Zhou, S., Kuang, L., and Pei, T. (2019). A novel model for predicting LncRNA-disease associations based on the LncRNA-MiRNA-disease interactive network. *Curr. Bioinform.* 14, 269–278. doi: 10.2174/1574893613666180703105258

Wang, Y., Chen, X., Liu, Z.-P., Huang, Q., Wang, Y., Xu, D., et al. (2013). *De novo* prediction of RNA–protein interactions from sequence information. *Mol. Biosyst.* 9, 133–142. doi: 10.1039/C2MB25292A

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 192–201. doi: 10.1109/TCBB.2013.146

Wei, L., Wan, S., Guo, J., and Wong, K. K. L. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Xiao, Q., Luo, J., and Dai, J. (2019). Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework. *IEEE J. Biomed. Health Inform.* 23, 2661–2669. doi: 10.1109/JBHI.2019.2891779

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Xiao, Q., Zhang, N., Luo, J., Dai, J., and Tang, X. (2020). Adaptive multi-source multi-view latent feature learning for inferring potential disease-associated miRNAs. *Brief. Bioinform.* bbaa028. doi: 10.1093/bib/bbaa028. [Epub ahead of print].

Yan, K., and Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors Actuat. B Chem.* 212, 353–363. doi: 10.1016/j.snb.2015.02.025

Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2013). NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res.* 42, D104–D108. doi: 10.1093/nar/gkt1057

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017a). Prediction and validation of disease genes using HeteSim scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34, 2425–2432. doi: 10.1093/bioinformatics/bty112

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to *in silico* drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi: 10.1039/C9SC04336E

Zhang, L., He, Y., Wang, H., Liu, H., Huang, Y., Wang, X., et al. (2019). Clustering count-based RNA methylation data using a nonparametric generative model. *Curr. Bioinform.* 14, 11–23. doi: 10.2174/1574893613666180601080008

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 283–291. doi: 10.1109/TCBB.2017.2776280

Zhang, Y., and Zou, Q. (2020). PPTPP: A novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. *Bioinformatics* 36, 3982–3987. doi: 10.1093/bioinformatics/btaa275

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018). IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10. doi: 10.1093/bib/bby090

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.