Predicting IncRNA–Protein Interactions With miRNAs as Mediators in a Heterogeneous Network Model

Yuan-Ke Zhou¹, Zi-Ang Shen¹, Han Yu¹, Tao Luo¹*, Yang Gao²* and Pu-Feng Du¹*

¹ College of Intelligence and Computing, Tianjin University, Tianjin, China, ² School of Medicine, Nankai University, Tianjin, China

Long non-coding RNAs (IncRNAs) play important roles in various biological processes, where IncRNA–protein interactions are usually involved. Therefore, identifying IncRNA–protein interactions is of great significance to understand the molecular functions of IncRNAs. Since the experiments to identify IncRNA–protein interactions are always costly and time consuming, computational methods are developed as alternative approaches. However, existing IncRNA–protein interaction predictors usually require prior knowledge of IncRNA–protein interactions with experimental evidences. Their performances are limited due to the number of known IncRNA–protein interactions. In this paper, we explored a novel way to predict IncRNA–protein interactions without direct prior knowledge. MiRNAs were picked up as mediators to estimate potential interactions between IncRNAs and proteins. By validating our results based on known IncRNA–protein interactions, our method achieved an AUROC (Area Under Receiver Operating Curve) of 0.821, which is comparable to the state-of-the-art methods. Moreover, our method achieved an improved AUROC of 0.852 by further expanding the training dataset. We believe that our method can be a useful supplement to the existing methods, as it provides an alternative way to estimate IncRNA–protein interactions in a heterogeneous network without direct prior knowledge. All data and codes of this work can be downloaded from GitHub (https://github.com/zyk2118216069/LncRNA-protein-interactions-prediction).

Keywords: heterogeneous network, IncRNA–protein interaction, IncRNA–miRNA interaction, miRNA–protein interaction, network similarity

INTRODUCTION

Non-coding RNAs (ncRNAs) refer to RNAs that do not encode proteins. These genes were once considered as “junk DNAs” or “dark matters” in the genome (Schaukowitch and Kim, 2014). However, over the last few years, more and more functioning ncRNAs have been discovered, such as ribosomal RNAs (rRNA), ribozymes, transfer RNAs (tRNA), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), micro RNAs (miRNAs), long noncoding RNAs (lncRNAs), and many others (Peculis, 2000; Henras et al., 2004; Okamura and Lai, 2008; Kung et al., 2013). All these ncRNAs can influence biological progress on various levels (Louro et al., 2009).
Long non-coding RNAs are ncRNAs with a length larger than 200 nt (Kapranov et al., 2007). Experiments show that lncRNA–protein interactions play important roles in many biological processes, such as splicing, polyadenylation, and translation (Singh, 2002; Lukong et al., 2008; Kishore et al., 2010; Licatalosi and Darnell, 2010). Therefore, studying interactions between lncRNAs and proteins makes great sense for us to understand a wide variety of biological processes.

Although we can now obtain RPIs (RNA–protein interactions) through large-scale experiments such as RNAcompete (Ray et al., 2009), RIP-Chip (Keene et al., 2006), HITS-CLIP (Licatalosi et al., 2008), and PAR-CLIP (Hafner et al., 2010), all these experiments are costly and time-consuming. Therefore, computational predictions have been recognized as an efficient alternative approach. Muppilala et al. (2011) proposed the RIPSeq method for predicting RNA–protein interactions using only sequence information. Wang et al. (2013) extracted sequence-based features to represent each protein–RNA pair and used naive-Bayes classifier to predicting protein–RNA interactions. Lu et al. (2013) introduced a new method named IncPro, which scored each RNA–protein pair by encoding RNA and protein sequences into numerical vectors. Suresh et al. (2015) presented an SVM-based method, named RPI-Pred, to predict protein–RNA interaction pairs based on their sequences and structures. Li et al. (2015) developed a heterogeneous network model (LPIHN) and a random walk with restart algorithm to predict novel lncRNA–protein interactions. Ge et al. (2016) constructed the lncRNA–protein bipartite network, and scored candidate proteins for each lncRNA based on the bipartite network projection algorithm. Yang et al. (2016) constructed another lncRNA–protein bipartite network, where the HeteSim algorithm was employed to evaluate the relevance between lncRNAs and proteins. Zheng et al. (2017) applied the HeteSim algorithm on the fusion of multiple protein–protein similarity networks to predict lncRNA–protein interactions. Hu et al. (2017) presented transformation-based semi-supervised link prediction (LPI-ETSLP) to predict lncRNA–protein interactions. Xiao et al. (2017) proposed a computational method named PLPIHS for predicting lncRNA–protein interactions using HeteSim Scores. Hu et al. (2018) presented a model named HLP+Ensemble integrated three mainstream machine learning algorithms for predicting human lncRNA–protein interaction. Zhang et al. (2018b) combined multiple similarities and features with a feature projection ensemble learning frame to predict lncRNA–protein interactions. Zhang et al. (2018a) proposed a linear neighborhood propagation method (LPLNP) to calculate the linear neighborhood similarity of lncRNA–protein interactions. Zhang et al. (2019c) proposed the KATZLGO method to predict lncRNA–protein interactions based on the KATZ measure, which utilize the information of all paths between pair of nodes.

All existing methods rely on known lncRNA–protein interactions to construct the predictor. However, the number of experimentally verified lncRNA–protein interactions is limited, which affects the prediction performances of all existing methods. To expand the spectrum of predictable lncRNA–protein interactions, we took miRNAs as intermediates in predicting lncRNA–protein interactions.

miRNAs are short RNA molecules with a length of 19 to 25 nucleotides (Lu and Rothenberg, 2018). Some miRNAs can regulate both lncRNAs and proteins. For example, PTEN (Phosphatase and TENSin homolog) is a kind of tumor suppressor gene, which is critical for maintaining cellular homeostasis (Poliseno et al., 2010). The miR-21 regulates the translation process of PTEN (Zhang et al., 2010), as well as the expression of PTENpg1, which is transcribed from PTEN pseudogene as an lncRNA (Yu et al., 2014). Meanwhile, the PTENpg1 alpha isoform affects the transcription process of PTEN by competing transcription factors (Johnsson et al., 2013). We assumed that this triangular regulation network can be common in the gene regulation system. To validate this assumption, we collected lncRNA–miRNA interactions and protein–miRNA interactions from the RAID v2.0 database. We found that the lncRNA–protein interactions are significantly enriched in the set of lncRNAs and proteins that are sharing a common set of interacting miRNAs (chi-square test, p-value < 10^-10).

In the light of this observation, miRNAs were taken as mediators to predict novel lncRNA–protein interactions in this work. Both lncRNA–miRNA interactions and miRNA–protein interactions were considered as the basis to predict lncRNA–protein interactions. In the cause of improving our prediction performance, the similarity of lncRNAs and proteins was calculated in various aspects, which is based on the assumption that similar lncRNAs or proteins tend to have similar interactions. Our methods provide a way to explore novel lncRNA–protein interactions without prior knowledge of direct lncRNA–protein interactions. Since existing methods always require direct lncRNA–protein interactions as training data, our method may provide a useful supplement to the state-of-the-arts methods.

**MATERIALS AND METHODS**

**Dataset Curation**

Biomolecule interactions have become a hot research topic in computational biology. RAID v2.0 is a large database for biomolecule interaction information, which contains more than 5.27 million RNA-associated interactions, including over 4 million RNA–RNA interactions and 1.2 million RNA–protein interactions, involving nearly 130,000 genes across 60 species (Yi et al., 2017). We downloaded the protein–miRNA interactions and lncRNA–miRNA interactions as our training dataset from this database. LncRNA–protein interactions were also obtained as our independent testing dataset simultaneously.

We downloaded 2,862 lncRNA–miRNA interactions and 2,521 protein–miRNA interactions, which are all experimentally verified, from the RAID v2.0 database (Yi et al., 2017). In order to ensure that each lncRNA has a protein linked to it via a miRNA, and vice versa, common miRNAs were extracted from these interactions. Altogether 360 miRNAs were included in our dataset.
Subsequently, the lncRNA–miRNA interactions and the protein–miRNA interactions were selected according to the common interacting miRNAs. We kept 1,356 lncRNA–miRNA interactions and 1,156 protein–miRNA interactions in our dataset. These interactions are among 331 lncRNAs, 360 miRNAs, and 103 proteins. The sequences of lncRNAs and proteins were obtained from NCBI Gene database (Brown et al., 2015) and the Uniprot database (The UniProt Consortium, 2017), respectively. For those lncRNAs, which cannot be found in the NCBI Gene database, the sequence was retrieved from the Ensemble database (Hunt et al., 2018).

In order to evaluate the performance of our predictive model, we obtained experimentally verified lncRNA–protein direct interactions from the RAID v2.0 database according to the lncRNAs and proteins in our dataset. Subsequently 1,925 lncRNA–protein interactions were chosen as our independent testing dataset, which are formed by 268 lncRNAs and 58 proteins. The interactions from the RAID database are listed in the Supplementary Materials (Table S1, Table S2 and Table S3).

### Similarity Measures

Previous studies (Gong et al., 2019; Zhang et al., 2019a; Zhang et al., 2019b) have demonstrated the usefulness of similarities for network models. For convenience, let \( L \) be the set of lncRNAs, \( M \) the set of miRNAs, and \( P \) the set of proteins, e.g. \( L = \{ l_1, l_2, \ldots, l_x \} \), \( M = \{ m_1, m_2, \ldots, m_y \} \) and \( P = \{ p_1, p_2, \ldots, p_z \} \), where \( x \) denotes the number of different lncRNAs, \( y \) the number of common miRNAs, and \( z \) the number of different proteins.

The lncRNA–miRNA interaction network can be represented using a bipartite graph \( G_1 \), as follows:

\[
G_1 = (L, M, E_1),
\]

where \( E_1 \) is the set of edges in this bipartite graph, and \( L \) and \( M \) as defined above. Each edge in \( E_1 \) represents an interaction between one lncRNA and one miRNA. A part of the lncRNA–miRNA interaction network is illustrated as Figure 1.

Similarly, we used another bipartite graph \( G_2 \) to represent the protein–miRNA interaction network, as follows:

\[
G_2 = (P, M, E_2),
\]

where \( E_2 \) is the edge set of the protein–miRNA interaction network, and \( P \) and \( M \) as defined above. Each protein–miRNA interaction corresponds to an edge in \( E_2 \). A part of the protein–miRNA interaction network is illustrated as Figure 2.

With the definition of two bipartite graphs, similarities between lncRNAs or proteins were both calculated in three different ways, which are elaborated in the following sections, respectively.

### Network Similarity

For a given miRNA, \( m_k \in M \) \((k = 1, 2, \ldots, y)\), we define the set of its interacting lncRNAs as \( L(m_k) \), which is a subset of \( L \):

\[
L(m_k) = \{ l(l, m_k) \in E_1, m_k \in M, l \in L \}.
\]

We also define \( P(m_k) \), which is a subset of \( P \), as follows:

\[
P(m_k) = \{ p(l, m_k) \in E_2, m_k \in M, p \in P \}.
\]
For miRNAs in \( M \), of which the network contribution in the lncRNA–miRNA interaction network or the protein–miRNA interaction network can be calculated respectively as follows:

\[
c_1(m_k) = -\ln \left( \frac{L(m_k)}{\sum_{k=1}^y L(m_k)} \right), \quad \text{and} \quad (5)
\]

\[
c_2(m_k) = -\ln \left( \frac{P(m_k)}{\sum_{k=1}^y P(m_k)} \right), \quad \text{where} \quad (6)
\]

where \( c_1(m_k) \) is the network contribution of miRNA \( m_k \) in the lncRNA–miRNA interaction network, \( c_2(m_k) \) the network contribution of miRNA \( m_k \) in the protein–miRNA interaction network, and \( \sum \) the cardinal operator on a set.

For convenience, \( M(l_i) \) and \( M(p_j) \), which are both subsets of \( M \), are defined as follows:

\[
M(l_i) = \{ m | (m, l_i) \in E_1 \}, l_i \in L, m \in M \}, \quad \text{and} \quad (7)
\]

\[
M(p_j) = \{ m | (m, p_j) \in E_2, p_j \in P, m \in M \}, \quad \text{where} \quad (8)
\]

where \( M(l_i) \) and \( M(p_j) \) represent the set of miRNAs that interact with a given lncRNA or a given protein.

With all definition above, the network similarity between two lncRNAs \( l_u \) and \( l_v \) can be defined as follows:

\[
n_1(l_u, l_v) = \sum_{m \in M(l_u) \cap M(l_v)} c_1(m_k), \quad \text{and} \quad (9)
\]

where \( n_1(l_u, l_v) \) is the network similarity between \( l_u \) and \( l_v \).

Similarly, given two proteins, \( p_u \) and \( p_v \), the network similarity between \( p_u \) and \( p_v \) can be defined as follows:

\[
n_2(p_u, p_v) = \sum_{m \in M(p_u) \cap M(p_v)} c_2(m_k), \quad \text{where} \quad (10)
\]

where \( n_2(p_u, p_v) \) is the network similarity between \( p_u \) and \( p_v \).

**Sequence Similarity**

The sequence similarity was calculated by the Smith-Waterman algorithm. Given two lncRNAs, the sequence similarity between two lncRNA sequences is defined as follows:

\[
e_1(l_u, l_v) = \frac{w(l_u, l_v)}{|l_u| + |l_v|}, \quad \text{where} \quad (11)
\]

where \( e_1(l_u, l_v) \) is the sequence similarity between \( l_u \) and \( l_v \) the Smith-Waterman score between \( l_u \) and \( l_v \) and \( |l_u| \) and \( |l_v| \) the length of the lncRNA \( l_u \) and \( l_v \), respectively.

Given two proteins, the sequence similarity between two protein sequences is defined similarly as follows:

\[
e_2(p_u, p_v) = \frac{w(p_u, p_v)}{|p_u| + |p_v|}, \quad \text{where} \quad (12)
\]

where \( e_2(p_u, p_v) \) is the sequence similarity, and \( w(p_u, p_v), |p_u|, \) and \( |p_v| \) the length of the protein \( p_u \) and \( p_v \), respectively.

**Statistical Feature Similarity**

Pseudo-amino acid composition (PseAAC), which was proposed by Chou in 2001 (Chou, 2001), has been widely applied in all branches of computational and functional proteomics (Chou, 2011; Chou, 2015). Pseudo-k nucleotides composition (PseKNC), which is a major advancement of the PseAAC concept in analyzing nucleotide sequences, has been introduced recently (Chen et al., 2014). Because of its simplicity and effectiveness, the PseKNC methods quickly penetrate into all major topics in functional genomics, in both genome and transcriptome levels (Chen et al., 2015a; Chen et al., 2015b). The computational procedures for PseAAC and PseKNC have been elaborated in many literatures (Chou, 2011; Chen et al., 2013; Qiu et al., 2017) and some recent reviews (Chen et al., 2015a; Zhao et al., 2018).

In this work, we employed pseudo di-nucleotide composition (PseDNC), which is a special form of PseKNC when \( k = 2 \), to represent lncRNA sequences, and PseAAC for protein sequences.

For simplicity, we do not describe the computational details of the PseDNC and PseAAC algorithms here. We only describe how we apply PseDNC and PseAAC in our work.

Given a lncRNA, its PseDNC representation can be described as a numerical vector with \( 16 + \lambda \) dimensions as follows:

\[
V_1(l, |\lambda, \omega_1, H) = [d_1 \ d_2 \ \ldots \ d_{16 + \lambda}]^T, \quad (13)
\]

where \( V_1(l, |\lambda, \omega_1, H) \) is the PseDNC representation of \( l, \lambda \) and \( \omega_1 \) two parameters in computing the PseDNC representation, and \( H \) a set of di-nucleotide physicochemical properties that are applied in computing the PseDNC representations.

The similarity between two lncRNAs can be defined as follows:

\[
f_1(l_u, l_v) = 1/ \| V_1(l_u, |\lambda, \omega_1, H) - V_1(l_v, |\lambda, \omega_1, H) \|^2, \quad (14)
\]

where \( f_1(l_u, l_v) \) is the feature similarity between two lncRNAs, and \( \| \| \) the operator that takes the length of a vector.

Similarly, given a protein, its PseAAC representation can be described as a numerical vector with \( 20 + \tau \) dimensions as follows:

\[
V_2(p, |\tau, \omega_2, H) = [r_1 \ r_2 \ \ldots \ r_{20 + \tau}]^T, \quad (15)
\]

where \( V_2(p, |\tau, \omega_2, H) \) is the PseAAC representation of \( p, \tau \) and \( \omega_2 \) two parameters in computing the PseAAC representation, and \( H \) a set of amino acid physicochemical properties that are used in computing the PseAAC representations.

The similarity between two proteins can be defined as follows:

\[
f_2(p_u, p_v) = 1/ \| V_2(p_u, |\tau, \omega_2, H) - V_2(p_v, |\tau, \omega_2, H) \|^2, \quad (16)
\]

where \( f_2(p_u, p_v) \) is the feature similarity between two proteins.

We utilized online webservice Pse-In-One (Liu et al., 2015) to generate PseDNC and PseAAC in our work.

**Heterogeneous Network Model**

By integrating the bipartite graph \( G_1 \) and \( G_2 \), we can construct a heterogeneous network model, where lncRNAs, miRNAs, and proteins are connected together. A part of this network is illustrated as Figure 3. Given a lncRNA \( l_i \) and a protein \( p_j \), a whole network correlation that is brought by the \( m_k \) can be defined as follows:
The whole network correlation function between lncRNA $l_i$ and protein $p_j$ can be defined as follows:

$$k(l_i, p_j) = -\sum_{m_k \in M(l_i) \cap M(p_j)} \ln \left( \frac{t(m_k)}{\sum_{k=1}^{y} t(m_k)} \right), \quad (18)$$

The whole network correlation matrix can be established as $K = \{k(l_i, p_j)\}$, $i = 1, 2, \ldots, x$ and $j = 1, 2, \ldots, z$.

For two given lncRNAs, the similarity between them can be noted as $s_1(l_u, l_v)$, where $l_u$ and $l_v$ are two lncRNAs. Similarly, for two given proteins, the similarity between them can be noted as $s_2(p_u, p_v)$. The similarity between two lncRNAs or two proteins can be measured in various aspects, which have been elaborated in the above section.

The similarity matrix for lncRNAs and proteins can be established as $S_1 = \{s_1(l_u, l_v)\}$, $u, v = 1, 2, \ldots, x$ and $S_2 = \{s_2(p_u, p_v)\}$, $u, v = 1, 2, \ldots, z$, respectively. We normalize the values in matrix $S_1$ and $S_2$ as follows:

$$q_1(l_u, l_v) = \begin{cases} \frac{s_1(l_u, l_v)}{\sum_{y=1}^{x} s_1(l_u, l_y)} & u \neq v \\ 1 & u = v \end{cases}, \quad u, v = 1, 2, \ldots, x \quad (19)$$

$$q_2(p_u, p_v) = \begin{cases} \frac{s_2(p_u, p_v)}{\sum_{y=1}^{z} s_2(p_u, p_y)} & u \neq v \\ 1 & u = v \end{cases}, \quad u, v = 1, 2, \ldots, z \quad (20)$$

where $q_1(l_u, l_v)$ and $q_2(p_u, p_v)$ are normalized value in $S_1$ and $S_2$. We note the normalized matrix as $Q_1$ and $Q_2$, respectively, where $Q_1 = \{q_1(l_u, l_v)\}$, $u, v = 1, 2, \ldots, x$ and $Q_2 = \{q_2(p_u, p_v)\}$, $u, v = 1, 2, \ldots, z$.

With all above definitions, we can establish the final scoring matrix as follows:

$$W = Q_1 K Q_2 \quad (21)$$
The prediction of lncRNA–protein interactions is made based on the scores in W. If a value in W were larger than a given threshold, the corresponding lncRNA and protein would be predicted to interact. Otherwise, no interaction would be predicted.

The whole flowchart of our method is illustrated in Figure 4. Three different similarity measures were applied to lncRNAs and proteins, respectively. Since they can be chosen independently to each other, there are nine different combinations of the similarity choices.

### Performance Evaluation

Given a threshold, a set of lncRNA–protein interactions can be predicted from the matrix W. By comparing this set against the testing dataset, the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) can be obtained, respectively. Five statistical measures can be calculated as follows:

\[
TPR = \frac{TP}{TP + FN}, \quad (22)
\]

\[
FPR = \frac{FP}{FP + TN}, \quad (23)
\]

\[
Pre = \frac{TP}{TP + FP}, \quad (24)
\]

\[
Rec = \frac{TP}{TP + FN}, \quad \text{and} \quad (25)
\]

\[
Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (26)
\]

where TPR is for true positive rate, FPR for false positive rate, Pre for precision, Rec for recall, and Acc for accuracy.

By varying the threshold from the maximum value to the minimal value in W, a receiver operating curve (ROC) can be plotted using the TPR and the FPR values. In the meantime, a precision-recall (PR) curve can be obtained using the precision and the recall values. Due to the nature that the negatives are far more than the positives in the current topic, the area under the ROC (AUROC) and the area under the PR curve (AUPR) are used both as primary performance measures of our method.

### Parameter Calibrations

In our work, there are parameters when the PseDNC and the PseAAC sequence representations are generated. We used a grid search strategy to find the optimal parameters in the PseDNC and PseAAC. The parameter \( \lambda \) varies from 10 to 20 with a step of 1, \( \omega_1 \) from 0.1 to 1 with a step of 0.1, \( \tau \) from 10 to 20 with a step of 1, and \( \omega_2 \) from 0.05 to 0.5 with a step of 0.05. We finally choose \( \lambda = 10, \omega_1 = 0.1, \tau = 11 \) and \( \omega_2 = 0.5 \). The physicochemical properties in the PseDNC are Rise, Tilt, Twist, Slide, Shift, and Roll, which are defined in Pse-In-One (Liu et al., 2015). The physicochemical properties in the PseAAC are HOPT810101, JOND750101, ZIMJ680104, KRIW790103, TAKK010101, ROSM880104, BLAS910101, and KRIW790101, which are all defined in AAIndex (Kawashima et al., 2008).

### RESULTS AND DISCUSSION

#### Performance Analysis

We compared the prediction performance under different combinations of similarity matrices. Figure 5 illustrates the ROC and PR curve of our method with nine different similarity combinations. The AUROC and AUPR values were collected in
According to these values, the prediction performance of our method is optimized when the network similarity measure was applied to both lncRNAs and proteins. Under this condition, the AUROC achieved 0.821, while the AUPR achieved 0.376. It seems like the AUPR is low. However, by analyzing the PR curve, we found that the precision is low when the recall is in the range of (0.1, 0.4). That is to say, some lncRNA–protein pairs with a high correlation score have no experimentally verified interaction between them. This may be because these interactions are not discovered yet.

Due to the nature that the negatives are far more than the positives in predicting lncRNA–protein interactions, the testing dataset is highly imbalanced. In order to provide a set of valuable prediction results in a practical application, a recommended threshold is 2.147, which will balance the TPR and FPR, and will produce 74.3% accuracy.

### Effects of the Two Similarity Matrices

In order to analyze the effect of every single similarity matrix individually, we combined every single similarity matrix solely with the whole network correlation matrix, respectively. In other words, either $Q_1$ or $Q_2$ is removed from Eq (21) to see the effect of the other matrix solely. The ROC and the PR curve of all six different configurations are illustrated in Figure 6. The network similarity for lncRNAs performs the best among three similarities for lncRNAs. For proteins, the best similarity measure is also the network similarity. This result consists with the other results in our work. Therefore, we can safely conclude that the network similarity best suits our method. Particularly, the network similarity matrix for protein achieved a very close prediction performance to the comprehensive form of our model. Since the number of proteins and lncRNAs are imbalanced in our dataset, the number of interactions from miRNAs to proteins is far more than that to lncRNAs on average. This may be the reason why the network similarity matrix for proteins can achieve a very promising performance solely with the whole network correlation matrix.

### Comparison With Existing Methods

HeteSim is a widely applied measure, which aims at quantifying the correlation of nodes in a heterogeneous network (Shi et al., 2014). It has been used in predicting various types of interactions and connections (Shi et al., 2014). Due to the mechanism difference between our method and existing methods, it is difficult to perform a completely fair comparison. We compared our method to Yang’s work (Yang et al., 2016), where HeteSim is employed to measure the correlation between lncRNAs and proteins. In order to perform a sufficiently fair comparison, we obtained protein–protein interaction from the STRING database (Mering et al., 2003) to satisfy the requirement of Yang’s work. Same testing datasets were applied to evaluate the prediction performance of Yang’s work and our method simultaneously. However, due to the different mechanisms between our method and Yang’s work, we tested our method using the independent testing dataset, while fivefold cross-validation was applied on Yang’s method with the same dataset. Since fivefold cross-validation may produce overestimated performance values, we believe that our method achieved a comparable performance in this comparison (Figure 7).

### Table 1

| Similarity matrix | AUROC$^a$ | AUPR$^b$ |
|------------------|----------|----------|
| Net + Net        | 0.821    | 0.376    |
| Net + Seq        | 0.752    | 0.287    |
| Net + Feat       | 0.748    | 0.310    |
| Seq + Net        | 0.815    | 0.393    |
| Seq + Seq        | 0.764    | 0.308    |
| Seq + Feat       | 0.758    | 0.345    |
|Feat + Net       | 0.818    | 0.392    |
|Feat + Seq        | 0.768    | 0.309    |
|Feat + Feat      | 0.762    | 0.344    |

$^a$AUROC, Area under receiver operating curve.

$^b$AUPR, Area under precision-recall curve.

$^c$Net, Network similarity.

$^d$Seq, Sequence similarity.

$^e$Feat, Statistical feature similarity.
Prediction of Novel Interactions

In order to evaluate the actual prediction effect of our method, we selected 20 interactions that are top ranked in our results. These interactions are recorded in Table 2. Fifteen out of 20 interactions in Table 2 had been verified by CLIP-seq (Ule et al., 2005) in RAID v2.0. Two of the remaining five had been verified by eCLIP (Van Nostrand et al., 2016) in NPInter database (Teng et al., 2019). Since our method does not require any prior knowledge of direct lncRNA–protein interactions, and all data in our method came only from the RAID v2.0 database, our method should have a good performance. Although other three interactions are not verified, it is possible that they are undiscovered interactions under certain conditions.

Prediction Based on Interactions of Whole Database

Since only experimentally verified interactions were obtained to compose our benchmarking dataset, a large number of predicted interactions in the RAID V2.0 database were discarded. We incorporated these predicted interactions to optimize our method. A total of 20,425 lncRNA–miRNA interactions and 1,349 protein–miRNA interactions were extracted while sharing a common set of miRNAs. These interactions are among 1,133 lncRNAs, 464 miRNAs, and 113 proteins. We also collected 2,803 lncRNA–protein interactions as our independent testing dataset. Altogether 615 lncRNAs and 65 proteins were included in this testing dataset. Our method achieved an AUROC of 0.852 on this dataset (Figure 8). Since our method can work with only known lncRNA–miRNA interactions and miRNA–protein interactions, it can be used as a supplement to state-of-the-art methods using direct lncRNA–protein interactions.

Database Coverage Analysis

Due to the mechanism of our method, we restricted the lncRNA–protein interactions within those lncRNAs and proteins, which can find a sharing miRNA interactor. This restriction narrowed the profile of applicable data in the database. There are 2,862 experimentally verified lncRNA–miRNA interactions in the RAID v2.0 database, including 358 lncRNAs and 1,208 miRNAs; 1,356 lncRNA–miRNA interactions between 331 lncRNAs and 360 miRNAs.
miRNAs were utilized in this work, accounting for 47.4%, 92.5%, and 29.8% of which in the whole database, respectively. There are 2,521 experimentally verified protein–miRNA interactions between 144 proteins and 1,032 miRNAs in the RAID v2.0 database; 1,156 protein–miRNA interactions of them were selected as our training data, composed by 103 proteins and 360 miRNAs. The protein–miRNA interactions, proteins, and miRNAs take up 45.8%, 71.5%, and 34.9% of the entire RAID database, respectively. There are 40,668 experimentally verified lncRNA–protein interactions in the RAID v2.0 database, including 3,066 lncRNAs and 10,224 proteins. The testing dataset in our work including 1,981 verified lncRNA–protein interactions between 266 lncRNAs and 58 proteins, taking up 4.87%, 8.7%, and 0.57% of which in the database, respectively. Due to the limited number of known miRNA–protein interactions and miRNA–lncRNA interactions, the coverage of proteins in the whole database is low.

We admit that this will limit the application scope of our method. However, we believe this will get better when the number of available miRNA–protein interactions is increased, because the statistical test has already shown that the lncRNA–protein interactions are significantly enriched in the set of lncRNAs and proteins that are sharing a common set of miRNAs.

**CONCLUSION**

LncRNAs can affect biological processes from various levels. It is of great importance to study the molecular functions of lncRNAs. In the meanwhile, lncRNAs perform their role mostly by their interaction with proteins. Therefore, lncRNA–protein interaction should be studied in detail. In this paper, we proposed a method to predict lncRNA–protein interactions without prior knowledge of existing lncRNA–protein interactions. Instead, we utilized the lncRNA–miRNA interactions and the miRNA–protein interactions as the basis of our prediction. The miRNAs are used as mediators to connect the realm of lncRNAs and the realm of proteins. This is based on the hypothesis that lncRNA and a protein may interact if they share interacting miRNAs. By quantitatively modelling the heterogeneous network that is formed by lncRNAs, miRNA, and

---

**TABLE 2** | 20 top-ranked predictions from this work.

| LncRNA* | Species | Protein* | Species | Verified? |
|---------|---------|----------|---------|-----------|
| XIST | Homo sapiens | DGCR8 | Homo sapiens | RAID04292086 |
| XIST | Homo sapiens | EIF4A3 | Homo sapiens | RAID05222952 |
| BCYRN1 | Homo sapiens | DGCR8 | Homo sapiens | ncRI-40427659 |
| XIST | Homo sapiens | LIN28A | Homo sapiens | RAID04539801 |
| XIST | Homo sapiens | FUS | Homo sapiens | RAID04292231 |
| XIST | Homo sapiens | DGCR8 | Homo sapiens | RAID04639969 |
| MALAT1 | Homo sapiens | DGCR8 | Homo sapiens | RAID05228988 |
| MCM3AP-AS1 | Homo sapiens | FUS | Homo sapiens | RAID04862787 |
| MCM3AP-AS1 | Homo sapiens | DGCR8 | Homo sapiens | RAID04862597 |
| MCM3AP-AS1 | Homo sapiens | LIN28A | Homo sapiens | None |
| MCM3AP-AS1 | Homo sapiens | LIN28B | Homo sapiens | ncRI-40454080 |
| OIP5-AS1 | Homo sapiens | DGCR8 | Homo sapiens | RAID05191621 |
| XIST | Homo sapiens | LIN28B | Homo sapiens | RAID05100914 |
| CTBP1-AS2 | Homo sapiens | FUS | Homo sapiens | RAID04330329 |
| MCM3AP-AS1 | Homo sapiens | EIF4A3 | Homo sapiens | RAID04868241 |
| XIST | Homo sapiens | RMR1 | Homo sapiens | RAID04486531 |
| MALAT1 | Homo sapiens | EIF4A3 | Homo sapiens | RAID05074375 |
| CRHR1-IT1 | Homo sapiens | DGCR8 | Homo sapiens | RAID05171544 |
| IGF2-AS | Homo sapiens | DGCR8 | Homo sapiens | None |
| CTBP1-AS2 | Homo sapiens | LIN28A | Homo sapiens | None |

*Species: The species names in the Gene or the Ensemble database.

**TABLE 2** | 20 top-ranked predictions from this work.

| LncRNA* | Species | Protein* | Species | Verified? |
|---------|---------|----------|---------|-----------|
| XIST | Homo sapiens | DGCR8 | Homo sapiens | RAID04292086 |
| XIST | Homo sapiens | EIF4A3 | Homo sapiens | RAID05222952 |
| BCYRN1 | Homo sapiens | DGCR8 | Homo sapiens | ncRI-40427659 |
| XIST | Homo sapiens | LIN28A | Homo sapiens | RAID04539801 |
| XIST | Homo sapiens | FUS | Homo sapiens | RAID04292231 |
| XIST | Homo sapiens | DGCR8 | Homo sapiens | RAID04639969 |
| MALAT1 | Homo sapiens | DGCR8 | Homo sapiens | RAID05228988 |
| MCM3AP-AS1 | Homo sapiens | FUS | Homo sapiens | RAID04862787 |
| MCM3AP-AS1 | Homo sapiens | DGCR8 | Homo sapiens | RAID04862597 |
| MCM3AP-AS1 | Homo sapiens | LIN28A | Homo sapiens | None |
| MCM3AP-AS1 | Homo sapiens | LIN28B | Homo sapiens | ncRI-40454080 |
| OIP5-AS1 | Homo sapiens | DGCR8 | Homo sapiens | RAID05191621 |
| XIST | Homo sapiens | LIN28B | Homo sapiens | RAID05100914 |
| CTBP1-AS2 | Homo sapiens | FUS | Homo sapiens | RAID04330329 |
| MCM3AP-AS1 | Homo sapiens | EIF4A3 | Homo sapiens | RAID04868241 |
| XIST | Homo sapiens | RMR1 | Homo sapiens | RAID04486531 |
| MALAT1 | Homo sapiens | EIF4A3 | Homo sapiens | RAID05074375 |
| CRHR1-IT1 | Homo sapiens | DGCR8 | Homo sapiens | RAID05171544 |
| IGF2-AS | Homo sapiens | DGCR8 | Homo sapiens | None |
| CTBP1-AS2 | Homo sapiens | LIN28A | Homo sapiens | None |

*Species: The species names in the Gene or the Ensemble database.

| Species | Protein* | Species | Verified? |
|---------|----------|---------|-----------|

*Verified: If the direct interaction had been verified by experiment in RAID or NPInter database, this column contains the RAID and NPInter interaction identifier; otherwise "None."}

![ROC curve](image-url)

**FIGURE 8** | The ROC curve including interactions without experimental evidences in the RAID v2.0 database. In other words, interactions of the whole RAID database were utilized to train our model. The network similarity of both lncRNAs and proteins were selected to generate our final scoring matrix, which performed the best in former experiment.
proteins, we developed a simple, yet effective, method to predict the lncRNA–protein interactions. The best similarity measure in our method is the network similarity, which does not rely on sequence information. This gives our method a unique capability to predict lncRNA–protein interaction without comprehensive sequence information of both interactors. By comparing our predictions to the known lncRNA–protein interactions, we can conclude that our method has, at least, a comparable prediction performance to the state-of-the-art methods. Since our method does not rely on prior knowledge of IncRNA–protein interactions, it is a helpful supplement to existing methods.

DATA AVAILABILITY STATEMENT
All datasets generated for this study are included in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS
Y-KZ curated the dataset, designed the algorithm, implemented the algorithm, and calibrated the parameters. Z-AS and HY performed the experiments and collected the results. TL, YG, and P-FD investigated the question, designed the whole study, conceptualized the algorithm, analyzed the results, and wrote the manuscript.

REFERENCES
Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermlaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 43, D36–D42. doi: 10.1093/nar/gku1055
Chen, W., Feng, P.-M., Lin, H., and Chou, K.-C. (2013). iSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68. doi: 10.1093/nar/gku1450
Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., and Chou, K.-C. (2014). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 456, 53–60. doi: 10.1016/j.ab.2014.04.001
Chen, W., Lin, H., and Chou, K.-C. (2015a). Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol. Biosyst. 11, 2620–2634. doi: 10.1039/C5MB00155B
Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2015b). PseKNC-Genera: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31, 119–120. doi: 10.1093/bioinformatics/btu602
Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255. doi: 10.1002/prot.1035
Chou, K.-C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024
Chou, K.-C. (2015). Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218–234. doi: 10.2174/1573406411666141229162834
Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding RNA–protein interactions. Genomics Proteomics Bioinformatics 14, 62–71. doi: 10.1002/gpbi.2016.01.004
Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA-disease association prediction. BMC Bioinf. 20, 468. doi: 10.1186/s12859-019-3063-3
Hafner, M., Landthaler, M., Burger, L., Khosrhadi, M., Hauser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141, 129–141. doi: 10.1016/j.cell.2010.03.009
Henras, A. K., Dez, C., and Henry, Y. (2004). RNA structure and function in C/D and H/ACA snoRNAs. Curr. Opin. Struct. Biol. 14, 335–343. doi: 10.1016/j.sbi.2004.05.006
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/c7mb0290d
Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). HLPI-Ensemble: Prediction of human IncRNA–protein interactions based on ensemble strategy. RNA Biol. 15, 797–806. doi: 10.1080/15476286.2018.1457935
Hunt, S. E., McLaren, W., Gil, L., Thomann, A., Schuhlenburg, H., Sheppard, D., et al. (2018). Ensembl variation resources. Database (Oxf.) 2018. doi: 10.1093/database/bay119
Johnson, P., Ackley, A., Vidarsdottir, L., Lui, W.-O., Corcoran, M., Grandé, D., et al. (2013). A pseudogene long noncoding RNA network regulates P TEN transcription and translation in human cells. Nat. Struct. Mol. Biol. 20, 440–446. doi: 10.1038/nsmb.2516
Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal novel RNA classes and a possible function for pervasive transcription. Science 316, 1484–1488. doi: 10.1126/science.1138341
Kawahisma, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanefusa, M. (2008). AAnindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36, D202–D205. doi: 10.1093/nar/gkm998
Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat. Protoc. 1, 302–307. doi: 10.1038/nprot.2006.47
Kishore, S., Suter, B., and Zavolan, M. (2010). Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief Funct. Genomics 9, 391–404. doi: 10.1093/bfgp/elq028
Kung, J. T. Y., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present and future. Genetics 193, 651–669. doi: 10.1534/genetics.112.146704
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
Licatalosi, D. D., and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. 11, 72–87. doi: 10.1038/nrg2673
Predicting IncRNA-Protein Interactions

Van Nostrand, E. L., Mele, A., Fak, J. J., Ule, J., Kaykci, M., Chi, S. W., et al. (2008). HTS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456, 464–469. doi: 10.1038/nature07488

Luo, R., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 43, W65–W71. doi: 10.1093/nar/gkv458

Louro, R., Smirnova, A. S., and Verjovski-Almeida, S. (2009). Long intronic noncoding RNA transcription: expression noise or expression choice? Genomics 93, 291–298. doi: 10.1016/j.ygeno.2008.11.009

Lu, T. X., and Rothenberg, M. E. (2018). MicroRNA. J. Allergy Clin. Immunol. 141, 1202–1207. doi: 10.1016/j.jaci.2017.08.034

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

Lukong, K. E., Chang, K., Khandjian, E. W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet. 24, 416–425. doi: 10.1016/j.tig.2008.05.004

Mering, C., von Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of functional associations between proteins. Nucleic Acids Res. 31, 258–261. doi: 10.1093/nar/gkg034

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

Okamura, K., and Lai, E. C. (2008). Endogenous small interfering RNAs in animals. Nat. Rev. Mol. Cell Biol. 9, 673–678. doi: 10.1038/nrm2479

Peculis, B. A. (2000). RNA-binding proteins: it looks like a sn(o)RNA. Curr. Biol. 10, R916–R918. doi: 10.1016/s0960-9822(00)00851-4

Polisen, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfo, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038. doi: 10.1038/nature09144

Qiu, W.-R., Jiang, S.-Y., Xu, Z.-C., Xiao, X., and Chou, K.-C. (2017). iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Pseudogene PTENP1 functions as a competing endogenous RNA to suppress clear-cell renal cell carcinoma progression. Mol. Cancer Ther. 13, 3086–3097. doi: 10.1158/1535-7163.MCT-14-0245

Zhang, J., Wang, J., Zhao, F., Liu, Q., Jiang, K., and Yang, G. (2010). MicroRNA-21 (miR-21) represses tumor suppressor PTEN and promotes growth and invasion in non-small cell lung cancer (NSCLC). Clin. Chim. Acta 411, 846–852. doi: 10.1016/j.cca.2010.02.074

Zhang, W., Qu, Z., Zhang, Y., and Wang, W. (2018a). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. Neurocomputing 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018b). SEPEL-LPI: sequence-based feature projection ensemble learning for predicting lncRNA-protein interactions. PloS Comput. Biol. 14, e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019a). SFLIN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. Inf. Sci. 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. IEEE/ACM Trans. Comput. Biol. Bioinform. 2931546. doi: 10.1109/TCBB.2019.2931546

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2019c). KATZLGO: large-scale prediction of lncRNA functions by using the KATZ measure based on multiple networks. IEEE/ACM Trans. Comput. Biol. Bioinform. 16, 407–416. doi: 10.1109/TCBB.2017.2704587

Zhao, W., Wang, L., Zhang, T.-X., Zhao, Z.-N., and Du, P.-F. (2018). A brief review on software tools in generating chou’s pseudo-factor representations for all types of biological sequences. Protein Pept. Lett. 25, 822–829. doi: 10.2174/092986652566618090311124

Zheng, X., Wang, Y., Tian, K., Zhou, J., Guan, J., Luo, L., et al. (2017). Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions. BMC Bioinf. 18, 420. doi: 10.1186/s12859-017-1819-1

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.

Copyright © 2020 Zhou, Shen, Yu, Luo, Gao and Du. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.