A Novel Method for LncRNA-Disease Association Prediction Based on an lncRNA-Disease Association Network

Pengyao Ping, Lei Wang, Linai Kuang, Songtao Ye, Muhammad Faisal Buland Iqbal, and Tingrui Pei

Abstract—An increasing number of studies have indicated that long-non-coding RNAs (lncRNAs) play critical roles in many important biological processes. Predicting potential lncRNA-disease associations can improve our understanding of the molecular mechanisms of human diseases and aid in finding biomarkers for disease diagnosis, treatment, and prevention. In this paper, we constructed a bipartite network based on known lncRNA-disease associations; based on this work, we proposed a novel model for inferring potential lncRNA-disease associations. Specifically, we analyzed the properties of the bipartite network and found that it closely followed a power-law distribution. Moreover, to evaluate the performance of our model, a leave-one-out cross-validation (LOOCV) framework was implemented, and the simulation results showed that our computational model significantly outperformed previous state-of-the-art models, with AUCs of 0.8825, 0.9004, and 0.9292 for known lncRNA-disease associations obtained from the LncRNADisease database, Lnc2Cancer database, and MNDR database, respectively. Thus, our approach may be an excellent addition to the biomedical research field in the future.

Index Terms—LncRNA-disease associations, bipartite network, computational model

1 Introduction

Large numbers of studies have shown that protein-coding genes account for only a small fraction of the human genome (~2 percent), and the remaining ~98 percent of the human genome does not encode protein sequences [1], [2], [3], [4], [5]. These non-coding genes were long regarded as transcriptional noise. However, in recent years, more evidence has shown that non-coding RNAs (ncRNAs), especially long non-coding RNAs (lncRNAs), ncRNAs with lengths >200 nucleotides, play significant roles in various biological processes, such as transcription, translation, epigenetic regulation, splicing, differentiation, immune responses, and cell cycle control [6], [7], [8], [9]. Mutations in and dysregulation of lncRNAs have been proven to be correlated with a broad range of human diseases. For example, the lncRNA HOTAIR is considered a potential biomarker of hepatocellular cancer recurrence for patients after liver transplantation [10] and the lncRNA UCA1 is regarded as a potential biomarker for bladder cancer diagnosis [11]. The lncRNA PCA3 is a well-known example of a potential cancer diagnostic biomarker because its increased expression level is greatly increased (approximately 60-fold) in prostate tumors compared with normal tissues [12], [13]. Therefore, it is necessary to discover additional potential lncRNA-disease associations to help understand the molecular mechanisms of human diseases at the lncRNA level and facilitate the identification of biomarkers for disease diagnosis, treatment, and prevention.

In the past few years, many computational models have been proposed to predict lncRNA-disease associations for further experimental validation. Such studies are becoming increasingly important because they can decrease the time and cost of biological experiments. For example, Chen et al. proposed a Laplacian regularized least squares method to predict novel human lncRNA-disease associations based on lncRNA expression profiles and the assumption that similar diseases will tend to be associated with functionally similar lncRNAs [14]. Based on the above work and assumption, two years later, Chen et al developed two novel lncRNA functional similarity calculation models and evaluated these new models by introducing similarity scores into the previous lncRNA-disease association prediction model [15]. Chen et al. also developed several other models to predict potential lncRNA-disease relationships. For instance, they developed a hypergeometric distribution model for lncRNA-disease relationship inference based on available miRNA-disease association and miRNA-lncRNA association information without any positive lncRNA-disease interactions [16]. They integrated disease similarity and lncRNA similarity through disease semantic similarity, lncRNA expression and function similarities, and diseases and lncRNAs Gaussian interaction profile kernel similarity; they also took the integrated similarity into account through the Katz measure to forecast probable interactions between diseases and lncRNAs [17]. The group also developed a fuzzy measure-based lncRNA functional similarity calculation model by combing information from known lncRNA-disease associations and diseases directed acyclic graphs (DAGs) with their previously proposed Laplacian regularized least squares model for predicting lncRNA-disease associations [18]; they later developed an improved lncRNA functional similarity calculation model that was combined with the previously proposed Laplacian regularized least squares model to further predict lncRNA-disease associations [19]. They also presented a model called Improved Random Walk with Restart for LncRNA-Disease Association prediction (IRWRLDA) to predict novel lncRNA-disease associations by integrating known lncRNA-disease associations, disease semantic similarity, and various lncRNA similarity measures [20]. In addition to the above works, they also summarized several computational models for identifying disease-related lncRNAs on a large scale and selecting promising disease-related lncRNAs for experimental validation [21]. Aside from the methods proposed by Chen et al., Yang et al. constructed a coding-non-coding gene-disease bipartite network based on known associations between diseases and disease-causing genes and then applied an algorithm to uncover possible lncRNA-disease interactions in that network [22]. Liu et al. built a protein-coding gene (PCG)-lncRNA bipartite network based on lncRNAs and PCG expression profiles and protein interaction datasets to predict cancer-related lncRNAs using a random walk method [23]. Zhou et al. proposed a rank-based model (RWRHLD) to identify potential lncRNA-disease associations by combing the miRNA-associated lncRNA-lncRNA crosstalk network, disease-disease similarity network and known lncRNA-disease association network into a heterogeneous network and applying a random walk with restart to the heterogeneous network [24].

As many computational models have been proposed over the years, several databases of experimentally verified lncRNA-disease interactions have been constructed and provided for free on the internet. For example, Chen et al. built the LncRNA and Disease Database (LncRNADisease), which integrates nearly 3,000 lncRNA-disease entries, including 914 lncRNA entries and 329 disease entries, from ~2000 publications. LncRNADisease also provided

- *P. Ping, S. Ye, M.F.B. Iqbal, and T. Pei are with the College of Information Engineering, Xiangtan University, Xiangtan 411105, P.R.China, and the Key Laboratory of Intelligent Computing & Information Processing, Xiangtan University, Xiangtan 411105, P.R.China. E-mail: {956272448, 985326696, 1146105816, 308407367}@qq.com.*
- *L. Wang and L. Kuang are with the College of Computer Engineering Applied Mathematics, Changsha University, Changsha 410001, P.R.China, the College of Information Engineering, Xiangtan University, Xiangtan 411105, P.R.China, and the Key Laboratory of Intelligent Computing & Information Processing, Xiangtan University, Xiangtan 411105, P.R.China. E-mail: wanglei@xtu.edu.cn, zhc680518@163.com.*

Manuscript received 20 Dec. 2016; revised 29 Mar. 2018; accepted 11 Apr. 2018. Date of publication 16 Apr. 2018; date of current version 29 Mar. 2019.
(Corresponding author: Lei Wang.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2827373

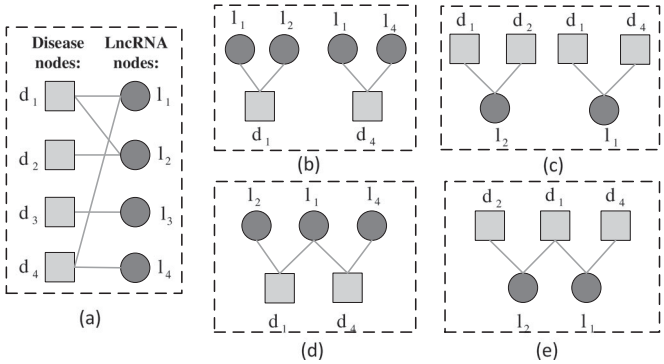

Fig. 1. (a) Original bipartite network. (b) l_1 and l_2, l_1 and l_4 are similar nodes since l_1 and l_2 have a common neighboring node d_1, l_1, and l_4 have a common neighboring node d_4. (c) d_1 and d_2, d_1 and d_4 are similar nodes since d_1 and d_2 have a common neighboring node l_2, d_1 and d_4 have a common neighboring node l_1. (d) l_2 and l_4 are similar nodes since their neighboring nodes d_1 and d_4 are similar nodes. (e) d_2 and d_4 are similar nodes since their neighboring nodes l_1 and l_2 are similar nodes.

1,564 predicted human disease-related lncRNAs [25]. Ning et al. constructed the Lnc2Cancer database, a manually curated database including 1,488 experimentally supported associations between 666 human lncRNAs and 97 human cancers collected from more than 2,000 published papers [26]. Wang et al. developed a comprehensive mammalian ncRNA-disease database (MNDR) that provided over 1,100 relationships between diseases and a variety of ncRNAs such as long non-coding (lncRNAs), microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs) and small nucleolar RNAs (snoRNAs), derived from a review of more than 370 published papers [27].

As described above, all of these existing computational models for identifying novel associations between lncRNAs and diseases were designed by integrating lncRNA similarity and disease similarity information based on different lncRNA-related data and disease-related resources, such as lncRNA expression profiles, gene-disease interactions and lncRNA-miRNA interactions. In contrast, in this article, we constructed a bipartite network to predict potential lncRNA-disease interactions based on known lncRNA-disease associations only. In addition, we considered the following assumption in our method: two nodes are similar if they have common neighbors or are connected to similar nodes. To illustrate the above assumption more intuitively, we provide an example in Fig. 1. Our newly proposed model relies only on topological information from known lncRNA-disease association networks for identifying potential disease-related lncRNAs. The flow chart of our method for predicting lncRNA-disease associations is shown in Fig. 2, where the blocks and circles represent diseases and lncRNAs, respectively. To evaluate the performance of our method, the Leave-one-out cross-validation (LOOCV) framework was implemented, and a series of experiments were performed based on the experimentally verified lncRNA-disease associations downloaded from the LncRNADisease database, Lnc2Cancer database and MNDR database. The simulation results demonstrated that our approach can achieve much better predictive performance than other state-of-the-art models. Moreover, our model can feasibly and efficiently predict lncRNA-disease associations on a large scale because it must consider only the topology information from known lncRNA-disease interaction networks.

2 Materials

To evaluate the performance of our newly proposed method, we collected three datasets from the LncRNADisease database, Lnc2Cancer database and MNDR database.

The first dataset is the set of known lncRNA-disease associations downloaded from the LncRNADisease database in June 2015. After eliminating duplicate samples that describe the same lncRNA-disease relationships based on evidence from different

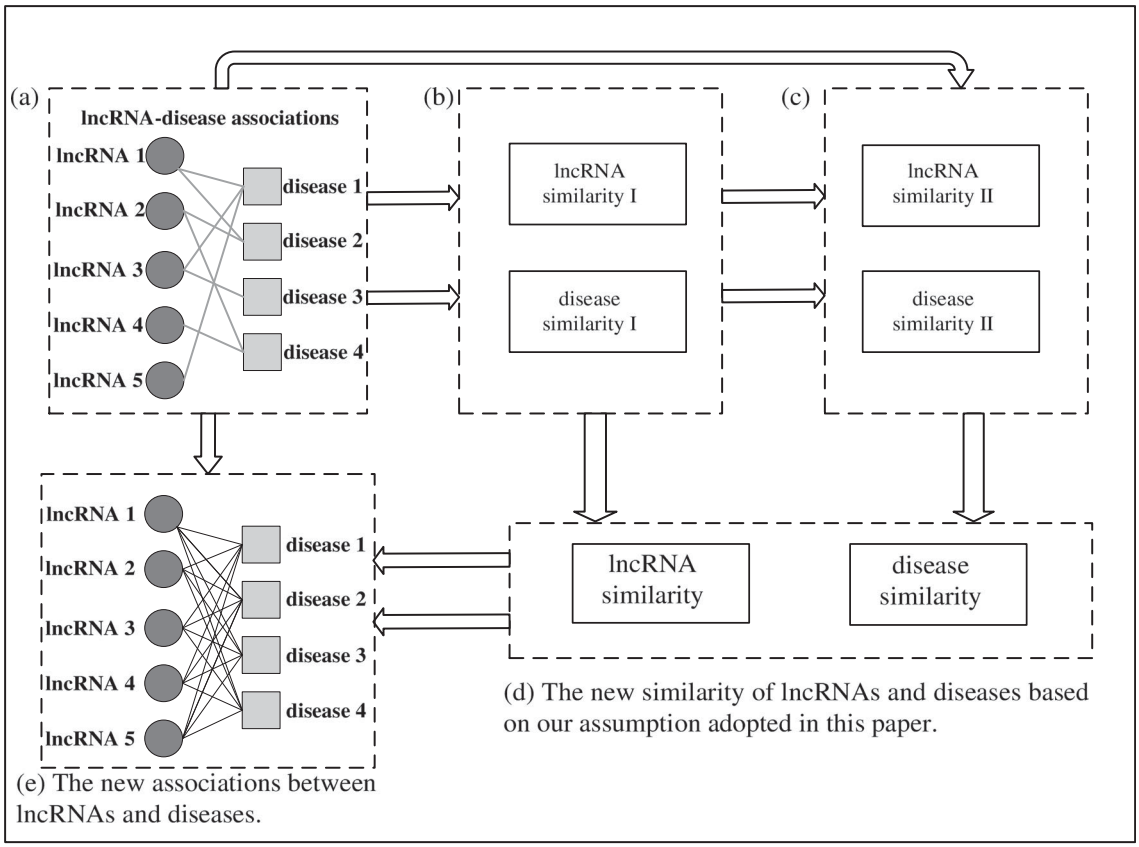

Fig. 2. The flowchart of our method.

TABLE 1
The Comparison Results of AUC Achieved by Our Model Based on Three Different Datasets While the Parameter α was Set to Different Values

lncRNADisease		Lnc2Cancer		MNDR	
α	AUC	α	AUC	α	AUC
0.2	0.8364	0.2	0.9004	0.2	0.9035
0.4	0.8720	0.4	0.8994	0.4	0.9248
0.5	0.8793	0.5	0.8983	0.5	0.9284
0.6	0.8825	0.6	0.8965	0.6	0.9292
0.8	0.8784	0.8	0.8933	0.8	0.9276

experiments, we obtained 554 human lncRNA-disease interactions involving 267 lncRNAs and 208 diseases (see Supplementary Table 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2018.2827373).

The second dataset is the set of experimentally supported lncRNA-cancer associations downloaded from the Lnc2Cancer database in July 2016. After removing the duplicate lncRNA-cancer associations based on different pieces of evidence, we obtained 1,103 distinct interactions involving 98 human cancers and 668 lncRNAs (see Supplementary Table 2, available online).

The third datastet is the assemblage of lncRNA-disease relationships obtained from the MNDR database in March 2015. In accordance with the data processing practices described above, we obtained 471 relationships involving 241 lncRNAs and 127 diseases (see Supplementary Table 3, available online).

3 Methods

Inspired by the concepts of complex networks [28], [29], [30] and bipartite networks [31], [32], we proposed a novel computational model for calculating the functional similarity of lncRNAs and diseases by using only the information from known lncRNA-disease associations. Barabasi et al. demonstrated that large networks are governed by robust self-organizing phenomena that surpass the particular details of individual systems [33]. Obviously, the similarity measure is a useful tool for determing the degree of similarity between objects and can be utilized in various fields. For example, Newman measured the probability of collaboration between scientists in two collaboration networks as a function of their number of mutual acquaintances in the network, their number of previous collaborations, and their number of previous collaborators [28]. Alaimo et al. proposed a technique for the prediction of new drug-target interaction via the similarity measure and resource transfer [32]. Distinct from these methods, our newly constructed bipartite network contains two different types of nodes, namely, lncRNAs and diseases, and similar nodes are assumed to exist only among nodes of the same type, whereas in these previous methods, all nodes were assumed to be the same. In addition, while computing the similarity measure between a pair of nodes in the newly constructed bipartite network, according to the assumption described above, we considered only those node pairs that have at least one path with length no larger than 4 hops because we can easily assume that the similarity of two nodes is inversely proportional to the length of path between them. For instance, in a social network, the degree of familiarity between two people is inversely proportional to the number of intermediaries that they need to establish a connection. Therefore, for the sake of time and resources, those node pairs with path lengths greater than 4 hops were not considered in our model. Moreover, it is easily assumed that two nodes will be more similar if they have more common neighboring nodes or similar nodes, just as in social networks, two people will be more familiar if they have more common friends or social circles. Therefore, in addition to the length of the path, we also considered only those paths with lengths no larger than 4 hops when computing the similarity measure between a pair of nodes.

TABLE 2
Performance Comparisons between Our Model and Seven State-of-the-Art of Models in Terms of AUC Based on Global LOOCV

Methods	AUC	References
Our model	0.8535	
LRLSLDA	0.7760	[14]
LNLSIM1-LRLSLDA	0.8130	[15]
LNLSIM2-LRLSLDA	0.8198	[15]
HGLDA	0.7621	[16]
KATZLDA	0.7886	[17]
LRLSLDA-FMLNCSIM	0.8266	[18]
IRWRLDA	0.7242	[20]

TABLE 3
Performance Comparisons between Our Model and Four State-of-the-Art of Models in Terms of AUC Based on Global LOOCV

Methods	AUC	References
Our Method	0.9292	
LRLSLDA	0.8850	[14]
LRLSLDA-ILNCSIM	0.9316	[19]
LRLSLDA-LNCSIM1	0.9135	[15], [19]
LRLSLDA-LNCSIM2	0.9169	[15], [19]

3.1 Bipartite Network

Let $L = \{l_1, l_2, \ldots, l_n\}$ be a set of lncRNAs and $D = \{d_1, d_2, \ldots, d_m\}$ be a set of diseases; then, the L-D network can be described as a bipartite graph $G(L, D, E)$, where $E = \{e_{ij}, l_i \in L, d_j \in D\}$. In addition, in $G(L, D, E)$, $\forall\, l_i \in L$, $d_j \in D$, the edge $e_{ij} \in E \Longleftrightarrow l_i$ is associated with d_j. Additionally, according to $G(L, D, E)$, an adjacency matrix $A = \{a_{ij}\}_{n\times m}$ can be constructed, where $a_{ij} = 1$ if l_i is connected to d_j; otherwise, $a_{ij} = 0$.

3.2 Similarities of lncRNAs and Diseases Based on Common Neighbors

In $G(L, D, E)$, $\forall\, l \in L$, let $\theta(l)$ denote the set of neighboring nodes of l, and $d(l)$ denote the degree of l (i.e., the number of neighboring nodes of l). For any two nodes l_i and l_j in L, we reasonably consider there to be some degree of similarity between them if they have common neighboring nodes, and their similarity score between them can be defined as follows:

$$SL1_{ij} = SL1(l_i, l_j) = exp(-S1_{ij}^{L}) \tag{1}$$

$$\begin{aligned} S1_{ij}^{L} &= \frac{1}{d(l_i)\times d(l_j)} \sum_{d_z \in (\theta(l_i)\cap\theta(l_j))} \frac{1}{d(d_z)} \\ &= \frac{1}{d(l_i)\times d(l_j)} \sum_{r=1}^{m} \frac{a_{ir}\times a_{jr}}{d(d_r)}. \end{aligned} \tag{2}$$

Specifically, if $i = j$, we set $SL1_{ij} = 1$. In other words, if l_i and l_j are the same node, then their similarity value is 1.

Similar to the above method for computing the similarity measure of lncRNA nodes, for any two nodes d_i and d_j in D, their similarity score can also be defined as follows:

$$SD1_{ij} = SD1(d_i, d_j) = exp(-S1_{ij}^{D}) \tag{3}$$

$$\begin{aligned} S1_{ij}^{D} &= \frac{1}{d(d_i)\times d(d_j)} \sum_{l_t \in (\theta(d_i)\cap\theta(d_j))} \frac{1}{d(l_t)} \\ &= \frac{1}{d(d_i)\times d(d_j)} \sum_{t=1}^{n} \frac{a_{ti}\times a_{tj}}{d(l_t)}. \end{aligned} \tag{4}$$

In particular, if $i = j$, we set $SD1_{ij} = 1$. In other words, if d_i and d_j are the same node, their similarity value is 1.

3.3 Similarities of lncRNAs and Diseases Based on SimRank Measure

According to the assumption that two nodes are similar if they are connected to similar nodes, in $G(L, D, E)$, if the two nodes l_i and l_j in L do not have common neighboring nodes but are connected to similar nodes, then we also consider that there is a degree of similarity between them. The similarity score between the two nodes l_i and l_j can be defined as follows, based on SimRank method [29], [30], [34].

$$SL2_{ij} = SL2(l_i, l_j) = exp(-S2^L_{ij}) \tag{5}$$

$$S2^L_{ij} = \frac{1}{d(l_i) \times d(l_j)} \sum_{p=1}^{n} \sum_{q=1}^{n} \frac{a_{pi} \times a_{qj} \times SD1_{pq}}{d(d_p) \times d(d_q)}. \tag{6}$$

In particular, if $i = j$, we set $SL2_{ij} = 1$. In other words, if l_i and l_j are the same node, their similarity value is 1.

As in the above methods, if the two nodes d_i and d_j in D, do not have common neighboring nodes but are connected to similar nodes, we consider there to be a degree of similarity between them, and the similarity score between these two nodes d_i and d_j can be defined as follows:

$$SD2_{ij} = SD2(d_i, d_j) = exp(-S2^D_{ij}) \tag{7}$$

$$S2^D_{ij} = \frac{1}{d(d_i) \times d(d_j)} \sum_{p=1}^{n} \sum_{q=1}^{n} \frac{a_{pi} \times a_{qj} \times SL1_{pq}}{d(l_p) \times d(l_q)}. \tag{8}$$

In particular, if $i = j$, we set $SD2_{ij} = 1$. In other words, if d_i and d_j are the same node, their similarity value is 1.

Furthermore, to integrate the similarity scores computed above, we defined a new similarity measurement between two lncRNA nodes, l_i and l_j as follows,

$$SL_{ij} = SL(l_i, l_j) = SL1_{ij} \times SL2_{ij}. \tag{9}$$

Correspindingly, a new similarity measurement between two disease nodes, d_i and d_j was defined as follows:

$$SD_{ij} = SD(d_i, d_j) = SD1_{ij} \times SD2_{ij}. \tag{10}$$

Therefore, based on matrices SL and A proposed above, we can construct a recommendation matrix $R1 = \{r1_{ij}\}_{n \times m}$ as follows:

$$R1 = SL \times A. \tag{11}$$

Here, $SL = \{SL_{ij}\}_{n \times n}$ is the similarity of two lncRNAs.

Likewise, we can also construct a recommendation matrix $R2 = \{r2_{ij}\}_{n \times m}$ as follows:

$$R2 = A \times SD. \tag{12}$$

Here, $SD = \{SD_{ij}\}_{m \times m}$ is the similarity of two diseases.

Thus, by integrating the above two recommendation matrices, we can construct a new similarity measurement between lncRNAs and diseases as follows:

$$R = \alpha \times R1 + (1 - \alpha) \times R2. \tag{13}$$

Here, $\alpha \in (0, 1)$ is a parameter utilized to tune the relative importance between the similarity of lncRNAs and the similarity of diseases [35].

4 Experimental Results

4.1 Analysis of the Bipartite Network Based on lncRNA-Disease Associations

In this work, the available relationships between lncRNAs and diseases were utilized to construct a bipartite network. A sub-network of this bipartite network is shown in Fig. 2a. There were two types of nodes in the constructed bipartite network: one class node corresponded to lncRNAs, the other class node corresponded to diseases. A link was constructed between one lncRNA and one disease if the lncRNA was associated with the disease. We analyzed the bipartite network constructed based on 554 human lncRNA-disease interactions, which included 267 lncRNAs and 208 diseases. We found that the degree distribution of the bipartite network closely followed a power-law distribution (see Supplementary Figure S1(a), available online). The degree value of each lncRNA node, that is, the number of diseases associated with the lncRNA, had a broad distribution (see Supplementary Figure S1 (b), available online). Many lncRNAs were connected to only a few diseases, whereas a small number of lncRNAs were connected to many diseases. For example, H19 was related to 47 diseases, including esophageal squamous cell cancer, bladder cancer and cervical cancer; HOTAIR was associated with 24 diseases, including lung cancer, gastric cancer and colorectal cancer. Similarly, the degree value of each disease node, that is, the number of lncRNAs associated with the disease, also had a broad distribution (see Supplementary Figure S1(c), available online). This indicated that many diseases were connected to a few lncRNAs, whereas a small number of diseases were related to many lncRNAs. For example, breast cancer was associated with 16 lncRNAs including LSINCT5, MALAT1 and MIR31HG; prostate cancer was related to 22 lncRNAs including ANRIL, GAS5 and MALAT1.

4.2 Leave-One-Out Cross-Validation Tests

To evaluate the performance of our newly proposed method for predicting the similarity between lncRNAs and diseases, a LOOCV framework was adopted based on the experimentally verified lncRNA-disease associations downloaded from the LncRNADisease database, Lnc2Cancer database and MNDR database. For the LOOCV, in each round, a known lncRNA-disease association is left out as a test sample and the remaining known lncRNA-disease associations are used as the training samples for model learning. All lncRNA-disease association pairs not supported by relevant evidence were also tested as candidate samples. Then, the rank of each candidate sample is obtained. The tested samples with a predicted rank higher than the given thresholds were considered true positives prediction. By setting different thresholds, we were able to obtain the corresponding true positive rates (TPR, sensitivity, recall) and false positive rates (FPR, 1-specificity). Sensitivity measures the proportion of positives that are correctly identified, while 1-specificity is the percentage of negative samples correctly identified to rank lower than the threshold. By plotting TPR versus FPR at different thresholds, receiver operating characteristics (ROC) curves can be obtained, and the areas under ROC curve (AUCs) are calculated. An AUC value of 1 indicates a perfect prediction, while an AUC value of 0.5 demonstrates a random performance. If an AUC value is much closer to 1 than to 0.5, then we can say that the prediction performance is much better than random.

As described in Section 3, the value of the parameter α may influence the prediction performance of our model, therefore, we implemented a series of experiments to evaluate the impact of α. By setting different values to α, different values of AUC are obtained in the framework of LOOCV, and simulation results were shown in Table 1 and Figure S2 (see Supplementary Figure, available online).

To further assess our method, we also compared the performance of our model to that of several other state-of-the-art models [14], [15], [16], [17], [18], [20] using a dataset of 293 known lncRNA-

TABLE 4
Prediction and Evaluation Results of lncRNA Associated with Colon Cancer, Osteosarcoma, and Cervical Cancer in top 20 Ranking Lists

Disease	lncRNA	Evidence(Database)	Rank
Colon cancer	GAS5	Lnc2cancer	4
Colon cancer	UCA1	MNDR	6
Colon cancer	TUG1	Lnc2cancer	12
Colon cancer	ANRIL	Lnc2cancer	13
Osteosarcoma	H19	MNDR/Lnc2cancer	1
Osteosarcoma	CDKN2B-AS1	MNDR/Lnc2cancer	2
Osteosarcoma	HOTAIR	Lnc2cancer	3
Osteosarcoma	MEG3	MNDR/Lnc2cancer	4
Osteosarcoma	UCA1	MNDR/Lnc2cancer	8
Osteosarcoma	ANRIL	MNDR/Lnc2cancer	13
Osteosarcoma	TUG1	MNDR/Lnc2cancer	14
Cervical cancer	CDKN2B-AS1	Lnc2cancer	1
Cervical cancer	MEG3	Lnc2cancer	2
Cervical cancer	PVT1	Lnc2cancer	3
Cervical cancer	UCA1	MNDR	5
Cervical cancer	GAS5	Lnc2cancer	9

disease associations collected from the LncRNADisease database (see Supplementary Table 4, available online) that has been used as a gold standard dataset in the evaluation of several other models. The simulation results, shown in Figure S3 (see Supplementary Figure, available online) and Table 2 clearly indicate that our model can achieve better performance in LOOCV than other state-of-the-art models, with AUC of 0.8535 ($\alpha = 0.8$). Furthermore, we compared the performance of our model to that of other methods proposed in [14], [15], [19] using a dataset of 471 lncRNA-disease interactions (see Supplementary Table 3, available online) downloaded from the MNDR database in, March 2015. The simulation results are illustrated in Table 3 and Figure S2(c) (see Supplementary Figure, available online) and show that our method achieve better performance in the framework of global LOOCV than previously developed models, with an AUC of 0.9292 ($\alpha = 0.6$). Finally, we compared our model with RWRHLD [24] based on a data set of 352 lncRNA-disease relationships (see Supplementary Table 5, available online) downloaded from LncRNADisease database. The simulation results are shown in Figure S4 (see Supplementary Figure, available online) and indicates that our method, with AUC of 0.8732 ($\alpha = 0.8$), achieved better performance than RWRHLD, which achieved an AUC of 0.77. We also compared our model with Yang et al.'s method [22] based on a data set of 554 lncRNA-disease relationships including 267 lncRNAs and 208 diseases (see Supplementary Table 1, available online). According to the description of Yang et al.'s method, we removed the nodes whose degree was one in LOOCV. Finally, we obtained 232 lncRNA-disease associations between 50 lncRNAs and 59 diseases that were to be utilized in LOOCV. The comparison results are shown in Figure S5 (see Supplementary Figure, available online) and indicates that our method, with AUC of 0.7678, achieved better performance than Yang et al.'s method, which achieved an AUC of 0.6773.

5 Case Studies

To further validate the effectiveness of our model, we applied it to predict three deadly types of cancer-related lncRNAs based on the dataset downloaded from the LncRNADisease databased. This approach is similar to the evaluation method adopted by most of the current prediction computational models. The 20 prediction results with the highest ranks were illustrated in Table 4 and verified based on the Lnc2Cancer database and MNDR database. Importantly, the predicted new associations presented in Table 4 do not exist in the training set.

Colon cancer is the third most commonly diagnosed cancer and the second leading cause of cancer deaths in men and women [36], [37]. With the development of cancer research, lncRNA has served as a promising target for cancer diagnosis and therapy [38], [39]. In this section, we applied our prediction method to identify potential lncRNAs related to colon cancer based on the dataset collected from the LncRNADisease database. As illustrated in Table 4, four of the top 20 predictions were proven to be related to colon cancer according to the Lnc2Cancer and MNDR databases. Furthermore, the experiments revealed that GAS5 overexpression significantly repressed cell proliferation both in vitro and in vivo, and GAS5 may therefore be a candidate prognostic biomarker in human colorectal cancer [40]. Other experiments revealed that the inhibition of TUG1 expression significantly blocked the cell migration ability of colon cancer cells, and TUG1 overexpression may contribute to enhanced cell proliferation and migration in colon cancer cells [41].

Osteosarcoma is the most prevalent primary malignant tumor in adolescents and is associated with poor prognosis and a high rate of disability in youth [42]. LncRNAs have received increasing attention due to their roles in many diseases, including osteosarcoma [43], [44]. Hence, we implemented our prediction method to identify potential osteosarcoma-related lncRNAs. As illustrated in Table 4, seven of the top 20 predictions were proven to be related to osteosarcoma according to Lnc2Cancer and MNDR databases. One study reported that H19 promoted metastasis through upregulation of ZEB1 and ZEB2 by competitively binding microRNAs in the miR-200 family, which suggests important roles for H19 in osteosarcoma metastasis and therapy [45]. The present study showed that HOTAIR silencing inhibited the growth, adhesion, migration and invasion of MG63 osteosarcoma cells, indicating that HOTAIR may serve as a potential tool for osteosarcoma therapy [46]. MEG3 has also been shown to have low expression in osteosarcoma cells, while its up-regulation of MEG3 can induce apoptosis in MG63 cells and inhibit cell proliferation, invasion and migration [47]. The study revealed that UCA1 was upregulated in osteosarcoma cells and promoted cell growth and caused cell cycle arrest through inactivation of the PTEN/AKT signaling pathway, indicating that UCA1 may be a potential prognostic marker and therapeutic target for osteosarcoma [48].

Cervical cancer contributed the second highest number of female cancer deaths, exceeded only by breast cancer. Many lncRNAs are considered pivotal regulators in various biological processes and play vital roles in the oncogenesis and progression of cervical cancer [49]. Hence, we applied our method to predict possible lncRNAs associated with cervical cancer. As illustrated in Table 4, five of the top 20 predictions were proven to be related to cervical cancer according to the Lnc2Cancer and MNDR databases. Specifically, most of these lncRNAs (CDKN2B-AS1, GAS5, MEG3 and PVT1) were recorded in the Lnc2Cancer database, and one lncRNA (UCA1) was recorded in the MNDR database. PVT1 is considered to play an oncogenic role in cervical cancer, and the overexpression of PVT1 can drive cervical carcinogenesis [50]. Moreover, MEG3 is a powerful tool for diagnosis and prognosis of patients with cervical cancer, and low expression of MEG3 is likely to be related to promoter hypermethylation in cervical cancer [51].

According to the above description, it is clear that our model can achieve reliable performance for predicting potential associations between lncRNAs and diseases. Therefore, our approach can be applied to prioritize all candidate lncRNA-disease pairs based on the lncRNA-disease associations recorded in the LncRNADisease, Lnc2Cancer and MNDR databases, and the obtained prediction results can be used for further research and experimental validation.

6 Discussion

Accumulating evidence has shown that identifying novel potential associations between lncRNAs and diseases can improve the understanding of disease pathogenesis at the lncRNA level, which is helpful for the prognosis, diagnosis, treatment and prevention of human diseases. Such studies have become greatly significant because they can decrease the time and cost of biological experiments. In this paper, we constructed a bipartite network based on known lncRNA-disease associations only. Based on assumptions about similar nodes, a novel prediction model is proposed to infer

potential lncRNA-disease associations by integrating two similarity calculation methods for lncRNAs and diseases. To evaluate the predictive performance of our method, LOOCV was implemented based on known lncRNA-disease associations collected from three databases. The validation results demonstrated the effectiveness of our model. Furthermore, we also compared the outcomes our method with those of several state-of-the-art models and found that our model was able to achieve better performance. Of course, there are still some deficiencies and limitations of our method. For example, currently, there is no effective way to choose the best value for parameter α to achieve the best predictive performance. Moreover, only known lncRNA-disease associations were considered in our method. There are many other known associations, such as miRNA-lncRNA associations and miRNA-disease associations; if we can integrate these different associations, then the predictive performance of our model may be improved significantly.

ACKNOWLEDGMENTS

The authors thank the anonymous referees for suggestions that helped improve the paper substantially. This project is partly sponsored by the National Natural Science Foundation of China (No.61640210, No.61672447).

REFERENCES

[1] P. Bertone, V. Stolc, T. E. Royce, et al., "Global identification of human transcribed sequences with genome tiling arrays," *Sci.*, vol. 306, no. 5705, pp. 2242–2246, 2004.
[2] P. Kapranov, J. Cheng, S. Dike, et al., "RNA maps reveal new RNA classes and a possible function for pervasive transcription," *Sci.*, vol. 316, no. 5830, pp. 1484–1488, 2007.
[3] J. E. Wilusz, H. Sunwoo, D. L. Spector, et al., "Long noncoding RNAs: Functional surprises from the RNA world," *Genes Develop.*, vol. 23, no. 13, pp. 1494–1504, 2009.
[4] R. J. Taft, K. C. Pang, T. R. Mercer, et al., "Noncoding RNAs: Regulators of disease," *J. Pathology*, vol. 220, no. 2, pp. 126–139, 2010.
[5] P. Carninci, A. Sandelin, B. Lenhard, et al., "Genome-wide analysis of mammalian promoter architecture and evolution," *Nature Genetics*, vol. 38, no. 6, pp. 626–635, 2006.
[6] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs," *Nature*, vol. 482, no. 7385, pp. 339–346, 2012.
[7] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding RNAs," *Mol. Cell*, vol. 43, no. 6, pp. 904–914, 2011.
[8] O. L. Wapinski and H. Y. Chang, "Long noncoding RNAs and human disease," *Trends Cell Biol.*, vol. 21, no. 6, pp. 354–361, 2011.
[9] T. Derrien, R. Johnson, G. Bussotti, et al., "The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression," *Genome Res.*, vol. 22, no. 9, pp. 1775–1789, 2012.
[10] R. A. Gupta, N. Shah, K. C. Wang, et al., "Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis," *Nature*, vol. 464, no. 7291, pp. 1071–1076, 2010.
[11] Z. Zhang, H. Hao, C. J. Zhang, et al., "Evaluation of novel gene UCA1 as a tumor biomarker for the detection of bladder cancer," *Nat. Med. J. China*, vol. 92, no. 6, pp. 384–387, 2012.
[12] R. Spizzo, M. I. Almeida, A. Colombatti, et al., "Long non-coding RNAs and cancer: A new frontier of translational research?" *Oncogene*, vol. 31, no. 43, pp. 4577–4587, 2012.
[13] H. Van Poppel , A. Haese, M. Graefen, et al., "The relationship between Prostate Cancer gene 3 (PCA3) and prostate cancer significance," *BJUI*, vol. 109, no. 3, pp. 360–366, 2012.
[14] X. Chen and G. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinf.*, vol. 29, no. 20, pp. 2617–2624, 2013.
[15] X. Chen, C. C. Yan, C. Luo, et al.,"Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Sci. Reports*, vol. 5, pp. 11338–11338, 2015.
[16] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Sci. Reports*, vol. 5, pp. 13186–13186, 2015.
[17] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Sci. Reports*, vol. 5, 2015, Art. no. 16840.
[18] X. Chen, Y. Huang, X. Wang, et al., "FMLNCSIM: Fuzzy measure-based lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 29, pp. 45948–45958, 2016.
[19] Y. Huang, X. Chen, Z. You, et al., "ILNCSIM: Improved lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, 2016.
[20] X. Chen, Z. H. You, G. Y. Yan, et al., "IRWRLDA: Improved random walk with restart for lncRNA-disease association prediction," *Oncotarget*, vol. 7, no. 36, 2016, Art. no. 57919.
[21] X. Chen, C. C. Yan, X. Zhang, et al., "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Briefings Bioinf.*, vol. 18, pp. 558–576, 2017.
[22] X. Yang, L. Gao, X. Guo, et al., "A network-based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases," *PLoS One*, vol. 9, no. 1, 2014, Art. no. e87797.
[23] Y. Liu, R. Zhang, F. Qiu, et al., "Construction of a lncRNA-PCG bipartite network and identification of cancer-related lncRNAs: A case study in prostate cancer," *Mol. BioSystems*, vol. 11, no. 2, pp. 384–393, 2015.
[24] M. Zhou, X. Wang, J. Li, et al., "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Mol. BioSystems*, vol. 11, no. 3, pp. 760–769, 2015.
[25] G. Chen, Z. Wang, D. Wang, et al., "LncRNADisease: A database for long-non-coding RNA-associated diseases," *Nucleic Acids Res.*, vol. 41, 2013, Art. no. D983-6.
[26] S. Ning, J. Zhang, P. Wang, et al., "Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers," *Nucleic Acids Res.*, vol. 44, 2016, Art. no. D980-5.
[27] Y. Wang, L. Chen, B. Chen, et al., "Mammalian ncRNA-disease repository: A global view of ncRNA-mediated disease network," *Cell Death Disease*, vol. 4, no. 8, 2013, Art. no. e765.
[28] M. E. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E*, vol. 64, no. 2, 2001, Art. no. 025102.
[29] D. Libennowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Assoc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.
[30] L. Lu and T. Zhou, "Link prediction in complex networks: A survey," *Physica A-Statistical Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2010.
[31] T. Zhou, J. Ren, M. Medo, et al., "Bipartite network projection and personal recommendation," *Phys. Rev. E*, vol. 76, no. 4, 2007, Art. no. 046115.
[32] S. Alaimo, A. Pulvirenti, R. Giugno, et al., "Drug-target interaction prediction through domain-tuned network-based inference," *Bioinf.*, vol. 29, no. 16, pp. 2004–2008, 2013.
[33] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Sci.*, vol. 286, no. 5439, pp. 509–512, 1999.
[34] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," *Knowl. Discovery Data Mining*, pp. 538–543, 2002.
[35] H. Peng, C. Lan, Y. Zheng, et al., "Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite," *BMC Bioinf.*, vol. 18, no. 1, 2017, Art. no. 193.
[36] S. Prenner, MD1 and J. Levitsky, MD, MS1, "Comprehensive review on colorectal cancer and transplant," *Amer. J. Transplantation*, vol. 17, pp. 2761–2774, 2017.
[37] J.-F. Xiang, Q.-F. Yin, T. Chen, Y. Zhang, X.-O. Zhang, Z. Wu, S. Zhang, H.-B. Wang, J. Ge, and X. Lu, "Human colorectal cancer-specifc CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus," *Cell Res.*, vol. 24, pp. 513–531, 2014.
[38] G. Yang, X. Lu, L. Yuan, et al., "LncRNA: A link between RNA and cancer," *Biochimica et Biophysica Acta*, vol. 1839, no. 11, pp. 1097–1109, 2014.
[39] B. Yue, S. Qiu, S. Zhao, et al., "LncRNA-ATB mediated E-cadherin repression promotes the progression of colon cancer and predicts poor prognosis," *J. Gastroenterology Hepatology*, vol. 31, no. 3, pp. 595–603, 2016.
[40] D. Yin, X. He, E. Zhang, et al., "Long noncoding RNA GAS5 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer," *Med. Oncology*, vol. 31, no. 11, 2014, Art. no. 253.
[41] H. Y. Zhai, M. H. Sui, X. Yu, et al., "Overexpression of long non-coding RNA TUG1 promotes colon cancer progression," *Med. Sci. Monitor Int. Med. J. Exp. Clinical Res.*, vol. 22, 2016, Art. no. 3281.
[42] W. Yong, Y. Tao, Z. Zhen, et al., "Long non-coding RNA TUG1 promotes migration and invasion by acting as a ceRNA of miR-335–5p in osteosarcoma cells," *Cancer Sci.*, vol. 108, pp. 859–867, 2017.
[43] N. Jiang, X. Wang, X. Xie, et al., "lncRNA DANCR promotes tumor progression and cancer stemness features in osteosarcoma by upregulating AXL via miR-33a-5p inhibition," *Cancer Lett.*, vol. 405, pp. 46–55, 2017.
[44] C. L. Zhang, K. P. Zhu, X. L. Ma, "Antisense lncRNA FOXC2-AS1 promotes doxorubicin resistance in osteosarcoma by increasing the expression of FOXC2," *Cancer Lett.*, vol. 396, pp. 66–75, 2017.
[45] M. Li, H. Chen, Y. Zhao, et al., "H19 functions as a ceRNA in promoting metastasis through decreasing miR-200s activity in osteosarcoma," *Dna Cell Biol.*, vol. 35, no. 5, 2016, Art. no. 235.
[46] E. Li, Z. Zhao, B. Ma, et al., "Long noncoding RNA HOTAIR promotes the proliferation and metastasis of osteosarcoma cells through the AKT/mTOR signaling pathway," *Experimental Therapeutic Med.*, vol. 14, no. 6, pp. 5321–5328, 2017.
[47] S. Z. Zhang, L. Cai, B. Li, "MEG3 long non-coding RNA prevents cell growth and metastasis of osteosarcoma," *Bratislavske Lekarske Listy*, vol. 118, no. 10, 2017, Art. no. 632.
[48] T. Li, Y. Xiao, T. Huang, "HIF-1-induced upregulation of lncRNA UCA1 promotes cell growth in osteosarcoma by inactivating the PTEN/AKT signaling pathway," *Oncology Reports*, vol. 39, no. 3, pp. 1072–1080, 2018.
[49] L. Peng, X. Yuan, B. Jiang, et al., "LncRNAs: Key players and novel insights into cervical cancer," *Tumor Biol.*, vol. 37, no. 3, pp. 2779–2788, 2015.
[50] M. Iden, S. Fye, K. Li, et al., "The lncRNA PVT1 contributes to the cervical cancer phenotype and associates with poor patient prognosis," *PLoS One*, vol. 11, no. 5, 2016, Art. no. e0156274.
[51] J. Zhang, Z. Lin, Y. Gao, et al., "Downregulation of long noncoding RNA MEG3 is associated with poor prognosis and promoter hypermethylation in cervical cancer," *J. Exp. Clinical Cancer Res.*, vol. 36, no. 1, 2017, Art. no. 5.