




## Research Article

# A Novel Approach for Predicting Disease-lncRNA Associations Based on the Distance Correlation Set and Information of the miRNAs

Haochen Zhao <sup>1,2</sup>, Linai Kuang <sup>1,2</sup>, Lei Wang <sup>1,2</sup> and Zhanwei Xuan<sup>1,2</sup>

<sup>1</sup>College of Information Engineering, Xiangtan University, Xiangtan 411105, China

<sup>2</sup>Key Laboratory of Intelligent Computing & Information Processing, Xiangtan University, Xiangtan 411105, China

Correspondence should be addressed to Lei Wang; [wanglei@xtu.edu.cn](mailto:wanglei@xtu.edu.cn)

Received 6 December 2017; Revised 4 April 2018; Accepted 17 April 2018; Published 26 June 2018

Academic Editor: Michele Migliore

Copyright © 2018 Haochen Zhao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recently, accumulating laboratorial studies have indicated that plenty of long noncoding RNAs (lncRNAs) play important roles in various biological processes and are associated with many complex human diseases. Therefore, developing powerful computational models to predict correlation between lncRNAs and diseases based on heterogeneous biological datasets will be important. However, there are few approaches to calculating and analyzing lncRNA-disease associations on the basis of information about miRNAs. In this article, a new computational method based on distance correlation set is developed to predict lncRNA-disease associations (DCSLDA). Comparing with existing state-of-the-art methods, we found that the major novelty of DCSLDA lies in the introduction of lncRNA-miRNA-disease network and distance correlation set; thus DCSLDA can be applied to predict potential lncRNA-disease associations without requiring any known disease-lncRNA associations. Simulation results show that DCSLDA can significantly improve previous existing models with reliable AUC of 0.8517 in the leave-one-out cross-validation. Furthermore, while implementing DCSLDA to prioritize candidate lncRNAs for three important cancers, in the first 0.5% of forecast results, 17 predicted associations are verified by other independent studies and biological experimental studies. Hence, it is anticipated that DCSLDA could be a great addition to the biomedical research field.

## 1. Introduction

For long time, RNA was just considered to be transcriptional noise and intermediary between a DNA sequence and its encoded protein [1, 2]. However, sequence analyses point out that more than 98% of the human genome does not encode protein sequences [3]. Furthermore, increasing studies based on biological experiments have indicated that ncRNAs play important roles in numerous critical biological processes such as chromosome dosage compensation, epigenetic regulation, and cell growth [4]. In particular, the lncRNAs, as a class of important ncRNAs with a length more than 200 nucleotides [5], have been found to be associated with a wide range of human diseases, such as breast cancer [6], colorectal cancer [7], lung cancer [8], and cardiovascular diseases [9]. Hence, the study of finding novel disease-lncRNA associations has captured the attention of a lot of

researchers and has been considered as one of the hottest topics in the research fields of diseases and lncRNAs. The identification of disease-lncRNA association can not only accelerate the understanding of human complex disease mechanism at the lncRNA level, but also serve as a biomarker identification for human disease diagnosis, treatment, and prevention [10]. So far, a lot of studies have generated a large amount of lncRNAs related biological data about sequence, expression, function, and so on [11–13]. However, compared with the rapidly increasing number of newly discovered lncRNAs, only few known lncRNA-disease associations have been reported. Hence, it is challenging and urgently needed to develop efficient and successful computational approaches to predict potential lncRNA-disease associations. In recent years, some computational methods have been proposed to predict novel lncRNA-disease associations, which can significantly decrease the time and cost of biological experiments

by calculating the association probability of lncRNA-disease pairs. For example, Chen G et al. presented the first prediction method (genomic locus based) and constructed a lncRNA-disease association database as well [14]. Liang et al. proposed a genetic mediator and key regulator model to unveil the subtle relationships between lncRNAs and lung cancer. Liu et al. developed a computational framework to accomplish this by combining human lncRNA expression profiles, gene expression profiles, and human disease-associated gene data. Applying this framework to available human long intergenic noncoding RNAs (lincRNAs) expression data, Chen et al. developed a semi-supervised learning method based on framework of Laplacian Regularized Least Squares, LRL-SLDA, to infer potential lncRNA-disease associations which did not need negative samples and could obtain a reliable AUC of 0.7760 in the leave-one-out cross-validations [15]. In 2014, Sun et al. constructed a lncRNA functional similarity network and applied random walk with restart (RWR) to infer potential lncRNA-disease associations [16]. In the same year, Li et al. presented a bioinformatics method based on genomic location to predict the lncRNAs associated with vascular disease [17]. Then, Zhao et al. developed a computational method based on the naïve Bayesian classifier to identify cancer-related lncRNAs by integrating genome, regulome, and transcriptome data [18]. In 2015 Zhou et al. proposed a novel rank-based method named RWRHLDA to prioritize candidate lncRNA-disease associations by integrating miRNA-associated lncRNA-lncRNA crosstalk network, disease-disease similarity network, and known lncRNA-disease association network into a heterogeneous network and implemented a random walk with restart on the newly generated heterogeneous network [19].

Nowadays, with advent of many biological datasets, such as lncRNADisease [14], lncRNADB [20], and NONCODE [13], the number of lncRNA-disease associations is still very limited. In 2015, Chen developed a method, named HGLDA, based on the information of miRNA [21], which predicted lncRNA-disease associations by integrating disease-miRNA associations with lncRNA-miRNA interactions and did not rely on known lncRNA-disease associations. Different from the method of HGLDA proposed by Chen et al., in this article, on the basis of experimentally reported lncRNA-disease associations collected from the HMDD database [22] and miRNA-lncRNA associations collected from the starBase database [23], a novel model based on distance correlation set is developed to predict potential lncRNA-disease associations by integrating known lncRNA-miRNA associations and known miRNA-disease associations. Compared with HGLDA, the advantage of DCSLDA lies in the introduction of the similarity of disease pairs and lncRNA pairs and distance correlation set. In addition, to optimize the prediction performance of DCSLDA, new methods to calculate the similarity of disease-disease pairs and lncRNA-lncRNA pairs are developed simultaneously. Finally, to evaluate the prediction performance of DCSLDA, LOOCV is implemented on the basis of the known lncRNA-disease associations and known lncRNA-cancer associations separately, and simulation results demonstrate that DCSLDA is superior to the state-of-the-art methods and can achieve a

reliable AUC of 0.8517 in the LOOCV when the pre-given threshold parameter  $r$  is set at 6. Additionally, to further evaluate the prediction performance of DCSLDA, case studies of breast cancer, colorectal cancer, and lung cancer are implemented for DCSLDA; as a result, among the first 0.5% of predictive results, 9, 6, and 2 predicted potential associations are confirmed by recent experimental reports, respectively. Hence, considering the excellent prediction performance of DCSLDA, it is obvious that DCSLDA can become a useful and efficient computational tool for biomedical researches.

## 2. Materials and Methods

**2.1. Disease-miRNA Associations.** We downloaded known disease-miRNA associations from the Human MicroRNA Disease Database (HMDD) in July 2017 (see Supplementary file 1), which included 10381 experimentally verified disease-miRNA associations (including 572 miRNAs and 383 diseases). After merging miRNAs which produce the same mature miRNA and eliminating duplicate data, we obtained *dataset1* including 5430 disease-miRNA associations (including 383 human diseases and 495 lncRNAs). Let  $D$  be the number of different diseases and  $M1$  be the number of different miRNAs collected from the *dataset1*, respectively,  $S_D = \{d_1, d_2, \dots, d_D\}$  represent the set of these  $D$  different diseases, and  $S_{M1} = \{m1_{D+1}, m1_{D+2}, \dots, m1_{D+M1}\}$  represent the set of these  $M1$  different miRNAs; then for any given  $d_i \in S_D$  and  $m1_j \in S_{M1}$ , we can define the *Association Strong Correlation (ASCI)* between  $d_i$  and  $m1_j$  as follows:

$$\begin{aligned} \text{ASCI}(d_i, m1_j) &= \begin{cases} 1, & \text{If } d_i \text{ is related to } m1_j \text{ in the dataset1} \\ 0, & \text{otherwise.} \end{cases} \quad (1) \end{aligned}$$

**2.2. miRNA-lncRNA Associations.** We downloaded known miRNA-lncRNA associations dataset from starBase v2.0 dataset in July 2017, which provided the most comprehensive experimentally confirmed lncRNA-miRNA interactions based on large scale CLIP-seq data. After data preprocessing (including elimination of duplicate values, erroneous data, disorganized data, and so on), *dataset2* (including 10195 lncRNA-miRNA associations, 275 miRNAs, and 1127 lncRNAs) was obtained from the starBase v2.0 (see Supplementary file 2). Let  $M2$  be the number of different miRNAs and  $L$  be the number of different lncRNAs collected from the *dataset2*,  $S_{M2} = \{m2_1, m2_2, \dots, m2_{M2}\}$  represent the set of these  $M2$  different miRNAs, and  $S_L = \{l_{M2+1}, l_{M2+2}, \dots, l_{M2+L}\}$  represent the set of these  $L$  different lncRNAs; then, for any given  $m2_i \in S_{M2}$  and  $l_j \in S_L$ , we can define the ASC2 between  $m2_i$  and  $l_j$  as follows:

$$\begin{aligned} \text{ASC2}(m2_i, l_j) &= \begin{cases} 1, & \text{If } m2_i \text{ is related to } l_j \text{ in the dataset2} \\ 0, & \text{otherwise.} \end{cases} \quad (2) \end{aligned}$$

**2.3. lncRNA-Disease Associations.** In order to evaluate the performance of DCSLDA, the newly lncRNA-disease associations were downloaded from LncRNADisease database, which integrated more than 1000 lncRNA-disease entries and 475 lncRNA interaction entries, including 321 lncRNAs and 221 diseases from ~500 publications. In this dataset, after duplicate associations and the lncRNA-disease associations involved in either diseases or lncRNAs which were not contained in the *dataset1* or *dataset2* were removed, 203 high-quality lncRNA-disease associations were obtained finally (see Supplementary file 3).

**2.4. Disease Functional Similarity Based on miRNAs.** For calculating the functional similarity between diseases, we introduced the concept of social network. In the social network, for any two nodes, we can calculate the similarities between them by comparing and integrating the similarities of nodes associated with these two nodes. In this section, based on the assumption that similar diseases tend to show a similar interaction and noninteraction pattern with the miRNAs, we calculated the disease similarity in the disease-miRNA interactive network. As illustrated in Figure 1, the calculation procedures of disease functional similarity based on miRNAs include 3 steps. First, we constructed miRNA-disease interactive network from known miRNA-disease associations (*dataset1*), whose topology can be abstracted as an undirected graph  $G_1 = (V_1, E_1)$ , where  $V_1 = S_D \cup S_{M1} = \{d_1, d_2, \dots, d_D, m1_{D+1}, m1_{D+2}, \dots, m1_{D+M1}\}$  is the set of vertices,  $E_1$  is the set of edges, and, for any two nodes  $a, b \in V_1$ , there is an edge between  $a$  and  $b$  in  $E_1$ , if and only if there are  $a \in S_D, b \in S_{M1}$ , and  $ASC1(a, b) = 1$ . However, since different miRNA terms in the *dataset1* may relate to different numbers of diseases, it is not suitable to assign the same contribution value to different miRNAs. Hence, we define the contribution value of each miRNA as follows:

$$C_D(m_i) = -\lg\left(\frac{\text{the number of } m_i \text{ - related edges in } E_1}{\text{the number of all edges in } E_1}\right). \quad (3)$$

Finally, we defined the functional similarity between diseases  $d_i$  and  $d_j$  by integrating the miRNAs related to  $d_i, d_j$ , or both of them as follows:

$$FSD(d_i, d_j) = \frac{\exp \sum_{m_k \in (D(d_i) \cap D(d_j))} C_D(m_k)}{|D(d_i)| + |D(d_j)| - |D(d_i) \cap D(d_j)|} \quad (4)$$

where  $FSD$  is the disease functional similarity matrix calculated based on miRNA and  $D(d_i)$  and  $D(d_j)$  are the number of  $d_i$  related edges and  $d_j$  related edges in  $E_1$ , respectively. As an example, in Figure 1, there is  $FSD(d_1, d_2) = \exp(C_D(m_1) + C_D(m_3) + C_D(m_4))/(4 + 5 - 3)$ .

**2.5. lncRNA Functional Similarity Based on miRNAs.** Based on the assumption that similar lncRNAs tend to show a similar interaction and noninteraction pattern with the miRNAs, we can calculate the lncRNA similarity in the lncRNA-miRNA interactive network. Similar to the calculation procedures of disease functional similarity, first, we constructed lncRNA-miRNA interactive network from known

lncRNA-miRNA associations (*dataset2*), whose topology can be abstracted as an undirected graph  $G_2 = (V_2, E_2)$ , where  $V_2 = S_{M2} \cup S_L = \{m2_1, m2_2, \dots, m2_{M2+1}, m2_{M2+2}, \dots, m2_{M2+L}\}$  is the set of vertices,  $E_2$  is the set of edges, and, for any two nodes  $a, b \in V_2$ , there is an edge between  $a$  and  $b$  in  $E_2$ , if and only if there are  $a \in S_{M2}, b \in S_L$ , and  $ASC2(a, b) = 1$ . Then, considering the number of lncRNA-miRNA associations, we defined the contribution value of each miRNA as follows:

$$C_L(m_i) = -\log_2\left(\frac{\text{the number of } m_i \text{ - related edges in } E_2}{\text{the number of all edges in } E_2}\right). \quad (5)$$

Additionally, we defined the functional similarity between lncRNA  $l_i$  and  $l_j$  by integrating the miRNAs related to  $l_i, l_j$ , or both of them as follows:

$$FSL(l_i, l_j) = \frac{\exp \sum_{m_k \in (D(l_i) \cap D(l_j))} C_L(m_k)}{|D(l_i)| + |D(l_j)| - |D(l_i) \cap D(l_j)|} \quad (6)$$

where  $FSL$  is the disease functional similarity matrix calculated based on miRNA and  $D(l_i)$  and  $D(l_j)$  are the number of  $l_i$  related edges and  $l_j$  related edges in  $E_2$ , respectively.

**2.6. Method for Predicting Potential Association between lncRNAs and Diseases.** Based on the assumptions that similar diseases tend to show a similar interaction and noninteraction pattern with the miRNAs and similar miRNAs tend to show a similar interaction and noninteraction pattern with the lncRNAs, we proposed a novel model, DCSLDA, based on miRNAs and distance correlation set to predict potential disease-lncRNA associations. As illustrated in Figure 2, the procedures of DCSLDA consist of the following 6 major steps.

*Step 1* (construction of the disease-miRNA-lncRNA interaction network). On the basis of the above descriptions and letting  $M = M1 \cap M2$ , we can construct a disease-miRNA-lncRNA interaction network based on *dataset1* and *dataset2*, whose topology can be abstracted to an undirected graph  $G_3 = (V_3, E_3)$ , where  $V_3 = S_D \cup S_M \cup S_L = \{d_1, d_2, \dots, d_D, m_{D+1}, m_{D+2}, \dots, m_{D+M}, l_{D+M+1}, l_{D+M+2}, \dots, l_{D+M+L}\}$  is the set of vertices,  $E_3$  is the edge set of  $G_3$ , and  $\forall l_i \in L, m_j \in M, d_k \in D$ . There is an edge between  $l_i$  and  $m_j$  in  $E_3$ , if and only if the lncRNA  $l_i$  relates to the miRNA  $m_j$ . Moreover, there is an edge between  $m_j$  and  $d_k$  in  $E_3$ , if and only if the miRNA  $m_j$  is related to the disease  $d_k$ . Then, for any given  $a, b \in V_3$ , we can define the  $ASC3$  between  $a$  and  $b$  as follows:

$$ASC3(a, b) = \begin{cases} 1, & \text{If there exists an edge between } a \text{ and } b \text{ in the } E_3 \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

In addition, although we did not use any known disease-lncRNA associations, the diseases and lncRNAs can still be linked by integrating edges between diseases node and miRNAs node and edges between miRNAs nodes and lncRNAs nodes in the  $G_3$ .

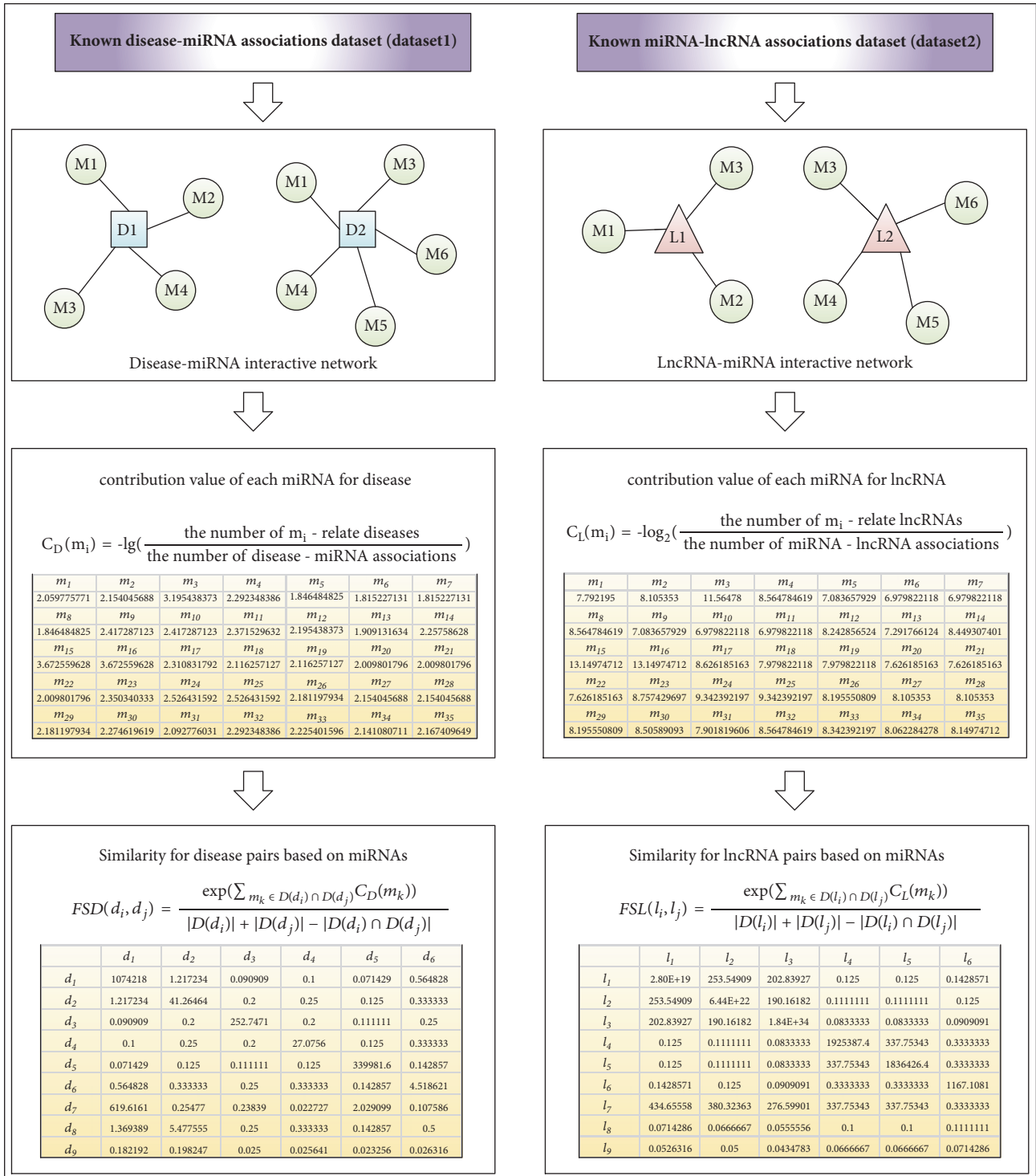


FIGURE 1: The flowchart of functional similarity calculation based on information of miRNA includes three steps: (1) constructing known disease-miRNA association and miRNA-lncRNA association network respectively; (2) obtaining contribution of each miRNA; (3) calculating functional similarity for diseases and lncRNAs, respectively.

*Step 2* (construction of the *Adjacency Matrix* based on the disease-miRNA-lncRNA interactive network). We can construct a  $(D + M + L) \times (D + M + L)$

dimensional *Adjacency Matrix* (*AM*) based on the disease-miRNA-lncRNA interactive network as follows:

$$AM(i, j) = \begin{cases} ASC3(d_i, d_j), & \text{if } i \in [1, D], j \in [1, D]. \\ ASC3(d_i, m_j), & \text{if } i \in [1, D], j \in [D, D + M]. \\ ASC3(d_i, l_j), & \text{if } i \in [1, D], j \in [D + M, D + M + L]. \\ ASC3(m_i, d_j), & \text{if } i \in [D, D + M], j \in [1, D]. \\ ASC3(m_i, m_j), & \text{if } i \in [D, D + M], j \in [D, D + M]. \\ ASC3(m_i, l_j), & \text{if } i \in [D, D + M], j \in [D + M, D + M + L]. \\ ASC3(l_i, d_j), & \text{if } i \in [D + M, D + M + L], j \in [1, D]. \\ ASC3(l_i, m_j), & \text{if } i \in [D + M, D + M + L], j \in [D, D + M]. \\ ASC3(l_i, l_j), & \text{if } i \in [D + M, D + M + L], j \in [D + M, D + M + L]. \end{cases} \quad (8)$$

where  $i \in [1, D + M + L]$  and  $j \in [1, D + M + L]$ .

*Step 3* (construction of the shortest distance matrix based on the disease-miRNA-lncRNA interactive network). Let  $r$  be a pre-given positive integer; then we can obtain  $r$  matrixes such as  $AM^1, AM^2, \dots, AM^r$  based on the *Adjacency Matrix*. Then, we can construct a  $(D+M+L) \times (D+M+L)$  dimensional Shortest Path Matrix (*SPM*) as follows:

$$SPM(i, j) = \begin{cases} 0, & \text{if } AM^r(i, j) = 0 \\ 1, & \text{if } AM(i, j) = 1 \\ k, & \text{otherwise} \end{cases} \quad (9)$$

where  $i \in [1, D + M + L], j \in [1, D + M + L], k \in [2, r]$ , and  $k$  satisfies  $AM^k(i, j) \neq 0$  while  $AM^1(i, j) = AM^2(i, j) = \dots = AM^{k-1}(i, j) = 0$ .

*Step 4* (collection of the *distance correlation sets* for nodes in the interactive network). In  $G = (V, E)$ , let  $V = \{d_1, d_2, \dots, d_D, m_{D+1}, m_{D+2}, \dots, m_{D+M}, l_{D+M+1}, l_{D+M+2}, \dots, l_{D+M+L}\} = \{v_1, v_2, \dots, v_D, v_{D+1}, v_{D+2}, \dots, v_{D+M}, v_{D+M+1}, v_{D+M+2}, \dots, v_{D+M+L}\}$ ; then for each node  $v_i \in V$ , we can obtain its distance correlation set  $DCS_i$  according to the shortest distance matrix as follows:

$$DCS_i = \{v_j \mid r \geq SPM(i, j) > 0, i \neq j\}. \quad (10)$$

For instance, in the disease-miRNA-lncRNA interaction network illustrated in Figure 3, supposing that we hope to collect the  $DCS_{D1}$ , then according to the above description, we can easily know that the *distance correlation sets* of  $D1$  will be  $\{M1, M2, M3, M4, L1, L2, L3, L4, L5\}$  when  $r = 2$ .

And thereafter, for any given node  $v_j \in DCS_i$ , where  $j \neq i$ , we can compute the distance correlation coefficient  $P(i, j)$  between the node  $v_i$  and  $v_j$  as follows:

$$P(i, j) = P(v_i, v_j) = \begin{cases} 1 - \frac{SPM(i, j)}{r + 1}, & \text{if } SPM(i, j) \neq 0 \\ 0, & \text{else.} \end{cases} \quad (11)$$

Hence, based on (11), we can further obtain a  $(D+M+L) \times (D+M+L)$  dimensional *Distance Correlation Coefficient Matrix* (*DCCM*) as follows:

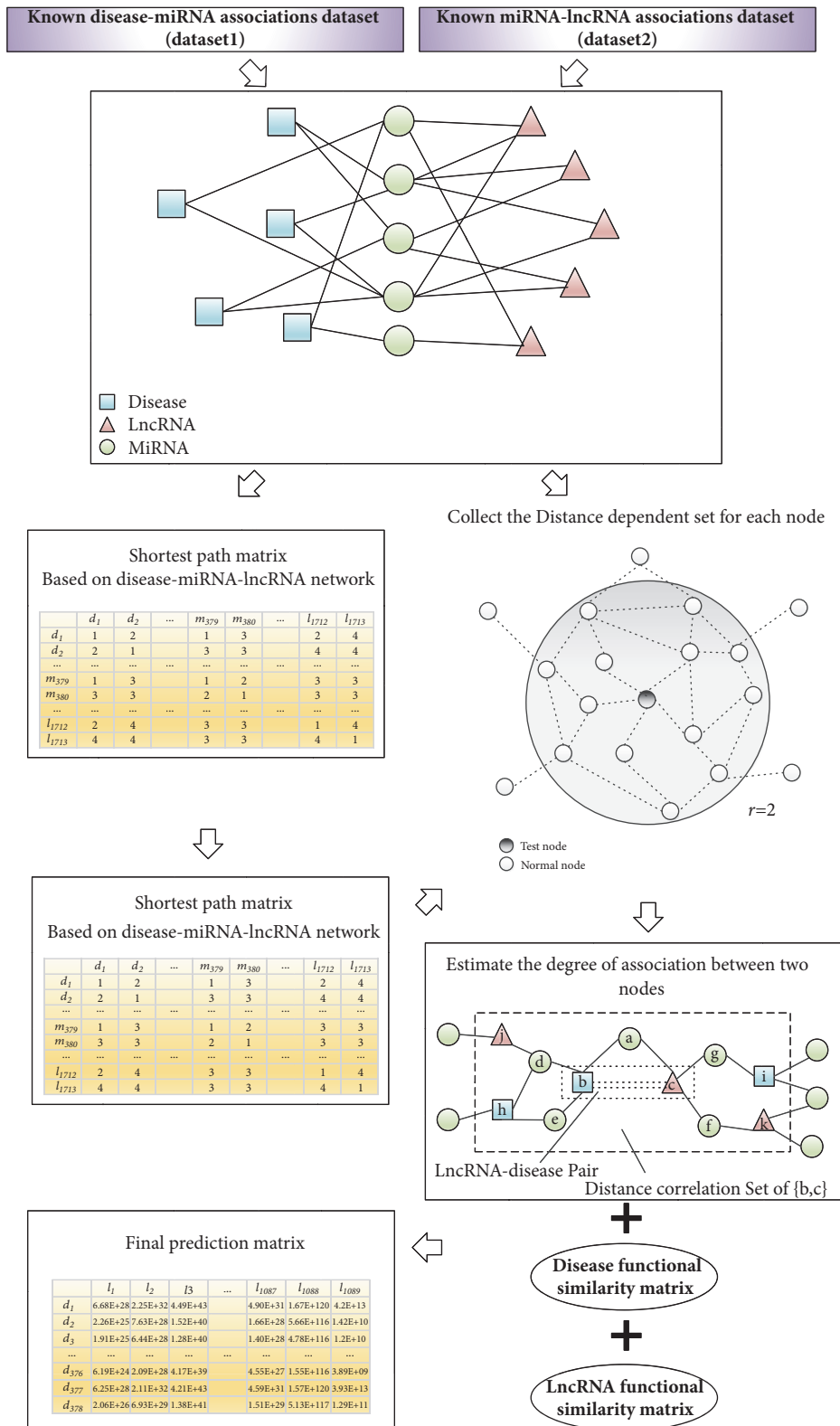
$$DCCM(i, j) = \begin{cases} \frac{r}{r + 1} & \text{if node } v_i = v_j \\ P(i, j), & \text{if node } v_j \in DCS_i \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

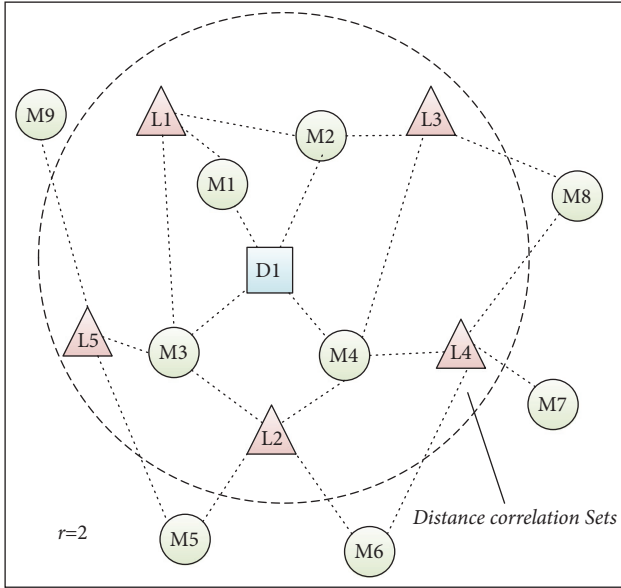
where  $i \in [1, D + M + L]$  and  $j \in [1, D + M + L]$ .

*Step 5* (estimation of association degree between a pair of nodes in the disease-miRNA-lncRNA interactive network). Based on (12), we can obtain distance correlation coefficient of each nodes pair. For any given nodes pair  $(v_i, v_j)$  in  $G = (V, E)$ , where  $V = \{d_1, d_2, \dots, d_D, l_{D+1}, l_{D+2}, \dots, l_{D+L}\} = \{v_1, v_2, \dots, v_D, v_{D+1}, v_{D+2}, \dots, v_{D+L}\}$  and  $\{v_i, v_j\} \subseteq V$ , we can obtain the association degree (*AD*) between them as follows:

$$AD(i, j) = \frac{\sum_{k \in DCS_i} DCCM(i, k) + \sum_{k \in DCS_j} DCCM(k, j)}{D + M + L} \quad (13)$$

where  $i \in [1, D + M + L]$  and  $j \in [1, D + M + L]$ .



FIGURE 3: Distance correlation set of D1 with  $r=2$ .

Step 6 (construction of the *Final Prediction Result Matrix*).

Based on (13), let  $AD = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$ , where  $C_{11}$  is a  $D \times D$  matrix,  $C_{12}$  is a  $D \times M$  matrix,  $C_{13}$  is a  $D \times L$  matrix,  $C_{21}$  is a  $M \times D$  matrix,  $C_{22}$  is a  $M \times M$  matrix,  $C_{23}$  is a  $M \times L$  matrix,  $C_{31}$  is a  $L \times D$  matrix,  $C_{32}$  is a  $L \times M$  matrix, and  $C_{33}$  is a  $L \times L$  matrix. It can be easily inferred that the matrix  $C_{13}$  will be our prediction results, which provided the association probability between each disease and lncRNA. Moreover, we can introduce disease functional similarity and lncRNA functional similarity for  $C_{13}$  as follows:

$$FAD = FSD \times C_{13} \times FSL \quad (14)$$

where the entity  $FAD(i, j)$  in row  $i$  column  $j$  reflects the probability that the lncRNA  $l(j)$  is related to the disease  $d(i)$ .

### 3. Results and Case Studies

To evaluate the prediction performance of DCSLDA, first of all, we implemented LOOCV (leave-one-out cross-validation) to compare DCSLDA with HGLDA [21] based on the lncRNA-disease association dataset downloaded from lncRNADisease database [14]. Next, LOOCV would be implemented to further evaluate the prediction performance of DCSLDA based on the known experimentally verified lncRNA-cancer associations. And then, the effects of the disease functional similarity and the lncRNA functional similarity to the prediction performance of DCSLDA would be analyzed also. Finally, experimental results about the prediction of associations between lncRNAs and three cancers were listed (see Table 1), and the performance comparisons between DCSLDA and HGLDA were implemented according to the rankings of these new disease-related lncRNAs in the case studies of three cancers (see Table 2).

TABLE 1: 17 predicted lncRNA-disease pairs with high predicted value while DCSLDA was applied to three important kinds of cancer (breast cancer, colorectal cancer, and lung cancer).

Cancer	lncRNA	PMID
Breast cancer	KCNQ1OT1	21304052; 26323944
Breast cancer	MALAT1	24525122; 19379481
Breast cancer	XIST	27248326
Breast cancer	NEAT1	25417700; 28034643
Breast cancer	LINC00657	26942882
Breast cancer	SNHG16	28232182
Breast cancer	CASP8AP2	28388918
Breast cancer	PPP1R9B	26387546
Breast cancer	TUG1	27791993
Colorectal cancer	KCNQ1OT1	16965397; 11340379
Colorectal cancer	MALAT1	25025966
Colorectal cancer	XIST	17143621
Colorectal cancer	NEAT1	26552600
Colorectal cancer	SNHG16	26823726
Colorectal cancer	CASP8AP2	22216762
Lung cancer	MALAT1	20937273; 24757675; 24667321
Lung cancer	XIST	27501756

TABLE 2: Performance comparisons between DCSLDA and HGLDA based on the rankings of ten lncRNA-disease associations related to three important kinds of cancer (breast cancer, colorectal cancer, and lung cancer).

Cancer	lncRNA	DCSLDA	HGLDA
Breast cancer	KCNQ1OT1	1	8
Breast cancer	MALAT1	4	30
Breast cancer	XIST	5	1
Breast cancer	NEAT1	8	12
Breast cancer	SNHG16	12	3
Colorectal cancer	KCNQ1OT1	1	5
Colorectal cancer	MALAT1	4	3
Colorectal cancer	XIST	5	1
Lung cancer	MALAT1	4	9
Lung cancer	XIST	5	1
Average ranks		4.9	7.3

3.1. *Performance Evaluation of Potential Disease-lncRNA Association Prediction.* According to the lncRNA-disease association datasets downloaded from lncRNADisease database, DCSLDA and HGLDA were applied in the framework of LOOCV, respectively. While the LOOCV was implemented for investigated diseases and lncRNAs, each known lncRNA-disease association would be left out in turn as test sample, and then we further evaluated how well this association ranked relatively to the candidate samples. Here, the candidate samples comprised all potential lncRNA-disease pairs without confirmed associations. Therefore, after the implementation of DCSLDA was completed, the rank of each left-out testing sample relative to the candidate samples

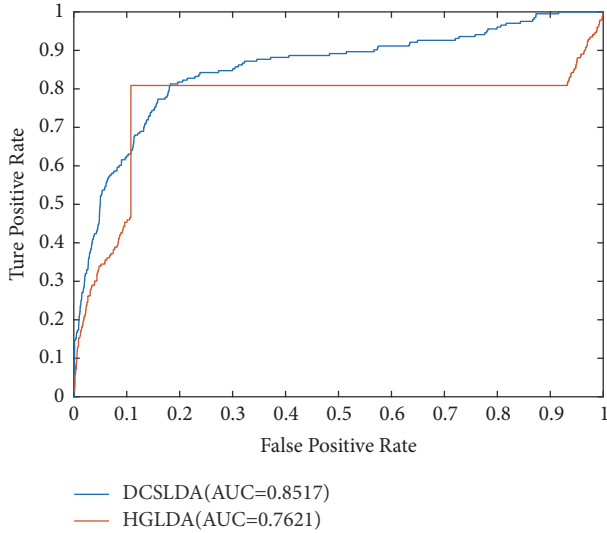


FIGURE 4: Performance comparisons between DCSLDA and HGLDA in terms of ROC curve and AUC based on LOOCV.

could be further obtained. And then, the testing samples with a prediction rank higher than the given threshold were considered successfully predicted. Thus, we could further obtain the corresponding true positive rates (TPR, sensitivity) and false positive rates (FPR, 1-specificity) by setting different thresholds. Here, sensitivity refers to the percentage of test samples that were predicted with ranks higher than the given threshold, and the specificity was computed as the percentage of negative samples with ranks lower than the threshold. Therefore, the receiver-operating characteristics (ROC) curves could be drawn by plotting TPR versus FPR at different thresholds. And then, the areas under ROC curve (AUC) would be further calculated to evaluate the prediction performance of DCSLDA. An AUC value of 1 represented a perfect prediction while an AUC value of 0.5 indicated purely random performance.

The results of the performance comparison between DCSLDA and HGLDA were shown in Figure 4. Since the HGLDA method predicts lncRNA-disease associations without relying on the information of known disease-lncRNA association, it was selected for performance comparison with our method DCSLDA. As a result, it is clear that our newly proposed method DCSLDA achieved the AUC of 0.8517 in the framework of LOOCV, which is much higher than the AUC of 0.7621 achieved by HGLDA [21]. Simulation results indicate that DCSLDA significantly improved the performance of HGLDA by at least 0.0896 in the term of AUC values and fully demonstrate the performance superiority of HGLDA.

**3.2. Performance Evaluation of Potential lncRNA-Cancer Association Prediction.** Cancer has become one of the most dangerous killers for human beings [24, 25], and there is a high incidence of cancer in both developed countries and developing countries. Therefore, to further evaluate the prediction performance of DCSLDA, LOOCV was implemented

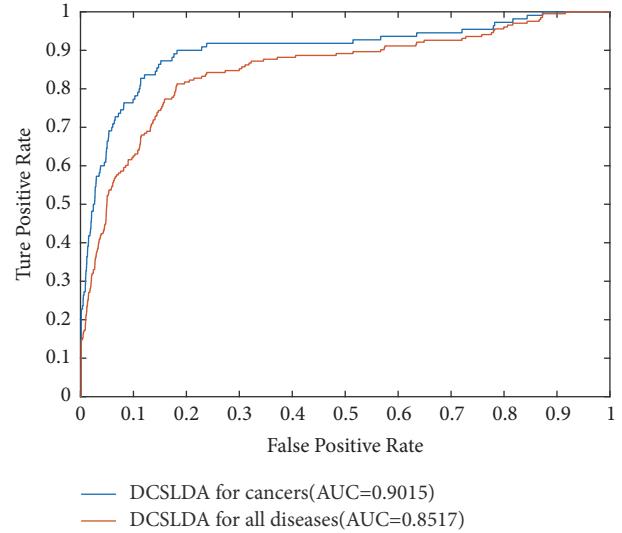


FIGURE 5: Performance evaluation of potential lncRNA-cancer association prediction in terms of ROC curve and AUC based on LOOCV.

on the basis of 117 lncRNA-cancer associations collected from the lncRNADisease dataset, and the simulation results were illustrated in Figure 5.

From Figure 5, it is easy to find that DCSLDA achieved the AUC of 0.9015 in the frameworks of LOOCV when  $r$  is set as 6, which indicates that our newly proposed method DCSLDA has a reliable predictive performance of cancers, and therefore it is a precise and high efficient method for the lncRNA-disease association prediction.

**3.3. Effects of the Disease Functional Similarity and lncRNA Functional Similarity.** In formula (14), we defined  $FAD = FSD \times C_{13} \times FSL$ . Then, in this section, we will analyze the effects of the disease similarity matrix  $FSD$  and the lncRNA similarity matrix  $FSL$  through comparing the prediction performances of DCSLDA in the framework of LOOCV while letting  $FAD = C_{13}$  and  $FAD = FSD \times C_{13} \times FSL$ , respectively. The simulation results are illustrated in Figure 6. It is obvious that DCSLDA achieved the AUCs of 0.8517 while matrixes  $FSD$  and  $FSL$  were considered, but the AUC achieved by DCSLDA is 0.8352 only when letting  $FAD = C_{13}$ . Simulation results indicated that the prediction performance of DCSLDA will be significantly improved by introducing the similarity matrixes  $FSD$  and  $FSL$ . Moreover, in Table 1, DCSLDA was applied to three important kinds of cancer (breast cancer, colorectal cancer, and lung cancer). As a result, 17 predicted lncRNA-disease pairs with high predicted value were publicly released to benefit the biological experimental validation.

**3.4. Case Studies.** Obviously, DCSLDA can predict all potential relationships between diseases and lncRNAs in *dataset1* and *dataset2* simultaneously. And of course, potential associations with high predicted value can be publicly released to benefit the biological experimental validation. It is anticipated that these potential disease-lncRNA associations that significantly share common miRNAs could be validated by



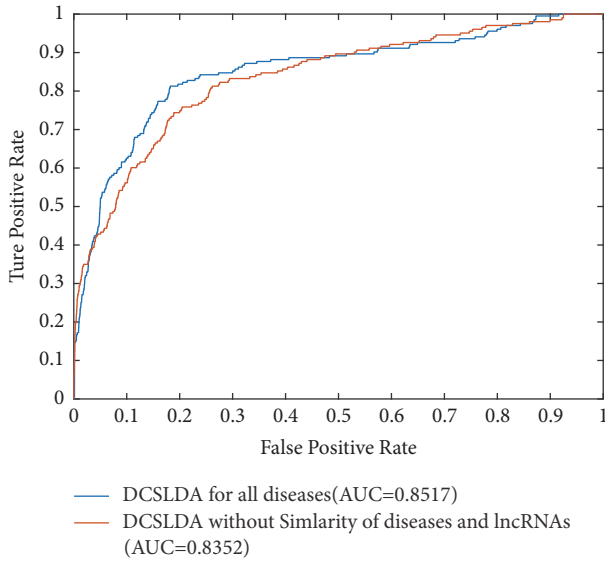


FIGURE 6: Comparison of effects of the disease functional similarity and lncRNA functional similarity to the prediction performance of DCPSLDA in the framework of LOOCV with  $r = 6$ .

biological experiments and provide important complement for experimental studies. Moreover, plentiful evidence has indicated that lncRNAs played important roles in various kinds of human cancers. The predicted results were sorted from best to worse, among which the first 0.5% results are selected to be analyzed (see Supplementary file 4). Case studies about three important kinds of cancers based on top 0.5% of predicted results were implemented to show the predictive performance of DCPSLDA. Prediction results were verified based on the recent updates in the LncRNADisease dataset and recently published experimental literature (ranking results have been listed in Table 1).

In the world, breast cancer is the most prevalent cancer in women and a major public health problem. Several studies have focused on studying this disease, but more are needed, especially at the genetic and molecular levels [26, 27]. Therefore, it is necessary to predict breast cancer-related lncRNAs and identify lncRNA biomarkers. DCPSLDA was implemented to prioritize candidate lncRNAs for breast cancer. Among the first 5% of predictive results, nine breast cancer-related lncRNAs have been confirmed based on recent experimental literature (see Table 1). For example, KCNQ1OT1, MALAT1, XIST, and NEAT1 are experimentally confirmed breast cancer-related lncRNAs, which have been ranked 2nd, 11th, 12th, and 19th in the predicted list based on the model of DCPSLDA, respectively. KCNQ1OT1 had significantly higher expression levels in invasive breast carcinoma and was induced by estrogen in estrogen receptor- $\alpha$  expressing breast cancer cells [28].  $17\beta$ -Estradiol treatment affects breast tumor or nontumor cells proliferation, migration, and invasion in an ER $\alpha$ -independent, but a dose-dependent, way by decreasing the MALAT1 RNA level [29]. XIST expression is significantly reduced in breast cancer cell lines and breast cancer samples [30]. Breast cancer patients with high level of NEAT1 expression show low survival rate [31].

Colorectal cancer (CRC) is a leading cause of cancer deaths worldwide, one of the fundamental processes driving the initiation and progression of CRC is the accumulation of a variety of genetic and epigenetic changes in colon epithelial cells. Colorectal cancer is usually caused by the combination of various factors, such as genetic and epigenetic changes [32, 33]. Specially, lncRNAs have been demonstrated to play a critical role in the development and progression of colon cancer [34]. As a result, six colorectal cancer-related lncRNAs were listed in Table 1. For example, Tanaka K et al. proved that Loss of imprinting of KCNQ1OT1 is considered as a useful marker for diagnosis of colorectal cancer because of its frequent occurrences in colorectal cancer samples [35]. Ji Q et al. findings implied that MALAT1 might be a potential predictor for tumor metastasis and prognosis [36]. Furthermore, the interaction between MALAT1 and SFPQ could be a novel therapeutic target for CRC. Lassmann S et al. proved that expression level change of or DNA amplification of XIST is associated with colorectal cancer [37].

Over the past 30 years, the morbidity and mortality of lung cancer have been increasing and the cancer has the highest incidence and mortality across the world [38]. Due to the early diagnosis of lung cancer and the lack of effective treatment, its survival rate is around 10% within five years, which seriously endangers human health. More and more evidence has shown that lncRNAs play a critical role in treatment of lung cancers. Among the first 5% of predictive results, three predicted lncRNAs have been confirmed by published experimental literature [39]. According to this literature, MALAT1 has been shown to be highly associated with metastasis of lung cancer and promote lung cancer cell motility by regulating motility related gene expression [40, 41]. Long noncoding RNA XIST acts as an oncogene in non-small cell lung cancer by epigenetically repressing KLF2 expression [42].

In addition, performance comparisons between DCPSLDA and HGLDA were implemented according to the rankings of these disease-related lncRNAs in the case studies of breast cancer, colorectal cancer, and lung cancer (see Table 2). By ranging the predicted results by HGLDA and our methods from good to bad, we selected the intersection of the underlying disease-lncRNA relationship predicted by HGLDA and the first 0.5 percent of the predicted results by our methods and listed the lncRNA items related to breast cancer, colorectal cancer, and lung cancer in this intersection in Table 2. As a result, DCPSLDA significantly improved the prediction ability of HGLDA with higher ranks for these new disease-related lncRNAs.

#### 4. Discussion and Conclusions

In recent years, plenty of studies have generated an enormous amount of biological data related to lncRNAs. Accumulating evidence shows that lncRNAs have played a very important role in the biological functions, and the study of lncRNA-disease association prediction is of great significance to human beings. However, there is a few computational models for predicting potential disease-lncRNA associations based on the information of miRNA. To utilize the wealth

of disease-miRNA, miRNA-lncRNA, and disease-lncRNA associations data collected from three datasets and recently published in experimental literature, in this article, the novel model of DCSLDA was developed to predict potential disease-lncRNA associations. We calculated distance correlation set of each node based on disease-miRNA-lncRNA interactive network first and then further integrated disease functional similarity and lncRNA functional similarity for DCSLDA. The important difference from previous computational model is that DCSLDA does not rely on any known disease-lncRNA associations and it predicts disease-lncRNA associations only based on disease-miRNA-lncRNA interactive network. In order to evaluate the prediction performance of DCSLDA, the validation frameworks of LOOCV were implemented based on known disease-lncRNA and cancer-related-lncRNA associations downloaded from LncRNADisease database. And case studies were further implemented to three important cancers (breast cancer, colorectal cancer, and lung cancer) based on recently published experimental literature. The simulation results show that DCSLDA can achieve reliable and excellent prediction performance and is superior to the state-of-the-art methods. Hence, it is anticipated that DCSLDA could play an important role in the prospective biomedical researches.

Disease functional similarity plays an important role in disease-related molecular function research. Functional associations between disease-related genes are often used to identify pairs of similar diseases from different perspectives. Calculating lncRNA functional similarity could benefit lncRNA function inference and disease-related lncRNA prioritization. Therefore, based on the two assumptions that (1) similar diseases tend to show a similar interaction and noninteraction pattern with the miRNAs and (2) similar lncRNAs tend to show a similar interaction and noninteraction pattern with the miRNAs, DCSLDA was developed to predict potential disease-related lncRNA by integrating lncRNA functional similarity and disease functional similarity. Simulation results indicated that the prediction performance of DCSLDA will be significantly improved by disease similarity and lncRNA similarity.

However, there are also some limitations in our method. Firstly, DCSLDA measures the correlations between lncRNAs and investigated diseases by integrating walks with different lengths in a lncRNA-miRNA-disease network, which is constructed by combining the known disease-miRNA network, miRNA-lncRNA network, and disease similarity network. The value of distance threshold parameters  $r$  is an important factor in DCSLDA, and how to select this parameter is not yet solved well. Secondly, although DCSLDA does not rely on any known experimentally verified lncRNA-disease relationships, the performance of DCSLDA was not very satisfactory compared with that of several existing methods. In the future, we will further integrate data of diseases and lncRNAs that do not rely on the lncRNA-disease interactive network, disease-miRNA interactive network, or miRNA-lncRNA interactive network; then these above problems may be well solved. Finally, introducing more reliable measure of disease similarity and lncRNA similarity and developing more reliable similarity integration method would improve

the performance of DCSLDA. In particular, disease similarity and lncRNA similarity in this model totally rely on known disease-miRNA and miRNA-lncRNA associations. The performance of DCSLDA would be further improved when sequence similarity of lncRNA and semantic similarity of disease are introduced.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The project is partly sponsored by the Natural Science Foundation of Hunan Province (No. 2018JJ4058, No. 2017JJ5036), the National Natural Science Foundation of China (No. 61640210, No. 61672447), and the CERNET Next Generation Internet Technology Innovation Project (No. NGII20160305).

## Supplementary Materials

Supplementary file 1: the known miRNA-disease associations for constructing the ASC1. We list 5430 known miRNA-disease associations which were collected from HMDD dataset to construct the ASC1. Supplementary file 2: the known lncRNA-miRNA associations for constructing the ASC2. We list 10195 known lncRNA-miRNA associations which were collected from starBase v2.0 database to construct the ASC2. Supplementary file 3: the known lncRNA-disease associations. We list 203 high-quality lncRNA-disease associations which were collected from LncRNADisease database to validate the performance of our method. Supplementary file 4: the top 0.5% results were listed to validate the performance of our method. (*Supplementary Materials*)

## References

- [1] Y. Okazaki, M. Furuno, T. Kasukawa et al., "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs," *Nature*, vol. 420, no. 6915, pp. 563–573, 2002.
- [2] F. H. C. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin, "General nature of the genetic code for proteins," *Nature*, vol. 192, no. 4809, pp. 1227–1232, 1961.
- [3] T. E. P. Consortium, "Identification and analysis of functional elements in 1% of the human genome by the encode pilot project," *Nature*, vol. 447, no. 7146, pp. 799–816, 2007.
- [4] F. F. Costa, "Non-coding RNAs: New players in eukaryotic biology," *Gene*, vol. 357, no. 2, pp. 83–94, 2005.
- [5] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.
- [6] A. Katarzyna, M. C. Brian, E. F. Plow, and S. A. Khalid, "Mir-31 and its host gene lncrna loc554202 are regulated by promoter hypermethylation in triple-negative breast cancer," *Molecular Cancer*, vol. 11, no. 1, p. 5, 2012.
- [7] X. He, X. Tan, X. Wang et al., "C-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion," *Tumor Biology*, vol. 35, no. 12, pp. 12181–12188, 2014.

- [8] Y. Yang, H. Li, S. Hou, B. Hu, J. Liu, and J. Wang, "The noncoding RNA expression profile and the effect of lncRNA AK126698 on cisplatin resistance in non-small-cell lung cancer cell," *PLoS ONE*, vol. 8, no. 5, Article ID e65309, 2013.
- [9] S. Uchida and S. Dimmeler, "Long noncoding RNAs in cardiovascular diseases," *Circulation Research*, vol. 116, no. 4, pp. 737–750, 2015.
- [10] R. Spizzo, M. I. Almeida, A. Colombatti, and G. A. Calin, "Long non-coding RNAs and cancer: a new frontier of translational research," *Oncogene*, vol. 31, no. 43, pp. 4577–4587, 2012.
- [11] J. Wang, R. Ma, W. Ma et al., "LncDisease: A sequence based bioinformatics tool for predicting lncRNA-disease associations," *Nucleic Acids Research*, vol. 44, no. 9, article no. e90, 2016.
- [12] M. E. Dinger, K. C. Pang, T. R. Mercer, M. L. Crowe, S. M. Grimmond, and J. S. Mattick, "NRED: A database of long noncoding RNA expression," *Nucleic Acids Research*, vol. 37, no. 1, pp. D122–D126, 2009.
- [13] D. Bu, K. Yu, S. Sun et al., "Noncode v3.0: integrative annotation of long noncoding rnas," *Nucleic Acids Research*, vol. 40, no. Database issue, pp. 210–215, 2012.
- [14] G. Chen, Z. Wang, D. Wang et al., "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, no. 1, pp. D983–D986, 2013.
- [15] X. Chen and G.-Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, 2013.
- [16] J. Sun, H. Shi, Z. Wang et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.
- [17] J. W. Li, G. Cheng, Y. C. Wang, M. Wei, T. Jian, J. P. Wang et al., "A bioinformatics method for predicting long noncoding rnas associated with vascular disease," *Science China Life Sciences*, vol. 57, no. 8, pp. 852–857, 2014.
- [18] T. Zhao, J. Xu, L. Liu et al., "Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features," *Molecular BioSystems*, vol. 11, no. 1, pp. 126–136, 2015.
- [19] 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," *Molecular BioSystems*, vol. 11, no. 3, pp. 760–769, 2015.
- [20] P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick, "LncRNADB: a reference database for long noncoding RNAs," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D146–D151, 2011.
- [21] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Scientific Reports*, vol. 5, Article ID 13186, 2015.
- [22] Y. Li, C. Qiu, J. Tu et al., "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, pp. D1070–D1074, 2014.
- [23] J. H. Li, S. Liu, H. Zhou, L. H. Qu, and J. H. Yang, "Starbase v2.0: decoding mirna-erna, mirna-ncrna and protein-rna interaction networks from large-scale clip-seq data," *Nucleic Acids Research*, vol. 42, no. Database issue, p. D92, 2014.
- [24] P. E. Spiess, J. Dhillon, A. S. Baumgarten, P. A. Johnstone, and A. R. Giuliano, "Pathophysiological basis of human papillomavirus in penile cancer: Key to prevention and delivery of more effective therapies," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 6, pp. 481–495, 2016.
- [25] M. K. Barton, "Local consolidative therapy may be beneficial in patients with oligometastatic non-small cell lung cancer," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 2, pp. 89–90, 2017.
- [26] M. Jin, P. Li, Q. Zhang, Z. Yang, and F. Shen, "A four-long non-coding rna signature in predicting breast cancer survival," *Journal of Experimental & Clinical Cancer Research*, vol. 33, no. 1, p. 84, 2014.
- [27] N. Xu, F. Wang, M. Lv, and L. Cheng, "Microarray expression profile analysis of long non-coding rnas in human breast cancer: a study of chinese women," *Sichuan Building Materials*, vol. 69, no. 3, pp. 221–227, 2010.
- [28] C. Lin, D. R. Crawford, S. Lin et al., "Inducible COX-2-dependent apoptosis in human ovarian cancer cells," *Carcinogenesis*, vol. 32, no. 1, pp. 19–26, 2011.
- [29] Z. Zhao, C. Chen, Y. Liu, and C. Wu, "17 $\beta$ -Estradiol treatment inhibits breast cell proliferation, migration and invasion by decreasing MALAT-1 RNA level," *Biochemical and Biophysical Research Communications*, vol. 445, no. 2, pp. 388–393, 2014.
- [30] Y.-S. Huang, C.-C. Chang, S.-S. Lee, Y.-S. Jou, and H.-M. Shih, "Xist reduction in breast cancer upregulates AKT phosphorylation via HDAC3-mediated repression of PHLPP1 expression," *Oncotarget*, vol. 7, no. 28, pp. 43256–43266, 2016.
- [31] H. Choudhry, A. Albukhari, M. Morotti et al., "Tumor hypoxia induces nuclear paraspeckle formation through HIF-2 $\alpha$  dependent transcriptional activation of NEAT1 leading to cancer cell survival," *Oncogene*, vol. 34, no. 34, pp. 4482–4490, 2015.
- [32] D. C. Chung, "The genetic basis of colorectal cancer: Insights into critical pathways of tumorigenesis," *Gastroenterology*, vol. 119, no. 3, pp. 854–865, 2000.
- [33] Y. Jia and M. Guo, "Epigenetic changes in colorectal cancer," *Chinese Journal of Cancer*, vol. 32, no. 1, pp. 21–30, 2013.
- [34] Y. Yang, L. Zhao, L. Lei, W. B. Lau, B. Lau, Q. Yang et al., "Lncrnas, the bridge linking rna and colorectal cancer," *Oncotarget*, vol. 8, no. 7, 2016.
- [35] K. Tanaka, G. Shiota, M. Meguro, K. Mitsuya, M. Oshimura, and H. Kawasaki, "Loss of imprinting of long QT intronic transcript 1 in colorectal cancer," *Oncology*, vol. 60, no. 3, pp. 268–273, 2001.
- [36] Q. Ji, L. Zhang, X. Liu et al., "Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex," *British Journal of Cancer*, vol. 111, no. 4, pp. 736–748, 2014.
- [37] S. Lassmann, R. Weis, F. Makowiec et al., "Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas," *Journal of Molecular Medicine*, vol. 85, no. 3, pp. 293–304, 2007.
- [38] T. Hensing, A. Chawla, R. Batra, and R. Salgia, "A personalized treatment for lung cancer: molecular pathways, targeted therapies, and genomic characterization," in *Systems Analysis of Human Multigene Disorders*, Springer, New York, USA, 2014.
- [39] W.-J. Gong, J.-Y. Yin, X.-P. Li et al., "Association of well-characterized lung cancer lncRNA polymorphisms with lung cancer susceptibility and platinum-based chemotherapy response," *Tumor Biology*, vol. 37, no. 6, pp. 8349–8358, 2016.
- [40] K. Tano, R. Mizuno, T. Okada et al., "MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes," *FEBS Letters*, vol. 584, no. 22, pp. 4575–4580, 2010.

- [41] G. Li, H. Zhang, X. Wan, X. Yang, C. Zhu, A. Wang et al., "Long noncoding rna plays a key role in metastasis and prognosis of hepatocellular carcinoma," *Biomed Research International*, vol. 5147, Article ID 780521, 2014.
- [42] J. Fang, C.-C. Sun, and C. Gong, "Long noncoding RNA XIST acts as an oncogene in non-small cell lung cancer by epigenetically repressing KLF2 expression," *Biochemical and Biophysical Research Communications*, vol. 478, no. 2, pp. 811–817, 2016.