A Novel Model for Predicting Associations between Diseases and LncRNA-miRNA Pairs Based on a Newly Constructed Bipartite Network

Shunxian Zhou,1,2 Zhanwei Xuan,2 Lei Wang,2 Pengyao Ping,2 and Tingrui Pei2

1College of Software and Communication Engineering, Xiangnan University, Chenzhou 423000, China
2College of Information Engineering, Xiangtan University, Xiangtan 411105, China

Correspondences should be addressed to Lei Wang; wanglei@xtu.edu.cn

Received 21 December 2017; Revised 16 March 2018; Accepted 26 March 2018; Published 6 May 2018

Motivation. Increasing studies have demonstrated that many human complex diseases are associated with not only microRNAs, but also long-non-coding RNAs (lncRNAs). LncRNAs and microRNA play significant roles in various biological processes. Therefore, developing effective computational models for predicting novel associations between diseases and lncRNA-miRNA pairs (LMPairs) will be beneficial to not only the understanding of disease mechanisms at lncRNA-miRNA level and the detection of disease biomarkers for disease diagnosis, treatment, prognosis, and prevention, but also the understanding of interactions between diseases and LMPairs at disease level. Results. It is well known that genes with similar functions are often associated with similar diseases. In this article, a novel model named PADLMP for predicting associations between diseases and LMPairs is proposed. In this model, a Disease-LncRNA-miRNA (DLM) tripartite network was designed firstly by integrating the lncRNA-disease association network and miRNA-disease association network; then we constructed the disease-LMPairs bipartite association network based on the DLM network and lncRNA-miRNA association network; finally, we predicted potential associations between diseases and LMPairs based on the newly constructed disease-LMPair network. Simulation results show that PADLMP can achieve AUCs of 0.9318 ± 0.0264, 0.9090 ± 0.0027 in the LOOCV, 2-fold, and 5-fold cross validation framework, respectively, which demonstrate the reliable prediction performance of PADLMP.

1. Introduction

MicroRNAs (miRNAs) are endogenous small and non-coding RNA molecules, which can regulate gene expression at the posttranscriptional level by combining the 3’ untranslated regions (UTRs) of target mRNAs (UTR) and lead the translation inhibited cleavage of the target mRNAs [1]. Moreover, long-non-coding RNAs (lncRNAs), as the biggest class of non-coding RNAs with length greater than 200 nt, can also regulate gene expression at different levels including transcriptional, posttranscriptional, and epigenetic regulation. Recently, increasing studies demonstrate that lncRNAs and miRNAs play a significant role in the cell proliferation and cell differentiation [2–5] and that the interactions between lncRNAs and microRNAs may have consequences for diseases, explain disease processes, and present opportunities for new therapies [6]. For example, Dey et al. proved that lncRNA H19 would give rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration [7]. Yao et al. discovered that knockdown of lncRNA XIST could exert tumor-suppressive functions in human glioblastoma stem cells by upregulating miR-152 [8]. Wang et al. demonstrated that silencing of lncRNA MALAT1 by miR-101 and miR-217 would inhibit proliferation, migration, and invasion of esophageal squamous cell carcinoma cells [9]. Zhang et al. presented that lncRNA ANRIL indicated a poor prognosis of gastric cancer and promoted tumor growth by epigenetically silencing of miR-99a/miR-449a [10]. You et al. found that miR-449a inhibited cell growth in lung cancer and regulated lncRNA NEAT1 [11]. Emmrich et al. discovered that lncRNAs MONC and MIR100HG would act as oncogenes in AMKL blasts [12]. Leung et al. found that miR-222 and miR-221 upregulated by Ang II were transcribed from a large transcript and knockdown of Lnc-Ang362 would
decrease expression of miR-221 and miR-222 and reduce cell proliferation [13]. Zhu et al. discovered that IncRNA H19 and H19-derived miRNA-675 were significantly downregulated in the metastatic prostate cancer cell line M22 compared with the non-meta-static prostate epithelial cell line [14]. Hirata et al. found that IncRNA MALAT1 was associated with miR-205 and promoted aggressive renal cell carcinoma [15]. Zhao and Ren demonstrated that TUG1 knockdown was significantly associated with decreased cell proliferation and promoted apoptosis of breast cancer cells through the regulation of miR-9 [16].

More and more researches have indicated that IncRNA-miRNA interactions are associated with the development of complex diseases, but until now, as far as we know, no prediction models have been proposed for large-scale forecasting of the associations between diseases and LMPairs. However, some prediction models have been reported to infer the associations between diseases and miRNA-miRNA pairs [17–21]. Moreover, there are researches showing that miRNA-miRNA pairs can work cooperatively to regulate an individual gene or cohort of genes that participate in similar processes [18, 22]. Inspired by these existing state-of-the-art methods and ideas for large-scale prediction of the associations between diseases and miRNA-miRNA pairs and based on the reasonable assumption that functionally similar LMPairs tend to be associated with similar diseases, in this paper, a new model named PADLMP is proposed to predict potential associations between diseases and LMPairs. To date, it is the first computational model used to predict disease-LMPairs associations. PADLMP can predict novel disease-LMPairs associations in a large scale by combining the known IncRNA-disease, miRNA-disease, and IncRNA-miRNA associations. To evaluate the prediction performance of the proposed model, evaluation frameworks of leave-one-out cross validation (LOOCV), 2-fold, and 5-fold cross validation were adopted based on the known disease-LMPairs. A series of comparison experiments were also implemented to evaluate the influence of the number of walks on prediction performance. As a result, PADLMP achieved its best performance when the number of walks was set as 2. Specifically, PADLMP achieved value of AUCs of 0.9318, 0.9090 and 0.8950 ± 0.0264, and 0.9090 ± 0.0900, respectively. The results of the prediction show that the PADLMP model is feasible and effective in predicting broad-scale disease-LMPairs associations by considering the topology information of the known disease-LMPairs dichotomous network.

## 2. Materials

### 2.1. IncRNA-Disease Associations

Known IncRNA-disease associations were downloaded from different databases such as the IncRNA-disease database IncRNADisease [23], MNDR [24], and Lnc2Cancer [25], respectively, and then, after preprocessing (getting rid of duplicate associations), 2048 distinct experimentally confirmed IncRNA-disease associations that including 1126 IncRNAs and 356 diseases were finally obtained (see Supplementary Table 1). Then we further constructed an adjacency matrix A1 of size $1126 \times 356$ as the information source.

### 2.2. miRNA-Disease Associations

We also downloaded known disease-miRNA associations from three different databases such as the miR2Disease [26], HMDD [27], and miR2Disease [28], respectively. And then, after preprocessing (getting rid of duplicate associations) and mapping these newly obtained miRNAs and diseases to databases of miRBase v21 [29] and Disease Ontology (DO) [30] separately, we finally obtained 4041 disease-miRNA associations including 438 miRNAs and 263 diseases from HMDD, 1839 disease-miRNA associations including 83 cancers and 327 miRNAs from miR2Disease, and 1487 disease-miRNA associations including 107 diseases and 276 miRNAs from miR2Disease (see Supplementary Table 2).

### 2.3. IncRNA-miRNA Associations

In this section, we downloaded two versions (2015 Version and 2017 Version) of IncRNA-miRNA association datasets from the starBase v2.0 database [31], which provided the most comprehensive experimentally confirmed IncRNA-miRNA interactions based on large-scale CLIP-Seq data. And then, after preprocessing (including elimination of duplicate values, erroneous data, and disorganized data), 20324 IncRNA-miRNA interactions including 494 miRNAs and 1127 IncRNAs were obtained finally (see Supplementary Table 3).

## 3. Methods

### 3.1. Methods Overview

In order to predict potential novel associations between diseases and LMPairs, a new model named PADLMP is proposed, which consists of three steps (Figure 1). First, the construction of association network and data integrate. Second, the similarities for lncRNAs, diseases, miRNAs, and IncRNA-miRNA pairs are calculated based on the association network. Finally, potential associations between disease and LMPairs are inferred.

### 3.2. Construct the Associated Network

#### 3.2.1. IncRNA-Disease Network, Disease-miRNA Network, and IncRNA-miRNA Network

Based on these newly obtained known IncRNA-disease associations, we constructed the IncRNA-disease bipartite network $G_1 = (V_1, E_1)$ according to the following steps.

**Step 1.** Let $V_{d1}$ be the set of newly collected 1126 IncRNAs, let $V_{d1}$ be the set of newly collected 356 diseases, and $V_1 = V_{d1} \cup V_{d1}$, then we can obtain the vertex set $V_1$ of $G_1$.

**Step 2.** $\forall l_i \in V_{d1}$, if there is $d_j \in V_{d1}$ satisfying the fact that the association between $l_i$ and $d_j$ belongs to the set of newly collected 2048 IncRNA-disease associations, then we define that there is an edge between $l_i$ and $d_j$ in $G_1$, and by this way, we can obtain the edge set $E_1$ of $G_1$. Obviously, $E_1$ is composed of these newly collected 2048 IncRNA-disease associations.
Similar to $G_1$, we constructed the disease-miRNA bipartite network $G_2 = (V_2, E_2)$ according to the following steps.

**Step 1.** Let $V_{m1}$ be the set of all these newly collected miRNAs, let $V_{d2}$ be the set of all these newly collected diseases, and $V_2 = V_{m1} \cup V_{d2}$, then we can obtain the vertex set $V_2$ of $G_2$.

**Step 2.** $\forall l_i \in V_{m1}$, if there is $d_j \in V_{d2}$ satisfying the fact that the association between $m_l$ and $d_j$ belongs to the set of all these newly collected disease-miRNA associations, then we define that there is an edge between $m_l$ and $d_j$ in $G_2$, and by this way, we can obtain the edge set $E_2$ of $G_2$. Obviously, $E_2$ is composed of all these newly collected disease-miRNA associations.

We also constructed the IncRNA-miRNA bipartite network $G_3 = (V_3, E_3)$ according to the following steps.

**Step 1.** Let $V_{12}$ be the set of newly collected 1127 IncRNAs, let $V_{m2}$ be the set of newly collected 494 miRNAs, and $V_3 = V_{m2} \cup V_{12}$, then we can obtain the vertex set $V_3$ of $G_3$.

**Step 2.** $\forall l_i \in V_{12}$, if there is $m_j \in V_{m2}$ satisfying the fact that the association between $l_i$ and $m_j$ belongs to the set of newly collected 18286 IncRNA-miRNA associations, then we define that there is an edge between $l_i$ and $m_j$ in $G_3$, and by this way, we can obtain the edge set $E_3$ of $G_3$. Obviously, $E_3$ is composed of these newly collected 20324 IncRNA-miRNA associations.

3.2.2. Disease-LncRNA-miRNA Network. Based on above newly constructed bipartite networks such as $G_1$, $G_2$, and $G_3$, we constructed a new tripartite network $G_4 = (V_4, E_4)$ according to the following steps.

**Step 1.** Let $V_{m3} = V_{m1} \cap V_{12}$, $V_{m'} = V_{m2} \cap V_{12}$, and $V_{d3} = V_{d1} \cap V_{d2}$. Let $d_k \in V_{d3}$, if there are $l_j \in V_{l'}$ and $m_k \in V_{m'}$ satisfying the fact that the association between $d_k$ and $l_j$ belongs to $E_1$, the association between $d_k$ and $m_k$ belongs to $E_2$, and the association between $l_j$ and $m_k$ belongs to $E_3$ simultaneously. Then we define that there is an edge between $d_k$ and $l_j$, an edge between $d_k$ and $m_k$, and an edge between $l_j$ and $m_k$ in $G_4$ separately, and by this way, we can obtain the edge set $E_4$ of $G_4$.

3.2.3. Disease-LMPairs Network. Based on above newly obtained tripartite Disease-LncRNA-miRNA network $G_4$, we constructed a new bipartite disease-LMPairs network $G = (V, E)$ according to the following steps.

**Step 1.** $\forall l_i \in V_l$ and $m_j \in V_m$, let $p_{ij} = (l_i, m_j)$ and $V_{p} = \{p_{ij}\}$ where $i \in [1, |V_l|]$ and $j \in [1, |V_m|]$, then we define $V = V_{d3} \cup V_{p}$, and by this way, we can obtain the vertex set $V$ of $G$.

**Step 2.** $\forall d_k \in V_{d3}$, there is $p_{ij} = (l_i, m_j) \in V_{p}$ satisfying the fact that the association between $d_k$ and $l_i$ belongs to $E_1$, the association between $d_k$ and $m_j$ belongs to $E_2$, and the association between $l_i$ and $m_j$ belongs to $E_3$ simultaneously. Then we define that there is an edge between $d_k$ and $p_{ij}$ in $G$, and by this way, we can obtain the edge set $E$ of $G$.

To make it easier to understand the construction of the network, we list in “The Meaning of Vertex and Edges in the Networks” each of the vertices, edges, and their meanings that appear in Sections 3.2.1, 3.2.2, and 3.2.3.
3.3. Calculation the Similarity of Disease

3.3.1. Calculation of the Disease Semantic Similarity (DisSemSim). Firstly, we downloaded MeSH descriptors from the National Library of Medicine and curated the names of diseases using the standard MeSH disease terms. Next, we represented the relationship of different diseases by a structure of directed acyclic graph (DAG) such as DAG(D) = (T(D), E(D)). Here, T(D) represented the node set including node D and its ancestor nodes, and E(D) denoted the edge set of corresponding direct links from a parent node to a child node, which represented the relationship between different diseases [32]. Then, based on the disease DAG, the contribution of an ancestor node to a child node, which represented the relationship between different diseases [32]. Based on these formulas presented above, we can finally define the semantic similarity between two diseases $d_i$ and $d_j$ as follows:

$$\text{DisSemSim}(d_i, d_j) = \sum_{t \in T(d_i) \cap T(d_j)} \left( D_{d_i} (t) + D_{d_j} (t) \right)$$

where $T(d_i)$ and $T(d_j)$ represented the node sets of the DAG of $d_i$ and $d_j$, respectively.

3.3.2. Calculation of the Gaussian Interaction Profile Kernel Similarity for Diseases (GIPSim). According to the assumption that functionally similar genes tend to be associated with similar diseases, we can integrate the topologic information of known miRNA-disease association network and lncRNA-disease association network to measure the disease similarity. Moreover, in this section, we will adopt Gaussian Interaction Profile Kernel to calculate the similarity of diseases. Firstly, based on the networks such as $G_1$ and $G_2$ constructed above, we can obtain two adjacency matrices such as $Y_1$ (or $Y_2$) as follows. For any given lncRNA $l_i$ (or miRNA $m_i$) and disease $d_j$, while $k$ takes 1 or 2, we define that

$$Y_k (i, j) = \begin{cases} 1 & \text{exist an edge between } l_i (m_i) \text{ and disease } d_j \text{ in } G_1 (G_2) \\ 0 & \text{otherwise.} \end{cases}$$

Hence, let $\text{IP}_k (d_j)$ denote the $i$th column of matrix $Y_k$, then we can calculate the Gaussian Kernel Similarity between the diseases $d_i$ and $d_j$ based on their interaction profiles as follows:

$$\text{GIP}_k (d_i, d_j) = \exp \left( -\gamma_k \left\| \text{IP}_k (d_i) - \text{IP}_k (d_j) \right\|^2 \right)$$

$$\gamma_k = \frac{1}{(1/n_k) \sum_{i=1}^{n_k} \left\| \text{IP}_k (d_i) \right\|^2}$$

where the parameter $n_k$ denotes the number of diseases in $G_k$ ($k = 1, 2$).

3.3.3. Calculation of the Integrated Similarity between Disease. Based on these formulas presented above, we can finally define the similarity measurement between diseases $d_i$ and $d_j$ as follows:

$$\text{DisSim}(d_i, d_j) = \begin{cases} \frac{\text{GIPSim}(d_i, d_j)}{2} + \text{DisSemSim}(d_i, d_j) & \text{if } \text{DisSemSim}(d_i, d_j) = 0 \\ \text{otherwise.} \end{cases}$$
3.4. Calculation of the Similarity between LncRNAs (miRNAs)

3.4.1. Calculation of the LncRNA (miRNA) Functional Similarity. For any given two LncRNAs (miRNAs) such as \( l_i(m_i) \) and \( l_j(m_j) \), let \( DT_1 = \{dt_{11}, dt_{12}, \ldots, dt_{1n}\} \) be all diseases related to \( l_i(m_i) \) in \( G_1(G_2) \) and let \( DT_2 = \{dt_{21}, dt_{22}, \ldots, dt_{2n}\} \) be all diseases related to \( l_j(m_j) \) in \( G_1(G_2) \), then we can define the functional similarity between \( l_i(m_i) \) and \( l_j(m_j) \) as follows (\( k = 1, \nu = I \) or \( k = 2, \nu = m \)):

\[
\text{FunSim}(v_i, v_j) = \sum_{1 \leq p \leq m_n} \text{SemSim}(dt_{1p}, DT_2) + \sum_{1 \leq p \leq m_n} \text{SemSim}(dt_{2p}, DT_1),
\]  

where

\[
\text{SemSim}(d_{1p}, DT_2) = \max_{1 \leq j \leq n} \text{DisSemSim}(d_{1p}, d_{2j}).
\]

3.4.2. Calculation of the Gaussian Interaction Profile Kernel Similarity for LncRNAs (miRNAs). For any given two LncRNAs (miRNAs) such as \( l_i(m_i) \) and \( l_j(m_j) \), in a similar way to the calculation of GIP, GIP2 can be obtained as follows (\( k = 1, \nu = I \) or \( k = 2, \nu = m \)):

\[
\text{GIP}_{LM_k}(v_i, v_j) = \exp \left( -y_k \frac{\| \text{IP}_k(v_i) - \text{IP}_k(v_j) \|^2}{(1/n_k) \sum_{i=1}^{n_k} \| \text{IP}_k(v_i) \|^2} \right),
\]

where \( \text{IP}_k(v_i) \) and \( \text{IP}_k(v_j) \) are the \( i \)th row and the \( j \)th row in matrix \( Y_k \), respectively, and \( n_k \) is the number of LncRNAs (miRNA) in \( G_k \).

LMPairSim\( (d_i, d_j) = \sqrt{(\text{IncSim}(l_i,l_a) - \text{AvgIncSim})^2 + (\text{miRSim}(m_j,m_b) - \text{AvgmiRSim})^2}, \)

where

\[
\text{AvgIncSim} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{IncSim}(l_i,l_j)}{n^2},
\]

\[
\text{AvgmiRSim} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{miRSim}(m_j,m_i)}{n^2}.
\]

3.4.3. Calculation of the Integrated Similarity between LncRNAs (miRNAs). Based on the formulas presented above, we can finally define the similarity measurement between LncRNAs \( l_i \) and \( l_j \) as follows:

\[
\text{IncSim}(l_i, l_j) = \frac{\text{FunSim}_1(l_i, l_j) + \text{GIP}_{LM_1}(l_i, l_j)}{2},
\]

\[
\text{miRSim}(m_i, m_j) = \frac{\text{FunSim}_2(m_i, m_j) + \text{GIP}_{LM_2}(m_i, m_j)}{2}.
\]

3.5. Similarity for LncRNA-miRNA Pairs (LMPairSim). Based on the bipartite disease-LMPairs network \( G \) constructed above, for any given two LncRNA-miRNA pairs \( p_{ij} = (l_i, m_j) \) and \( p_{ab} = (l_a, m_b) \), we can calculate the similarity between them according to the following three different ways:

1. Average Approach

\[
\text{LMPairSim}(P_{ij}, P_{ab}) = \frac{(\text{IncSim}(l_i, l_a) + \text{miRSim}(m_j, m_b))}{2}.
\]

2. Squared Root Approach

\[
\text{LMPairSim}(P_{ij}, P_{ab}) = (\text{IncSim}(l_i, l_a) \times \text{miRSim}(m_j, m_b))^{1/2}.
\]

3. Centre Distance Approach

\[
\text{LMPairSim}(d_i, d_j) = \sqrt{(\text{IncSim}(l_i,l_a) - \text{AvgIncSim})^2 + (\text{miRSim}(m_j,m_b) - \text{AvgmiRSim})^2},
\]

where \( \text{LncRNAs in } G, n_i \) denote the number of miRNAs in \( G \), respectively, then we can obtain a \( N_i \times N_j \) dimensional matrix \( \text{DisSim} \) and \( N_i \times N_j \) dimensional matrix \( \text{PairSim} \). Moreover, we can construct \( N_i \times N_j \) dimensional adjacency matrices \( \text{DP} \) as follows:

\[
\text{DP}(i, j) = \begin{cases} 1 & \text{exist an edge between } d_i \text{ and } p_j \text{ in } G \\ 0 & \text{otherwise} \end{cases}
\]

where \( d_i \) denotes the \( i \)th disease in \( G \) and \( p_j \) denotes the \( j \)th LMPair in \( G \).

Hence, inspired by the approach based on KATZHMDA [38] and KATZ [35], we can construct an integrated matrix
DP* for further predicting the potential associations between diseases and LMPairs as follows:

$$DP^* = \begin{bmatrix} PairSim & DP \\ DP^T & DisSim \end{bmatrix}. \quad (17)$$

Based on the integrated matrix $DP^*$ constructed above and letting $V_p = \{P_1, P_2, \ldots, P_{N_p}\}$, then, for any given lncRNA-miRNA pair $p_i \in V_p$ and diseases node $d_j \in V_d$, the probability of potential association between $p_i$ and $d_j$ can be obtained as follows:

$$S(i, j) = \sum_{l=1}^{K} \times DP^* (i, j), \quad (18)$$

where the parameter $K$ is an integer bigger than 1 and the parameter $\gamma$ satisfies $0 < \gamma < 1$.

Additionally, according to the above formula (18), it is obvious that the $(N_p + N_d) \times (N_p + N_d)$ dimensional matrix $S$ depicts the possibilities of all associations between diseases and LMPairs in $G$, and it can be further modified into the following form:

$$S = \sum_{l=1}^{K} \times DP^* (i, j) = (I - \gamma \times DP^*)^{-1} - I = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}. \quad (19)$$

where $S_{11}$ is $N_p \times N_p$ dimensional matrix, $S_{12}$ is $N_p \times N_d$ dimensional matrix, $S_{21}$ is $N_d \times N_p$ dimensional matrix, and $S_{22}$ is $N_d \times N_d$ dimensional matrix.

From formula (19), it is easily to know that $S_{12}$ is exactly the final prediction result matrix, which includes all of the potential associations between diseases and LMPairs in $G$. In addition, considering that a long walker in a sparse network may be less meaningful, it will disrupt association prediction, so we set $K$ to 2, 3, and 4 here. Then, final prediction result matrix could be represented by matrix DP, PairSim, and DisSim based on aforementioned equation (19).

While $K = 2$, there is

$$S_{122} = \gamma \times DP + \gamma^2 \times (PairSim \times DP + DP \times DisSim). \quad (20)$$

While $K = 3$, there is

$$S_{123} = S_{122} + \gamma^3 \times \left\{ DP \times DP^T \times DP + PairSim^2 \times DP \\
+ PairSim \times DP \times DisSim + DP \times DisSim^2 \right\}. \quad (21)$$

While $K = 4$, there is

$$S_{124} = S_{123} + \gamma^4 \times \left\{ PairSim^3 \times DP + DP \times DP^T \\
\times PairSim \times DP + PairSim \times DP \times DP^T \times DP \\
+ DP \times DisSim \times DP^T \times DP \right\} + \gamma^4 \times \left\{ DP \times DP^T \\
\times DP \times DisSim + PairSim^2 \times DP \times DisSim \\
+ PairSim \times DP \times DisSim^2 + DP \times DisSim^3 \right\}. \quad (22)$$

### 4. Results

In order to estimate the prediction performance of our newly proposed model PADLMP, the leave-one-out cross validation (LOOCV) procedure was adopted based on the positive samples of disease-LMPair associations. In the LOOCV validation framework, each known disease-LMPair association is used as a test sample, and the remaining disease-LMPairs association is used as a training sample for model learning. In particular, all the disease-LMPairs without known relevance proofs will be considered as candidate samples. In the LOOCV, we can obtain the rank of each left-out testing sample relative to candidate samples, and if the test samples are with a prediction level higher than a given threshold, then it will be considered to be successfully predicted. The corresponding true positive rates (TPR, sensitivity) and false positive rates (FPR, 1 – specificity) could be obtained by setting different thresholds. Here, sensitivity measures the percentage of test samples which are predicted with a higher rank than given threshold, specificity is calculated as the percentage of negative samples ranked below a given threshold. The receiver operating characteristics (ROC) curves can be drawn by plotting TPR versus FPR by different thresholds. In order to evaluate the predictive performance of PADLMP, the areas under the ROC curve (AUC) were further calculated. 1 of the AUC value showed a perfect prediction, while 0.5 of the AUC value represented purely random performance.

From the above, we can find that there are some parameters such as $K, \gamma$ adopted in our prediction model PADLMP. It is obvious that these parameters are critical to the prediction performance of our model. Moreover, in Section 3.5, three different ways have been proposed to calculate the similarity for lncRNA-miRNA pairs (LMPairSim), then we need to further evaluate the performances of these three different ways also. Hence, in this section, based on the validation framework of LOOCV, we implemented a series of comparison experiments to evaluate the influence of these parameters, and the simulation results were shown in Figure 2. As a result, from Figure 2, it is easy to see that PADLMP can achieve the best prediction performance while $K$ was set to 2. Additionally, as for other parameters $\gamma$, during simulations, we will set $\gamma$ as 0.01 based on the empirical values given by previous state-of-the-art works [37, 39–41]. Moreover, in the LOOCV, for the similarity calculation of LMPairSim, we use formulas (12), (13), and (14) in order and then select the formula that obtains the maximum AUC value. As a result, the AUC value of 0.9318, 0.9262, and 0.9247 were obtained when selecting formulas (12), (14), and (13), respectively.

Furthermore, we also compared the performance of our prediction model PADLMP with that of the RLSMDA [42], WBSMDA [39], and LRLSLDA [41] in LOOCV, since negative samples were not required in PADLMP, RLSMDA, WBSMDA, and LRLSLDA. The simulation results were shown in Figure 3. It is easy to see that PADLMP can achieve a reliable AUC of 0.9318, which is much higher than the AUC of 0.8104 and 0.9281 achieved by RLSMDA, WBSMDA, LRLSLDA, respectively. In addition, we can clearly see that the AUC value of the model LRLSLDA is less than 0.5, which
is obviously unreasonable. So based on prior knowledge [43], we subtract this value less than 0.5 from 1 and then we get the AUC value of LRSLDA being 0.5254.

Moreover, in order to further evaluate the prediction performance of PADLMP, the k-fold cross validation was also implemented, in which all the known disease-LMPair association samples were randomly equally divided into k parts, and k – 1 parts were then used as training samples for model learning while the rest part was used as testing samples for model evaluation. Specifically, in this section, considering time complexity and costs, we would only implement 2-fold and 5-fold cross validation to evaluate the prediction performance of PADLMP. In a similar way to that of LOOCV, all the disease-LMPairs without known relevance evidences would be considered as candidate samples in the k-fold cross validation. Next, in case of the prediction performance bias caused by random division of the testing samples, we would repeat the random division of the testing samples and our simulations for 100 times, and then, the corresponding ROC curves and AUCs would be obtained in a similar way to that of LOOCV. Simulation results were shown in Table 1, and as a result, from the Table 1, it is easy to see that PADLMP can achieve the best prediction performance with average AUCs of 0.9090 and 0.8950 with Standard Deviation (STD) of 0.0264 and 0.0027 in the 2-fold and 5-fold cross validation, respectively, while setting K = 2.

From the above descriptions, it is obvious that the newly proposed model PADLMP can achieve a reliable and effective prediction performance in both LOOCV and k-fold cross validation. Therefore, we released the potential disease-LMPair associations with higher predicted relevance scores publicly (see Supplementary Table 4) and anticipated that these disease-LMPair associations may offer valuable information and clues for corresponding biological experiments and would be confirmed by experimental observations in the future.

### 5. Case Studies

Colon cancer is a malignant tumor that is usually found at the borders of rectum and sigmoid colon [44]. This is the third most common cancer and the third leading cause of cancer death in men and women in the United States [45]. However, patients with early colon tumors only suffer from subtle symptoms [46], which make the disease difficult to be detected. In addition, worse, it is reported that its incidence has an upward trend in recent years [47]. Therefore, there is an urgent need to predict potential miRNAs and lncRNAs associated with colon tumors. With the help of modern medicine, many miRNAs have been shown to be associated with colon tumors. For example, miRNA-145 targets the insulin receptor substrate-1 and thus inhibits the growth of colon cancer cells [48].

Moreover, as the second largest cause of cancer deaths in women, breast cancer accounts for the total number of cancers in women 22% [49, 50]. Breast cancer is caused by a variety of molecular changes, traditionally diagnosed by histopathological features such as tumor size, grading, and lymph node status [49]. Studies have shown that lncRNAs and
Table 2: PADLMP was applied to three kinds of important cancer.

| Disease       | LncRNA | miRNA          | Evidence |
|---------------|--------|----------------|----------|
| Colon cancer  | MALAT1 | hsa-miR-145-5p | #, $, !  |
| Colon cancer  | MALAT1 | hsa-miR-181a-5p| #, $, +  |
| Colon cancer  | MALAT1 | hsa-miR-155-5p | #, $, !  |
| Colon cancer  | MALAT1 | hsa-miR-101-3p | #, $, !  |
| Colon cancer  | MALAT1 | hsa-miR-25-3p  | #, $, +  |
| Colon cancer  | MALAT1 | hsa-miR-143-3p | #, $, !  |
| Colon cancer  | MALAT1 | hsa-miR-200c-3p| #, $, !  |
| Colon cancer  | MALAT1 | hsa-miR-429    | #, $, +  |
| Colon cancer  | MALAT1 | hsa-miR-22-3p  | #, $, !  |
| Breast cancer | XIST   | hsa-let-7b-5p  | #, $, !  |
| Breast cancer | XIST   | hsa-let-7a-5p  | #, $, !  |
| Breast cancer | XIST   | hsa-miR-146a-5p| #, $, !  |
| Breast cancer | XIST   | hsa-miR-27a-3p | #, $, !  |
| Breast cancer | XIST   | hsa-miR-7c-3p  | #, $, !  |
| Breast cancer | XIST   | hsa-miR-181b-3p| #, $, !  |
| Breast cancer | XIST   | hsa-miR-181a-5p| #, $, !  |
| Breast cancer | XIST   | hsa-miR-34a-5p | #, $, !  |
| Breast cancer | XIST   | hsa-miR-25-3p  | #, $, !  |
| Breast cancer | XIST   | hsa-miR-30a-5p | #, $, !  |
| Prostate cancer| XIST | hsa-let-7b-5p | #, $, & |
| Prostate cancer| XIST | hsa-miR-146a-5p| $, $, & |
| Prostate cancer| XIST | hsa-miR-27a-3p| $, $, & |
| Prostate cancer| XIST | hsa-miR-181b-5p| $, $, & |
| Prostate cancer| XIST | hsa-miR-7a-5p | $, $, & |
| Prostate cancer| XIST | hsa-miR-30a-5p| $, $, & |
| Prostate cancer| XIST | hsa-miR-34a-5p| $, $, & |
| Prostate cancer| XIST | hsa-miR-155-5p| $, $, + |
| Prostate cancer| XIST | hsa-miR-124-3p| $, $, + |
| Prostate cancer| XIST | hsa-miR-181b-5p| $, $, & |
| Prostate cancer| XIST | hsa-miR-25-3p | $, $, & |

miRNAs play important role in many biological processes and are closely related to the formation of various cancers, including breast cancer [51, 52]. In order to better diagnose and treat breast cancer, it is necessary to predict breast cancer-related lncRNA or miRNAs and to identify lncRNA and miRNA biomarkers [52]. In addition, prostate cancer is a malignant tumor derived from prostate epithelial cells [53]. There are many factors, including age, family history of disease, and race, which may increase the risk of prostate neoplasms [54]. So far, many miRNAs and lncRNAs, such as miRNA has-let-7a-5p and lncRNA XIST in the prostate, have been found to be associated with prostate tumors.

As described previously, PADLMP has been demonstrated that it can achieve a reliable and effective prediction performance. Hence, in this section, case studies about above three kinds of important cancers based on top 5% of predicted results will be implemented to show the prediction performance of PADLMP. As illustrated in Table 2, the prediction results have been verified based on the recent updates in the databases such as lncRNA Disease, MNDR v2.0, starBase v2.0, HMDD, miR2Disease, and miRCancer.

In Table 2, “#” and “∗” stand for databases of lncRNA-disease and MNDR v2.0, respectively, which consist of known disease-lncRNA associations. “$” stands for starBase v2.0 database, which consists of known lncRNA-miRNA associations. “!,” “&,” and “+” stand for databases of HMDD, miR2Disease, and miRCancer, respectively, which consist of known disease-miRNA associations.

6. Discussion and Conclusion

Accumulating evidences show that the interaction of lncRNA-miRNAs is involved in the formation of many complex human diseases, such as breast cancer [16]; however, to our knowledge, there are no prediction models proposed for large scale forecasting the associations between diseases and LMPairs. Hence, based on the existing miRNA-disease associations, lncRNA-disease associations, lncRNA-miRNA interactions, and the assumption that genes with similar functions are often associated with similar diseases, we proposed a novel prediction model PADLMP to infer potential associations between diseases and LMPairs.
In this paper, we achieved the following contributions mainly: (1) we proposed the first computational model PADLMP for large-scale prediction of disease-LMPair associations, which can predict potential associations between diseases and lncRNA-miRNA pairs effectively. (2) We transformed the tripartite Disease-lncRNA-miRNA network into a bipartite disease-LMPair network, which greatly reduced the complexity of our prediction model. (3) Negative samples were not required in our prediction model.

However, although PADLMP is a powerful tool to infer novel associations between diseases and lncRNA-miRNA pairs, there are some limitations still existing in our method. For example, firstly, although we introduced semantic similarity for diseases and LMPairs, but the calculation of Gaussian Interaction Profile Kernel Similarity greatly relied on known disease-lncRNA associations, disease-miRNA associations, and disease-LMPairs associations. Therefore, it would cause inevitable bias towards those well-investigated diseases and LMPairs. Secondly, PADLMP could not be applied to unknown diseases and LMPairs, which were poorly investigated and had not any known associations. In the future, we will try to design new methods that do not rely on the topological information of disease-LMPair association network to solve these limitations.

### The Meaning of Vertex and Edges in the Networks

- **$G_1$:** LncRNA-disease bipartite network
- **$G_2$:** Disease-miRNA bipartite network
- **$G_3$:** LncRNA-miRNA bipartite network
- **$G_4$:** Disease-LncRNA-miRNA tripartite network
- **$G$:** Disease-LMPairs bipartite network
- **$V_l$:** LncRNA in lncRNA-disease association
- **$V_m$:** Disease in miRNA-disease association
- **$V_m'$:** miRNA in miRNA-disease association
- **$V_d$:** Disease in LncRNA-miRNA association
- **$V_d'$:** Disease in miRNA-disease association
- **$V_{m'}$:** miRNA in miRNA-disease association
- **$V_{l'}$:** lncRNA in lncRNA-disease association
- **$V_{m'}':$** LMPair
- **$E_{ij}$:** Edge of $G_1$, lncRNA $l_i$ associated with disease $d$, if $\langle l_i, d \rangle \in E_1$
- **$E_{ij}$:** Edge of $G_2$, miRNA $m_i$ associated with disease $d$, if $\langle m_i, d \rangle \in E_2$
- **$E_{ij}$:** Edge of $G_3$, lncRNA $l_i$ associated with miRNA $m_i$, if $\langle l_i, m_i \rangle \in E_3$
- **$E_{ij}$:** Edge of $G_4$, lncRNA $l_i$ associated with disease $d$, if $\langle l_i, d \rangle \in E_4$
- **$E_{ij}$:** Edge of $G_4$, miRNA $m_i$ associated with disease $d$, if $\langle m_i, d \rangle \in E_4$
- **$E_{ij}$:** Edge of $G_4$, miRNA $m_i$ associated with disease $d$, if $\langle m_i, d \rangle \in E_4$
- **$E_{ij}$:** Edge of $G_4$, disease $d$, if $\langle d, p_{jk} \rangle \in E_5$.

### Conflicts of Interest

The authors declare no conflicts of interest in this work.

### Acknowledgments

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

### Supplementary Materials

There are four supplementary tables in this manuscript. And among them, Supplementary Table 1 is a description of the lncRNA-disease associations, in which there are 2048 lncRNA-disease associations included, and the 1st column represents the lncRNAs and the 2nd column represents the diseases. Supplementary Table 2 is a description of the miRNA-disease associations, which contains the ID of diseases, the ID of miRNAs, and the associations between diseases and miRNAs. Supplementary Table 3 is a description of the lncRNA-miRNA associations, which contains 20343 lncRNA-miRNA associations and in which the first column represents the lncRNAs and the second column represents the miRNAs. Supplementary Table 4 is a description of the predictive results of associations between diseases and lncRNA-miRNA pairs while adopting PADLMP to execute prediction. (Supplementary Materials)

### References

[1] E. Berezikov and E. Al, “Approaches to microrna discovery,” Nature Genetics, vol. 38, p. 52, 2006.
[2] P. J. Batista and H. Y. Chang, “Long noncoding RNAs: cellular address codes in development and disease,” Cell, vol. 152, no. 6, pp. 1298–1307, 2013.
[3] K. C. Wang and H. Y. Chang, “Molecular mechanisms of long noncoding RNAs,” Molecular Cell, vol. 43, no. 6, pp. 904–914, 2011.
[4] C. P. Ponting, P. L. Oliver, and W. Reik, “Evolution and functions of long noncoding RNAs,” Cell, vol. 136, no. 4, pp. 629–641, 2009.
[5] J. Zheng, H. Peng, and L. Wang, “Similarities/dissimilarities analysis of protein sequences based on recurrence quantification analysis,” Current Bioinformatics, vol. 10, no. 1, pp. 112–119, 2015.
[6] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, “A ceRNA hypothesis: the rosetta stone of a hidden RNA language?” Cell, vol. 146, no. 3, pp. 353–358, 2011.
[7] B. K. Dey, K. Pfeifer, and A. Dutta, “The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration,” Genes & Development, vol. 28, no. 5, pp. 491–501, 2014.
S. Yoon and G. D. Micheli, Y. Wang, L. Chen, B. Chen, X. Li, J. Kang, K. Fan et al., "Computational and Mathematical Methods in Medicine

X. B. Zhao and G. S. Ren, "LncRNA promotes breast cancer metastasis via targeting TGFBI," FEBS Journal, vol. 281, no. 16, pp. 3766–3775, 2014.

M. Zhu, Q. Chen, X. Liu et al., "LncRNA H19/miR-675 axis represses prostate cancer metastasis by targeting TGFBI," FEBS Journal, vol. 281, no. 16, pp. 3766–3775, 2014.

H. Hirata, Y. Hinoda, V. Shahryari et al., "Long noncoding RNA MALAT1 promotes aggressive renal cell carcinoma through Ezh2 and interacts with miR-205," Cancer Research, vol. 75, no. 7, pp. 1322–1331, 2015.

X. B. Zhao and G. S. Ren, "Lncrna tgl promotes breast cancer cell proliferation via inhibiting mir-9," Cancer Biomarkers, pp. 1–8, 2016.

H. Peng, C. Lan, Y. Zheng, G. Huvtagnier, D. Tao, and J. Li, "Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite," BMC Bioinformatics, vol. 18, no. 1, article193, 2017.

X. Li, J. Xu, Y. Li et al., "Dissection of the potential characteristic of miRNA-miRNA functional synergistic regulations," Molecular Biosystems, vol. 9, no. 2, pp. 217–224, 2013.

X. Yun, C. Xu, J. Guan, Y. Ping, H. Fan, Y. Li et al., "Discovering dysfunction of multiple lncrnas cooperation in disease by a conserved lncrna co-expression network," PLoS ONE, vol. 7, no. 2, e32201, 2012.

S. Yoon and G. D. Micheli, Prediction of Regulatory Modules Comprising microRNAs and Target Genes, Oxford University Press, Oxford, UK, 2005.

B. Wu, C. Li, P. Zhang, Q. Yao, J. Wu, J. Han et al., "Dissection of mirna-mirna interaction in esophageal squamous cell carcinoma," PLoS ONE, vol. 8, no. 9, e73911, 2013.

X. Lai, U. Schmitz, S. K. Gupta et al., "Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs," Nucleic Acids Research, vol. 40, no. 18, pp. 8818–8834, 2012.

G. Chen, Z. Wang, D. Wang et al., "LncRNA Disease: a database for long-non-coding RNA-associated diseases," Nucleic Acids Research, vol. 41, no. 1, pp. D983–D986, 2013.

Y. Wang, L. Chen, B. Chen, X. Li, J. Kang, K. Fan et al., "Mammalian ncrna-disease repository: a global view of ncrna-mediated disease network," Cell Death and Disease, vol. 4, no. 8, p. e765, 2013.

S. Ning, J. Zhang, P. Wang et al., "Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers," Nucleic Acids Research, vol. 44, no. 1, pp. D980–D985, 2016.

Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," Nucleic Acids Research, vol. 37, no. 1, pp. D98–D104, 2009.

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.

B. Xie, Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA cancer association database constructed by text mining on literature," Bioinformatics, vol. 29, no. 5, pp. 638–644, 2013.

A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," Nucleic Acids Research, vol. 42, pp. D68–D73, 2014.

L. M. Schriml, C. Arze, S. Nadendla et al., "Disease ontology: a backbone for disease semantic integration," Nucleic Acids Research, vol. 40, no. 1, pp. D940–D946, 2012.

J. H. Li, S. Liu, H. Zhou, L. H. Qu, and J. H. Yang, "Starbase v2.0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data," Nucleic Acids Research, vol. 42, p. D92, 2014.

D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," Bioinformatics, vol. 26, no. 13, pp. 1644–1650, 2010.

X. Chen, Z. You, G. Yan, and D. Gong, "IRWRLDA: improved random walk with restart for IncRNA-disease association prediction," Oncotarget, vol. 7, no. 36, pp. 57919–57931, 2016.

Y. A. Huang, X. Chen, Z. H. You, D. S. Huang, and K. C. Chan, "ILNCSIM: improved IncRNA functional similarity calculation model," Oncotarget, vol. 7, no. 18, pp. 25902–25914, 2016.

L. Katz, "A new status index derived from sociometric analysis," Psychometrika, vol. 18, no. 1, pp. 39–43, 1953.

X. Yang, L. Gao, X. Guo, X. Shi, H. Wu, F. Song et al., "A network based method for analysis of lncRNA-disease associations and prediction of Incrnas implicated in diseases," PLoS ONE, vol. 9, no. 1, e87797, 2014.

X. Chen, "KATZLDA: KATZ measure for the IncRNA-disease association prediction," Scientific Reports, vol. 5, Article ID 16840, 2014.

X. Chen, Y. A. Huang, Z. H. You, G. Y. Yan, and X. S. Wang, "A novel approach based on Katz measure to predict associations of human microbiota with non-infectious diseases," Bioinformatics, vol. 33, no. 5, pp. 733–739, 2016.

X. Chen, C. C. Yan, X. Zhang et al., "WBSDMA: within and between score for MiRNA-disease association prediction," Scientific Reports, vol. 6, Article ID 21106, 2016.

X. Chen, Y. C. Clarence, X. Zhang, Z. H. You, Y. A. Huang, and G. Y. Yan, "HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction," Oncotarget, vol. 7, no. 40, pp. 65257–65269, 2016.

X. Chen and G.-Y. Yan, "Novel human IncRNA-disease association inference based on IncRNA expression profiles," Bioinformatics, vol. 29, no. 20, pp. 2617–2624, 2013.

X. Chen and G. Y. Yan, "Supervised learning for potential human microrna-disease associations inference," Scientific Reports, vol. 4, p. 5501, 2014.

T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[44] A. I. Phipps, N. M. Lindor, M. A. Jenkins et al., “Colon and rectal cancer survival by tumor location and microsatellite instability: The colon cancer family registry,” Diseases of the Colon & Rectum, vol. 56, no. 8, pp. 937–944, 2013.

[45] F. Liu, D. Yuan, Y. Wei, W. Wang, L. Yan, T. Wen et al., “Systematic review and meta-analysis of the relationship between ephx1 polymorphisms and colorectal cancer risk,” PLoS ONE, vol. 7, no. 8, p. e43821, 2012.

[46] S. Pita-Fernández, S. Pertega-Díaz, B. López-Calviño et al., “Diagnostic and treatment delay, quality of life and satisfaction with care in colorectal cancer patients: a study protocol,” Health and Quality of Life Outcomes, vol. 11, no. 1, article 117, 2013.

[47] V. H. Chong, M. S. Abdullah, P. U. Telisinghe, and A. Jalilhal, “Colorectal cancer: incidence and trend in Brunei Darussalam,” Singapore Medical Journal, vol. 50, no. 11, pp. 1085–1089, 2009.

[48] B. Shi, L. Sepp-Lorenzino, M. Prisco, P. Linsley, T. Deangelis, and R. Baserga, “Micro RNA 145 targets the insulin receptor substrate-1 and inhibits the growth of colon cancer cells,” The Journal of Biological Chemistry, vol. 282, no. 45, pp. 32582–32590, 2007.

[49] H. J. Donahue and D. C. Genetos, “Genomic approaches in breast cancer research,” Briefings in Functional Genomics, vol. 12, no. 5, pp. 391–396, 2013.

[50] K. Karagoz, R. Sinha, and K. Y. Arga, “Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways,” OMICS: A Journal of Integrative Biology, vol. 19, no. 2, pp. 115–130, 2015.

[51] J. Meng, P. Li, Q. Zhang, Z. Yang, and S. Fu, “A four-long non-coding rna signature in predicting breast cancer survival,” Journal of Experimental and Clinical Cancer Research, vol. 33, no. 1, article 84, 2014.

[52] 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,” Biomedicine Pharmacotherapy, vol. 69, no. 3, pp. 221–227, 2015.

[53] G. A. Gmyrek, M. Walburg, C. P. Webb et al., “Normal and malignant prostate epithelial cells differ in their response to hepatocyte growth factor/scatter factor,” The American Journal of Pathology, vol. 159, no. 2, pp. 579–590, 2001.

[54] P. C. Walsh and A. W. Partin, “Family history facilitates the early diagnosis of prostate carcinoma,” Cancer, vol. 80, no. 9, pp. 1871–1874, 1997.