frontiers in Genetics

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fgene.2020.00089

# Multiview Consensus Graph Learning for lncRNA–Disease Association Prediction

*Haojiang Tan[1], Quanmeng Sun[1], Guanghui Li[2], Qiu Xiao[3], Pingjian Ding[4], Jiawei Luo[5] and Cheng Liang[1]**

*[1] School of Information Science and Engineering, Shandong Normal University, Jinan, China, [2] School of Information Engineering, East China Jiaotong University, Nanchang, China, [3] College of Information Science and Engineering, Hunan Normal University, Changsha, China, [4] School of Computer Science, University of South China, Hengyang, China, [5] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China*

OPEN ACCESS

**Edited by:**
*Wen Zhang, Huazhong Agricultural University, China*

**Reviewed by:**
*Wei Lan, Guangxi University, China*
*Yang Li, The Ohio State University, United States*
*Jin-Xing Liu, Qufu Normal University, China*

***Correspondence:**
*Cheng Liang*
*alcs417@sdnu.edu.cn*

**Specialty section:**
*This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics*

**Received:** *15 December 2019*
**Accepted:** *27 January 2020*
**Published:** *21 February 2020*

**Citation:**
*Tan H, Sun Q, Li G, Xiao Q, Ding P, Luo J and Liang C (2020) Multiview Consensus Graph Learning for lncRNA–Disease Association Prediction. Front. Genet. 11:89. doi: 10.3389/fgene.2020.00089*

Long noncoding RNAs (lncRNAs) are a class of noncoding RNA molecules longer than 200 nucleotides. Recent studies have uncovered their functional roles in diverse cellular processes and tumorigenesis. Therefore, identifying novel disease-related lncRNAs might deepen our understanding of disease etiology. However, due to the relatively small number of verified associations between lncRNAs and diseases, it remains a challenging task to reliably and effectively predict the associated lncRNAs for given diseases. In this paper, we propose a novel multiview consensus graph learning method to infer potential disease-related lncRNAs. Specifically, we first construct a set of similarity matrices for lncRNAs and diseases by taking advantage of the known associations. We then iteratively learn a consensus graph from the multiple input matrices and simultaneously optimize the predicted association probability based on a multi-label learning framework. To convey the utility of our method, three state-of-the-art methods are compared with our method on three widely used datasets. The experiment results illustrate that our method could obtain the best prediction performance under different cross validation schemes. The case study analysis implemented for uterine cervical neoplasms further confirmed the utility of our method in identifying lncRNAs as potential prognostic biomarkers in practice.

**Keywords: lncRNA–disease association, multiple similarity matrices, consensus graph learning, multi-label learning, survival analysis**

## INTRODUCTION

With the completion of ENCODE project, researchers have found that only 2% of genes in the human genome encode proteins, while approximately 75% of the human genome is involved in the process of primary transcripts (Djebali et al., 2012; Li and Chang, 2014; Zhang et al., 2018b). The discovery of extensive transcription of large RNA transcripts which do not code for proteins, termed long noncoding RNAs (lncRNAs), provides a new perspective in understanding the centrality of RNA in gene regulation (Rinn and Chang, 2012). Evidences have shown that lncRNAs are key regulators for many cellular functions, including splicing, gene regulation, and hormone-like activity (Gao et al., 2019a; Mongelli

et al., 2019). Moreover, the dysregulation of lncRNAs has been proved to be closely related with various human diseases, such as types of cancer, neurological as well as cardiovascular diseases (Feng et al., 2018; Zhang et al., 2019b). Consequently, identifying potential disease-related lncRNAs is of great importance and might shed new light on the understanding of the pathogenesis of complex diseases.

As a powerful complementary tool for biological and clinical experiments, many computational approaches have been developed to effectively predict the lncRNA-disease associations (Zou et al., 2016; Chen et al., 2017; Zhang et al., 2018c; Gong et al., 2019; Yue et al., 2019). Under the assumption that similar diseases are more likely to be associated with functionally similar lncRNAs, Chen et al. proposed Laplacian regularized least squares for lncRNA-disease association in terms of a semi-supervised learning framework(Chen and Yan, 2013). Liu et al. combined the gene expression profiles, lncRNA expression profiles and disease-associated genes to infer the potential associated diseases for human lncRNAs globally (Liu et al., 2014). In addition to the aforementioned datasets, Chen also incorporated the Gaussian interaction profile kernel similarity into their model and adopted the KATZ measure for lncRNA–disease association (Chen, 2015). Zhou et al. first constructed a heterogeneous network in terms of three sub-networks and then ranked the relevant lncRNAs for a given disease by applying the random walk with restart on the constructed network (Zhou et al., 2015). Chen et al. further improved the random walk with restart framework by initializing the probability vector according to the integration of lncRNA expression similarity and disease semantic similarity (Chen et al., 2016). Fu et al. decomposed the data matrices of heterogeneous data sources into low-rank matrices *via* matrix tri-factorization to explore the intrinsic as well as the shared structure, and then used the optimized low-rank matrices to obtain the potential associations (Fu et al., 2018). Lu et al. extracted a set of primary feature vectors and used the inductive matrix completion framework to infer the lncRNA-disease association (Lu et al., 2018). Lan et al. constructed a web server for lncRNA-disease association prediction by integrating multiple biological data resources (Lan et al., 2017). Xiao et al. obtained the association probability for a given lncRNA-disease association according to the lengths of the paths linking them in the constructed heterogeneous network (Xiao et al., 2018). Hu et al. adopted the bi-random walk algorithm to construct a linear model for the lncRNA-disease association prediction (Hu et al., 2019). Yu et al. applied a collaborative filtering model together with the Naive Bayesian Classifier on a constructed lncRNA-miRNA-disease tripartite network to effectively predict novel lncRNA-disease associations (Yu et al., 2019). Both Xie et al. and Chen et al. first fused different similarity matrices for lncRNAs and diseases based on a similarity kernel fusion model and then applied different classification frameworks to predict potential associations (Chen et al., 2019; Xie et al., 2019). Cui et al. developed a novel computational framework based on bipartite local model with nearest profile-based association inferring for prediction (Cui et al., 2019). Recently, Guo et al. employed the autoencoder to obtain the optimal feature space from the original feature set which was constructed from different types of similarities (Guo et al., 2019). The newly constructed features were then fed to a rotating forest to classify the lncRNA–disease associations and achieved remarkable performance.

Although the methods mentioned above have made great contributions to discover potential disease-related lncRNAs, the prediction accuracy is still limited in several ways. For example, in spite of the multiple biological data sources used in existing methods, the integration of the similarity matrices constructed from these data sources was simply performed by averaging them, which might be suboptimal. Furthermore, since the lncRNA–disease association data was relatively sparse, how to fully take advantage of the existing information during the prediction process remains challenging. To solve these issues, we here propose a multiview consensus graph learning method for disease-related lncRNAs prediction. Concretely, a set of similarity matrices for lncRNAs and diseases are first constructed by leveraging the known lncRNA–disease associations, respectively. We then iteratively learn a consensus graph from the multiple similarity matrices and obtain the final association probabilities between lncRNAs and diseases using a multi-label learning framework. To confirm the utility of our method, we compare the proposed method with several state-of-the-art methods on three widely used datasets under different evaluation metrics. The experimental results of various cross validation schemes clearly indicate that our method could achieve better prediction performance compared to the other three methods. Furthermore, we illustrate the potential of our method in identifying prognostic biomarkers for uterine cervical neoplasms in a case-study analysis.

# MATERIALS AND METHODS

## Human lncRNA–Disease Associations

The LncRNADisease database is used as the data of known lncRNA–disease associations (Chen et al., 2013; Bao et al., 2019). We used three versions of LncRNADisease, June-2012 Version (marked as Dataset1), January-2014 Version (marked as Dataset2), and June-2015 Version (marked as Dataset3) in our experiments (Li et al., 2019). After filtering the lncRNA-disease associations with irregular disease names or lncRNA names and merging duplicate items, we obtained 276 interactions between 150 diseases and 112 lncRNAs for dataset1, 319 interactions between 169 diseases and 131 lncRNAs for dataset2, and 621 interactions between 226 diseases and 285 lncRNAs for dataset3, respectively (**Table 1**). For convenience, we use $Y \in \mathbb{R}^{p \times q}$ to represent the known lncRNA–disease association matrix, where $p$ and $q$ denote the number of lncRNAs and diseases, respectively. If disease $j$ has an association with lncRNA $i$, then $Y_{ij} = 1$, otherwise $Y_{ij} = 0$.

**TABLE 1** | Details of the three datasets used in this study.

| Dataset | lncRNAs# | diseases# | interactions# |
|---|---|---|---|
| Dataset1 | 112 | 150 | 276 |
| Dataset2 | 131 | 169 | 319 |
| Dataset3 | 285 | 226 | 621 |

## Disease Semantic Similarity

To calculate the disease semantic similarity, we followed the same approach as described in previous work (Wang et al., 2010). Specifically, each disease $d$ can be described by a Directed Acyclic Graphs (*DAG*s) that consists of three items $DAG = (d, T(d), E(d))$, where $T(d)$ and $E(d)$ are all the parent nodes of $d$ including itself and all links from the ancestor nodes to child nodes, respectively. The contribution of disease $t$ to the semantic value of disease $d$ is defined as:

$$\begin{cases} D_d(t) = 1 & \text{if } t = d \\ D_d(t) = \max\{0.5 * D_d(t') | t' \in \text{children of } t\} & \text{if } t \neq d \end{cases} \quad (1)$$

The overall semantic value of a given disease $d$ can then be calculated as:

$$D(d) = \sum_{t \in T(d)} D_d(t) \quad (2)$$

As a result, given a pair of diseases $i$ and $j$, their semantic similarity is defined as:

$$S(i,j) = \frac{\sum_{t \in T(i) \cap T(j)} (D_i(t) + D_j(t))}{\sum_{t \in T(i)} D_i(t) + \sum_{t \in T(j)} D_j(t)} \quad (3)$$

We use $AD^{(1)} \in \mathbb{R}^{q \times q}$ to denote the obtained disease semantic similarity matrix and $AD^{(1)}_{ij}$ stands for the semantic similarity for a disease pair $i$ and $j$.

## lncRNA Functional Similarity

Similarly, the lncRNA functional similarity was also calculated according to previous studies (Wang et al., 2010; Liang et al., 2019). For each lncRNA pair, we measured their similarity as follows:

$$LFS(i,j) = \frac{\sum_{d \in D(l_j)} S(d, D(l_i)) + \sum_{d \in D(l_i)} S(d, D(l_j))}{m + n} \quad (4)$$

$$S(d, D(l_i)) = \max_{d_1 \in D(l_i)} (S(d, d_1)) \quad (5)$$

where $m$ and $n$ are the number of diseases related to lncRNA $l_i$ and $l_j$, and D($l$) represents the disease set related to lncRNA $l$. We use $AL^{(1)} \in \mathbb{R}^{p \times p}$ to denote the obtained lncRNA functional similarity matrix and $AL^{(1)}_{ij}$ stands for the functional similarity for a pair of lncRNAs $i$ and $j$.

## Gaussian Interaction Profile Kernel Similarity

Gaussian interaction profile kernel similarity is widely used in various semi-supervised prediction tasks for measuring similarities (Zou et al., 2016; Zhang et al., 2017; Zhu et al., 2018; Pan et al., 2019; Yin et al., 2019). Here we also adopted this similarity measure to construct the similarity matrices for lncRNAs and diseases, respectively. Concretely, given two lncRNAs $l_i$ and $l_j$, their Gaussian interaction profile kernel similarity is defined as:

$$KL(l_i, l_j) = \exp(-\beta_l \| IP(l_i) - IP(l_j) \|^2) \quad (6)$$

$$\beta_l = \beta'_l / \left( \frac{1}{p} \sum_{i=1}^{p} \|IP(l_i)\|^2 \right) \quad (7)$$

where $IP(l_i)$ is in essence the $i$-th row of matrix $Y$, $\beta'_l$ is a parameter controlling the kernel bandwidth and $p$ is the number of lncRNAs. Similarly, for a pair of diseases $d_i$ and $d_j$, we have:

$$KD(d_i, d_j) = \exp(-\beta_d \left\| IP(d_i) - IP(d_j) \right\|^2) \quad (8)$$

$$\beta_d = \beta'_d / \left( \frac{1}{q} \sum_{i=1}^{q} \|IP(d_i)\|^2 \right) \quad (9)$$

where $IP(d_i)$ is in essence the $i$-th column of matrix $Y$, $\beta'_d$ controls the kernel bandwidth and $q$ is the number of diseases. Finally, we use $AD^{(2)} \in \mathbb{R}^{q \times q}$ and $AL^{(2)} \in \mathbb{R}^{p \times p}$ to denote the kernel similarity matrices for diseases and lncRNAs, respectively.

## Cosine Similarity

Cosine similarity is another effective method for measuring similarities and is widely used in recommender systems (Gao et al., 2019b; Zhang et al., 2019a). Therefore, we also adopted cosine similarity to build the similarity matrices for lncRNAs and diseases. The cosine similarity for a pair of lncRNAs or diseases is calculated as:

$$CL(l_i, l_j) = \frac{IP(l_i) \cdot IP(l_j)}{\|IP(l_i)\| \times \left\|IP(l_j)\right\|} \quad (10)$$

$$CD(d_i, d_j) = \frac{IP(d_i) \cdot IP(d_j)}{\|IP(d_i)\| \times \left\|IP(d_j)\right\|} \quad (11)$$

where the definition of $IP(\cdot)$is the same as that in the previous section. As a result, we use $AD^{(3)} \in \mathbb{R}^{q \times q}$ and $AL^{(3)} \in \mathbb{R}^{p \times p}$ to record the cosine similarities for disease pairs and lncRNA pairs, respectively.

# METHODS

## Notations

We first briefly introduce the notations used throughout the paper. All the matrices are denoted by italic uppercase letters while vectors are expressed in bold lowercase letters. The transpose, the trace and the Frobenius norm of a given matrix $M$ are denoted by $M^T$, $Tr(M)$ and $||M||_F$, respectively. $M_{ij}$ represents the element at the $i$-th row and $j$-th column of $M$. $\mathbf{1}$ is a column vector with all elements equal to 1. For a given similarity matrix $S$, its degree matrix $D_S$ is a diagonal matrix whose main diagonal entry is $\sum_j (S_{ij} + S_{ji})/2$, and its Laplacian matrix $L_S$ is defined as $L_S = D_S - (S^T + S)/2$.

## Multiview Consensus Graph Learning for LncRNA–Disease Association Prediction

Given a set of similarity matrices for both lncRNAs and diseases, our aim is to find an optimal consensus graph based on these similarity matrices for subsequent prediction. Specifically, suppose we have $n$ similarity matrices $AD^{(1)}, AD^{(2)}, \ldots, AD^{(n)} \in \mathbb{R}^{q \times q}$ constructed for diseases, and $m$ similarity matrices $AL^{(1)}$,

$AL^{(2)},\ldots,AL^{(m)} \in \mathbb{R}^{p \times p}$ for lncRNAs, we propose to learn a consensus graph for the disease space and lncRNA space from multiple views by the following objective function respectively (Han et al., 2018; Wang et al., 2020):

$$\min_{SD,w_D^{(v)},F} \left\| SD - \sum_{v=1}^{n} w_D^{(v)} AD^{(v)} \right\|_F^2 + 2\alpha Tr(FL_{SD}F^T), \quad s.t. \sum_{j=1}^{q} SD_{ij} = 1, SD_{ij} \geq 0, \sum_{v=1}^{n} w_D^{(v)} = 1, w_D^{(v)} \geq 0 \tag{12}$$

$$\min_{SL,w_L^{(u)},F} \left\| SL - \sum_{u=1}^{m} w_L^{(u)} AL^{(u)} \right\|_F^2 + 2\beta Tr(F^T L_{SL}F), \quad s.t. \sum_{j=1}^{p} SL_{ij} = 1, SL_{ij} \geq 0, \sum_{u=1}^{m} w_L^{(u)} = 1, w_L^{(u)} \geq 0 \tag{13}$$

The weight parameters $w_L = [w_L^{(1)},\ldots,w_L^{(m)}]^T$ and $w_D = [w_D^{(1)},\ldots,w_D^{(n)}]^T$ added for each view guarantee that the objective functions in Eq. (12) and (13) adaptively learn an optimal consensus graph in terms of the importance of each view (Liu et al., 2018b). Finally, we integrate the optimization process from two spaces into one framework with graph-based multi-label learning and obtain the final objective function as follows:

$$\begin{aligned} &\min_{SD,w_D^{(v)},SL,w_L^{(u)},F} \left\| SD - \sum_{v=1}^{n} w_D^{(v)} AD^{(v)} \right\|_F^2 + 2\alpha Tr(FL_{SD}F^T) + \\ &\left\| SL - \sum_{u=1}^{m} w_L^{(u)} AL^{(u)} \right\|_F^2 + 2\beta Tr(F^T L_{SL}F) + \|F - Y\|_F^2, \\ &s.t. \sum_{j=1}^{q} SD_{ij} = 1, SD_{ij} \geq 0, \sum_{j=1}^{p} SL_{ij} = 1, SL_{ij} \geq 0, \sum_{v=1}^{n} w_D^{(v)} = 1, \\ &w_D^{(v)} \geq 0, \sum_{u=1}^{m} w_L^{(u)} = 1, w_L^{(u)} \geq 0, F \in \mathbb{R}^{p \times q} \end{aligned} \tag{14}$$

where $L_{SD}$ and $L_{SL}$ are the Laplacian matrices for the similarity matrices $SD$ and $SL$, $SD_{ij}$ and $SL_{ij}$ denote the $(i,j)$-th elements in $SD$ and $SL$, respectively. The constraints imposed on both $SD$ and $SL$ ensures that the learned similarities have explicit meanings. $Y$ is the known binary lncRNA-disease association matrix defined above. Specifically, the objective proposed in Eq. (14) has two advantages in predicting lncRNA–disease associations. First of all, it incorporates multiple data resources to learn a reliable similarity matrix and could be well adapted to arbitrary number of input similarity matrices. Moreover, the predicted label matrix $F$ and the learned consensus graph can collaboratively guide the learning process of each other and thus lead to better results (Zhang et al., 2018a). We propose an efficient method to solve Eq. (14) in the following subsection.

## Optimization

In this section, we derive an efficient algorithm to solve the objective function in Eq. (14) in an iterative manner.

i) Updating $SD$ and $SL$. For clarity, we only give the derivation for solving $SD$ and the optimization for $SL$ can be obtained similarly. By fixing the other variables in the objective function, Eq. (14) degenerates to Eq. (12). It can be rewritten in the following form:

$$\min_{SD} \sum_{i,j=1}^{q} \left\| SD_{ij} - \sum_{v=1}^{n} w_D^{(v)} AD_{ij}^{(v)} \right\|_F^2 + \alpha \sum_{i,j=1}^{q} \left\| F_i - F_j \right\|_2^2 SD_{ij}, \quad s.t. \sum_{j=1}^{q} SD_{ij} = 1, 0 \leq SD_{ij} \leq 1 \tag{15}$$

Since different rows of $SD$ are independent, we can then solve each row separately:

$$\min_{SD} \sum_{j=1}^{q} \left\| SD_{ij} - \sum_{v=1}^{n} w_D^{(v)} AD_{ij}^{(v)} \right\|_F^2 + \alpha \sum_{j=1}^{q} \left\| F_i - F_j \right\|_2^2 SD_{ij}, \quad s.t. \sum_{j=1}^{q} SD_{ij} = 1, 0 \leq SD_{ij} \leq 1 \tag{16}$$

Denoting $h_i$ as a vector whose $j$-th element is $h_{ij} = \| F_i - F_j \|_2^2$, Eq. (16) can then be converted to:

$$\min_{SD_i} \left\| SD_i + \left( \frac{\alpha}{2} h_i - \sum_{v=1}^{n} w_D^{(v)} AD_i^{(v)} \right) \right\|_2^2, \quad s.t. \; SD_i \mathbf{1} = 1, 0 \leq SD_{ij} \leq 1 \tag{17}$$

Eq. (17) could be solved by an efficient iterative algorithm proposed in (Huang et al., 2015).

ii) Updating $w_D$ and $w_L$. When $SD$, $SL$, $F$ and $w_L$ are fixed, Eq. (14) becomes:

$$\min_{w_D} \left\| SD - \sum_{v=1}^{n} w_D^{(v)} AD^{(v)} \right\|_F^2, \quad s.t. \; w_D^{(v)} \geq 0, \sum_{v=1}^{n} w_D^{(v)} = 1 \tag{18}$$

To solve Eq. (18), we first convert the target graph $SD$ into a column vector $a \in \mathbb{R}^{q^2 \times 1}$ by stacking its columns together. Similarly, we convert the multiple input similarity matrices $AD^{(v)}(v = 1,2\ldots,n)$ into a set of vectors $G^{(1)}, G^{(2)}, \ldots, G^{(n)} \in \mathbb{R}^{q^2 \times 1}$ and denote a matrix $G$ as $G = [G^{(1)}, G^{(2)}, \ldots, G^{(n)}] \in \mathbb{R}^{q^2 \times n}$. Then Eq. (18) can be transformed into:

$$\min_{w_D} \|a - Gw_D\|_2^2, \quad s.t. \; w_D^{(v)} \geq 0, \sum_{v=1}^{n} w_D^{(v)} = 1 \tag{19}$$

Eq. (19) can also be solved by the algorithm proposed in (Huang et al., 2015; Liu et al., 2018a). The optimization for $w_L$ could be derived in a similar way.

iii) Update $F$. By fixing the other variables, Eq. (14) is reduced to the following problem:

$$\min_F \quad 2\alpha Tr(FL_{SD}F^T) + 2\beta Tr(F^T L_{SL}F) + \|F - Y\|_F^2, \quad s.t. \; F \in \mathbb{R}^{p \times q} \tag{20}$$

Taking the derivative of Eq. (20) with respect to $F$ and setting it to zero, we have:

$$(2\beta L_{SL} + I)F + 2\alpha FL_{SD} = Y \tag{21}$$

Eq. (21) could be solved easily as a Sylvester equation (Zha et al., 2009; Shi et al., 2018).

The whole optimization process is summarized in Algorithm 1 and **Figure 1** illustrates the overall workflow of our method. Moreover, the source code of our method can be freely downloaded at: https://github.com/hjtan516/MCGLLDA.

**Algorithm 1**.

**Input:** Known association matrix $Y \in \mathbb{R}^{p \times q}$, lncRNA similarity matrices $\{AL^{(1)}, AL^{(2)},\ldots,AL^{(m)}\}$ from $m$ views, disease similarity matrices $\{AD^{(1)},AD^{(2)},\ldots,AD^{(n)}\}$ from $n$ views, parameters $\alpha$ and $\beta$.
**Output:** Final association matrix $F$.

1. For each view of lncRNAs and diseases, initialize the weights as $w_D^{(v)} = 1/n, w_L^{(u)} = 1/m$;
2. While not converge do
3. While not converge do
4. Update $SD$ according to Eq. (12);
5. Update $SL$ according to Eq. (13);
6. Update $F$ according to Eq. (21);
7. end while
8 Update $w_D^{(v)}, w_L^{(u)}$ according to Eq. (19);
9. end while
10. return $F$

# RESULTS

## Performance Evaluation

In this section, we compared the proposed method with three state-of-the-art methods i.e. BiwalkLDA (Hu et al., 2019), SIMCLDA (Lu et al., 2018) and KATZLDA (Chen, 2015) on the aforementioned three datasets. Firstly, two evaluation metrics Leave-One-Out Cross Valuation (LOOCV) and five-fold Cross Validation (CV) were conducted to systematically evaluate the prediction performance of each method. Both LOOCV and five-fold CV take part of the known lncRNA–disease associations as test samples and use the remaining as the training samples. However, LOOCV only takes one association at a time as the test sample while in five-fold CV all the known associations are randomly divided into five parts and one part was used as the test set each time. The Receiver Operating Characteristic (ROC) Curve was plotted in terms of the cross validation results and the Area Under the ROC Curve (AUC) was calculated to measure the prediction accuracy. As shown in **Figures 2** and **3**, our method reached the highest AUCs on all three datasets in both LOOCV and five-fold CV.

Next, we adopted Leave-One-Disease-Out Cross Validation (LODOCV) to test the ability of all methods in predicting the potential related lncRNAs for diseases without known associations. Specifically, for each disease, we removed all its associated lncRNAs and made predictions by leveraging the information from other diseases and lncRNAs. As a result, we could obtain a list of AUC values for each method and we used density plots to demonstrate the comparison results. As shown in **Figure 4**, compared with the other methods, our method obtained the highest numbers of AUC values greater than 0.9 on all three datasets. The Wilcoxon signed rank test also validated the significance of our method over the other three methods in terms of LODOCV (**Table 2**). In summary, these results clearly indicated that our method outperformed the other three methods in predicting reliable lncRNA–disease associations.

## Parameter Analysis

In Eq. (14), we used two parameters $\alpha$ and $\beta$ to balance the importance between the similarity graph learning and the predicted association matrix learning. We investigated the impacts of $\alpha$ and $\beta$ on the prediction performance of our method. Specifically, $\alpha$ was tested in the range from 0.0001 to 1 and $\beta$ was tested from 0.0001 to 10. To determine the best combination of $\alpha$ and $\beta$, five-fold cross validation was carried out on Dataset3. As a result, when both $\alpha$ and $\beta$ were set to 0.0001, our method achieved the best performance (**Figure 5**).

## Convergence Analysis

We also studied the practical convergence speed of our method. Specifically, **Figure 6** illustrated the value variations of Eq. (14) with the number of iterations on Dataset3. As can be seen from the figure, the objective function value of Eq. (14) became stable in 5 iterations, indicating that our method converges rapidly and can be used in practice.

## Case Study

To demonstrate the potential of our method in identifying lncRNAs as meaningful biomarkers for a given disease, we carried out a case-study analysis on Uterine CErvical Neoplasms (UCEC). Uterine Cervical Neoplasms is one of the most frequent causes of death in women and its early detection can significantly decrease its death rate (Jeong et al., 2003). To make reliable predictions, we applied our method on a newer version (July-2017) of lncRNA–disease associations from LncRNADisease database. In particular, associations with lncRNAs that were not recorded in BioMart and diseases that were not included in the MeSH Category C for diseases were excluded during the implementation. The predicted associations were then validated by another two widely used databases recording disease-related lncRNAs, *i.e.* Lnc2Cancer (Ning et al., 2016) and MNDR (Wang et al., 2013). As expected, the two databases confirmed that 9 out of the top 10 predicted lncRNAs were verified to be related with UCEC (**Table 3**). The only unconfirmed lncRNA is MIR7-3HG. To evaluate whether this lncRNA might be involved in UCEC, we further downloaded the lncRNA expression profile of 316 UCEC samples from TANRIC (Li et al., 2015) and performed the Kaplan–Meier survival analysis by using MIR7-3HG as the biomarker accordingly (**Figure 7**). The statistical significance in the survival analysis was calculated using the log rank test (Bewick et al., 2004). Notably, the results demonstrated that the higher expression level of MIR7-3HG was related with significantly decreased survival rates of UCEC patients, indicating that MIR7-3HG might play an important role in the pathogenesis of UCEC.

# CONCLUSION

Increasing evidences have shown that lncRNAs accomplish a remarkable variety of biological functions and thus the aberrant expression or dysfunction of lncRNA might lead to various

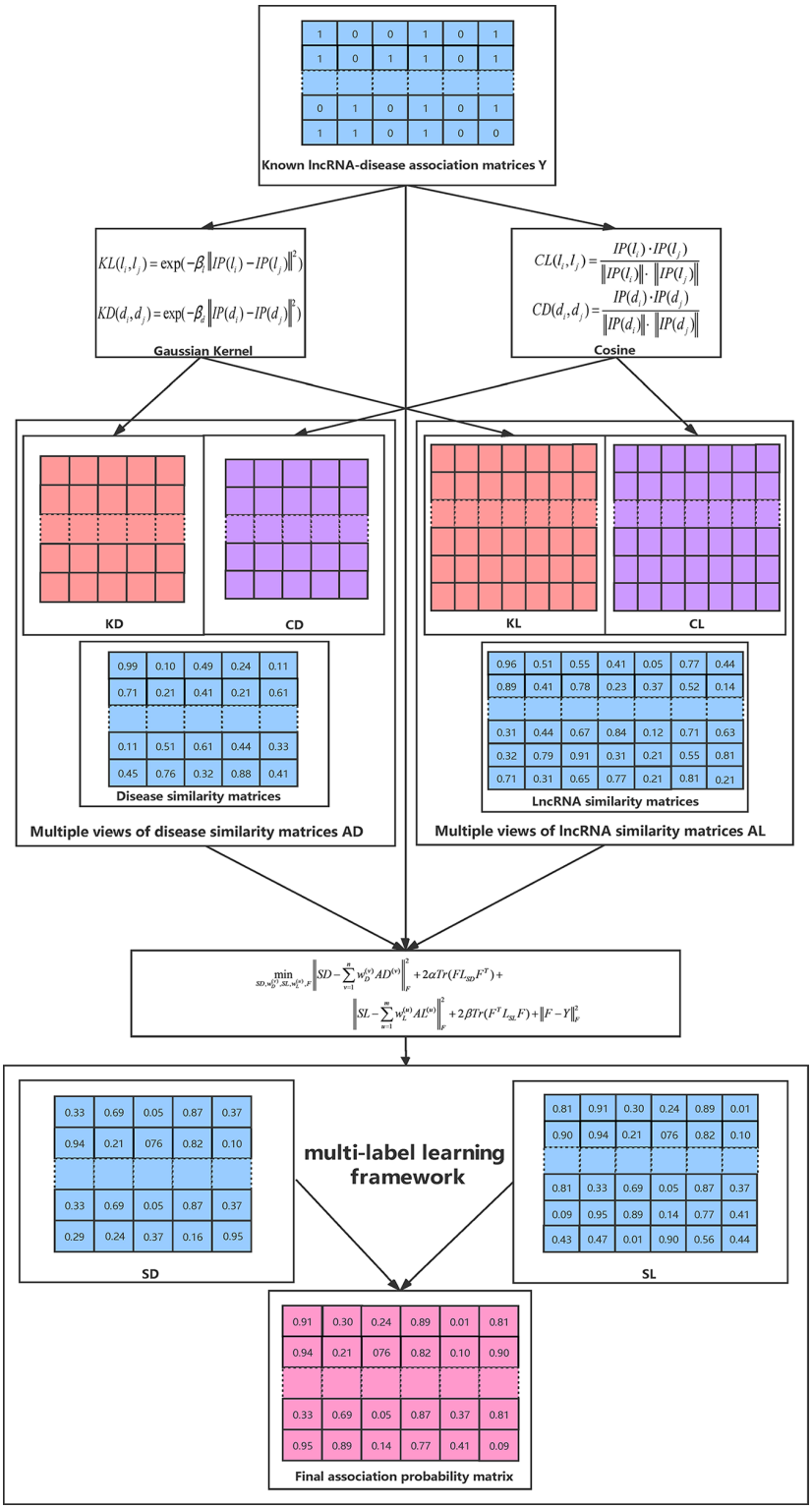


**FIGURE 1 |** An overall workflow of our method.

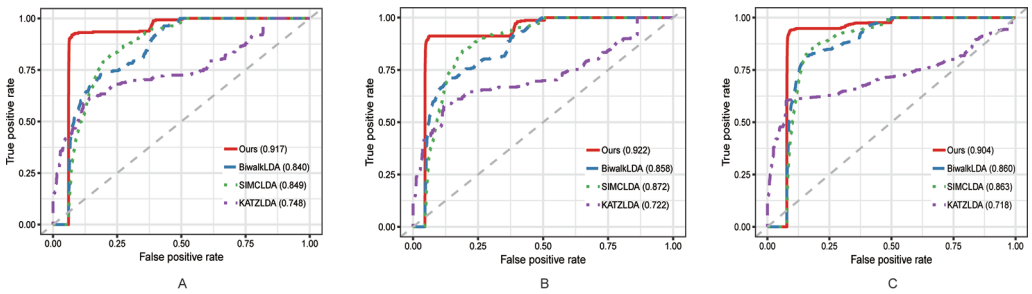


**FIGURE 2 |** The comparison results between our method and the other three methods in terms of LOOCV using **(A)** Dataset1; **(B)** Dataset2; **(C)** Dataset3.

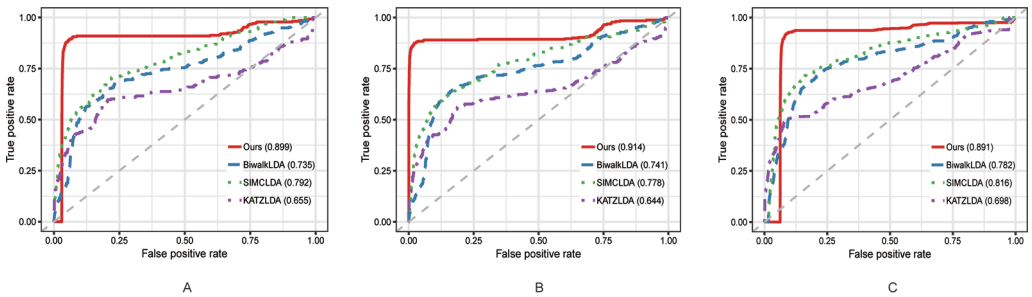


**FIGURE 3 |** The comparison results between our method and the other three methods in terms of five-fold CV using **(A)** Dataset1; **(B)** Dataset2; **(C)** Dataset3.

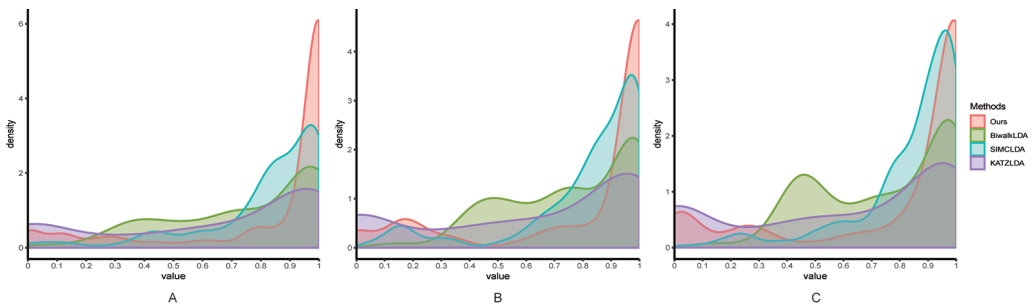


**FIGURE 4 |** The comparison results between our method and the other three methods in terms of LODOCV using **(A)** Dataset1; **(B)** Dataset2; **(C)** Dataset3.

diseases. As a result, discovering newly disease-related lncRNAs might deepen our understanding of the biological roles of lncRNAs in carcinogenesis. In this work, a novel multiview consensus graph learning method for predicting lncRNA–disease associations was proposed. We first constructed a set of similarity matrices for lncRNAs and diseases by leveraging the known lncRNA–disease associations. We then learned a consensus graph for lncRNAs and diseases from the multiple similarity matrices and predicted the association probability between lncRNAs and diseases based on a multi-label learning framework. The results of LOOCV, five-fold CV as well as LODOCV on three widely used datasets all confirmed the superiority of our method. Moreover, the convergence analysis indicates that our method has a fast convergence rate and could be well adapted in practice. Lastly, the case study conducted for UCEC indicated that the expression level of MIR7-3HG was significantly related with the survival rate of patients and thus it might play important roles in the pathogenesis of UCEC. In summary, our method could reliably predict potential lncRNA–disease associations and could be easily extended to incorporate more data sources.

The success of our method is mainly two-fold. First, the known lncRNA–disease associations were leveraged to construct multiple kernel similarity matrices to better characterize the lncRNA similarities as well as disease similarities. Second, the view weights imposed for each view during the learning process guaranteed that more reliable similarity matrices have higher impacts on the final consensus graph. Despite the commendable

**TABLE 2 |** Comparison of different methods based on LODOCV using Wilcoxon signed rank test.

| Dataset | BiwalkLDA | SIMCLDA | KATZLDA |
|---|---|---|---|
| Dataset1 | 8.41e−10 | 1.84e−09 | 3.74e−12 |
| Dataset2 | 4.57e−09 | 1.22e−07 | 8.07e−13 |
| Dataset3 | 5.981e−09 | 7.49e−07 | 5.54e−14 |

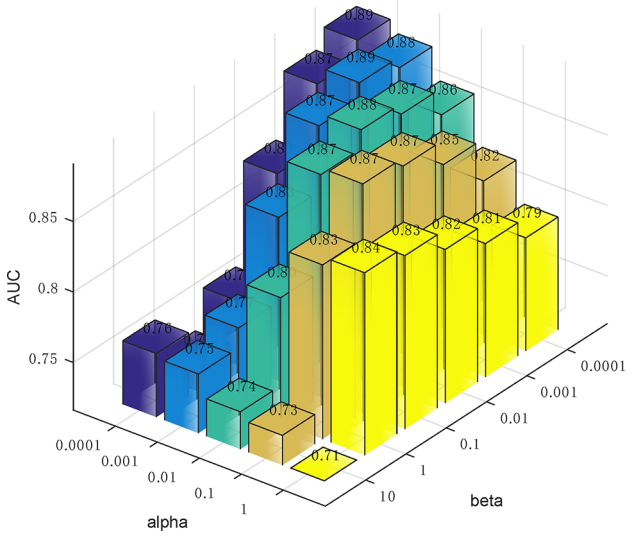


**FIGURE 5 |** The influence of the two parameters $\alpha$ and $\beta$ on the prediction accuracy of five-fold cross-validation.

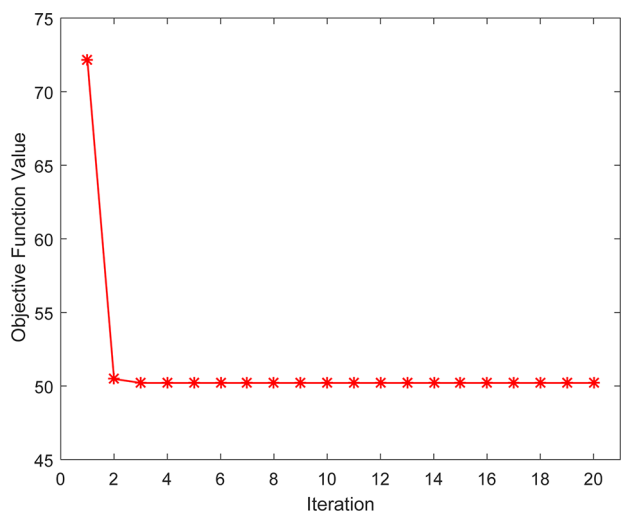


**FIGURE 6 |** The convergence rate of our method.

results obtained, our method could still be improved in several ways. For example, the optimal values of the two parameters $\alpha$ and $\beta$ might be searched by dynamic objective genetic algorithms. Besides, the integration of lncRNA expression data in our model should also be considered in the future.

**TABLE 3 |** The top 10 predicted lncRNAs to be associated with cervical uterine neoplasms by our method.

| Rank | lncRNA | Evidence |
|---|---|---|
| 1 | UCA1 | Lnc2Cancer;MNDR |
| 2 | TUG1 | Lnc2Cancer;MNDR |
| 3 | MIR99AHG | MNDR |
| 4 | MIR7-3HG | Unknown |
| 5 | HIF1A-AS1 | MNDR |
| 6 | HOXC-AS1 | MNDR |
| 7 | LINC-ROR | Lnc2Cancer |
| 8 | NEAT1 | Lnc2Cancer;MNDR |
| 9 | GSEC | MNDR |
| 10 | HOTTIP | MNDR |

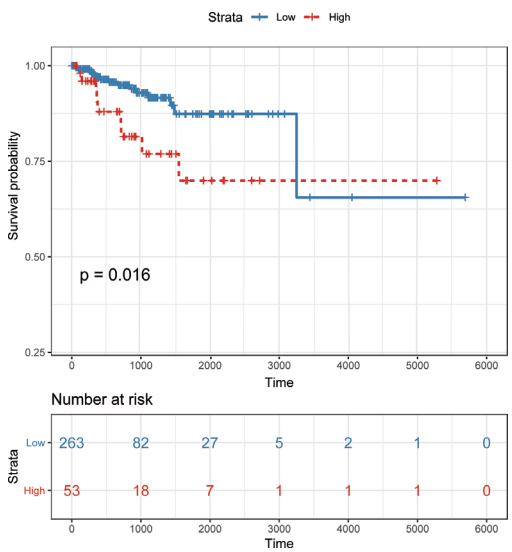


**FIGURE 7 |** Kaplan–Meier survival analysis using MIR7-3HG as a prognostic biomarker in uterine cervical neoplasms. Patients are divided into "high" and "low" groups according to their expression level of MIR7-3HG against the mean expression level across all patients.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## AUTHOR CONTRIBUTIONS

CL conceived the study. HT and QS developed the algorithm and analyzed the results. CL and HT wrote the paper. CL and JL supervised the study. All authors have read and approved the final manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61602283, 61873089, U1836216) and the Major Fundamental Research Project of Shandong Province (Grant No. ZR2019ZD03).

## REFERENCES


Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037. doi: 10.1093/nar/gky905

Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 12: survival analysis. *Crit. Care* 8, 389–394. doi: 10.1186/cc2955

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi: 10.1093/nar/gks1099

Chen, X., You, Z. H., Yan, G. Y., and Gong, D. W. (2016). IRWRLDA: improved random walk with restart for lncRNA–disease association prediction. *Oncotarget* 7, 57919–57931. doi: 10.18632/oncotarget.11141

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y. P., et al. (2019). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2936476

Chen, X. (2015). KATZLDA: KATZ measure for the lncRNA–disease association prediction. *Sci. Rep.* 5, 16840. doi: 10.1038/srep16840

Cui, Z., Liu, J. X., Gao, Y. L., Zhu, R., and Yuan, S. S. (2019). LncRNA–disease associations prediction using bipartite local model with nearest profile-based association inferring. *IEEE J. BioMed. Health.* doi: 10.1109/JBHI.2019.2937827

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233

Feng, Y. M., Chen, S., Xu, J. R., Zhu, Q., Ye, X. L., Ding, D. F., et al. (2018). Dysregulation of lncRNAs GM5524 and GM15645 involved in high-glucose-induced podocyte apoptosis and autophagy in diabetic nephropathy. *Mol. Med. Rep.* 18, 3657–3664. doi: 10.3892/mmr.2018.9412

Fu, G. Y., Wang, J., Domeniconi, C., and Yu, G. X. (2018). Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Gao, M. M., Cui, Z., Gao, Y. L., Liu, J. X., and Zheng, C. H. (2019a). Dual-network sparse graph regularized matrix factorization for predicting miRNA–disease associations. *Mol. Omics* 15, 130–137. doi: 10.1039/C8MO00244D

Gao, Y. L., Cui, Z., Liu, J. X., Wang, J., and Zheng, C. H. (2019b). NPCMF: Nearest Profile-based Collaborative Matrix Factorization method for predicting miRNA–disease associations. *BMC Bioinf.* 20, 353. doi: 10.1186/s12859-019-2956-5

Gong, Y., Niu, Y., Zhang, W., and Li, X. (2019). A network embedding-based multiple information integration method for the MiRNA–disease association prediction. *BMC Bioinf.* 20, 468. doi: 10.1186/s12859-019-3063-3

Guo, Z. H., You, Z. H., Wang, Y. B., Yi, H. C., and Chen, Z. H. (2019). A learning-based method for lncRNA–disease association identification combing similarity information and rotation forest. *iScience* 19, 786–795. doi: 10.1016/j.isci.2019.08.030

Han, Y., Zhu, L., Cheng, Z., Li, J., and Liu, X. (2018). Discrete optimal graph clustering. *IEEE Trans. Cybern.* doi: 10.1109/TCYB.2018.2881539

Hu, J., Gao, Y., Li, J., Zheng, Y., Wang, J., and Shang, X. (2019). A novel algorithm based on bi-random walks to identify disease-related lncRNAs. *BMC Bioinf.* 20, 569. doi: 10.1186/s12859-019-3128-3

Huang, J., Nie, F. P., and Huang, H. (2015). "A new simplex sparse learning model to measure data similarity for clustering," in *The twenty-fourth International Conference on Aritificial Intelligence*(Buenos Aires, Argentina), 3569–3575.

Jeong, Y. Y., Kang, H. K., Chung, T. W., Seo, J. J., and Park, J. G. (2003). Uterine cervical carcinoma after therapy: CT and MR imaging findings. *Radiographics* 23, 969–981. discussion 981. doi: 10.1148/rg.234035001

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F. X., Pan, Y., et al. (2017). LDAP: a web server for lncRNA–disease association prediction. *Bioinformatics* 33, 458–460. doi: 10.1093/bioinformatics/btw639

Li, L. J., and Chang, H. Y. (2014). Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends Cell Biol.* 24, 594–602. doi: 10.1016/j.tcb.2014.06.003

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., et al. (2015). TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* 75, 3728–3737. doi: 10.1158/0008-5472.CAN-15-0273

Li, G. H., Luo, J. W., Liang, C., Xiao, Q., Ding, P. J., and Zhang, Y. J. (2019). Prediction of lncRNA–disease associations based on network consistency projection. *IEEE Access* 7, 58849–58856. doi: 10.1109/ACCESS.2019.2914533

Liang, C., Yu, S., and Luo, J. (2019). Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. *PloS Comput. Biol.* 15, e1006931. doi: 10.1371/journal.pcbi.1006931

Liu, M. X., Chen, X., Chen, G., Cui, Q. H., and Yan, G. Y. (2014). A computational framework to infer human disease-associated long noncoding RNAs. *PloS One* 9, e84408. doi: 10.1371/journal.pone.0084408

Liu, H., Xu, B., Lu, D. J., and Zhang, G. J. (2018a). A path planning approach for crowd evacuation in buildings based on improved artificial bee colony algorithm. *Appl. Soft Comput.* 68, 360–376. doi: 10.1016/j.asoc.2018.04.015

Liu, R., Wang, H., and Yu, X. (2018b). Shared-nearest-neighbor-based clustering by fast search and find of density peaksR. *Inform Sci.* 450, 200–225. doi: 10.1016/j.ins.2018.03.031

Lu, C. Q., Yang, M. Y., Luo, F., Wu, F. X., Li, M., Pan, Y., et al. (2018). Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics* 34, 3357–3364. doi: 10.1093/bioinformatics/bty327

Mongelli, A., Martelli, F., Farsetti, A., and Gaetano, C. (2019). The dark that matters: long non-coding RNAs as master regulators of cellular metabolism in non-communicable diseases. *Front. In Physiol.* 10, 369. doi: 10.3389/Fphys.2019.00369

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., et al. (2016). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* 44, D980–D985. doi: 10.1093/nar/gkv1094

Pan, Z., Zhang, H., Liang, C., Li, G., Xiao, Q., Ding, P., et al. (2019). Self-weighted multi-kernel multi-label learning for potential miRNA–disease association prediction. *Mol. Ther. Nucleic Acids* 17, 414–423. doi: 10.1016/j.omtn.2019.06.014

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

Shi, D., Zhu, L., Cheng, Z., Li, Z., and Zhang, H. (2018). Unsupervised multi-view feature extraction with dynamic graph learning. *J. Visual Commun. Image Representation* 56, 256–264. doi: 10.1016/j.jvcir.2018.09.019

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., et al. (2013). Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis.* 4, e765. doi: 10.1038/cddis.2013.292

Wang, Q., Chen, M., Nie, F., and Li, X. (2020). Detecting coherent groups in crowd scenes by multiview clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 46–58. doi: 10.1109/TPAMI.2018.2875002

Xiao, X., Zhu, W., Liao, B., Xu, J., Gu, C., Ji, B., et al. (2018). BPLLDA: predicting lncRNA–disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. In Genet.* 9, 411. doi: 10.3389/fgene.2018.00411

Xie, G., Meng, T., Luo, Y., and Liu, Z. (2019). SKF-LDA: similarity kernel fusion for predicting lncRNA–disease association. *Mol. Ther. Nucleic Acids* 18, 45–55. doi: 10.1016/j.omtn.2019.07.022

Yin, M. M., Cui, Z., Gao, M. M., Liu, J. X., and Gao, Y. L. (2019). LWPCMF: logistic weighted profile-based collaborative matrix factorization for predicting miRNA–disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2019.2937774

Yu, J., Xuan, Z., Feng, X., Zou, Q., and Wang, L. (2019). A novel collaborative filtering model for lncRNA–disease association prediction based on the naive Bayesian classifier. *BMC Bioinf.* 20, 396. doi: 10.1186/s12859-019-2985-0

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2019). Graph embedding on biomedical networks: methods, applications, and evaluations. *Bioinformatics.* doi: 10.1093/bioinformatics/btz718

Zha, Z. J., Mei, T., Wang, J., Wang, Z., and Hua, X. S. (2009). Graph-based semi-supervised learning with multiple labels. *J. Vis. Commun. Image Represent.* 20, 97–103. doi: 10.1016/j.jvcir.2008.11.009

Zhang, W., Zhu, X., Fu, Y., Tsuji, J., and Weng, Z. (2017). Predicting human splicing branchpoints by combining sequence-derived features and multi-label learning methods. *BMC Bioinf.* 18, 464. doi: 10.1186/s12859-017-1875-6

Zhang, B., Lei, Z., Sun, J., and Zhang, H. (2018a). Cross-media retrieval with collective deep semantic learning. *Multimedia Tools Appl.* 7, 1–20.

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

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

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

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

Zhou, M., Wang, X. J., Li, J. W., Hao, D. P., Wang, Z. Z., Shi, H. B., et al. (2015). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network, *Mol. Biosyst* 11, 760–769. doi: 10.1039/C4MB00511B

Zhu, L., Huang, Z., Li, Z., Xie, L., and Shen, H. T. (2018). Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 5264–5276. doi: 10.1109/TNNLS.2018.2797248

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA–disease network: a survey. *Briefings In Funct. Genomics* 15, 55–64. doi: 10.1093/bfgp/elv024

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