IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 17, 2019, accepted April 16, 2019, date of publication April 25, 2019, date of current version May 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912945

IIRWR: Internal Inclined Random Walk With Restart for LncRNA-Disease Association Prediction

LEI WANG[1,2], YUBIN XIAO[1], JIECHEN LI[1], XIANG FENG[1,2], QIAN LI[3], AND JIALIANG YANG[3]

[1]Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China
[2]College of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410001, China
[3]Genesis Beijing Co., Ltd., Beijing 100102, China

Corresponding authors: Lei Wang (wanglei@xtu.edu.cn) and Jialiang Yang (yangjl@geneis.cn)

The authors thank the anonymous referees for suggestions that helped improve the paper substantially. The project is partly sponsored by the National Natural Science Foundation of China (No. 61873221, No. 61672447), the Natural Science Foundation of Hunan Province (No. 2018JJ4058, No. 2019JJ70010, No. 2017JJ5036), and the CERNET Next Generation Internet Technology Innovation Project (No. NGII20160305, No. NGII20170109).

ABSTRACT Experimental studies have demonstrated that long-non-coding RNAs (lncRNAs) are closely related to human disease. However, due to the complexity of diseases and high costs of bio-experiments, associations between diseases and lncRNAs are still unclear. Hence, it is essential to establish effective computational models to predict the potential relationships between diseases and lncRNAs. In this paper, different from traditional prediction models based on random walk with restart (RWR), a novel prediction model based on internal inclined random walk with restart (IIRWR) has been established to infer potential lncRNA-disease associations and compared to the state-of-the-art RWR-based prediction models. One major novelty of our IIRWR-based prediction model is the introduction of the concept of disease clique, which makes the process of the random walk to possess an internal tendency. The other major novelty of our model lies in the addition of the weights of disease linkages to the traveling network, which guarantees our model can achieve excellent prediction performance while the number of known lncRNA-disease associations is limited. The simulation results show that our model can achieve reliable AUCs of 0.8080, 0.8363, and 0.8745 under the frameworks of five-fold cross-validation (CV), ten-fold CV, and leave-one-out cross validation (LOOCV), respectively. Moreover, in case studies of cervical cancer and leukemia, the experimental results show that eight and ten out of the top ten predicted lncRNAs can be confirmed by related literature, which demonstrates that our method is effective in predicting novel diseases associated lncRNAs.

INDEX TERMS lncRNA, lncRNA-disease associations, internal inclined random walk with restart.

I. INTRODUCTION

Accumulating evidence suggests that lncRNAs play critical roles in lots of biological processes such as epigenetic regulation, cell cycle regulation, cell differentiation, transcriptional regulation, and so on [1]–[5]. With more and more lncRNAs being identified by newly developed sequencing technologies, their functions especially their contributions to various diseases have received much attention recently. More and more biological experiments have shown that mutations and dysregulations of lncRNA are associated with diseases [6]–[8], such as leukemia [9], [10], cardiovascular diseases [11], Alzheimer's disease [12] and various kinds of cancers [13]. Therefore, effectively identifying the association between disease and lncRNA cannot only help us understand the molecular mechanism of disease but also provide biomarkers for disease treatment and find drug target. Many lncRNA-related databases including LncRNADisease [14], NRED [15], lncRNAdb [16] and NONCODE [17] have been established successively. However, since traditional bio-experiments in testing the relationships between lncRNAs and diseases are costly and time-consuming, the numbers of known lncRNA-disease associations in these databases are still small [18]–[20]. Hence, it is necessary to develop high-throughput computational models to discover potential lncRNA-disease associations. Various computational models have been proposed for this purpose.

According to their implementation strategy, these methods can be roughly classified into three categories to predict novel lncRNA-disease associations [6]. The first category

The associate editor coordinating the review of this manuscript and approving it for publication was Chee Keong Kwoh.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

is to build a machine learning model that predicts potential lncRNA-Disease association based on known lncRNA-disease associations. For instance, Yu and Wang et al. develop a Naive Bayesian Classifier to predict lncRNA-disease associations [21], which builds two global networks by integrating multiple biological information to predict potential lncRNA-disease associations. However, this kind of supervised classifier model needs negative sample information, which is usually unavailable. So they randomly selected unlabeled lncRNA-disease pairs as negative samples, which would affect the prediction performance. In 2013, Chen et al. establish a model of Laplacian Regularized Least Squares for LncRNA-Disease Association (LRLSLDA) to predict potential lncRNA-disease association base on a semi-supervised learning framework [22]. This model does not require negative sample information, and meanwhile significantly improves the performance of previous predictions. However, there are also some limitations, for example, how to choose the optimal parameters have not been solved.

The second category are to integrate known lncRNA-disease association network, disease similar network, lncRNA similar network to establish a heterogeneous network, and implement propagation algorithm. In 2014, Sun et al. propose a prediction model named RWRlncD by adopting random walk with restart (RWR) in the lncRNA functional similarity network [23]. As a matter of fact, the random walk algorithm has been widely used in bioinformatics and other fields, and has achieved good performance. For example, Yang and Li et al. develop a model called RWPCN (Random Walker on Protein Complex Network) for predicting and prioritizing disease genes [24]. In the same year, Yang et al. establish coding-non-coding gene-disease bipartite network by integrating known disease genes with lncRNA-disease associations [25], and implement propagation algorithms on this bipartite network to predict potential lncRNA-disease associations. Although this model achieved good predictive performance, there is some limitation such as the lack of lncRNA function annotation would have an influence on its performance.

Due to the confirmed lncRNA-disease associations are still limited, some researchers began to predict potential lncRNA-disease associations by other means, rather than the above two categories method based on known lncRNA-disease association pairs. For example, in 2014, Liu et al. established the first computational model that does not rely on known lncRNA-disease association [26], it integrates disease genes and the relationship between genes and lncRNA, successfully avoided the limitations of using limited lncRNA-disease correlation samples, however, this model can not be applied to lncRNA that without related gene records.

In this paper, a novel Internal Inclined Random Walk with Restart (IIRWR) was proposed to predict potential lncRNA-disease associations by integrating known lncRNA-disease associations, disease semantic similarity, disease weight and Gaussian interaction profile kernel similarity for lncRNAs. Different from the traditional RWR method, IIRWR in the addition of the weights of disease linkages to the traveling network, solved some limitations of the traditional RWR method such as aimlessness, ensuring that our models get better predictive performance (MATLAB code can be downloaded at https://github.com/xiaoyubin123/code.git). In order to better estimate the prediction performance of our newly proposed IIRWR-based model, several methods including 5-fold Cross Validation (5-fold CV), 10-fold Cross Validation (10-fold CV) and Leave-One-Out Cross Validation (LOOCV) have been implemented. As a result, our IIRWR-based model can achieve reliable AUCs of 0.8080, 0.8363, 0.8745 in the 5-fold CV, 10-fold CV and LOOCV, respectively, outperforming those of several state-of-the-art traditional RWR-based models. Moreover, we also validated novel lncRNA-disease associations predicted by our method by literature mining.

II. MATERIALS AND METHODS

A. CONSTRUCTION OF THE ADJACENCY MATRIX OF KNOWN LNCRNA-DISEASE ASSOCIATIONS

In this paper, we downloaded lncRNA-disease associations from the latest version of the LncRNADisease database (http://www.cuilab.cn/lncrnadisease) [14]. After removing non-human data and repeated associations supported by multiples evidences, we finally got 1695 unique experimentally verified human lncRNA-disease associations (see Supplementary material 1), which including 828 unique lncRNAs and 314 unique diseases. The data can be expressed as an 828 × 314 lncRNA-disease association adjacency matrix A, in which $A(i, j) =$, if and only if there is a known association between lncRNA l_i and disease d_j, and $A(i, j) = 0$ otherwise. In addition, we defined $N_L = 828$ and $N_D = 314$, thus, the newly obtained adjacency matrix A is of dimension $N_L \times N_D$.

B. CONSTRUCTION OF THE DISEASE SEMANTIC SIMILARITY MATRIX

For each of the 314 diseases, we further downloaded its corresponding MESH descriptor from the National Library of Medicine (http://www.nlm.nih.gov) [27], which categorizes and provides semantic information for various diseases. In the light of previous knowledge, relationships between different diseases can be illustrated as a structure of Directed Acyclic Graph (DAG) [27], [28]. For example, a disease d can be described as DAG(d) = ($D(d)$, $E(d)$), in which $D(d)$ is a node set consist of d itself and its ancestor nodes, while $E(d)$ is the set of directed edges from parents to child nodes. Therefore, for a disease d and one of its ancestor nodes t in $D(d)$, we can define the contribution of disease t to the semantic value of disease d as follow:

$$D_d(t) = \begin{cases} 1 & if\ t = d \\ max\left\{\Delta * D_d(t') | t' \in children\ of\ t\right\} & if\ t \neq d \end{cases} \tag{1}$$

Δ is the semantic contribution factor with value between 0 and 1, according to experimental results done by

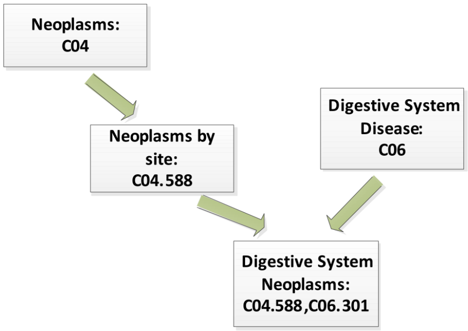


FIGURE 1. The DAG corresponding to the digestive system neoplasms.

predecessors [27], it is better to set Δ to 0.5. Hence, by combining contributions from all disease terms in DAG(d), we can define the semantic value of d as follow:

$$D(d) = \sum_{t_i \in DAG(d)} D_d(t_i) \tag{2}$$

According to formula (1) and formula (2), the calculation of the semantic value of Digestive-System Neoplasms D (DSN) was shown in Fig. 1, There is $D(DSN) = 1$ (the contribution of Digestive System Neoplasm) + 0.5 (the contribution of Neoplasms by site) + 0.5 (the contribution of Digestive system Disease) + 0.5*0.5 (the contribution of Neoplasms) = 2.25.

Moreover, based on the concept of DAG, it is reasonable to assume that two diseases with more common ancestor nodes will have higher semantic similarity, the disease semantic similarity (DSS) between disease d_i and disease d_j can be calculated as follow:

$$DSS(i,j) = \frac{\sum_{t \in (DAG(d_i) \cap DAG(d_j))} (D_{d_i}(t) + D_{d_j}(t))}{D(d_i) + D(d_j)} \tag{3}$$

Therefore, an $ND \times ND$ dimensional DSS matrix can be constructed(see Supplementary material 2).

C. CONSTRUCTION OF THE GAUSSIAN INTERACTION PROFILE KERNEL SIMILARITY MATRIX FOR LNCRNAS

For any two given lncRNAs l_i and l_j, the Gaussian interaction profile kernel similarity (KL) between them can be defined as follows [29]:

$$KL(i,j) = exp(-\gamma_l ||IP(l_i) - IP(l_j)||^2) \tag{4}$$

$$\gamma_l = \frac{\gamma_l'}{\sum_{i=1}^{N_l} ||IP(l_i)||^2} \tag{5}$$

$IP(l_i)$ and $IP(l_j)$ denote the i^{th} and j^{th} column in the adjacency matrix $A\ (i, j)$ separately, γ_l control kernel bandwidth base on normalizing the new bandwidth parameter γ_l', the best choice of γ_l' is 1 depend on previous data [22]. Combining formula (4) and formula (5), an $N_L \times N_L$ dimensional Gaussian interaction profile kernel similarity matrix KL is established.

D. CONSTRUCTION OF THE IIRWR-BASED PREDICTION MODEL

As illustrated in Fig. 2, the IIRWR-based prediction model consists of three major steps. Step 1: Constructing the roaming network (lncRNAs network) based on diseases similarity and lncRNAs similarity. Step 2: Implementing random walk on the newly constructed roaming network. Step 3: Ranking candidate lncRNAs after obtaining stable random walk probability.

1) STEP 1: CONSTRUCTION OF THE ROAMING NETWORK AND INITIAL PROBABILITY

Step 1.1 (Revision the Disease Semantic Similarity Matrix: Although a DSS matrix can be constructed through formula (3), the matrix is very sparse. Moreover, for any disease d_i, it is obvious that elements with a value not equal to zero in the i^{th} row of DSS matrix represent all collected diseases that are related to d_i. lncRNA with more associations to disease clique will have greater relevance to the target disease. Thus, we defined the set of elements with the value not equal to zero in the i^{th} row of DSS matrix as the *Disease Clique* of disease d_i. Thereafter, for any two diseases d_i and d_j, the similarity between them can be re-calculated as follows:

$$\begin{aligned} RDSS(i,j) = {} & \beta_{ij} * DSS(i,j) \\ & + (1-\beta_{ij}) * \alpha * \sum_{t \in DS(i)} DSS(i,t) * DSS(t,j) \end{aligned} \tag{6}$$

$$\beta_{ij} = \begin{cases} 1; & if\ DSS(i,j) > 0 \\ 0; & Otherwise \end{cases} \tag{7}$$

Here, α is the penalty factor with a value between 0 and 1, $DS(i)$ denotes the set consisting of the nonzero elements in the i^{th} row of DSS. Along with rectification by formula (6) and formula (7), the problem of the sparsity of the DSS matrix is solved efficiently.

Step 1.2 (Construction of the Disease Weight Matrix): For any given lncRNA l_i and disease d_j, it is reasonable to assume that there exists a potential association between them if l_i has more known associations with diseases in the Disease Clique of d_j. Hence, based on the above assumption, we obtained an $N_L \times N_D$ dimensional disease weight matrix DW as follows:

$$DW(i,j) = \frac{SDW(i,j)}{max(SDW(j))} \tag{8}$$

$$SDW(i,j) = \sum_{m \in DL(i)} RDSS(m,j) \tag{9}$$

Here, $DL(i)$ represents a set of diseases that have known associations with l_i, $SDW(j)$ denotes the j^{th} row of the matrix SDW. The calculation process of the DW matrix is illustrated in detail. As showed in Fig. 3 (1), there are three diseases d_3, d_4 and d_5 related with lncRNA l_1, then we can obtain that $RDSS$ (3,1) = 0.68, $RDSS$ (4,1) = 0.9, $RDSS$ (5,1) = 0.7 through Fig. 3 (2) and (3). Therefore, SDW(1,1) = $RDSS$(3,1) + $RDSS$ (4,1) + $RDSS$ (5,1) = 0.68 + 0.9 + 0.7 = 2.28. Finally, DW(1,1) = 1 according to formula (8), Fig. 3 (4).

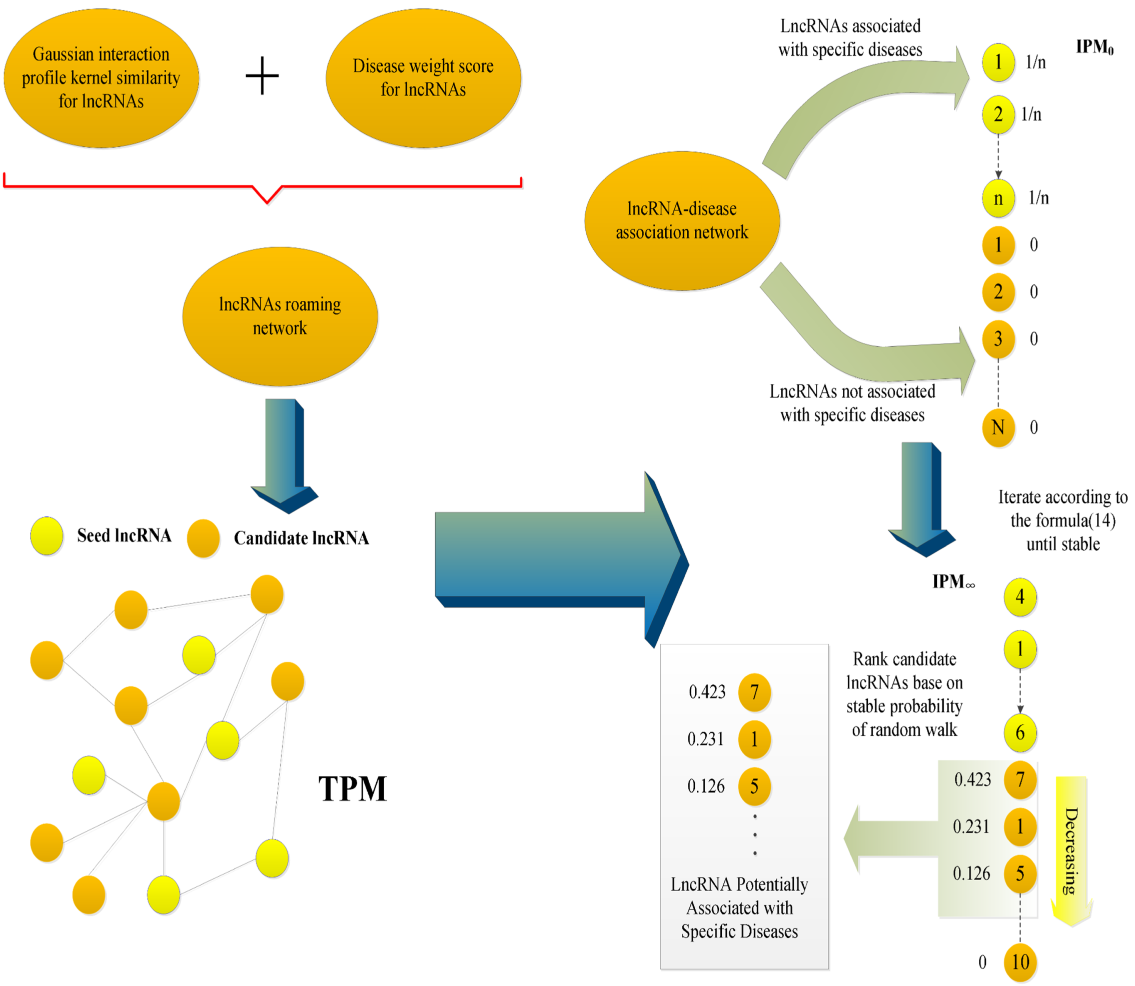
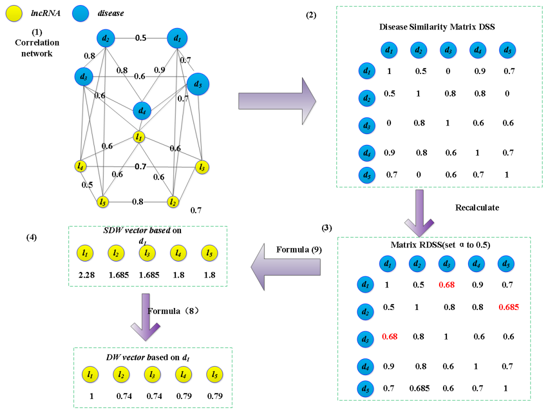

Gaussian interaction profile kernel similarity for lncRNAs
+
Disease weight score for lncRNAs
lncRNAs roaming network
Seed lncRNA
Candidate lncRNA
TPM
LncRNAs associated with specific diseases
IPM₀
1 1/n
2 1/n
n 1/n
1 0
2 0
3 0
N 0
lncRNA-disease association network
LncRNAs not associated with specific diseases
Iterate according to the formula(14) until stable
IPM∞
4
1
6
0.423 7
0.231 1
0.126 5
0 10
Decreasing
Rank candidate lncRNAs base on stable probability of random walk
0.423 7
0.231 1
0.126 5
LncRNA Potentially Associated with Specific Diseases

FIGURE 2. Flow chart of IIRWR-based prediction model.



Disease Similarity Matrix DSS

	d_1	d_2	d_3	d_4	d_5
d_1	1	0.5	0	0.9	0.7
d_2	0.5	1	0.8	0.8	0
d_3	0	0.8	1	0.6	0.6
d_4	0.9	0.8	0.6	1	0.7
d_5	0.7	0	0.6	0.7	1

Matrix RDSS(set α to 0.5)

	d_1	d_2	d_3	d_4	d_5
d_1	1	0.5	0.68	0.9	0.7
d_2	0.5	1	0.8	0.8	0.685
d_3	0.68	0.8	1	0.6	0.6
d_4	0.9	0.8	0.6	1	0.7
d_5	0.7	0.685	0.6	0.7	1

SDW vector based on d_1

l_1	l_2	l_3	l_4	l_5
2.28	1.685	1.685	1.8	1.8

DW vector based on d_1

l_1	l_2	l_3	l_4	l_5
1	0.74	0.74	0.79	0.79

FIGURE 3. Illustration of the construction of the disease weight matrix.

Step 1.3 (Construction of the Roaming Network): For any given disease d_m, as a walker, it can move from the current node to the next node according to a transition probability matrix and obtain the migration probability vector of the next hop node simultaneously. Therefore, based on formula (4) and formula (8), for any two given lncRNAs l_i and l_j, an $N_L \times N_L$ dimensional transition probability matrix TPM can be structured as follows:

$$TPM(i,j) = \frac{STPM(i,j)}{\sum_{k=1}^{N_L} STPM(k,j)} \tag{10}$$

$$STPM(i,j) = \gamma_{jm} * DW(j,m) + (1-\gamma_{jm}) * KL(i,j) \tag{11}$$

$$\gamma_{jm} = \begin{cases} 1; & if\ DW(j,m) > 0 \\ 0; & Otherwise \end{cases} \tag{12}$$

Based on the matrix TPM, a roaming network can be obtained easily, in which, the node-set consists of $N_L = 828$ kinds of lncRNAs. For any two given lncRNA nodes l_i and l_j in the roaming network, there is an edge between them when $TPM(i,j) > 0$.

Additionally, let the disease d_m be a walker, then for each lncRNA l_i ($i \in [1, N_L]$), we can structure a N_L dimensional initial probability vector IPM_0 for d_m as follows:

$$IPM_0(i) = \frac{A(i,m)}{\sum_{i=1}^{N_L} A(i,m)} \tag{13}$$

2) STEP 2: PROCESS OF THE RANDOM WALK

For any given walker disease d_m, its random walk is initially started from any given lncRNA node in the roaming network and can be carried on from the current node to the next hop node according to the TPM matrix and the IPM_0 indicated in step 1.3. Moreover, during the period of disease $d_m{}'$s random walk, IIRWR will restart its random walk in each step with a probability of r. Hence, supposing the walker d_m has currently arrived at the lncRNA node l_i after having gone through t steps in the roaming network and $IPM_t = (IPM_t(1), (IPM_t(2), \ldots, (IPM_t(N_L))^T$ also been obtained by d_m at present, then it is evident that d_m can further obtain an updated probability vector IPM_{t+1} as follow:

$$IPM_{t+1} = (1-r) * TPM * IPM_t + r * IPM_0 \qquad (14)$$

3) STEP 3: OBTAIN THE RANKS OF CANDIDATE LNCRNAS

For any given disease node d_m, while it walks in the roaming network, all lncRNAs have known associations with d_m will be regarded as seed lncRNAs, while other lncRNAs without associations with d_m will be considered as the candidate lncRNAs. According to formula (14), it is easy to deduce that the probability vector IPM_t obtained by d_m will be stable in final as long as t is big enough. Additionally, in view of time efficiency and accuracy requirement, IPM_t will be considered stable if the difference between IPM_t and IPM_{t+1} is less than 10^{-10}. So far, we can get the ranking for all lncRNAs based on IPM_t, for any given lncRNA l_i, the higher the ranking the more probability it would be associated with the disease walker d_m.

III. RESULTS

A. PERFORMANCE EVALUATION

In order to assess the prediction performance of IIRWR, the framework of LOOCV was firstly implemented based on our previously obtained 1695 known lncRNA-disease associations. During the experiment, for a given disease d, each lncRNA has known association with disease d was left out as a test sample in turn, and all other lncRNAs have associations with disease d was retained as seed lncRNAs or training samples for our model learning. Test samples and those lncRNAs have no associations with disease d were considered as candidate lncRNAs. Thus the ranking of the left-out test sample relative to candidate samples could be evaluated. If the ranking of the test sample is greater than the given threshold, then it would be regarded as a successful prediction, otherwise, it is an unsuccessful prediction. Moreover, upon different given thresholds, their corresponding true positive rates (TPR, sensitivity) and false positive rate (FPR, 1-specificity) could be figured out, in which sensitivity denotes the percentage of test samples with ranking higher than the corresponding threshold, while specificity represents the percentage of test samples that rank below the corresponding threshold. Hence, the Receiver-Operating Characteristics (ROC) curve can be drawn through plotting TPR versus FPR at different thresholds, and the area below ROC curve (AUC) could be applied to evaluate the prediction performance of our prediction model IIRWR. In general, the AUC value of 1 represents a perfect prediction while the AUC value of 0.5 means a random guess. However, in the experimental process of LOOCV, there would be a situation that the left-out lncRNA is the only lncRNA has known association with a specific disease, which would lead to both the initial probability vector IPM_0 turning to be a zero vector and some points in the ROC curve focus on the location of FPR $= 0.5$. In order to ensure the fairness of our forecasted results, we have randomly set an initial vector in this kind of situation.

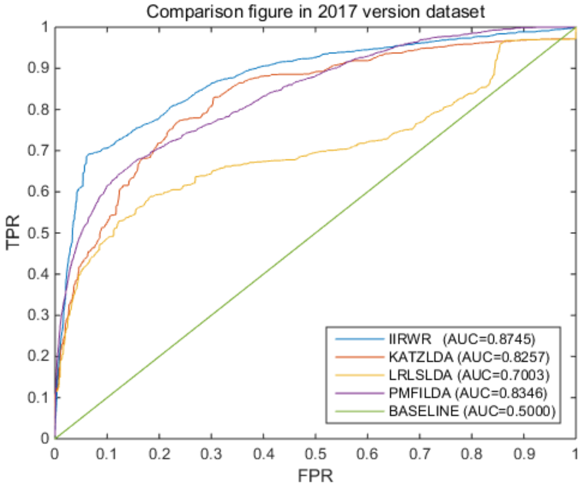


FIGURE 4. The AUCs achieved by IIRWR, KATZLDA, PMFILDA, and LRLSLDA in LOOCV based on the dataset downloaded from the 2017 version of the lncRNAdisease database.

Through simulation, we first compared IIRWR with three state-of-the-art lncRNA-disease association prediction models KATZLDA [30], LRLSLDA [22] and PMFILDA [31] in the framework of LOOCV. The comparison results were shown in Fig. 4. IIRWR achieves a reliable AUC of 0.8745, which is higher than that of 0.8257 acquired by KATZLDA, the AUC of 0.7003 obtained from LRLSLDA and the AUC of 0.8346 from PMFILDA. Besides, in the validation experiment, it took about 3 seconds for our IIRWR model to validate a lncRNA-disease association, while it cost more than 10 seconds for both KATZLDA and LRLSLDA.

Next, the performance of our IIRWR model was further compared with several others state-of-the-art lncRNA-disease association prediction models such as LRLSLDA [22], RWRLNCD [23] and NRWRH [32] and in the framework of LOOCV. Considering that different data versions may lead to different prediction performance, we downloaded the same dataset adopted by RWRLNCD, NRWRH and LRLSLDA from the 2012 version of the lncRNADisease database, which consists of 293 known lncRNA-disease associations including 167 different diseases and 118 various lncRNAs. The comparison results were listed in Table. 1. IIRWR achieves reliable AUC of 0.6796, which far outweighs the AUC of 0.5024 from RWRLNCD,

TABLE 1. The AUC achieved by IIRWR, RWRLNCD, NRWRH, and LRLSLDA in LOOCV based on the 2012 version of the lncRNAdisease database.

Method	IIRWR	RWRLNCD	NRWRH	LRLSLDA
AUC	0.6796	0.5024	0.6363	0.6585

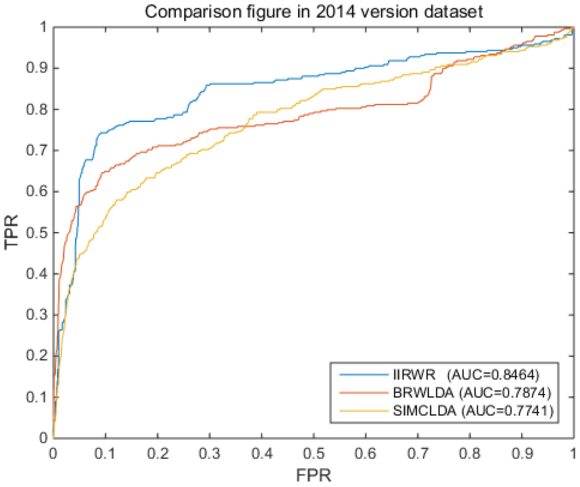


FIGURE 5. The AUCs achieved by IIRWR, BRWLDA, and SIMCLDA in LOOCV based on the dataset downloaded from the 2014 version of the lncRNAdisease database.

and also higher than the AUC of 0.6363 in NRWRH and 0.6585 achieved by LRLSDA as well. Then, IIRWR was compared with the recently proposed two state-of-the-art computational models such as BRWLDA [33] and SIMCLDA [34] in the framework of LOOCV. Consistent with the above comparison, we downloaded the same database adopted by BRWLDA and SIMCLDA from the 2014 version of lncRNADisease database,which consists of 319 known lncRNA-disease associations including 169 different diseases and 131 various lncRNAs. As shown in Fig. 5, we can see that IIRWR achieves a reliable AUC of 0.8464,which is significantly higher than AUCs of others (0.7874 from BRWLDA and 0.7741 from SIMCLDA).

In addition, by comparing Table. 1, Fig. 4 and Fig. 5, it is easy to find that the AUC achieved by IIRWR based on the 2012 version of the lncRNADisease database is not as good as that based on the 2014 version, Even far below the 2017 version. This difference is owing to the known lncRNA-disease associations in the dataset of 2012 version is sparse than in the dataset of 2014 version and 2017 version. Therefore, for the purpose of further assess the prediction performance of IIRWR, frameworks of *K*-fold Cross Validation including 5-fold Cross Validation (5-fold CV) and 10-fold Cross Validation (10-fold CV) were applied. In the framework of *K*-fold Cross Validation, all these known lncRNA-disease associations were equally divided into K distinct groups, each group was ruled out as a test group in turn, while others were retained as training groups. Fig. 6 and Fig. 7 shows that IIRWR can achieve reliable AUC of 0.8080 and 0.8363 under the framework of 5-fold CV and 10-fold CV,respectively.

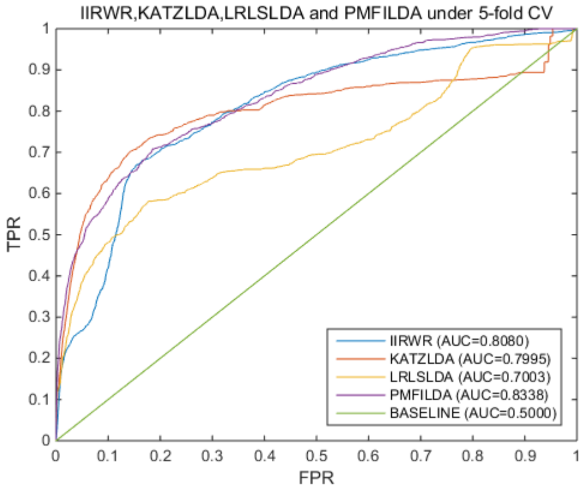


FIGURE 6. AUC achieved by IIRWR, KATZLDA, LRLSLDA, and PMFILDA under the framework of 5-fold cross validation.

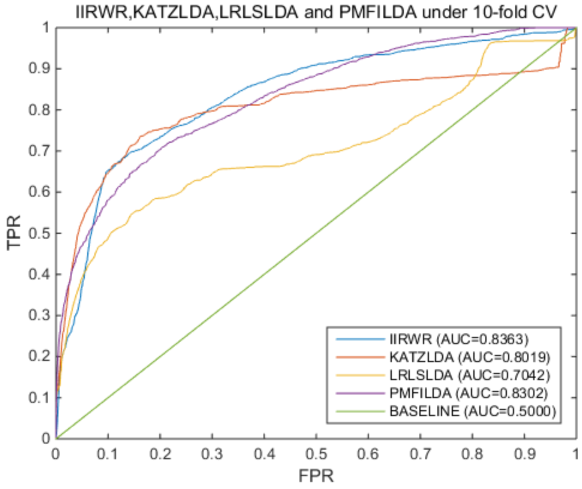


FIGURE 7. AUC achieved by IIRWR, KATZLDA, LRLSLDA, and PMFILDA under the framework of 10-fold cross validation.

B. SENSITIVITY ANALYSIS OF PARAMETERS

According to above descriptions, there are two major parameters in our prediction model IIRWR, one is the parameter α in formula (6), and the other is the parameter r in the formula (13). In this section, the effects of both parameters to the performance of our prediction model were estimated. For testing the effect of the parameter α, a series of AUCs in the framework of LOOCV with α varying from 0 to 1 were calculated. The simulation results in Table. 2 demonstrates that IIRWR achieves the best prediction performance when $\alpha = 0.6$.

Similarly, with r varying from 0.1 to 0.9. IIRWR achieve the best prediction performance when $r = 0.4$ (Table. 3).

TABLE 2. AUCs achieved by IIRWR in LOOCV with different values of α.

α	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.6**
AUC	0.8243	0.8733	0.8734	0.8736	0.8738	0.8744
α	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
AUC	0.8745	0.8744	0.8743	0.8743	0.8742	

TABLE 3. AUCs achieved by IIRWR in LOOCV with different values of r.

r	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
AUC	0.8654	0.8717	0.8729	0.8740	0.8745	0.8741
r	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
AUC	0.8718	0.8719	0.8724	0.8718	0.5124	

TABLE 4. Top 10 potential cervical cancer-related lncRNAs predicted by IIRWR and confirmed by PubMed unique identifier.

Disease	lncRNA	Evidence(PMID)	RANK
Cervical cancer	UCA1	29620291;28414550;17416635	1
Cervical cancer	NEAT1	29416780;29207151	2
Cervical cancer	LINC-ROR	26314857	3
Cervical cancer	CCAT1	28849215;28978096;27775802	4
Cervical cancer	TUG1	29029428;28088836	5
Cervical cancer	XIST	29909347	6
Cervical cancer	CASC2	28495512	7
Cervical cancer	FER1L4	unknown	8
Cervical cancer	SNHG12	29533945	9
Cervical cancer	FAM212B-AS1	unknown	10

C. CASE STUDIES

Cancer and leukemia have threatened human beings for hundreds of years [35], [36]. In order to further confirms the practical prediction performance of our IIRWR model, cervical cancer and leukemia were selected as case studies. while simulation, only those leukemia-related lncRNAs and cervical cancer-related lncRNAs which have not been included in the data set of 1695 known lncRNA-disease associations would be considered as validation candidates. And moreover, all predicted lncRNAs associated with leukemia and cervical cancer would be ranked according to their scores respectively. Top 10 disease-related lncRNAs predicted by the IIRWR were confirmed by experiments and articles downloaded from NCBI, and the corresponding evidence was listed in Table. 4 and Table. 5.

A large number of evidences proves that lncRNA plays a key role in the development of cervical cancer [37], [38]. As exhibited in Table. 4, when implementing IIRWR to predict cervical cancer-related lncRNAs, there are 8 out of the top 10 predicted candidate lncRNAs having been confirmed by biomedical literature. Rankings of the other two lncRNAs without being confirmed were 8th and 10th, respectively. Similarly, abnormalities of some lncRNAs have been reported closely related to the development of leukemia [39], [40]. As displayed in Table. 5, 10 out of the top 10 predicted candidate lncRNAs having been confirmed by biomedical literature.

In summary, IIRWR achieves satisfactory and reliable prediction performance in prediction of potential lncRNA-disease associations.

TABLE 5. Top 10 potential leukemia-related lncRNAs predicted by IIRWR and confirmed by pubMed unique identifier.

Disease	lncRNA	Evidence(PMID)	RANK
Leukemia	H19	15645136; 28765931; 29643943	1
Leukemia	HOTAIR	26622861; 26261618;25979172	2
Leukemia	MEG3	19595458	3
Leukemia	MALAT1	28713913	4
Leukemia	PVT1	29510227	5
Leukemia	UCA1	29762824; 26053097;29663500	6
Leukemia	GAS5	27951730	7
Leukemia	TUG1	29654398	8
Leukemia	XIST	7981672	9
Leukemia	CCAT1	26923190	10

IV. DISCUSSIONS

Accumulating evidence manifested that lncRNAs play critical roles in various biological processes and relate to the pathological change of human diseases. However, verification of lncRNA-disease associations through bio-experiments is costly and time-consuming. Therefore, it is necessary and feasible by using an effective computational model to infer potential lncRNA-disease associations. In this study, a novel computational model called IIRWR was constructed based on random walk with internal tendency. In order to assess the prediction performance of IIRWR, various experiments have been carried out, simulation results show that IIRWR achieves reliable AUCs of 0.8080, 0.8363 and 0.8745 under the framework of 5-fold CV, 10-fold CV and LOOCV, respectively. Comparing with traditional state-of-the-art computational models, IIRWR outperforms more as well. In addition, in case studies, experimental results further verify that 8 out of the top 10 predicted lncRNAs and 10 out of top 10 predicted lncRNAs are associated with cervical cancer and leukemia separately, which strongly supports that IIRWR greatly improves the recognition of potential lncRNA-disease associations.

Even so, the current version of IIRWR has limitations. For example, only 1695 known lncRNA-disease associations have been adopted by IIRWR, the prediction accuracy of IIRWR will improve higher if more known lncRNA-disease associations are added. Additionally, when applied IIRWR in the situation that a disease has no association with any lncRNA, we tried to change the initial probability vector to the weight probability vector of lncRNAs related to the clique of the specific disease, however, the effect is not very satisfactory. Certainly, all problems in the current version of IIRWR will be the focuses of our following researches.

REFERENCES

[1] M. Guttman *et al.*, "Chromatin signature reveals over a thousand highly conserved large non-coding Rnas in mammals," *Nature*, vol. 458, no. 7235, pp. 223–227, 2009.

[2] L. Es *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, p. 860, 2001.

[3] 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 Res.*, vol. 39, pp. D146–D151, Jan. 2011.

[4] D. Bu, K. Yu, and S. Sun, "NONCODE v3. 0: Integrative annotation of long noncoding RNAs," *Nucleic Acids Res.*, vol. 40, pp. D210–D215, Nov. 2012.

[5] A. M. Khalil *et al.*, "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 28, pp. 11667–11672, 2009.

[6] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Briefings Bioinformatics*, vol. 18, no. 4, pp. 558–576, 2016.

[7] 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.

[8] O. Wapinski and H. Y. Chang, "Long noncoding RNAs and human disease," *Trends Cell Biol.*, vol. 21, no. 6, pp. 354–361, 2011.

[9] G. A. Calin *et al.*, "Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas," *Cancer Cell*, vol. 12, no. 3, pp. 215–229, 2007.

[10] X. Zhang *et al.*, "A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster," *Blood*, vol. 113, no. 11, pp. 2526–2534, 2009.

[11] A. Congrains *et al.*, "Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B," *Atherosclerosis*, vol. 220, no. 2, pp. 449–455, 2011.

[12] M. A. Faghihi *et al.*, "Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase," *Nature Med.*, vol. 14, no. 7, pp. 723–730, 2008.

[13] G. Yang, X. Lu, and L. Yuan, "LncRNA: A link between RNA and cancer," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mech.*, vol. 1839, no. 11, pp. 1097–1109, 2014.

[14] G. Chen *et al.*, "LncRNADisease: A database for long-non-coding RNA-associated diseases," *Nucleic Acids Res.*, vol. 41, pp. D983–D986, Jan. 2013.

[15] 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 Res.*, vol. 37, pp. D122–D126, Jan. 2009.

[16] 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 Res.*, vol. 39, pp. D146–D151, Jan. 2011.

[17] D. Bu *et al.*, "NONCODE v3.0: Integrative annotation of long noncoding RNAs," *Nucleic Acids Res.*, vol. 40, pp. D210–D215, Jan. 2012.

[18] G. Borsani *et al.*, "Characterization of a murine gene expressed from the inactive X chromosome," *Nature*, vol. 351, no. 6324, pp. 325–329, 1991.

[19] C. I. Brannan, E. C. Dees, R. S. Ingram, and S. M. Tilghman, "The product of the H19 gene may function as an RNA," *Mol. Cellular Biol.*, vol. 10, no. 1, pp. 28–36, 1990.

[20] N. Brockdorff *et al.*, "The product of the mouse Xist gene is a 15 Kb inactive X-specific transcript containing no conserved ORF and located in the nucleus," *Cell*, vol. 71, no. 3, pp. 515–526, 1992.

[21] J. Yu, P. Ping, L. Wang, L. Kuang, X. Li, and Z. Wu, "A novel probability model for LncRNA–disease association prediction based on the Naïve Bayesian classifier," *Genes*, vol. 9, no. 7, p. 345, 2018.

[22] 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.

[23] J. Sun *et al.*, "Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network," *Mol. BioSyst.*, vol. 10, no. 8, pp. 2074–2081, 2014.

[24] P. Yang, X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Inferring gene-phenotype associations via global protein complex network propagation," *PLoS ONE*, vol. 6, no. 7, 2011, Art. no. e21502.

[25] X. Yang *et al.*, "A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases," *PLoS ONE*, vol. 9, no. 1, 2014, Art. no. e87797.

[26] M.-X. Liu, X. Chen, G. Chen, Q.-H. Cui, and G.-Y. Yan, "A computational framework to infer human disease-associated long noncoding RNAs," *PLOS ONE*, vol. 9, no. 1, 2014, Art. no. e84408.

[27] 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.

[28] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Sci. Rep.*, vol. 5, Aug. 2015, Art. no. 13186.

[29] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[30] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Sci. Rep.*, vol. 5, no. 1, 2015, Art. no. 16840.

[31] Z. Xuan, J. Li, J. Yu, X. Feng, B. Zhao, and L. Wang, "A probabilistic matrix factorization method for identifying lncRNA-disease associations," *Genes*, vol. 10, no. 2, p. 126, 2019.

[32] M. Zhou *et al.*, "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Mol. BioSyst.*, vol. 11, no. 3, pp. 760–769, 2015.

[33] G. Yu, G. Fu, C. Lu, Y. Ren, and J. Wang, "BRWLDA: Bi-random walks for predicting lncRNA-disease associations," *Oncotarget*, vol. 8, no. 36, pp. 60429–60446, 2017.

[34] C. Lu *et al.*, "Prediction of lncRNA–disease associations based on inductive matrix completion," *Bioinformatics*, vol. 34, no. 19, pp. 3357–3364, 2018.

[35] P. E. Spiess, J. Dhillon, and A. S. Baumgarten, "Pathophysiological basis of human papillomavirus in penile cancer: Key to prevention and delivery of more effective therapies," *CA, Cancer J. Clinicians*, vol. 66, no. 6, pp. 481–495, 2016.

[36] A. Omer, P. Singh, N. K. Yadav, and R. K. Singh, "MicroRNAs: Role in leukemia and their computational perspective," *Wiley Interdiscipl. Rev.,*, vol. 6, no. 1, pp. 65–78, 2015.

[37] S. Cao, W. Liu, and F. Li, "Decreased expression of lncRNA GAS5 predicts a poor prognosis in cervical cancer," *Int. J. Clin. Exp. Pathol.*, vol. 7, no. 10, pp. 6776–6783, 2014.

[38] M. Iden, S. Fye, K. Li, T. Chowdhury, R. Ramchandran, and J. S. Rader, "The lncRNA PVT1 contributes to the cervical cancer phenotype and associates with poor patient prognosis," *PLoS One*, vol. 11, no. 5, 2016, Art. no. e0156274.

[39] T. R. Fernando *et al.*, "LncRNA expression discriminates karyotype and predicts survival in B-lymphoblastic leukemia," *Mol. Cancer Res.*, vol. 13, no. 5, pp. 839–851, 2015.

[40] Y. Wang, P. Wu, R. Lin, L. Rong, Y. Xue, and Y. Fang, "LncRNA NALT interaction with NOTCH1 promoted cell proliferation in pediatric T cell acute lymphoblastic leukemia," *Sci. Rep.*, vol. 5, Sep. 2015, Art. no. 13749.

Authors' photographs and biographies not available at the time of publication.