IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 22, 2019, accepted April 28, 2019, date of publication May 3, 2019, date of current version May 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914533

Prediction of LncRNA-Disease Associations Based on Network Consistency Projection

GUANGHUI LI[1], JIAWEI LUO[2], CHENG LIANG[3], QIU XIAO[4], PINGJIAN DING[2], (Student Member, IEEE), AND YUEJIN ZHANG[1]

[1]School of Information Engineering, East China Jiaotong University, Nanchang 330013, China
[2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[3]College of Information Science and Engineering, Shandong Normal University, Jinan 250000, China
[4]College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China

Corresponding author: Jiawei Luo (luojiawei@hnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61862025, Grant 61873089, Grant 61602283, Grant 11862006, Grant 61861017, and Grant 61862023, in part by the Jiangxi Provincial Natural Science Foundation under Grant 20181BAB211016, Grant 2018ACB21032, Grant 20181BAB211013, and Grant 20181BAB202007, in part by the Scientific and Technological Research Project of Education Department in Jiangxi Province under Grant GJJ170383, Grant GJJ170381, and Grant GJJ170414, and in part by the Hunan Provincial Natural Science Foundation under Grant 2018JJ2024.

ABSTRACT A growing body of research has uncovered the role of long noncoding RNAs (lncRNAs) in multiple biological processes and tumorigenesis. Predicting novel interactions between diseases and lncRNAs could help decipher disease pathology and discover new drugs. However, because of a lack of data, inferring disease-lncRNA associations accurately and efficiently remains a challenge. In this paper, we present a novel network consistency projection for LncRNA-disease association prediction (NCPLDA) model by integrating the lncRNA-disease association probability matrix with the integrated disease similarity and lncRNA similarity. The lncRNA-disease association probability matrix is calculated based on known lncRNA-disease associations and disease semantic similarity. The integrated disease similarity and lncRNA similarity are computed based on disease semantic similarity, lncRNA functional similarity and Gaussian interaction profile kernel similarity. In leave-one-out cross validation experiments, NCPLDA achieved outstanding AUCs of 0.8900, 0.8996, and 0.9012 for three datasets. Furthermore, prostate cancer and ovarian cancer case studies demonstrated that the NCPLDA can effectively infer undiscovered lncRNAs.

INDEX TERMS Disease-related lncRNAs, lncRNA-disease association, network consistency projection, similarity measure.

I. INTRODUCTION

Long noncoding RNAs (lncRNAs) have a length greater than 200 nucleotides and were initially thought to be transcriptional noise; they are a class of important regulators of various cellular processes [1]–[3], such as cell cycle control, translational and post-translational regulation, and chromatin modification. Not surprisingly, aberrant lncRNA expression can cause the initiation and progression of numerous human diseases [4]–[6]. Therefore, inferring disease-related lncRNAs can contribute to understanding the complex mechanisms underlying carcinogenesis at the lncRNA level and uncover new prognostic markers for disease diagnosis and therapy. However, experimentally verified lncRNA-disease relationships are still comparatively limited. Moreover, most biological experiments are laborious and costly. Accordingly, it is important to compute the association scores between diseases and lncRNAs using computational methods.

Over the past few years, some prediction models have been proposed to quantify the lncRNA-disease association probability based on multiple data types and sources [7], [8]. Recently, proposed models can be predominantly classified into three types. The first type of model makes use of known disease-related lncRNAs to infer new associations. Related studies are built on the basic view that similar diseases are likely to be linked with functionally similar lncRNAs. Chen and Yan [9] presented a computational method called LRLSLDA to make prediction for lncRNA-disease pairs based on a Laplacian regularized least squares framework, which achieved reliable prediction by utilizing known interactions and lncRNA expression profiles. Sun *et al.* [10] developed

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Huei Cheng.

RWRlncD, a global network-based method that constructs a lncRNA-lncRNA functional similarity network and performs random walks restarting to detect potential links. Afterwards, Chen *et al*. [11] proposed an improved model called IRWRLDA based on random walk, which sets the initial probability vector of lncRNAs by combining disease semantic similarity, lncRNA expression similarity, and known disease-lncRNA associations. Ping *et al*. [12] devised a new model based on a constructed bipartite network that relies on available disease-lncRNA network topological information. Yang *et al*. [13] presented an iterative model that integrates disease-lncRNA interactions and coding gene-disease interactions into a coding-noncoding gene-disease bipartite network; the model ranks the potential candidates for all diseases by applying a propagation algorithm to the newly constructed bipartite network. Ding *et al*. [14] applied a resource allocation algorithm on a gene-disease-lncRNA tripartite graph, integrating disease-gene interactions with disease-lncRNA interactions. However, most models of this type fail to identify interactions for new diseases or lncRNAs. The second type of model mines underlying disease-lncRNA pairs based on known disease (or lncRNA)-associated miRNAs or genes. Based on a hypergeometric distribution, Chen [15] designed the model of HGLDA, which scored each disease-lncRNA pair by testing whether this disease notably shared common microRNAs with lncRNA. Liu *et al*. [16] constructed a lncRNA prioritization model by integrating gene-disease interactions, gene expression profiles, and lncRNA expression profiles to prioritize novel disease-associated lncRNAs. Alaimo *et al*. [17] used the resource propagation technique to compute the weight between each disease-ncRNA pair by integrating disease-target interactions with ncRNA-target interactions. Later, Mori *et al*. [18] extended Alaimo's ncPred to make predictions, which integrated biological sequence information with weights computed by ncPred. The third type of models predicts possible associations between diseases and lncRNAs by combining multiple data sources. For instance, Chen [19] released a computational model termed KATZLDA to excavate latent lncRNA-disease associations by integrating known lncRNA-disease interactions and different types of lncRNA and disease similarity into a heterogeneous network. Fu *et al.* [20] presented a matrix factorization-based data fusion model to prioritize potential disease-related lncRNAs, which was capable of selecting and weighing different data sources. Lu *et al*. [21] developed a computational model named SIMCLDA using inductive matrix completion aiming to complete the missing lncRNA-disease interaction based on known interactions, lncRNA similarity data, and disease similarity data. Subsequently, a probabilistic model named NBCLDA was proposed by Yu *et al*. [22]. In NBCLDA, multiple heterogeneous kinds of biological data were combined to generate a tripartite network and a quadruple network, in which a naïve Bayesian classifier was applied for the prediction of latent disease-lncRNA interactions. Recently, Xiao *et al*. [23] developed a new computational path weighted method to compute the association score of each lncRNA-disease pair based on paths connecting them in a heterogeneous network, which was composed of known lncRNA-disease interactions, lncRNA similarity data, and disease similarity data. In addition, Lan *et al*. [24] integrated multiple data sources and utilized a bagging SVM classifier to mine latent relationships between diseases and lncRNAs. However, effectively fusing multiple heterogenous data sources is still a big challenge. Moreover, many models fail to identify interactions for new diseases, and some models need negative samples, which usually are unknown. Additionally, because of a lack of known association data, excavating the latent disease-lncRNA interactions accurately and efficiently remains a challenge.

In this paper, we propose the use of network consistency projection for lncRNA-disease association prediction (NCPLDA). NCPLDA computes the association score for each lncRNA-disease pair by integrating the lncRNA-disease association probability matrix with the integrated disease similarity and lncRNA similarity. To estimate the prediction accuracy of NCPLDA, leave-one-out cross validation was carried out on three datasets downloaded from the LncRNADisease database [25]. Moreover, we conducted two types of case studies to examine the practical ability of NCPLDA, including association prediction for diseases based on known interactions and for new diseases without any known related lncRNAs. NCPLDA performed well in the above experiments, which suggests that NCPLDA is effective in inferring latent interactions between lncRNAs and diseases.

Unlike the previous models, we predict the lncRNA-disease relationships according to a simple and effective algorithm (i.e., network consistency projection), which does not require negative instances and can greatly reduce the prediction time. In addition, a preprocessing procedure is adopted to derive the intermediate interaction probability of non-associated lncRNA-disease pairs that may be missed in the current databases. Such a consideration is often better for improving the prediction precision and enhancing predictions in the new disease cases.

II. MATERIALS AND METHODS

A. HUMAN LncRNA-DISEASE ASSOCIATIONS

The data of known lncRNA-disease associations were gathered from the LncRNADisease database. Three versions (June-2012 Version, January-2014 Version, and June-2015 Version) of LncRNADisease were used in the experiments. A few lncRNA-disease interactions with irregular lncRNA names or disease names were filtered out, and all repeating records were merged. As a result, the June-2012 Version (marked as DS1) consisted of 276 interactions between 112 lncRNAs and 150 diseases, the January-2014 Version (marked as DS2) consisted of 319 interactions between 131 lncRNAs and 169 diseases, and the June-2015 Version (marked as DS3) consisted of 621 interactions between 285 lncRNAs and 226 diseases. For convenience, we used an adjacency matrix $A \in R^{m \times n}$ to encode the lncRNA-

disease interactions with m lncRNAs as rows and n diseases as columns, where $A(i, j) = 1$ if lncRNA i has association with disease j and 0 if not.

B. DISEASE SEMANTIC SIMILARITY

Recent research has increasingly demonstrated that disease semantic similarity aids in predicting disease-related ncRNAs [26]–[28]. Here, the calculation of the disease semantic similarity was identical to the method proposed by Wang *et al*. [29], in which diseases are organized as directed acyclic graphs (DAGs). According to their corresponding DAGs, semantic similarities among all diseases were computed and the calculating process was illustrated by the DOSE software package [30]. Therefore, disease semantic similarity matrix SS can be obtained, where the element $SS(d_i, d_j)$ denotes the value of the semantic similarity between disease d_i and disease d_j.

C. LncRNA FUNCTIONAL SIMILARITY

It is observed that functionally similar lncRNAs are often linked with similar diseases [10], [26]. Here, the calculation of the lncRNA functional similarity was identical to that in the previous study [10], which computed functional similarity of two lncRNAs by estimating the semantic similarity of two disease sets that are associated with these two lncRNAs. Specifically, we supposed that lncRNA l_i and lncRNA l_j were related to m and n diseases, respectively. Thus, the similarity between lncRNA l_i and lncRNA l_j can be calculated by equations (1) and (2) as follows:

$$FS(l_i, l_j) = \frac{\sum_{d \in D(l_j)} S(d, D(l_i)) + \sum_{d \in D(l_i)} S(d, D(l_j))}{m + n} \quad (1)$$

$$S(d1, D(l_i)) = \max_{d \in D(l_i)} (SS(d1, d)) \quad (2)$$

where FS is the lncRNA functional similarity matrix, $D(l_i)$ indicates the disease set related to lncRNA l_i.

Note that disease similarity matrix SS and lncRNA similarity matrix FS are both sparse. Therefore, we further introduced the Gaussian interaction profile kernel similarity to alleviate this weakness.

D. GAUSSIAN INTERACTION PROFILE KERNEL SIMILARITY FOR LncRNAS AND DISEASES

Based on the notion that functionally similar lncRNAs tend to have similar association patterns with similar diseases and vice versa, a Gaussian interaction profile kernel similarity was constructed to measure lncRNA similarity and disease similarity. Firstly, we defined the association profile of lncRNA l_i, which is a binary vector specifying the presence or absence of association with each disease. In fact, the association profile of lncRNA l_i is the i-th row vector of the adjacency matrix A, i.e., $A(i, :)$. Then, the similarity between lncRNA l_i and l_j can be computed by utilizing a Gaussian kernel function:

$$KL(l_i, l_j) = \exp(-\gamma_l \|A(i, :) - A(j, :)\|^2) \quad (3)$$

$$\gamma_l = \gamma / (\frac{1}{m} \sum_{i=1}^{m} \|A(i, :)\|^2) \quad (4)$$

where γ_l is charged with controlling the kernel bandwidth, which could be obtained by normalizing the original bandwidth γ. Here, γ is simply set to 1.

Similarly, disease Gaussian interaction profile kernel similarity can be defined as follows:

$$KD(d_i, d_j) = \exp(-\gamma_d \|A(:, i) - A(:, j)\|^2) \quad (5)$$

$$\gamma_d = \gamma / (\frac{1}{n} \sum_{i=1}^{n} \|A(:, i)\|^2) \quad (6)$$

where $A(:, i)$ denotes the association profile of disease d_i, and the parameter γ_d is defined similarly as γ_l.

E. INTEGRATED SIMILARITY FOR LncRNAS AND DISEASES

We combined lncRNA functional similarity FS with Gaussian interaction profile kernel similarity for lncRNA KL to construct the final lncRNA similarity matrix (LS). Specifically, for lncRNA l_i and lncRNA l_j, if $FS(l_i, l_j) = 0$, we have $LS(l_i, l_j) = KL(l_i, l_j)$, otherwise $LS(l_i, l_j) = FS(l_i, l_j)$. The combination is presented as follows:

$$LS(l_i, l_j) = \begin{cases} KL(l_i, l_j) & if \quad FS(l_i, l_j) = 0 \\ FS(l_i, l_j) & otherwise \end{cases} \quad (7)$$

Accordingly, we integrated semantic similarity SS and Gaussian interaction profile kernel similarity KD for diseases, and the final disease similarity matrix (DS) can be combined in the following manner:

$$DS(d_i, d_j) = \begin{cases} KD(d_i, d_j) & if \quad SS(d_i, d_j) = 0 \\ SS(d_i, d_j) & otherwise \end{cases} \quad (8)$$

F. ASSOCIATION PROBABILITY MATRIX

As we know, existing lncRNA-disease interactions are very sparse [9], [25]. In fact, many of the non-associated lncRNA-disease pairs in adjacency matrix A are unknown interactions [21]. Inspired by the solutions to new disease cases, we applied the WKNKN algorithm [31] as a preprocessing step to calculate the temporary interaction probability for these non-associated pairs according to their known neighbors. For instance, we estimate the interaction probability that lncRNA l_i and disease d_j interact. Firstly, we select K nearest known diseases as the neighbors of d_j based on their semantic similarity to d_j. Then, we compute the interaction probability profile for disease d_j by using the weighted average of its neighbors' interaction profiles. We formulate the WKNKN as follows:

$$A_d(:, d_j) = \frac{1}{Q_d} \sum_{i=1}^{K} w_i A(:, d_i) \quad (9)$$

where d_1 to d_K denote K nearest known neighbors of d_j sorted in descending order; $w_i = T^{i-1} SS(d_i, d_j)$ is the weight coefficient, where T is a decay term with $T \leq 1$, and $Q_d = \sum_{i=1}^{K} SS(d_i, d_j)$ is the normalization term.

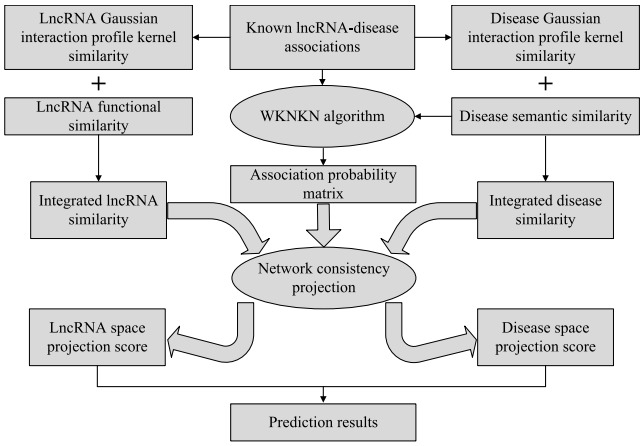


FIGURE 1. The overall workflow of NCPLDA method.

Finally, assuming $A(l_i, d_j)$ equals zero, we replace it with an intermediate interaction probability $A_d(l_i, d_j)$.

G. NCPLDA METHOD

In this paper, we generated a novel computational model NCPLDA to infer lncRNA-disease relationships by using network consistency projection [32]. The implementation process of NCPLDA could be summarized in three steps (see Fig. 1). Firstly, we constructed the integrated disease similarity and lncRNA similarity by using disease semantic similarity, lncRNA functional similarity as well as known disease-lncRNA interactions. Secondly, we calculated the disease-lncRNA association probability matrix based on known disease-lncRNA interactions and disease semantic similarity. Lastly, we implemented the network consistency projection on lncRNA space and disease space respectively, and then we combined the results from these two spaces to obtain the final predictions.

NCPLDA measures the relevance between lncRNA l_i and disease d_j by combining two separate network consistency projection scores, i.e., the lncRNA space projection score and the disease space projection score. The projection of the lncRNA similarity network (denoted as matrix LS) on the lncRNA-disease association probability network (denoted as matrix A) represents the lncRNA space projection. Thus, the lncRNA space projection can be formulated in vector form as:

$$LSP(i,j) = \frac{LS(i,:) \times A(:,j)}{|A(:,j)|} \tag{10}$$

where $LS\ (i,:)$ is a row vector representing the similarities between lncRNA l_i and all other lncRNAs; $A\ (:,j)$ is a column vector encoding the interactions between disease d_j and all lncRNAs; $|A(:,j)|$ denotes the length of vector $A\ (:,j)$; and $LSP\ (i,j)$ is the projection score of $LS\ (i,:)$ on $A\ (:,j)$. Obviously, the smaller angle between $LS\ (i,:)$ and $A(:,j)$, the more similar lncRNAs and lncRNA l_i are, and the more lncRNAs related to disease d_j, the higher the projection score $LSP\ (i,j)$ is.

Algorithm 1 NCPLDA

Input: the known lncRNA-disease association matrix A, disease semantic similarity matrix SS, the nearest neighbor number K and the decay term T

Output: the final network consistency projection score matrix NCP

1: Calculate the lncRNA functional similarity matrix FS by eq. (1) and eq. (2);
2: Calculate the lncRNA Gaussian interaction profile kernel similarity matrix KL by eq. (3) and eq. (4);
3: Calculate the disease Gaussian interaction profile kernel similarity matrix KD by eq. (5) and eq. (6);
4: Construct the final lncRNA similarity matrix LS and the final disease similarity matrix DS by eq. (7) and eq. (8), respectively;
5: Compute the temporary lncRNA-disease association probability matrix A by eq. (9);
6: Calculate the lncRNA space projection score matrix LSP by eq. (10);
7: Calculate the disease space projection score matrix DSP by eq. (11);
8: Integrate $LSP$and DSP by eq. (12) to obtain the final network consistency projection score matrix NCP;
9: **return** NCP.

Similarly, the projection of the disease similarity network (denoted as matrix DS) on the lncRNA-disease association probability network (denoted as matrix A) can be defined in a similar way:

$$DSP(i,j) = \frac{A(i,:) \times DS(:,j)}{|A(i,:)|} \tag{11}$$

where $DSP\ (i,j)$ is the projection score of $DS\ (:,j)$ on $A(i,:)$.

Finally, the lncRNA space projection score and the disease space projection score can be integrated and normalized as follows:

$$NCP(i,j) = \frac{LSP(i,j) + DSP(i,j)}{|LS(i,:)| + |DS(:,j)|} \tag{12}$$

where NCP is the final network consistency projection score matrix, which quantifies the relevance between each lncRNA-disease pair.

Algorithm 1 describes the implementation process of NCPLDA for disease-lncRNA association prediction. MATLAB code and datasets of NCPLDA can be accessed at https://github.com/ghli16/NCPLDA.

III. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETTINGS

We conducted leave-one-out cross validation (LOOCV) to evaluate and compare the predictive accuracy of NCPLDA against other competitive methods, including SIMCLDA [21], LRLSLDA [9], and RWRlncD [10]. In LOOCV, for a given disease d_i, each labeled d_i-associated lncRNA is selected in turn as the testing sample, while other labeled

lncRNA-disease interactions are deemed as the training set. All the unlabeled d_i-associated lncRNAs, including the testing sample, comprise the candidate samples. After performing prediction, we ranked the association probability of the testing sample with the other candidate samples to judge whether the rank of the testing sample exceeded a given threshold. After each known association has been tested, the receiver operating characteristic (ROC) curve can be obtained, which plots the true positive rate (TPR) versus the false positive rate (FPR) at various cutoff points. From the ROC curve, the area under the curve (AUC) can be used to measure the overall performance of the model. Given that the Gaussian interaction profile kernel similarity and the association probability matrix are connected with known lncRNA-disease interactions, they need to be recalculated in each round of cross validation.

B. PARAMETER ANALYSIS

We investigated the impacts of the nearest neighbor number K and the decay term T on the performance of NCPLDA, where K and T range from 10 to 50 with step 10 and 0.1 to 0.9 with step 0.1, respectively. To search for proper values of K and T, we performed cross validation experiments using our three datasets. As shown in Fig. 2, the AUC scores for the DS1 dataset were relatively robust for the change of K and T values. For instance, the maximal and minimal AUC values were 0.8900 and 0.8796 respectively. It means that there is no significant difference between them. In general, the AUC values just fluctuated within a 1.04% interval. Similar results were obtained for the DS2 and DS3 datasets for all K and T, which are illustrated in Supplementary Figs. S1 and S2.

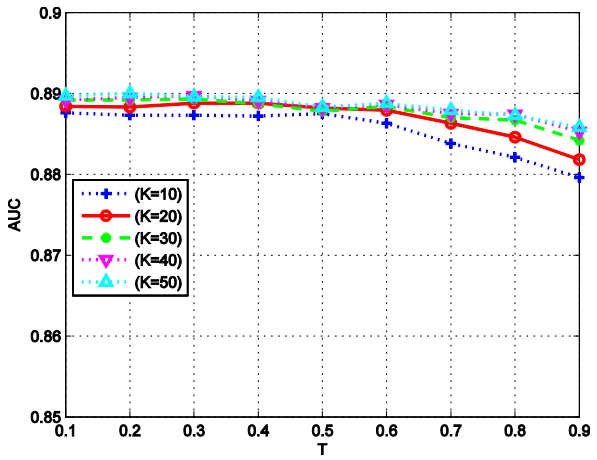


FIGURE 2. The effects of different values of K and T under the DS1 dataset.

C. PERFORMANCE EVALUATION

To demonstrate the effectiveness of NCPLDA in predicting disease-lncRNA interactions, we implemented three representative methods (i.e., SIMCLDA, LRLSLDA, and RWRlncD) using the same three datasets. Fig. 3 shows the ROC curves for each method and reports their corresponding AUC scores for the DS1 dataset. The ROC curve of NCPLDA was clearly superior to those of the other methods in most cases, and NCPLDA had the best AUC (0.8900) among the approaches, whereas the AUCs of SIMCLDA, LRLSLDA, and RWRlncD were 0.8237, 0.6444, and 0.6886, respectively. Additionally, for the DS2 dataset, NCPLDA almost always achieved the highest TPR for the same FPR and obtained an AUC value of 0.8996, which was better than those of the other models (SIMCLDA: 0.8526; LRLSLDA: 0.6407; RWRlncD: 0.6803); the results are presented in Fig. 4. We also provide the results for the DS3 dataset in Fig. 5. Consistent with DS1 and DS2, NCPLDA yielded the highest AUC. The AUCs obtained for NCPLDA were 4.34%, 17.12%, and 28.91% higher than those obtained for SIMCLDA, LRLSLDA, and RWRlncD, respectively. In conclusion, these experimental results suggest that NCPLDA is an effective lncRNA-disease interaction prediction tool.

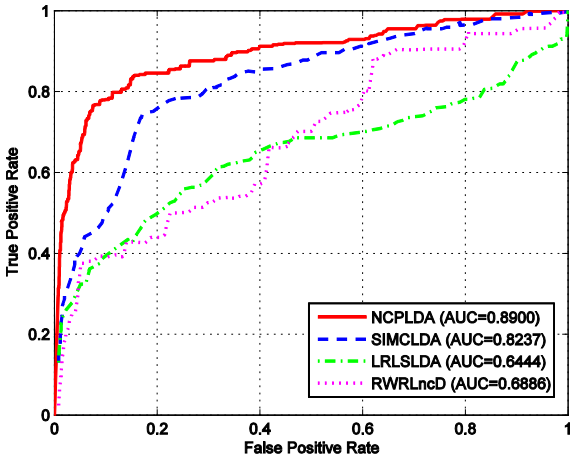


FIGURE 3. Comparison of different prediction models using LOOCV under the DS1 dataset.

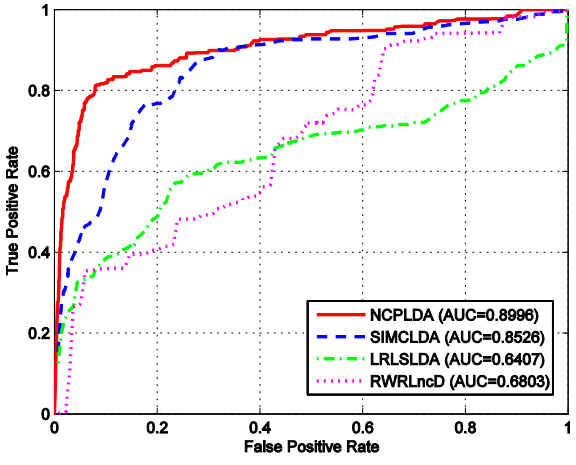


FIGURE 4. Comparison of different prediction models using LOOCV under the DS2 dataset.

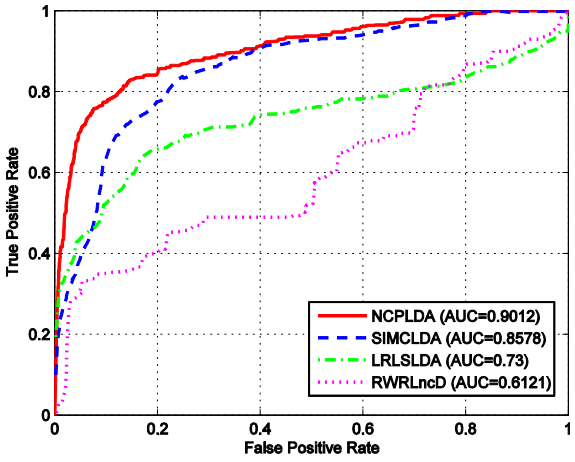


FIGURE 5. Comparison of different prediction models using LOOCV under the DS3 dataset.

D. CASE STUDIES

Two different types of case studies, on prostate cancer and ovarian cancer, were implemented to examine the practicability of NCPLDA in predicting novel disease-lncRNA interactions. For the first type of study, we used all known disease-lncRNA associated pairs from the DS3 dataset as training samples and selected the top 10 candidate lncRNAs as prediction lists for each investigated disease. The predicted associations were then checked by two other disease-lncRNA association databases: MNDR [33] and Lnc2cancer [34].

We implemented NCPLDA for analysis of prostate cancer, and 9 out of the top 10 predicted lncRNA candidates were supported by either MNDR or Lnc2cancer (see Table 1). For instance, H19 has been found to be expressed in lower quantities in prostatic carcinoma cell lines (Du-145 and PC-3) than in normal cell lines [35]. Besides, the only unsupported lncRNA was BOK-AS1, which had been reported to be associated with prostate cancer in the literature [36].

The top 10 ovarian cancer-associated candidates ranked by NCPLDA and the corresponding evidence are illustrated in Table 2. The results showed 8 of the top 10 potential lncRNAs were found in public resources. As reported in [37], lncRNA-HOTAIR plays a role in regulating cell invasion, migration, and proliferation of ovarian cancer through promoting the expression of PIK3R3. Moreover, we found that lncRNA-HOTAIR is also differentially expressed [38].

For the second type of study, we constructed the NCPLDA model by eliminating all known association information of the watched disease from the DS3 dataset and then used the model to predict disease-lncRNA interactions. The top 10 predicted lncRNAs for prostate cancer and ovarian cancer and the evidence supporting these candidates are described in Table 3 and Table 4, respectively. As the results show, all of the top 10 predictions for the two investigated diseases were verified as true by the DS3 dataset and/or the other two databases. Moreover, with all known disease-lncRNA associations of each disease considered as positive samples, NCPLDA was used to predict its interactions and achieved an average AUC score of 0.8951, suggesting that our method can predict latent disease-lncRNA interactions for new diseases with confidence.

TABLE 1. The top 10 lncRNA candidates predicted by NCPLDA for prostate cancer.

Rank	lncRNAs	Evidences
1	H19	Lnc2cancer, MNDR
2	HOTAIR	Lnc2cancer, MNDR
3	MALAT1	Lnc2cancer, MNDR
4	PVT1	Lnc2cancer, MNDR
5	MEG3	Lnc2cancer, MNDR
6	BOK-AS1	PMID: 25257554
7	XIST	Lnc2cancer, MNDR
8	CDKN2B-AS1	MNDR
9	GAS5	Lnc2cancer, MNDR
10	PCAT1	MNDR

TABLE 2. The top 10 lncRNA candidates predicted by NCPLDA for ovarian cancer.

Rank	lncRNAs	Evidences
1	HOTAIR	Lnc2cancer, MNDR
2	MALAT1	Lnc2cancer, MNDR
3	MEG3	Lnc2cancer
4	CDKN2B-AS1	MNDR
5	HOST2	Lnc2cancer, MNDR
6	GAS5	Lnc2cancer, MNDR
7	XIST	Lnc2cancer, MNDR
8	TUSC8	Unknown
9	BOK-AS1	Unknown
10	UCA1	Lnc2cancer, MNDR

TABLE 3. The top 10 lncRNA candidates predicted by NCPLDA for prostate cancer by hiding all association information of the watched disease from the DS3 dataset.

Rank	lncRNAs	Evidences
1	H19	Lnc2cancer, MNDR
2	MALAT1	Lnc2cancer, MNDR
3	HOTAIR	Lnc2cancer, MNDR
4	XIST	Lnc2cancer, MNDR
5	BOK-AS1	PMID: 25257554
6	PVT1	Lnc2cancer, MNDR
7	MEG3	Lnc2cancer, MNDR
8	MIR17HG	Lnc2cancer
9	CDKN2B-AS1	MNDR
10	GAS5	Lnc2cancer, MNDR

TABLE 4. The top 10 lncRNA candidates predicted by NCPLDA for ovarian cancer by hiding all association information of the watched disease from the DS3 dataset.

Rank	lncRNAs	Evidences
1	H19	DS3, Lnc2cancer
2	HOTAIR	Lnc2cancer, MNDR
3	MALAT1	Lnc2cancer, MNDR
4	DNM3OS	DS3
5	CDKN2B-AS1	MNDR
6	BCYRN1	DS3, Lnc2cancer, MNDR
7	MEG3	Lnc2cancer
8	PVT1	DS3, Lnc2cancer, MNDR
9	HOST2	Lnc2cancer, MNDR
10	XIST	Lnc2cancer, MNDR

IV. CONCLUSION

Exploring disease-lncRNA relationships is not only the key to deciphering the mechanism of lncRNA-influencing diseases, but it is also important for curing diseases. In this study, we introduced a novel method, NCPLDA, for disease-lncRNA interaction prediction based on network consistency projection. In NCPLDA, the lncRNA space projection score and the disease space projection score were combined to compute the relevance score of each candidate lncRNA-disease pair. Compared with three previous methods, NCPLDA had higher accuracy in terms of AUC for three tested datasets. We also implemented two types of case studies, on prostate cancer and ovarian cancer, and found that more than 80% of lncRNA candidates in their top 10 predictions were validated by previous experimental reports. These results imply that NCPLDA is a promising tool for discovering more underlying disease-lncRNA associations.

The good prediction performance obtained by NCPLDA could be due to several reasons. First of all, we utilized the nearest-neighbor information to construct an intermediate association probability matrix, which could fill the incompleteness and sparsity of known associations. Secondly, NCPLDA could fully make use of the integrated similarity data of both the diseases and the lncRNAs, which can further enhance its detection results and make it applicable to isolated nodes. In the end, NCPLDA could mine hidden lncRNAs for all queried diseases on a large scale as a global ranking model.

Despite the commendable results obtained by NCPLDA, there are also several limitations that need to be further investigated. For example, NCPLDA depends on the quality of lncRNA-similarity and disease-similarity matrices. More biological information about lncRNAs and disease, such as gene-lncRNA and gene-disease interactions, could be integrated to further expand the model. Furthermore, we simply treated the disease space projection and the lncRNA space projection as being of equal importance, which might not be optimal. In addition, there are merely hundreds of known available disease-lncRNA interactions. The prediction accuracy of NCPLDA could be further enhanced when more experimentally verified relationships between diseases and lncRNAs are confirmed.

REFERENCES

[1] S. Washietl, M. Kellis, and M. Garber, "Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals," *Genome Res.*, vol. 24, no. 4, pp. 616–628, Apr. 2014.

[2] 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, Mar. 2009.

[3] 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, Mar. 2013.

[4] M. Huarte, "The emerging role of lncRNAs in cancer," *Nature Med.*, vol. 21, no. 11, pp. 1253–1261, Nov. 2015.

[5] A. M. Schmitt and H. Y. Chang, "Long noncoding RNAs in cancer pathways," *Cancer Cell*, vol. 29, no. 4, pp. 452–463, Apr. 2016.

[6] Q. Xiao *et al.*, "Identifying lncRNA and mRNA co-expression modules from matched expression data in ovarian cancer," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published. doi: 10.1109/TCBB.2018.2864129.

[7] W. Lan, L. Huang, D. Lai, and Q. Chen, "Identifying interactions between long noncoding RNAs and diseases based on computational methods," in *Computational Systems Biology* (Methods in Molecular Biology), vol. 1754. New York, NY, USA: Humana Press, 2018, pp. 205–221.

[8] 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 Bioinf.*, vol. 18, no. 4, pp. 558–576, Jul. 2017.

[9] 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, Oct. 2013.

[10] 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, Apr. 2014.

[11] X. Chen, Z.-H. You, G.-Y. Yan, and D.-W. Gong, "IRWRLDA: Improved random walk with restart for lncRNA-disease association prediction," *Oncotarget*, vol. 7, no. 36, pp. 57919–57931, Sep. 2016.

[12] P. Ping, L. Wang, L. Kuang, S. Ye, M. F. B. Iqbal, and T. Pei, "A novel method for lncRNA-disease association prediction based on an lncRNA-disease association network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 2, pp. 688–693, Mar./Apr. 2018.

[13] 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, Jan. 2014, Art. no. e87797.

[14] L. Ding, M. Wang, D. Sun, and A. Li, "TPGLDA: Novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph," *Sci. Rep.*, vol. 8, no. 1, Jan. 2018, Art. no. 1065.

[15] 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.

[16] 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, Jan. 2014, Art. no. e84408.

[17] S. Alaimo, R. Giugno, and A. Pulvirenti, "ncPred: ncRNA-disease association prediction through tripartite network-based inference," *Frontiers Bioeng. Biotechnol.*, vol. 2, Dec. 2014, Art. no. 71.

[18] T. Mori, H. Ngouv, M. Hayashida, T. Akutsu, and J. C. Nacher, "ncRNA-disease association prediction based on sequence information and tripartite network," *BMC Syst. Biol.*, vol. 12, Apr. 2018, Art. no. 37.

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

[20] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncRNA–disease associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529–1537, May 2017.

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

[22] 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, Jul. 2018, Art. no. 345.

[23] X. Xiao *et al.*, "BPLLDA: Predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network," *Frontiers Genet.*, vol. 9, Oct. 2018, Art. no. 411.

[24] W. Lan *et al.*, "LDAP: A Web server for lncRNA-disease association prediction," *Bioinformatics*, vol. 33, no. 3, pp. 458–460, Feb. 2017.

[25] G. Chen *et al.*, "LncRNADisease: A database for long-non-coding RNA-associated diseases," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D983–D986, Jan. 2013.

[26] X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Sci. Rep.*, vol. 5, Jun. 2015, Art. no. 11338.

[27] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, Jan. 2018.

[28] H. Zhao *et al.*, "Prediction of microRNA-disease associations based on distance correlation set," *BMC Bioinf.*, vol. 19, Art. no. 141, Apr. 2018.

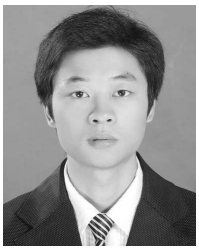
[29] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, Jul. 2010.

[30] G. Yu and L. G. Wang, "Disease ontology semantic and enrichment analysis," Bioconductor version 3.9, Oct. 2014. [Online]. Available: https://www.bioconductor.org/packages/release/bioc/html/DOSE.html
[31] G. Li, J. Luo, Q. Xiao, C. Liang, and P. Ding, "Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity," *J. Biomed. Inform.*, vol. 82, pp. 169–177, Jun. 2018.
[32] C. Gu, B. Liao, X. Li, and K. Li, "Network consistency projection for human miRNA-disease associations inference," *Sci. Rep.*, vol. 6, Oct. 2016, Art. no. 36054.
[33] Y. Wang *et al.*, "Mammalian ncRNA-disease repository: A global view of ncRNA-mediated disease network," *Cell Death Disease*, vol. 4, no. 8, Aug. 2013, Art. no. e765.
[34] S. Ning *et al.*, "Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D980–D985, Jan. 2016.
[35] H. Song *et al.*, "Long non-coding RNA expression profile in human gastric cancer and its clinical significances," *J. Transl. Med.*, vol. 11, no. 1, Sep. 2013, Art. no. 225.
[36] H. Chen *et al.*, "Cisplatin and paclitaxel target significant long noncoding RNAs in laryngeal squamous cell carcinoma," *Med. Oncol.*, vol. 31, no. 11, Sep. 2014, Art. no. 246.
[37] L. Dong and L. Hu, "HOTAIR promotes proliferation, migration, and invasion of ovarian cancer SKOV3 cells through regulating PIK3R3," *Med. Sci. Monitor, Int. Med. J. Exp. Clin. Res.*, vol. 22, pp. 325–331, Jan. 2016.
[38] A. R. Özeş, Y. Wang, X. Zong, F. Fang, J. Pilrose, and K. P. Nephew, "Therapeutic targeting using tumor specific peptides inhibits long non-coding RNA HOTAIR activity in ovarian and breast cancer," *Sci. Rep.*, vol. 7, no. 1, Apr. 2017, Art. no. 894.

GUANGHUI LI received the Ph.D. degree in computer science from Hunan University, in 2015. He is currently an Assistant Professor with the School of Information Engineering, East China Jiaotong University. His research interests include data mining and computational biology.

JIAWEI LUO received the Ph.D. degree in computer science from Hunan University, in 2008, where she is currently a Professor with the College of Computer Science and Electronic Engineering. She has authored about 50 research papers in various international journals and the proceedings of conferences. Her research interests include graph theory, data mining, computational biology, and bioinformatics.

CHENG LIANG received the Ph.D. degree in computer science from Hunan University, in 2015. She was with the Donnelly Centre, University of Toronto, from 2012 to 2014, as a joint Ph.D. Student. She is currently an Assistant Professor with the College of Information Science and Electronic Engineering, Shandong normal University. Her research interests include data mining and computational biology.



QIU XIAO received the Ph.D. degree in computer science from Hunan University, in 2017. He is currently an Assistant Professor with the College of Information Science and Engineering, Hunan Normal University. His research interests include data mining and computational biology.

PINGJIAN DING (S'16) received the master's degree in computer science from Hunan University, in 2015, where he is currently pursuing the Ph.D. degree with the College of Computer Science and Electronic Engineering. His research interests include data mining and computational biology.

YUEJIN ZHANG received the Ph.D. degree in biomedical engineering from the Huazhong University of Science and Technology, in 2017. He is currently a Professor with the School of Information Engineering, East China Jiaotong University. He has authored about 20 research papers in various international journals and the proceedings of conferences. His research interests include computational biology, bioinformatics, mechanical biotechnologies, computer application technology, and image processing technology.