Integrating Multiple Heterogeneous Networks for Novel LncRNA-Disease Association Inference

Jingpu Zhang, Zuping Zhang, Zhigang Chen, and Lei Deng

Abstract—Accumulating experimental evidence has indicated that long non-coding RNAs (lncRNAs) are critical for the regulation of cellular biological processes implicated in many human diseases. However, only relatively few experimentally supported lncRNA-disease associations have been reported. Developing effective computational methods to infer lncRNA-disease associations is becoming increasingly important. Current network-based algorithms typically use a network representation to identify novel associations between lncRNAs and diseases. But these methods are concentrated on specific entities of interest (lncRNAs and diseases) and they do not allow to consider networks with more than two types of entities. Considering the limitations in previous computational methods, we develop a new global network-based framework, LncRDNetFlow, to prioritize disease-related lncRNAs. LncRDNetFlow utilizes a flow propagation algorithm to integrate multiple networks based on a variety of biological information including lncRNA similarity, protein-protein interactions, disease similarity, and the associations between them to infer lncRNA-disease associations. We show that LncRDNetFlow performs significantly better than the existing state-of-the-art approaches in cross-validation. To further validate the reproducibility of the performance, we use the proposed method to identify the related lncRNAs for ovarian cancer, glioma, and cervical cancer. The results are encouraging. Many predicted lncRNAs in the top list have been verified by the biological studies.

Index Terms—LncRNA-disease, flow propagation, heterogeneous network

1 Introduction

In the past few years, accumulating studies have demonstrated that over 90 percent of the eukaryotic genomes are transcribed into numerous of small and long non-coding RNAs (such as miRNA, siRNA, piRNA) [1], [2], [3], [4], [5]. Long non-coding RNAs (lncRNAs), usually defined as >200 nucleotides (nt) in length, make up a class of important ncRNAs [6], [7]. Evidence of regulation has shown that lncRNAs may play an essential role in various biological processes, such as epigenetic regulation, cell cycle control, nuclear and cytoplasmic trafficking, splicing, cell differentiation, and others [2], [5], [8], [9], [10], [11]. However, many lncRNAs do not have any precise functional annotations even now, and there are still many gaps about the function of long non-coding RNA in our current understanding. The further studies on long non-coding RNAs could help to explain the biological role and illustrate the contribution of those mutations in lncRNA genes to the pathogenesis of disease [5].

Growing evidences based on biological experiments have revealed that the distinct types of mutations and dysregulations in lncRNA genes are correlated with a broad range of diverse human diseases [3], [5], [6], such as cardiovascular diseases [12], neurodegenerative disorders [13] and various kinds of cancers [14]. For example, the expression levels of the lncRNA H19 are remarkably associated with liver cancer, bladder cancer and pancreatic cancer [15], [16], [17]. However, the precise mechanism of action for lncRNAs which contributes to the pathogenesis of disease remains unclear. More recently, researchers have made great effort to identify the relationships between lncRNAs and diseases. The studies may not only help us interpret molecular mechanisms of human diseases, but also facilitate biomarker identification for human disease diagnosis, treatment and prevention at lncRNA level [18]. A number of databases, such as lncRNAdb [19], NONCODE [20], LNCipedia [21], have been developed to store the generating lncRNA-related biological data including lncRNA sequence, expression profiles, genomic annotations and so on. But, only a few lncRNA-disease associations have been confirmed by experiments and reported openly. Therefore, it would be necessary to develop powerful computational methods based on the available datasets to predict potential associations between lncRNAs and diseases.

In recent years, several prediction methods to infer the disease-related lncRNAs have been developed. Yang et al. [22] built a coding-non-coding disease-gene bipartite network using known disease-gene associations.Two relevant biological networks, lncRNA-disease network (lncDN) and disease-

- *J. Zhang is with the School of Information Science and Engineering, Central South University, Changsha 410083, China, and the School of Computer (Software), Ping Ding Shan University, Pingdingshan 467000, China. E-mail: zhangjp@csu.edu.cn.*
- *Z. Zhang is with the School of Information Science and Engineering, Central South University, Changsha 410083, China. E-mail: zpzhang@csu.edu.cn.*
- *Z. Chen is with the School of Software, Central South University, Changsha 410075, China. E-mail: czg@csu.edu.cn.*
- *L. Deng is with the School of Software, Central South University, Changsha 410075, China, and the Shanghai Key Laboratory of Intelligent Information Processing, Shanghai 200433, China. E-mail: leideng@csu.edu.cn.*

Manuscript received 30 Aug. 2016; accepted 1 May 2017. Date of publication 4 May 2017; date of current version 29 Mar. 2019.
(Corresponding author: Lei Deng.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2017.2701379

associated lncRNA network (DlncN), were derived. Then a propagation algorithm was used to uncover the potential lncRNA-disease associations in the network. The method greatly depends on the number of edges in the bipartite network. Therefore, it can't effectively predict potentially related lncRNAs for the diseases without known related lncRNA records. Based on the hypothesis that related diseases are usually linked with lncRNAs which have similar functions, Chen et al. [23] applied the model of Laplacian Regularized Least Squares for lncRNA-Disease Association (LRLSLDA) to predict potential disease-related lncRNAs. Chen et al. [24] proposed two lncRNA functional similarity estimate models (LNCSIM) based on the assumption that the functional similarity between two lncRNAs can be obtained by quantitatively calculating the similarity of the associated two groups of diseases. Then, the lncRNA functional similarity was introduced into the model of LRLSLDA for lncRNA-disease association prediction. Some researchers used the model of random walk to infer potential lncRNA-disease associations by combining three networks (lncRNA-lncRNA similarity network, disease-disease similarity network and known lncRNA-disease association network) into an integrated network [25], [26], [27]. The methods based on the random walk are sensitive to the cutoff values of lncRNA and disease similarity. So, the performance will fluctuate wildly when the cutoff values change. Recently, Biswas et al. [28] used nonnegative matrix factorization algorithm to infer disease associations of the lncRNAs. A multi-label classification framework was used to predict disease associations of lncRNAs by integrating secondary-based features and composition-based features [29]. For the secondary structures of non-coding RNA were predicted by CentroidFold package, the overall predictive performance and effectiveness rely on the correctness of the secondary structures and the extracting of new features. Chen adopted the KATZ measure for lncRNA-Disease Association prediction in the heterogeneous network [30]. Most of these approaches cannot be applied to more than two types of networks, and these methods do not allow to integrate information of multiple types of biological entities to improve the performance.

In this paper, in order to predict potential lncRNA-disease associations, we assume that functionally similar lncRNAs tend to be associated with phenotypically similar diseases [23], [24], [25]. On the other hand, the "Gilt-by-association" (GBA) principle states that biological entities having the same or related behavior tend to be related. Based the above assumption and the GBA principle, we propose a novel computational framework, LncRDNetFlow, to infer potential lncRNA-disease associations. LncRDNetFlow is based on a generic network-based prioritization model [31], [32] that integrating and propagating information in an arbitrary number of heterogeneous data networks, including an lncRNA similarity network, a protein-protein interaction network, a disease similarity network and interactions or association networks between the three biological entity networks. We show that the prioritization model with multiple heterogeneous networks can boost the performance of lncRNA-disease association prediction according to leave-one-out and 5-fold cross validation. Furthermore, LncRDNetFlow is more robust than other existing approaches for a wide range of parameter values.

2 METHOD

The flowchart of LncRDNetFlow is illustrated in Fig. 1. Based on multiple heterogeneous data sources, three similarity/interaction networks (lncRNA, disease and protein) and three different interconnected association networks (lncRNA-disease, disease-protein and lncRNA-protein) are built. Then the global network is constructed by integrating the heterogeneous networks. LncRDNetFlow employs heterogeneous networks of interaction or similarities between biological entities (e.g., diseases, proteins, lncRNAs) to prioritize the nodes in the networks for a given query set. A flow propagation algorithm considering the network topological information is implemented to compute the global distance measurements and predict the potential lncRNA-disease associations.

The global network is regarded as an undirected graph $G_i = (V_i, E_i)$ where V_i is a set of nodes and E_i is a set of weighted undirected edges. Each node in the network represents a biological entity (e.g., a lncRNA, a disease or a protein) and each weighted edge represents a relationship, similarity or interaction between the connected pair of entities. There are two types of edges: Edges which connect the nodes of the same type (e.g., lncRNA-lncRNA networks, disease-disease networks) and edges which connect the nodes of two different types (e.g., lncRNA-disease associations, protein-disease associations).

Assume W is the adjacency matrix of a network in global network G. Considering the network topological information [33], [34], [35], [36], we first normalize W as: $W' = D_G^{-1/2} W D_G^{-1/2}$, where D_G is a diagonal matrix such that $D_G(i,i)$ is the sum of row i of W, namely $D_G(i,i) = \sum_j W_{i,j}$. Therefore, W' is a symmetric matrix where $W'(i,j) = W(i,j)/\sqrt{D(i,i)D(j,j)}$.

We define the problem of lncRNA-disease association prediction as measurement of association between a set of nodes Q called the query set (e.g., a set of diseases of interest) and a set of nodes T called the target set (e.g., a set of lncRNAs of interest) in the global network. The initial values for the nodes in the query and target sets are set to 1, and the others in global network G are set to 0. We calculate the correlation between the query set Q which are as the input of the algorithm and the target set T by performing the information flow propagation algorithm. The obtained correlation value measures the degree of association between the query set and the target set, namely a set of diseases of interest and a set of lncRNAs of interest in our problem. There is maybe more than one path which connects the query network to the target network. For example, there exist two paths in the scenario of inferring lncRNA-disease associations, as is illustrated in Fig. 2. The information is transmitted in each path. Therefore, the propagation process includes two steps which are "propagation inside a network" and "propagation between adjacent networks" respectively, and the two propagation steps are performed alternately. Fig. 3 shows an example of propagation process including two steps which are represented by the green and purple dashed arrow lines, respectively. The method of "propagation inside a network" utilizes a network's global information to propagate the values to the nodes inside the network [33], [34], [37], [38]. For a node v in the network, we

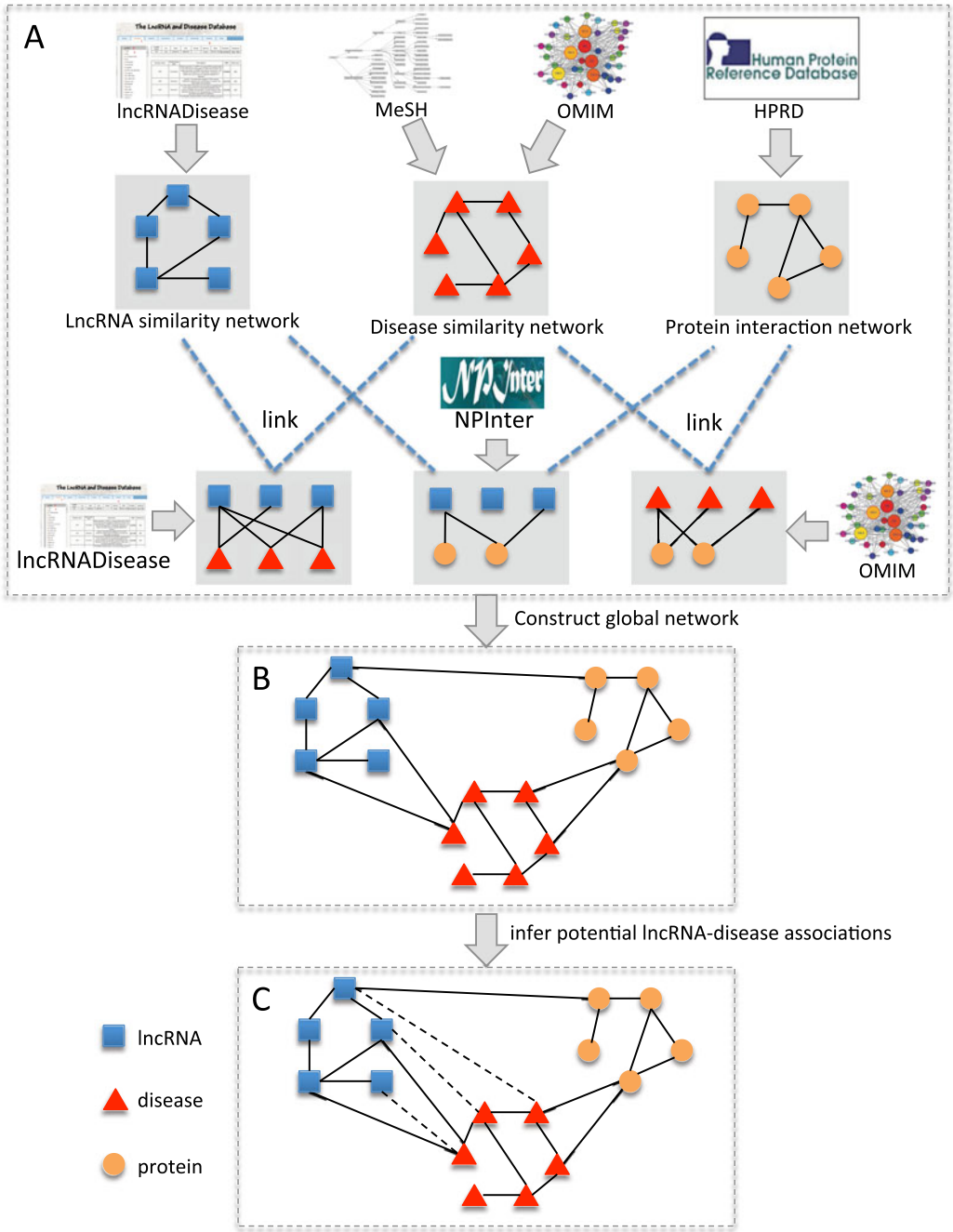


Fig. 1. Flowchart of LncRDNetFlow. (A) Three similarity/interaction networks (lncRNA, disease, and protein) and three different interconnected association networks (lncRNA-disease, disease-protein, and lncRNA-protein) are built based on multiple heterogeneous data sources. (B) The global network is constructed by integrating the six heterogeneous networks. (C) The potential lncRNA-disease associations are predicted with a flow propagation algorithm. The known lncRNA-disease associations are represented as the solid edge, and the dashed edges represent the predicted potential lncRNA-disease associations.

define a function F that is both smooth over the network and also respects the prior knowledge, and express the requirements on F as a combination of the two conditions

$$F(v) = a\left[\sum_{u \in N(v)} F(u)w'(v,u)\right] + (1-a)Y(v).$$

Here, w' is a normalized matrix whose values are given by the adjacency matrix of the network. $\alpha \in (0,1)$ is a balancing parameter which determines the importance of the prior information in the network. Y represents a prior knowledge function and $N(v)$ denotes the direct neighborhood of v. The requirements on F can be expressed in linear form as follows:

$$F = \mathrm{a}W'F + (1-a)Y \Leftrightarrow F = (I - aW')^{-1}(1-a)Y.$$

An iterative algorithm can efficiently calculate the closed-form solution of the method with the following update rule at each time step i,

$$F^i = aW'F^{i-1} + (1-a)Y,$$

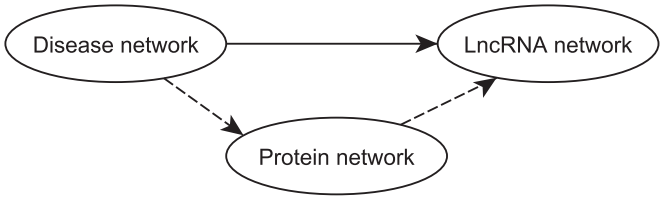


Fig. 2. Illustration of paths in global heterogeneous network. There are two paths: One path represented by the solid arrow line and the other represented by the dashed arrow line. The latter connects the disease network and the lncRNA network via the protein network.

where $F^1 = Y$. The update rule can be looked upon as simulating a process where nodes having prior information pump the information to their neighbors. Furthermore, each node transmits the information received in the previous iterative step to its neighbors. In Fig. 3, the node in query set is initially set to 1, then each node in the query network is assigned a value by the method of "propagation inside a network".

In the "propagation between networks", the value of each node n from the next network are given by the nodes from the current network which are directly connected to the former by the formula as follows:

$$\varphi(n) = \frac{\sum_{x \in neigbor(n)} \varphi(x)}{|neigbor(n)|},$$

where $neigbor(n)$ is the set of nodes in the current network which are directly connected to the node n, $\varphi(x)$ and $\varphi(n)$ are the values of node x and n respectively. By the propagation process between adjacent networks, the information will flow from the current network to the following network in one path. In Fig. 3, the two nodes in upper left corner in the network represented by red triangles are assigned values by the method of "propagation between networks", then the propagation inside the network is performed.

Assume there are l different paths in the global heterogeneous network which connect the query network to the target network. The propagation process along the l paths does not end until the nodes from all the networks next to the target network derive information. After the propagation process finishing through one path, the nodes from the network next to the target network in the path are given values. The values of the nodes measure their degree of relationship to the query set. Then, the values expressed in terms of a vector are multiplied by the normalized adjacency matrix of the association between the adjacent network and the target network. L vectors are obtained when the propagation process ends through l different paths. We concatenate the l vectors into one vector $\hat{y}$. All the nodes in the target network are given values expressed as a vector t by the operation of propagation inside the target network. We also concatenate the vector t l times into one vector $\hat{t}$. The degree of relationship between the query set and the target set is measured by correlating vectors $\hat{y}$ and $\hat{t}$

$$s = corr(\hat{y}, \hat{t}),$$

where corr is Pearson's Correlation. It is worth to point out that the correlation between diseases and lncRNAs measures the topological similarity in the networks, not the absolute values of correlation coefficient.

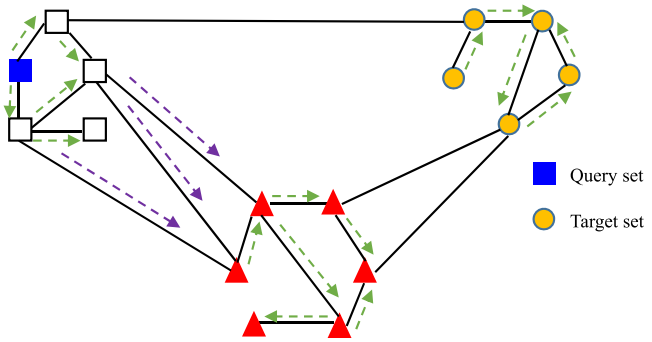


Fig. 3. Example of information flow propagation in the global heterogeneous network. The solid lines represent edges, and the dashed arrow lines denote the process of information propagation. The global network involves three networks with the elements of each network represented by squares, triangles, and circles.

For a given disease regarded as a query set in the query network, we compute the correlation value between the disease and each lncRNA, i.e., a target set, utilizing the above method in turn. The all lncRNAs in the target network are sorted by the correlation values in decreasing order after the values are obtained. The greater the values, the stronger association between the disease and the lncRNAs. The pseudocode is described in Algorithm 1.

Algorithm 1. Network Flow Propagation Algorithm of LncRDNetFlow

Require: G: Global network, Q: query set, G_q: Query network, G_t: Target network
Ensure: R: Priorizing list
  propagate values within G_q
  t: Propagate values within G_t
  L: compute the list of paths from G_q to G_t in G
  for each path $l_i = \{l_{i1}, \ldots, l_{ij}, \ldots, l_{ik}\}$ in L
    for each network l_{ij} in the path l_i from l_{i1} to $l_{i(k-1)}$
      propagate values from l_{ij} to $l_{i(j+1)}$
      propagate values within $l_{i(j+1)}$
    end for
      store the values of $G_{i(k-1)}$ after propagation through path l_i as $\hat{y}_{i(k-1)}$
  end for
  $\hat{t}$: Concatenate $|L|$ times the vector t
  Compute correlation coefficient $s = corr(\hat{y}, \hat{t})$
  R: Sort all entities $e \in G_t$ by the s values in descending order
  Return R

3 Results

In this section, we first discuss the network preparation and the construction of lncRNA similarity network. Then, different validation tests are employed to evaluate the performance of our method: a leave-one-out cross validation (LOOCV) and a 5-fold cross validation applied with two different similarity processing of lncRNA and disease. Also, LncRDNetFlow is tested on two different network configurations: 1) consider lncRNA entity network and disease entity network (LncRDNetFlow-2N); 2) consider lncRNA entity network, protein entity network and disease entity network (LncRDNetFlow-3N). Finally, we demonstrate the robustness of the approach.

3.1 Datasets and Pre-Processing

3.1.1 LncRNA-Disease Associations

We download 1,102 experimental lncRNA-disease association dataset from the LncRNADisease database (http://cmbi.bjmu.edu.cn/lncrnadisease) in December, 2015 [18]. The dataset provides a gold standard resource of experimentally confirmed lncRNA-disease associations which were manually collected. There exist the same lncRNA-disease association based on different experimental evidences. Also, some non-human lncRNA-disease associations are included in the database. We obtained 392 different high-quality experimentally validated lncRNA-disease associations among 178 lncRNAs and 169 diseases to construct lncRNA-disease association network after removing the duplicate and non-human entries.

3.1.2 Disease Similarity Matrix

We construct the disease-disease similarity matrix by utilizing phenotype information. But the disease names in the lncRNA-disease associations downloaded from the LncRNADisease database are not normative. Therefore, we can't find the index by using standard naming (e.g., ENSEMBL, RefSeq etc.). We make a list of all the disease names from the lncRNA-disease associations, then utilize an OMIM API function call [39] to obtain closely matched phenotype IDs (i.e., the IDs prefixed with character percent, # or none) [28]. The disease names that can not be matched with any valid OMIM phenotype ID are deleted. Also, the lncRNA-disease associations related to the diseases are removed at the same time. The disease-disease similarity is extracted according to the matrix calculated with use of text mining method developed by van Driel and his collaborators [40]. Finally, of all the disease names in the lncRNA-disease associations, 169 diseases are successfully mapped into the matrix and the disease phenotype network is composed of 5,080 vertexes. The disease-disease similarity values represent the edge weights of disease phenotype network. As suggested by Oron Vanunu et al. [41], the similarity values in the range [0, 0.3] are not informative, while the similarity scores falling in the range [0.6, 1] means informative similarity which denotes potentially relevant phenotypic similarity [42]. Two measures are taken to deal with the raw similarity matrix. 1) By setting a similarity score cutoff (e. g. 0.3 or 0.4), we update the similarity scores to 0 when they are less than the threshold. 2) The disease phenotype network is pruned by taking the five nearest neighbors (5 nn) of each node [33].

3.1.3 LncRNA Functional Similarity Matrix

Since functionally similar lncRNAs tend to be correlated with phenotypically similar diseases [23], [24], [25], [43], we quantitatively measure functional similarity between lncRNAs by calculating the phenotypic similarity of their related two sets of diseases. The calculation of lncRNA functional similarity is elucidated by the example of the similarity calculation between lncRNA1 and lncRNA2. Assuming that lncRNA1 is associated with a group of m diseases denoted by $D_1 = \{d_1, d_2, \ldots d_m\}$ and lncRNA2 is associated with a group of n diseases denoted by $D_2 = \{d_1, d_2, \ldots d_n\}$, the similarity score between one disease d and one group of k diseases D is computed as follows:

$$Sim(d, D) = \max_{1 \le i \le k}(Sim(d, d_i)),$$

where $d_i \in D$. $Sim(d, D)$ is the maximum similarity score between one disease d and one group of k diseases D. The functional similarity of lncRNA1 and lncRNA2 is defined as follows:

$$LncRSim(\ln cRNA1, \ln cRNA2) = \frac{\sum_{1 \le i \le m} Sim(d_{1i}, D_2) + \sum_{1 \le j \le n} Sim(d_{2j}, D_1)}{m + n},$$

where $d_{1i} \in D_1$ and $d_{2j} \in D_2$. LncRsim represents the lncRNA functional similarity matrix. We also construct the lncRNA function similarity matrix according to the disease similarity matrix differently pre-processed, respectively.

There are 178 lncRNAs in total according to the downloaded 392 distinct lncRNA-disease associations and the lncRNAs are looked upon as vertexes in the lncRNA similarity network.

3.1.4 Protein Interaction Network and Gene-Disease Relationship

We construct the protein-protein interaction (PPI) network from the Human Protein Reference Database (HPRD) [44]. The PPI network has 8,919 proteins and 32,331 interactions between the proteins. The edges in the interaction network are unweighted, hence Element $G(i, j)$ at row i and column j of G is 1 if gene i and gene j interacts with each other, otherwise $G(i, j)$ is 0. The gene-disease associations represented by an unweighted undirected graph with edges connecting disease vertexes with their causative gene vertexes is obtained from OMIM. The matrix constructed according to the gene-disease associations is an 8,919*5,080 matrix containing 1,393 relationships.

3.1.5 LncRNA-Gene Interactions

The NPInter database (http://www.bioinfo.org/NPInter/) contains experimentally verified functional interactions between noncoding RNAs (excluding tRNAs and rRNAs) and biomolecules (proteins, RNAs and DNAs) [45]. The interactions from many different species (including Homo sapiens, Mus musculus, etc.) are collected in the database. We download the dataset and filter to focus on the interactions between the lncRNAs from the lncRNA-disease associations and the gene (proteins) from the PPI network. Finally, we extract 1,052 interactions between 178 lncRNAs and 8,919 genes.

3.1.6 Construction of the lncRNA Functional Similarity Network

We design a new global method based on multiple heterogeneous entity networks, which integrates protein-protein interaction information, to infer potential lncRNA-disease associations. In these entity networks, the lncRNA functional similarity network is constructed according to the phenotypic similarity of their associated disease groups. The number of edges in the network are subjected to the similarity threshold β. Namely, The edges will decrease

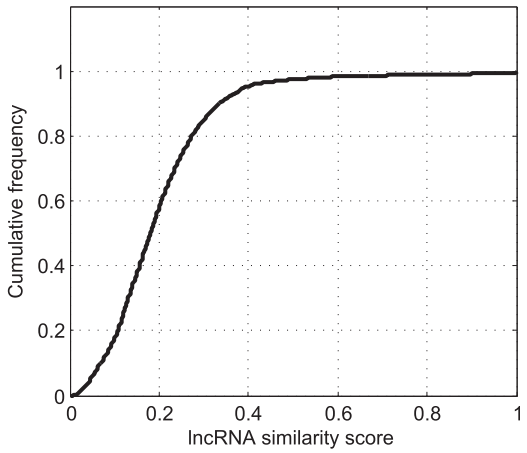


Fig. 4. Cumulative distribution of the edges in the lncRNA similarity network when the similarity cutoff varies. The curve starts to slow when the cutoff is 0.4, and the curve is relative stable when the cutoff is larger than 0.5.

when the similarity threshold β increase. In order to evaluate the influence of the threshold β when lncRNA similarity scores are calculated according to the raw disease similarity matrix, we calculated the cumulative distribution and found that the number of the edges is relatively stable when the cutoff β is set to equal or larger 0.5 (Fig. 4). Hence, we select 0.5 as the cutoff to construct the lncRNA similarity network, and there will be an edge between two lncRNAs if their similarity score is equal or greater 0.5. We obtain 760 associations altogether between 178 lncRNAs.

3.2 Evaluation Measures

The performance of the proposed method is evaluated by using leave-one-out cross validation (LOOCV) and 5-fold cross validation [46], [47]. LOOCV is implemented based on the 392 experimentally verified lncRNA-disease associations between the 178 lncRNAs and 169 diseases. At each turn an explicit lncRNA-disease association in the global graph is used as a test case while all the remaining lncRNA-disease associations are used as the training data. For the 5-fold cross validation, all the known associations in the global graph are randomly divided into 5 subsets with an approximately equal number of associations. For each time, 4 subsets with known associations are used as training data, and the remaining subset is used as test data.

We use receiver operating characteristics (ROC) curve to measure the performance. The ROC curve is plotted with true positive rates (TPR or sensitivity) as a function of false positive rates (FPR or 1-specificity) for various classification thresholds. For each threshold, the corresponding TPR and FPR are calculated. The receiver operating characteristics curve is created by varying the threshold. The area under the ROC curve (AUC) to measure the overall performance is also calculated.

3.3 The Effect of Changing Network Pre-Processing Characteristics

As described in datasets and pre-processing section, the raw disease similarity matrix and the lncRNA similarity matrix are pre-processed by two measures before using our method. One measure is taken by setting cutoffs for the disease similarity score and the lncRNA similarity score computed according to the raw disease similarity matrix, the other is adopted by extracting the five nearest neighbors (5 nn) of each disease while the similarity calculation of lncRNAs by use of the 5 nn matrix. LOOCV is performed to evaluate the performance of the methods by changing the two categories of network pre-processing characteristics. The cutoff values of disease similarity score are 5 nn, 0.3 and 0.4, respectively. For the lncRNA similarity score, the cutoff values are 0.5, 0.6, 0.1(5 nn), 0.2(5 nn) and 0.3(5 nn), respectively. As shown in Table 1, LncRDNetFlow (LncRDNetFlow-3N/LncRDNetFlow-2N) achieves the highest AUC score when the disease similarity network is built with the five nearest neighbors and the cutoff of lncRNA similarity score is set to 0.5 or 0.6. We select a moderate cutoff combination [Disease-cutoff = 5 nn, lncRNA-cutoff = 0.2(5 nn)] for further comparison.

TABLE 1
The LOOCV AUC Values of lncRDNetFlow (LncRDNetFlow-3N and LncRDNetFlow-2N) Considering Different Pre-Processing and Cutoffs

Disease-cutoff	lncRNA-cutoff	LncRDNetFlow-3N	LncRDNetFlow-2N
5 nn	0.5	0.906	0.894
5 nn	0.6	0.906	0.894
5 nn	0.1(5 nn)	0.849	0.830
5 nn	0.2(5 nn)	0.841	0.821
5 nn	0.3(5 nn)	0.817	0.790
0.3	0.5	0.898	0.889
0.3	0.6	0.896	0.883
0.3	0.1(5 nn)	0.835	0.818
0.3	0.2(5 nn)	0.829	0.814
0.3	0.3(5 nn)	0.806	0.786
0.4	0.5	0.902	0.891
0.4	0.6	0.900	0.886
0.4	0.1(5 nn)	0.844	0.825
0.4	0.2(5 nn)	0.836	0.816
0.4	0.3(5 nn)	0.811	0.787

The similarity scores in lncRNA and disease matrices are pre-processed by different measures. The first column shows the cutoff values of disease similarity score (5 nn, 0.3, and 0.4). The second column lists the cutoff values of lncRNA similarity score according to the full disease similarity matrix (0.5 and 0.6) and the 5 nn matrix of disease [0.1(5 nn), 0.2(5 nn), and 0.3(5 nn)].

3.4 The Benefit of Protein Interaction Network

To demonstrate the influence of considering the protein interaction network for prioritization of disease-related lncRNAs, LncRDNetFlow is tested on two different network configurations: The global network composed of lncRNA and disease entity networks (LncRDNetFlow-2N), and the global network composed of lncRNA, protein and disease entity networks (LncRDNetFlow-3N). We carry out various comparisons considering the different pre-processing of the similarity matrices and the different cutoffs. The AUC scores of the prioritizing for the two different network configurations are calculated. As shown in Table 1, LncRDNetFlow-3N outperforms LncRDNetFlow-2N across all the cutoff settings. The AUC values in LOOCV are 0.841 for LncRDNetFlow-3N and 0.821 for LncRDNetFlow-2N when the similarity scores are pre-processed according to the 5 nn similarity matrix (Fig. 5). The values are 0.838 versus 0.819 in 5-fold cross validation (Fig. 6). The proposed method can

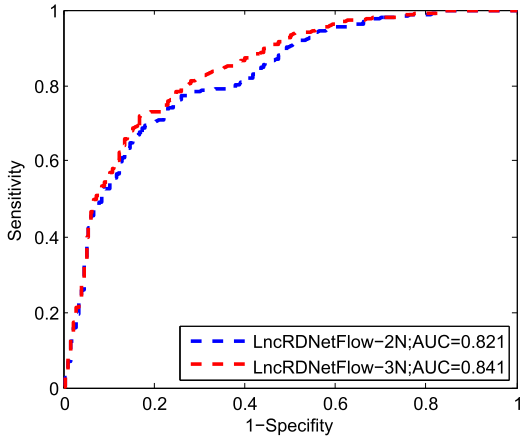


Fig. 5. ROC curves and AUC values for LOOCV tests on two different network configurations.

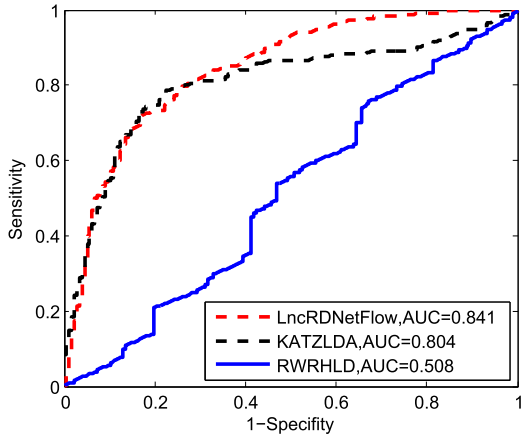


Fig. 7. ROC curves and AUC values of LncRDNetFlow, KATZLDA, and RWRHLD. The ROC curves are plotted and AUC values are computed by LOOCV.

benefit from the addition of protein interaction network. The results prove that there is an improvement in the performance by integrating information about other types of biological entities (e.g., protein).

3.5 Comparison with State-of-the-Art Methods

In order to further evaluate the performance of our method, we compare LncRDNetFlow with the following two state-of-the-art computational methods: RWRHLD [27] and KATZLDA [30]. RWRHLD and KATZLDA are recent methods which were developed to infer potential disease-lncRNA associations. RWRHLD is based on a random walking model on this heterogeneous network, and KATZLDA is based on the model of KATZ measure. Unfortunately, the scripts of the two methods have not been provided. Therefore, we implement these methods and set the parameters to the values suggested by the authors in the corresponding papers. To perform a fair comparison with the two methods, we carried out the following different experiments by using the same dataset downloaded recently (392 different experimentally validated lncRNA-disease associations as described in Section 3.1.1). As shown in Fig. 7, the LncRDNetFlow algorithm outperforms the KATZLDA algorithm in LOOCV, while it outperforms the RWRHLD algorithm enormously. The performance of the RWRHLD algorithm is largely influenced by the number of the edges in the global network. Therefore, there is sharp drop when the edges decrease. Also, LncRDNetFlow is compared with the two methods in the framework of 5-fold cross validation (Fig. 8). The results show that our method achieves the best performance in the term of both LOOCV and 5-fold cross validation.

3.6 Robustness Evaluation

There are two main factors which may impact the performance of our method. The first is the parameter α, which determines the importance of prior information in information propagation. The previous researchers have demonstrated that it hasn't great effect on performance [31], [33], [34]. The similar performance for LncRDNetFlow is obtained when the parameter vary from 0.5 to 0.9, and we choose α = 0.9. The second is the pre-processing of the similarity scores, which controls the number and weight of edges in disease and lncRNA similarity networks. In order to assess its effect on the performance of the method, we take different measures and set different score cutoffs to perform the above LOOCV and 5-fold cross validation. To demonstrate the comparison, a line chart including 15 markers is drawn, as shown in Fig. 9. Each marker represents one pre-processing, which is listed in Table 1. LncRDNetFLow outperform others and the AUC values for different pre-processing of the

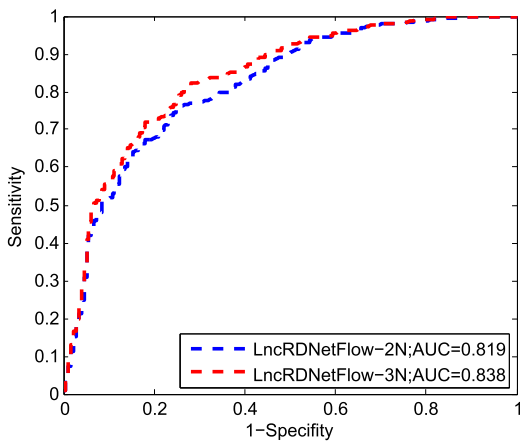


Fig. 6. ROC curves and AUC values for 5-fold cross valiadation tests on two different network configurations.

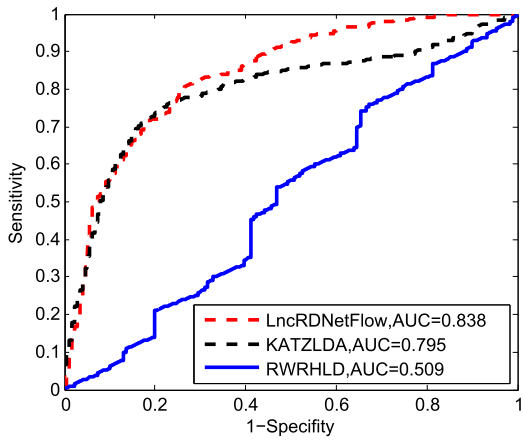


Fig. 8. ROC curves and AUC values of LncRDNetFlow, KATZLDA, and RWRHLD. The ROC curves are plotted and AUC values are computed by 5-fold validation.

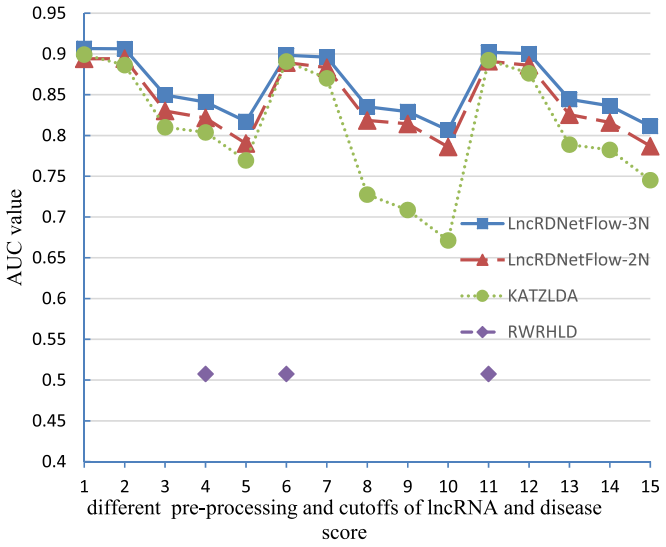


Fig. 9. Comparison of different methods when the similarity scores are differently pre-processed.

similarity scores in LncRDNetFLow are all greater than 0.8. It demonstrates that our LncRDNetFlow can achieve a relatively stable performance, and that the results are robust to different pre-processing.

TABLE 2
The lncRNAs in the Top 10 for Three Case Studies

lncRNA	rank	lncRNA	rank
ovarian cancer (MIM: 60,430)			
DNM3OS#	1	XIST	6
LSINCT5#	2	MALAT1	7
SRA1#	3	BCAR4	8
BCYRN1#	4	HOTAIR	9
H19#	5	MIR31HG	10
glioma (MIM: 137,800)			
ADAMTS9-AS2#	1	H19#	6
CCDC26#	2	CDKN2B-AS1#	7
CDKN2B-AS12#	3	MALAT1	8
anti-NOS2A	4	HOTAIR	9
MEG3#	5	RMST	10
cervical cancer (MIM: 211,980)			
TUSC8#	1	H19#	6
BLACAT1	2	UCA1	7
BCYRN1#	3	MEG3	8
MALAT1#	4	SNHG16	9
HOTAIR#	5	TUG1	10

In the table, the lncRNAs and the rank for three case studies (ovarian cancer, melanoma, and cervical cancer) are listed. LncRNAs related to the disease of interest in our dataset are marked with character '#'.

3.7 Case Studies

To further demonstrate the predictive performance and elucidate the application of our method on real case stuides, we have examined it for lncRNA-disease ranking of three multifactorial diseases (i.e., ovarian cancer (MIM: 60,430), glioma (MIM: 137,800), cervical cancer (MIM: 211,980)). For each case, the lncRNAs ranked in the top 10 are listed in Table 2 and the full list can be obtained from the additional file.

Results for ovarian cancer. Ovarian is the most lethal gynecological cancer without an effective prognosis for women [48], [49]. It accounts for 4 percent of deaths from cancer in women [50]. More and more biological experiments are performed to identify ovarian cancer-related lncRNAs. In LncRNADisease database, there are 6 experimentally validated lncRNAs which are related to ovarian cancer, as shown in Table 2. LncRDNetFow is implemented to infer ovarian cancer-related lncRNAs. As a result, the lncRNAs which have connected to ovarian cancer according to LncRNADisease database are: DNM3OS, LSINCT5, SRA1, BCYRN1, H19, and PVT1. Most of them are ranked in the top 10 of the prioritized list and marked with character '#' except for PVT1 (ranked 14th). Three lncRNAs of the other five lncRNAs have been confirmed by the MNDR database [51] and biological experiment literature. For example, MALAT1 and XIST were validated by MNDR database. HOTAIR plays an import role in proliferation, migration and invasion of ovarian cancer SKOV3 by regulating PIK3R3 [52]. MEG3, ranked 15th, is downregulated in epithelial ovarian cancer by using RT-PCR and western blotting, and is associated with epithelial ovarian cancer [53]. The evidences suggest the relationship between the four lncRNAs and ovarian cancer.

Results for Glioma. Glioma is responsible for the great majority of primary tumors in the brain [54]. Recent studies demonstrate that lncRNAs play significant roles in glioma pathogenesis [55]. Evidence indicates that lncRNAs may regulate certain tumorigenic processes in glioma such as cellular proliferation and apoptosis [56]. Our method is applied to glioma for related lncRNA prediction. In the top-10 ranked lncRANs, previously known lncRNAs related to this disease according to LncRNADisease database are: ADAMTS9-AS2, CCDC26, CDKN2B-AS12, MEG3, H19 and CDKN2B-AS1. MALAT1 and HOTAIR have been confirmed by biological experiments, which are ranked 8th and 9th, respectively. For example, Ma et al. [57] identified the role of MALAT1 and found that MALAT1 was increased in glioma tissues than in paired adjacent brain normal tissues. HOTAIR promotes cell cycle progress in glioma as a result of the binding of its 5′ domain to the PRC2 complex [58]. Other related lncRNAs were found below in the top list, such as TUG1 (ranked 12th) [59].

Results for Cervical cancer. Cervical cancer contributes 10-15 percent of cancer-related deaths in women worldwide, exceeded only by breast cancer [60]. By the widespread using of new technologies, many lncRNAs have been proved to be new regulators in various biological processes and play a vital role in the oncogenesis and progression of cervical cancer and other cancers [61]. By the method of LncRDNetFlow, the LncRNAs related to cervical cancer ranked in the top 10 are listed in Table 2. The lncRNAs related to cervical cancer according to LncRNADisease datasets are TUSC8, BCYRN1, MALAT1, HOTAIR and H19. New relations not definitely represented in the datasets are found in the top of the prioritized list. UCA1 and MEG3 are ranked 7th and 8th respectively. MEG3 are downregulated in cervical cancer tissues, compared to the adjacent normal tissues [62]. Also, MEG3 acts as a tumor suppressor by regulating miR-21-5p, resulting in the inhibition of tumor growth in cervical cancer [62]. UCA1 is

validated by the MNDR database. The expressions of UCA1 are higher in group of cervical intraepithelial neoplasia than those in groups of normal cervical tissues. It may promote the occurrence of cervical intraepithelial neoplasia.

4 Discussion and Conclusion

Inferring novel lncRNA-disease relationships by integrating varieties of biological datasets is emerging as a tool for understanding disease mechanism at the lncRNA level and disease biomarker detection. In this work, we first calculate lncRNA functional similarity by integrating known lncRNA-disease associations and disease phenotypic similarity. Then, LncRDNetFlow, an information propagation flow algorithm, is developed to infer novel lncRNA-disease associations by integrating a variety of biological information including lncRNA similarity, protein-protein interactions, disease similarity and the associations between them. Our method could also be used to validate novel associations between complex diseases and lncRNAs without known associations. In other words, integrating information of proteins and their associations with lncRNAs and diseases enables potential relations to be inferred that could not be obtained using lncRNA-disease associations alone. Furthermore, the method is able to perform both lncRNA-disease and disease-lncRNA prioritization. Namely, it can also be applied to predict lncRNA-related diseases and vice versa.

LOOCV and 5-fold cross validation are implemented based on different pre-processing of similarity scores and different network configurations to evaluate the performance of our method. Results show that LncRDNetFlow gains better performance than RWRHLD and KATZLDA. Also, LncRDNetFlow is reliable and relatively stable when the effects of different parameters and cutoffs are considered. By applying this method to case studies of ovarian cancer, glioma and cervical cancer, we have ranked potential disease-related lncRNAs for these three types of diseases. Some potential lncRNA-disease associations ranked in the top 10 of the prioritized list have been experimentally confirmed in recent studies. In addition, the method enables a new network of biological entities to be integrated into the global network as long as there are connections between the new network and two other networks in the global network. For example, the disease pathways can be integrated into the global network.

It is worth to point out that our method has some biases. First, we have to retrieve the closely matched phenotype since the disease names in the LncRNADisease dataset are not canonical. Therefore, maybe there is a bias against the disease similarity score. Second, the lncRNA functional similarity network is constructed depending on known lncRNA-disease relations. It would affect the performance of the method. Hence, calculating lncRNA similarity by integrating more data would benefit the improvement of predictive ability. In the future, we will add new networks such as microRNA entity network and construct lncRNA similarity including more data integration to predict the potential lncRNA-disease associations.

Acknowledgments

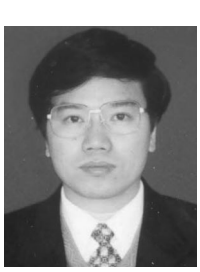
This work was supported by the National Natural Science Foundation of China under grants No. 61672541, No. 61309010, No. 60970095, and No. 61379057, the China Postdoctoral Science Foundation under grant No. 2015T80886, the Specialized Research Fund for the Doctoral Program of Higher Education of China under grant No. 20130162120073, and the Shanghai Key Laboratory of Intelligent Information Processing under grant No. IIPL-2014-002.

References

[1] A. C. Marques and C. P. Ponting, "Catalogues of mammalian long noncoding RNAs: Modest conservation and incompleteness," *Genome Biol.*, vol. 10, no. 11, 2009, Art. no. R124.
[2] G. Mitchell, 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.
[3] C. P. Ponting, P. L. Oliver, and R. Wolf, "Evolution and functions of long noncoding RNAs," *Cell*, vol. 136, no. 4, pp. 629–641, 2009.
[4] Z. Jing, et al., "Genome-wide identification of polycomb-associated RNAs by RIP-SEQ: Molecular cell," *Molecular Cell*, vol. 40, no. 6, pp. 939–953, 2010.
[5] O. Wapinski and H. Y. Chang, "Long noncoding RNAs and human disease," *Trends Cell Biol.*, vol. 21, no. 6, pp. 354–361, 2011.
[6] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: Insights into functions," *Nature Rev. Genetics*, vol. 10, no. 3, pp. 155–159, 2009.
[7] M. Lina, V. B. Bajic, and Z. Zhang, "On the classification of long non-coding RNAs," *RNA Biol.*, vol. 10, no. 6, pp. 924–933, 2013.
[8] I. A. Qureshi, J. S. Mattick, and M. F. Mehler, "Long non-coding RNAs in nervous system function and disease," *Brain Res.*, vol. 1338, no. 2, pp. 20–35, 2010.
[9] J. E. Wilusz, S. Hongjae, and D. L. Spector, "Long noncoding RNAs: Functional surprises from the RNA world," *Genes Develop.*, vol. 23, no. 13, pp. 1494–1504, 2009.
[10] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Human Molecular Genetics*, vol. 15, no. 8, pp. R17–29, 2006.
[11] J. S. Mattick, "The genetic signatures of noncoding RNAs," *Plos Genetics*, vol. 5, no. 4, pp. e1 000 459–x, 2009.
[12] C. Klattenhoff, et al., "Braveheart, a long noncoding RNA required for cardiovascular lineage commitment," *Cell*, vol. 152, no. 3, pp. 570–583, 2013.
[13] Y. Nishimoto, et al., "The long non-coding RNA nuclear-enriched abundant transcript 1_2 induces paraspeckle formation in the motor neuron during the early phase of amyotrophic lateral sclerosis," *Molecular Brain*, vol. 6, no. 1, pp. 1–18, 2013.
[14] G. Yang, X. Lu, and L. Yuan, "LncRNA: A link between rna and cancer," *Biochimica Et Biophysica Acta*, vol. 1839, no. 11, pp. 1097–1109, 2014.
[15] W. Tsang and T. Kwok, "Riboregulator H19 induction of MDR1-associated drug resistance in human hepatocellular carcinoma cells," *Oncogene*, vol. 26, no. 33, pp. 4877–81, 2007.
[16] L. Ming, Z. Li, W. Wei, Y. Zeng, Z. Liu, and J. Qiu, "Upregulated h19 contributes to bladder cancer cell proliferation by regulating id2 expression," *Febs J.*, vol. 280, no. 7, pp. 1709–1716, 2013.
[17] M. Chenchao, et al., "H19 promotes pancreatic cancer metastasis by derepressing let-7s suppression on its target HMGA2-mediated EMT," *Tumor Biol.*, vol. 35, no. 9, pp. 9163–9169, 2014.
[18] G. Chen, et al., "LncRNADisease: A database for long-non-coding RNA-associated diseases," *Nucleic Acids Res.*, vol. 41, pp. D983–D986, 2013.
[19] 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, 2011.
[20] B. Dechao et al., "Noncode v3.0: Integrative annotation of long noncoding RNAs," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D210–D215, 2012.
[21] P. J. Volders, et al., "LNCipedia: A database for annotated human lncRNA transcript sequences and structures," *Nucleic Acids Res.*, vol. 41, pp. D246–D251, 2012.
[22] Y. Xiaofei, 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.
[23] X. Chen and G. Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinf.*, vol. 29, no. 20, pp. 2617–2624, 2013.

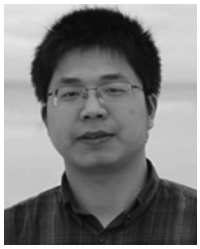
[24] X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Sci. Rep.*, vol. 5, p. 11338, 2015.
[25] S. Jie, et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular Biosyst.*, vol. 10, no. 8, pp. 2074–2081, 2014.
[26] G. U. Ganegoda, M. Li, W. Wang, and Q. Feng, "Heterogeneous network model to infer human disease-long intergenic non coding RNA associations," *IEEE Trans. Nanobiosci.*, vol. 14, no. 2, pp. 175–183, Mar. 2015.
[27] M. Zhou, et al., "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Molecular Biosyst.*, vol. 11, no. 3, pp. 760–769, 2014.
[28] A. K. Biswas, et al., "Inferring disease associations of the long non-coding RNAs through non-negative matrix factorization," *Netw. Modeling Anal. Health Inform. Bioinf.*, vol. 4, no. 1, pp. 1–17, 2015.
[29] A. K. Biswas, B. Zhang, X. Wu, and J. X. Gao, *A Multi-Label Classification Framework to Predict Disease Associations of Long Non-coding RNAs (lncRNAs)*. Berlin, Germany: Springer, 2015.
[30] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Sci. Rep.*, vol. 5, p. 16840, 2015.
[31] V. Martínez, C. Cano, and A. Blanco, "Prophnet: A generic prioritization method through propagation of information," *BMC Bioinf.*, vol. 15, no. 1, pp. 1506–1526, 2014.
[32] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "Drugnet: Network-based drug-disease prioritization by integrating heterogeneous data," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 41–49, 2015.
[33] T. H. Hwang, Z. Wei, M. Xie, J. Liu, and K. Rui, "Inferring disease and gene set associations with rank coherence in networks," *Bioinf.*, vol. 27, no. 19, pp. 2692–2699, 2011.
[34] O. Vanunu and R. Sharan, "A propagation-based algorithm for inferring gene-disease associations," in *Proc. Ger. Conf.*, 2008, pp. 54–52.
[35] X. Liu, M. Dong, K. Ota, P. Hung, and A. Liu, "Service pricing decision in cyber-physical systems: Insights from game theory," *IEEE Trans. Serv. Comput.*, vol. 9, no. 2, pp. 186–198, Mar./Apr. 2016.
[36] M. Dong, K. Ota, L. T. Yang, A. Liu, and M. Guo, "LSCD: A low-storage clone detection protocol for cyber-physical systems," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 35, no. 5, pp. 712–723, May 2016.
[37] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinf.*, vol. 26, no. 8, pp. 1057–1063, 2010.
[38] J. I. Garzón, L. Deng, D. Murray, S. Shapira, D. Petrey, and B. Honig, "A computational interactome and functional annotation for the human proteome," *Elife*, vol. 5, 2016, Art. no. e18715.
[39] J. S. Amberger, C. A. Bocchini, S. Fran Ois, A. F. Scott, and H. Ada, "Omim.org: Online mendelian inheritance in man (omim?), an online catalog of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 43, pp. D789–D798, 2015.
[40] M. A. V. Driel, B. Jorn, V. Gert, G. B. Han, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Humangenetics*, vol. 14, no. 5, pp. 535–542, 2006.
[41] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *Plos Computat. Biol.*, vol. 6, no. 1, pp. e1 000 641–e1 000 641, 2010.
[42] Y. F. Huang, H. Y. Yeh, and V. W. Soo, "Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation," *BMC Med. Genomics*, vol. 6, no. 46, pp. 5369–5376, 2013.
[43] L. Deng and Z. Chen, "An integrated framework for functional annotation of protein structural domains," *IEEE/ACM Trans. Computat. Biol. Bioinf.*, vol. 12, no. 4, pp. 902–913, Jul./Aug. 2015.
[44] S. Peri, et al., "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Res.*, vol. 13, no. 10, pp. 2363–2371, 2003.
[45] Y. Jiao, W. Wei, X. Chaoyong, Z. Guoguang, Z. Yi, and C. Runsheng, "NPInter v2.0: An updated database of ncRNA interactions," *Nucleic Acids Res.*, vol. 42, D104–D108, 2014.
[46] L. Deng, Q. C. Zhang, Z. Chen, Y. Meng, J. Guan, and S. Zhou, "PredHS: A web server for predicting protein-protein interaction hot spots by using structural neighborhood properties," *Nucleic Acids Res.*, vol. 42, pp. 290–5, 2014.
[47] C. Fan, D. Liu, R. Huang, Z. Chen, and L. Deng, "Predrsa: A gradient boosted regression trees approach for predicting protein solvent accessibility," vol. 17, no. Suppl. no. 1, 2016, Art. no. 8.
[48] R. C. Bast, H. Bryan, and G. B. Mills, "The biology of ovarian cancer: New opportunities for translation," *Nature Rev. Cancer*, vol. 9, no. 6, pp. 415–428, 2009.
[49] Q. Guo, et al. "Comprehensive analysis of lncRNA-mRNA co-expression patterns identifies immune-associated lncRNA biomarkers in ovarian cancer malignant progression," *Sci. Rep.*, vol. 5, p. 17683, 2015.
[50] A. Roshan and S. B. Kaye, "Ovarian cancer: Strategies for overcoming resistance to chemotherapy," *Nature Revi. Cancer*, vol. 3, no. 7, pp. 502–16, 2003.
[51] Y. Wang, et al., "Mammalian ncRNA-disease repository: A global view of ncRNA-mediated disease network," *Cell Death Disease*, vol. 4, no. 8, 2013, Art. no. e765.
[52] L. Dong and L. Hui, "Hotair promotes proliferation, migration, and invasion of ovarian cancer skov3 cells through regulating pik3r3," *Med. Sci. Monitor Int. Med. J. Exp. Clinical Res.*, vol. 22, pp. 325–331, 2016.
[53] S. Xiujie, et al., "Promoter hypermethylation influences the suppressive role of maternally expressed 3, a long non-coding RNA, in the development of epithelial ovarian cancer," *Oncology Rep.*, vol. 32, no. 1, pp. 277–85, 2014.
[54] W. Patrick Y and K. Santosh, "Malignant gliomas in adults," *New England J. Med.*, vol. 359, no. 5, 2008, Art. no. 877.
[55] B. Er-Bao, L. Jia, X. Yong-Sheng, Z. Gang, L. Jun, and Z. Bing, "LncRNAs: New players in gliomas, with special emphasis on the interaction of lncRNAs with EZH2," *J. Cellular Physiology*, vol. 230, no. 3, pp. 496–503, 2015.
[56] P. Wang, Z. Ren, and P. Sun, "Overexpression of the long non-coding RNA meg3 impairs in vitro glioma cell proliferation," *J. Cellular Biochemistry*, vol. 113, no. 6, pp. 1868–74, 2012.
[57] K. X. Ma, et al., "Long noncoding RNA malat1 associates with the malignant status and poor prognosis in glioma," *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.*, vol. 36, no. 5, pp. 3355–3359, 2015.
[58] Z. Kailiang, et al., "Long non-coding RNA hotair promotes glioblastoma cell cycle progression in an EZH2 dependent manner," *Oncotarget*, vol. 6, no. 1, pp. 537–546, 2014.
[59] J. Li, M. Zhang, G. An, and Q. Ma, "LncRNA TUG1 acts as a tumor suppressor in human glioma by promoting cell apoptosis," *Exp. Biol. Med.*, vol. 241, no. 6, pp. 644–649, 2016.
[60] A. I. Ojesina, et al., "Landscape of genomic alterations in cervical carcinomas," *Nature*, vol. 506, no. 7488, pp. 371–375, 2014.
[61] P. Li, X. Yuan, B. Jiang, Z. Tang, and G. C. Li, "LncRNAs: Key players and novel insights into cervical cancer," *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.*, vol. 37, no. 3, pp. 2779–2788, 2015.
[62] J. Zhang, T. Yao, Y. Wang, J. Yu, Y. Liu, and Z. Lin, "Long noncoding RNA MEG3 is downregulated in cervical cancer and affects cell proliferation and apoptosis by regulating MIR-21," *Cancer Biol. Therapy*, vol. 17, no. 1, pp. 104–113, 2015.

Jingpu Zhang received the BS degree in computer science from the School of Information Science and Engineering, Northeastern University, Shenyang, China, in 2003 and the MS degree in computer science from the School of Computer Science, Xidian University, Xi'an, China, in 2010. He is currently working toward the PhD degree in the School of Information Science and Engineering, Central South University, ChangSha, China. His current research interests include data mining and bioinformatics.



Zuping Zhang received the BS degree in foundation of mathematics, Hunan Normal University, in 1989, the MS degree in foundation of mathematics, Jilin University, in 1992, and the PhD degree in computer application technology, Central South University, in 2005. He is now a professor in the School of Information Science and Engineering, Central South University, ChangSha, China. His current research interests include information fusion and information system, parameter computing, and biology computing.

Zhigang Chen received the BS, MS, and PhD degrees in computer science from the School of Information Science and Engineering, Central South University, Changsha, China, in 1984, 1987, and 1998, respectively. From 1997 to 1998, he was a visiting PhD student at Kanazawa University, Kanazawa, Japan. From 1998-1999, he worked in NTT Data as an employee of JCS, Tokyo, Japan. He is currently a professor and dean of the School of Software, Central South University, Changsha, China. His current research interests include computer network, distributed systems, and data mining. He is a member of the CCF Council and the IEEE.



Lei Deng received the BS and MS degrees in computer science from the School of Information Science and Engineering, Central South University, Changsha, China, in 2005 and 2008, respectively, and the PhD degree from the Department of Computer Science and Technology, Tongji University, Shanghai, China, in 2012. From 2010 to 2011, he was a visiting PhD student in the Center for Computational Biology and Bioinformatics, Columbia University, New York, USA. From 2013 to 2014, he was a postdoctoral research scientist in the Center for Computational Biology and Bioinformatics, Columbia University, New York, USA. He is currently an associate professor in the School of Software, Central South University, Changsha, China. His research interests include data mining, bioinformatics, and systems biology.