A novel algorithm based on bi-random walks to identify disease-related IncRNAs

Jialu Hu\textsuperscript{1,2}, Yiqun Gao\textsuperscript{1}, Jing Li\textsuperscript{3}, Yan Zheng\textsuperscript{1}, Jingru Wang\textsuperscript{1} and Xuequn Shang\textsuperscript{1,*}

From Biological Ontologies and Knowledge bases workshop at IEEE BIBM 2018
Madrid, Spain. 3-6 December 2018

Abstract

Backgrounds: There is evidence to suggest that IncRNAs are associated with distinct and diverse biological processes. The dysfunction or mutation of IncRNAs are implicated in a wide range of diseases. An accurate computational model can benefit the diagnosis of diseases and help us to gain a better understanding of the molecular mechanism. Although many related algorithms have been proposed, there is still much room to improve the accuracy of the algorithm.

Results: We developed a novel algorithm, BiWalkLDA, to predict disease-related IncRNAs in three real datasets, which have 528 IncRNAs, 545 diseases and 1216 interactions in total. To compare performance with other algorithms, the leave-one-out validation test was performed for BiWalkLDA and three other existing algorithms, SIMCLDA, LDAP and LRLSLDA. Additional tests were carefully designed to analyze the parameter effects such as $\alpha$, $\beta$, $I$ and $r$, which could help user to select the best choice of these parameters in their own application. In a case study of prostate cancer, eight out of the top-ten disease-related IncRNAs reported by BiWalkLDA were previously confirmed in literatures.

Conclusions: In this paper, we develop an algorithm, BiWalkLDA, to predict IncRNA-disease association by using bi-random walks. It constructs a IncRNA-disease network by integrating interaction profile and gene ontology information. Solving cold-start problem by using neighbors' interaction profile information. Then, bi-random walks was applied to three real biological datasets. Results show that our method outperforms other algorithms in predicting IncRNA-disease association in terms of both accuracy and specificity.

Availability: https://github.com/screamer/BiwalkLDA

Keywords: LncRNA-disease association, Bi-random walks, Gene ontology, Interaction profile

Background

It suggests that only 1.5% of genes in the human genome were protein-coding genes, which are twice as many as that of worm and fruit fly [1]. However, 74.7% of the human genome is involved in the process of primary transcripts [2]. It implies that non-coding RNAs play major roles in the regulation of gene expression. The presence or absence of some non-coding RNAs could down- or up-regulate a cascade of gene expression, which could be drug targets for medical therapy of a disease. Many researchers put efforts in to the discovery of the long non-coding RNAs function. Recent studies have found strong association between IncRNA and diseases. It shows that many IncRNAs play as some functional roles in diverse biological processes, such as cell proliferation, RNA binding complexes, immune surveillance, neuronal processes, morphogenesis and gametogenesis [3]. Their dysfunction may cause various diseases. For example, HOTAIR would induce androgen-independent (AR) activation, which plays a central role in establishing an oncogenic cascade that drives prostate cancer progression. It is also a causal reason for AR-mediated transcription programs in the absence of androgen [4]. Therefore, the prediction of IncRNA function would give us a new way to

\textsuperscript{*}Correspondence: shang@nwpu.edu.cn; jhu@nwpu.edu.cn
\textsuperscript{1}School of Computer Science, Northwestern Polytechnical University, 710072 Xi'an, China
Full list of author information is available at the end of the article

© The author(s). 2019 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
understand the regulation mechanism and disease pathology. There is an urgent demand for the development of fast and accurate algorithm to predict IncRNA-disease association.

Many computational tools have recently been developed to predict potential IncRNA-disease association and functional patterns in biological networks [5–10]. Functional patterns in biological networks. These computational methods are majorly in three categories. One of them is based on the idea of matrix factorization. Matrix factorization can be seen as a linear model of latent factors. In these methods, a corresponding latent factor is generated for each IncRNA and disease. Then, it uses a dot product of the latent factors to represent their similarity. The objective function of matrix factorization is to learn the optimal latent factors which can minimize the prediction error. Recently, these methods have been widely used in the prediction of IncRNA-disease relationship. For example, MFLDA reduces the high dimension of heterogeneous data sources into low-rank matrices via matrix tri-factorization, which can help to explore and exploit their intrinsic and shared structure [11]. SIMCLDA translates the IncRNA-disease association prediction problem into a recommendation, which can be solved with inductive matrix completion (IMC) [12]. However, matrix factorization may also bear the risk of over-fitting and the problem of costing-time complexity. Another type of methods is based on the idea of "guilt-by-associate". They are intuitively guided by the assumption that similar disease or IncRNA have similar connection patterns. If disease (A) and IncRNA (A) are known to be related, and disease (A) and disease (B) are very similar. We can infer disease (B) may also related to IncRNA (A). Obviously, the performance of these algorithms heavily depends on the accuracy of the similarity measures. Many "guilt-by-association" algorithms have been proposed. For example, RWRIncD infers potential human IncRNA-disease associations by implementing the random walk with restart method on a IncRNA functional similarity network [13]. IRWRLDA predicts novel IncRNA-disease associations by integrating known IncRNA-disease associations, disease semantic similarity, and various IncRNA similarity measures and make prediction based on improved Random Walk with Restart [14]. The third type of methods focus on classification. Feature extraction was performed on the complex network. Binary classifiers could be applied in the following step to predict whether there exists a connection between IncRNAs and diseases. Another typical prediction algorithm is LRLSLDA, which constructs a cost function in IncRNA and disease space and makes prediction by combining several classifiers in the IncRNA and disease space into a single classifier [15]. LDAP predicts potential IncRNA-disease associations by using a bagging SVM classifier based on IncRNA similarity and disease similarity [16].

In this paper, we proposed a novel algorithm, BiWalkLDA, to predict potential IncRNA-disease associations. The design of BiWalkLDA was intuitively guided by the assumption of "guilt-by-associate". In order to construct more accurate similarity network, we integrate two types of data from interaction profiles and gene ontology. Furthermore, our method was designed to solve the cold-start problem. BiWalkLDA uses bi-random walks algorithm to predict IncRNA-disease association base on a similarity network we constructed. The experiments were carried out on three real datasets downloaded from the LncRNADisease database [17]. Algorithm performance were evaluated by using Leave-one-out cross validation (LOOCV). Results show that BiWalkLDA outperforms other four state-of-art algorithms, meanwhile it is robust on different datasets and parameters in predicting novel IncRNA-disease associations.

**Methods**

**Construction of disease similarity networks**

Association patterns were commonly used to calculate disease similarity [14, 18]. In that case, disease similarity will depend on known LncRNA and disease association. Because of the lack of these prior knowledge in IncRNA-disease association, we considered to use gene ontology as an additional information. Gene ontology informations are obtained from previous work [12], which downloaded association between genes and gene ontology terms of human being from Ensemble database [19] and derived disease-gene associations from DisGeNet database [20]. For each disease, we can get the corresponding GO set. Then we use jaccard similarity to measure the similarity between the two sets. The calculation process is shown in the following formula:

\[
S_{GO}(d_i, d_j) = \frac{|GO_{d_i} \cap GO_{d_j}|}{|GO_{d_i} \cup GO_{d_j}|}
\]

where \(GO_{d_i}\) and \(GO_{d_j}\) are two sets of gene ontology terms of disease \(d_i\) and \(d_j\), respectively. Like previous algorithms, we also construct disease similarity networks by using known disease and LncRNA associations. The construction process can be divided into two steps: (1) construction of an adjacency matrix \(A_{n_d \times n_d}\), where \(n_d\) is the number of IncRNA and \(n_d\) is the number of diseases. \(A_{ij} = 1\) represent that the \(i^{th}\) IncRNA is associated with \(d_j\), otherwise \(A_{ij} = 0\). (2) With the matrix \(A\), we referred \(IP(d(i))\) to the \(i^{th}\) column of \(A\), which is the interaction profile of disease \(d_i\). \(IP(d(i))\) is a binary vector of length \(n_d\) and represents an association pattern of disease \(d(i)\). Then we calculate the similarity between two diseases based on the gaussian
linear kernel,
\[ S_{GK}(d_i, d_j) = \exp(-\gamma_d ||IP(d(i)) - IP(d(j))||^2) \]
where \(-\gamma_d\) is the bandwidth of kernel which is calculated as follow:
\[ \gamma_d = 1/\left(1/\sum_{i=1}^{n_d} ||IP(d(i))||^2\right) \]

Here \(n_d\) is the number of diseases. Up to now, we have constructed \(S_{GK}\) based on known association between IncRNA and disease and \(S_{GO}\) based on disease-related GO set. Then we use a simple linear model to fuse the two similarity networks.
\[ S_d = \alpha S_{GO} + (1 - \alpha) S_{GK} \]

Here \(\alpha\) is a hyperparameter that control the proportion of \(S_{GK}\) and \(S_{GO}\). If \(\alpha = 1\), disease similarity only be calculated base on gene ontology information. If \(\alpha = 0\), disease similarity only be calculated base on known disease-IncRNA associations. When the matrix is sparse, it would be better to give a large \(\alpha\) so that similarity rewards can be obtained from geneontology. This technique makes the algorithm more robust.

**Construction of IncRNA similarity network**

Similar to the previous process, we calculate IncRNA gaussian similarity based on known disease-IncRNA association. First, we use \(IP(l(i))\) which is the \(i^{th}\) row of \(A\) to represent the interaction profile of IncRNA \(l(i)\). \(IP(l(i))\) is a binary vector of length \(n_d\) and represents an association pattern of IncRNA \(l(i)\). Then IncRNA gaussian similarity was calculated base on the following formula:
\[ S_{GKL}(l_i, l_j) = \exp(-\gamma_l ||IP(l(i)) - IP(l(j))||^2) \]
\[ \gamma_l = 1/\left(1/\sum_{i=1}^{n_l} ||IP(l(i))||^2\right) \]
where \(\gamma_l\) is the bandwidth of kernel, \(n_l\) is the number of the IncRNA.

**Calculation of interaction profiles for new IncRNAs**

In the prediction process, if an IncRNA only knows very few diseases associated with it, this IncRNA is difficult to predict accurately. This is a common problem in industry, such as the difficulty of recommending products to a new user. This problem is also known as cold-start problem. There are two ways to solve it. The first way is to consider additional information in the definition of node similarity. The other one is to use prior information, e.g., diseases with many connections are more likely to interact with a new unknown IncRNA. We had considered using additional data such as IncRNA sequence information to measure similarity between IncRNAs. But on the one hand, the length of IncRNA sequence is very long(> 300 bp), and it is difficult to find an appropriate algorithm to measure their similarity. On the other hand, it is difficult for some new IncRNA to collect their sequence information. So in this paper, we mainly deal with the cold start problem through the second ways. We will describe this process in detail. First, we calculate the interaction profile for a new IncRNA using the mean of its neighbors’ interaction profile. Taken IncRNA \(l(i)\) as an example, the neighbors of IncRNA \(l(i)\) should be satisfied with the following formula:
\[ ||IP(l(j)) - IP(l(i))||^2 \geq \frac{\sum_{k=1}^{n_l} ||IP(l(k)) - IP(l(i))||^2}{n_l} \]
Here, \(n_l\) is the number of IncRNA. In another words, if similarity between \(l(i)\) and \(l(j)\) were larger than the mean of the similarity, \(l(j)\) can be defined as the neighbors of \(l(i)\). \(IP(l(i))\) was the mean of its neighbors’ interaction profile.
\[ IP(l(i)) = \frac{\sum_{k \in N(l(i))} IP(l(k))}{|N(l(i))|} \]
Here \(N(l(i))\) is the set of the neighbors of IncRNA \(l(i)\) and \(|N(l(i))|\) is the size of \(N(l(i))\). Notice that our approach here is different from the traditional approach to dealing with cold-start problem. Typically, the traditional method uses the mean of other IncRNAs interaction profile to fill in the new IncRNA. This is actually based on the popularity to make prediction. In contrast, BiwalkLDA uses local topological structure to predict missing interactions. Given a new IncRNA, we first find all its similar (or nearest) IncRNAs, which are likely to share common disease interactors with our node of interest. So, the key point is the definition of similarity function. Unlike all other algorithms, we assume that these IncRNAs sparsely connected to diseases would contribute more to the given node. It means they are likely to share common disease nodes. For example, an inactive user didn’t buy Harry Potter, although the book is one of the best seller. How likely does a new user would choose to buy the book. In our model, new users would more likely to learn from inactive users.

**The algorithm of Bi-random walk**

Based on the construction of IncRNA similarity network and disease similarity network, we use the bi-random walk algorithm to predict potential IncRNA-disease associations. First, I will explain the idea of bi-random walk algorithm. BiwalkLDA mainly make prediction base on the assumption that similar diseases or IncRNA have similar connection patterns. For example, if we know that IncRNA \((i)\) is associated with disease \((j)\) and IncRNA \((i)\) is very similar to IncRNA \((j)\), it is obvious that we can infer that IncRNA \((j)\) may also associate with disease \((j)\). So far we have constructed disease similarity networks and
IncRNA similarity networks. Bi-random walk algorithm actually constructed a linear model based on similarity. Suppose we want to predict the relationship between IncRNA \((i)\) and disease \((j)\).

\[
a_{ij} = \sum_{k=1}^{n_i} a_{i,k} \cdot \text{sim}_d(k,j)
\]

Here \(a_{ij}\) represents the possibility that IncRNA \((i)\) and disease \((j)\) are related. \(\text{sim}_d(k,j)\) represents the similarity of disease \((k)\) and disease \((j)\). So the process of calculation is actually to traverse every disease \(k\) and add \(a_{i,k} \cdot \text{sim}_d(k,j)\) up. It can be seen as a linear model based on similarity. Considering that we want to keep part of the original \(a_{ij}\), the formula can be written as below:

\[
a_{ij} = (1 - \beta) \cdot a_{ij} + \beta \sum_{k=1}^{n_i} a_{i,k} \cdot \text{sim}_d(k,j)
\]

Note that we need to normalize the similarity to ensure that \(a_{ij}\) is always less than 1. The above formula is based on disease similarity to make predictions. Similarly, we can make predictions based on the similarity of IncRNA and then combine the two results together to make final prediction. So the whole process of the algorithm can be divided into three steps: (1) First, we predict new scores based on disease similarity and IncRNA similarity according to random walk algorithm. (2) Then, we use the mean of two scores as the result of this round of prediction. (3) The two steps are repeatedly performed until maximum number of iterations. Let's go into the details of the algorithm. We do row normalization on both IncRNA similarity network and disease similarity. This is because random walk is actually a linear prediction model based on similarity. The similarity should be normalized so that the prediction results are between 0 and 1.

\[
D_k(i) = S_d(i,j)/\sqrt{D_d(i,i) \cdot D_d(j,j)}
\]

Here \(D_d(i,i)\) is the sum of the \(i\)th row of \(D_d\). Similarly, we normalized the similarity of IncRNA as following formula:

\[
S_l(i,j) = S_{GKL}(i,j)/\sqrt{D_{GKL}(i,i) \cdot D_{GKL}(j,j)}
\]

Here \(D_{GKL}(i,i)\) is the sum of the \(i\)th row of \(S_{GKL}\). Adjacent matrix \(A\) also needs to be initialized. Scores of all known IncRNA-disease association are set to \(1/n\) where \(n\) is the total number of known IncRNA and disease associations. Scores of Other unobserved associations are set to zero.

\[
S_{ini} = \frac{A}{\text{sum}(A)}
\]

Here \(S_{ini}\) represent the initial probability and the sum of initial probabilities is 1. Because the importance of predicting results based on different similarity networks may be different. We introduce two parameter \(l\) and \(r\) as the numbers of maximal iterations in the left and right random walks on these two networks. The more iterations, the more important the prediction through this similarity network is. The iterative process can be described by the following formula:

\[
R_d = \beta S_{ini}^{t-1} \cdot S_d + (1 - \beta) S_{ini}^0
\]

\[
R_l = \beta S_l \cdot S_{ini}^{t-1} + (1 - \beta) S_{ini}^0
\]

\[
S_{ini}^t = \frac{R_d + R_l}{2}
\]

Here \(S_{ini}, S_t\) represent initial score of all disease-IncRNA association. \(\beta\) is the decay factor which control the degree of retention of initial information. \(R_l\) represents the score of random walk on the IncRNA similarity network and \(R_d\) represents the score of random walk on the disease similarity network. In the iterative function, we use the averaged value of \(R_d\) and \(R_l\) as \(S_{ini}^t\) in step \(t\). This process can be seen as a combination of IncRNA similarity and disease similarity to make predictions. When the number of iterations reached \(\max(l, r)\), \(S_{ini}^t\) is the final result which represents the possibilities of all IncRNA-disease association. The pseudocode of bi-random walk algorithm can be seen in Algorithm 1.

**Algorithm 1** The specific process of bi-random walk:

**Input**: Disease similarity: \(S_d\); IncRNA similarity: \(S_{GKL}\); Adjacency matrix: \(A\); \(l\); \(r\)

**Output**: Final score: \(S_{ini}^t\)

1. Normalizing \(S_d, S_l\)
2. \(S_{ini}^0 = \frac{A}{\text{sum}(A)}\)
3. \(t = 1\)
4. while \(t <= \max(l, r)\) do
5. if \(t <= l\) then
6. \(R_d = \beta S_{ini}^{t-1} \cdot S_d + (1 - \beta) S_{ini}^0\)
7. end if
8. if \(t <= r\) then
9. \(R_l = \beta S_l \cdot S_{ini}^{t-1} + (1 - \beta) S_{ini}^0\)
10. end if
11. \(S_{ini}^t = \frac{R_d + R_l}{2}\)
12. end while
13. return \(S_{ini}^t\)

**Data and materials**

Known IncRNA and disease associations were downloaded from the LncRNADisease database [17], which is commonly used as the gold standard in predicting IncRNA-disease associations question. In the
experiment, we used three databases which are downloaded from three different versions of 2012, 2014 and 2015 (Table 1). In 2012, experiments confirm there exists 276 interactions between 112 lncRNAs and 150 diseases. In 2014, this number has increased to 319 interactions between 131 lncRNAs and 169 diseases. In 2015, it suggests that there are 621 interactions between 285 lncRNAs and 226 diseases.

### Results

We use leave-one-out cross validation (LOOCV) to test the performance of BiwalkLDA. LOOCV is a widely-used strategy to evaluate the quality of the algorithms. In each turn, one known association was set as a test sample. All other lncRNA-disease association were set to training set to train model. All associations that are not observed will be considered as a candidate set and will be scored by BiwalkLDA. A correspond rankList can be generated based on the predicted results. Then true positive rates (TPR, sensitivity) and false positive rates (FPR, 1-specificity) can be calculated by giving different thresholds. Based on the calculated values of TPR and FPR, the receiver-operating characteristics (ROC) curves can be plotted. Then we use the areas under ROC curve (AUC) as evaluation criteria of algorithmic performance which reflects the global prediction accuracy in different situation. The value of AUC closed to one means a perfect prediction, while the AUC value of 0.5 indicates purely random performance.

### The effects of parameters

#### The effects of $\alpha$

In the section of disease similarity, we use a linear model to fuse $S_{GO}$ and $S_{GKD}$. Here $\alpha$ is a hyperparameter that control the proportion of $S_{GO}$ and $S_{GKD}$. If $\alpha = 1$, disease similarity only be calculated base on gene ontology information. If $\alpha = 0$, disease similarity only be calculated base on known disease-lncRNA associations. BiWalkLDA use gene ontology information as a supplement to $S_{GKD}$, which makes the generalization ability of the algorithm stronger. To test the performance of the algorithm under different $\alpha$ values, we changed $\alpha$ from 0 to 1 and increased 0.1 per time. Then we use BiwalkLDA to make prediction. The experimental results are shown in Fig. 1. When $\alpha = 0.1$, BiwalkLDA obtain the best results on dataset1 and dataset2. On dataset3, it reaches the peak when $\alpha = 0.3$. It can see that small changes in $\alpha$ do not have much impact on the results. Therefore, we recommend the region of $\alpha$ could be set between 0.1 and 0.3 for using BiwalkLDA. The experimental results show that the fusion of $S_{GKD}$ and $S_{GO}$ can improve the accuracy of the algorithm. Meanwhile, the algorithm can achieve good performance even if we only use the GO similarity network. It indicates that the algorithm

| Datasets | Version | No. of lncRNA | No. of disease | No. of interaction |
|----------|---------|---------------|---------------|--------------------|
| Dataset1 | 2012    | 112           | 150           | 276                |
| Dataset2 | 2014    | 131           | 169           | 319                |
| Dataset3 | 2015    | 285           | 226           | 621                |
still works in the absence of disease-lncRNA association information.

**The effects of β**

β is a decay factor in bi-random walk algorithm. β determines the degree of retention of initial information in each iteration. If β = 0, all initial information will be retained. If β = 1, all initial information will be used to predict new score in each turn. Obviously, it will result in a poor performance if β is either 0 or 1 are inappropriate and will result in a poor performance. To test the performance of the algorithm under different β values, we increased β from 0 to 1 in 10 steps, and run BiwalkLDA. The value of β was changed from 0 to 1 and increased 0.1 each time and then using BiwalkLDA to make prediction. The experimental results are shown in Fig. 2. When 0.1 ≤ β ≤ 0.9, the results of the algorithm varied slightly. It indicts that BiWalkLDA is robust to β. BiWalkLDA performs the best AUC when β = 0.8 in dataset1 and dataset2 and performs the best AUC when β = 0.7 in dataset3. Intuitively, if the initial data is sufficient, a smaller β is more appropriate. Because dataset3 contains more known lncRNA-disease associations, the optimal β in dataset3 is less than the other dataset. Finally, we set β = 0.8 as default in three datasets.

**The effects of l and r**

In bi-random walk algorithm, the parameters l and r are used to limit the number of random walk steps in the disease and lncRNA similarity network respectively. l and r can actually be understood as the importance of disease similarity and lncRNA similarity. If the value of l is larger, it means that we will predict more based on disease similarity. Values of different l and r make the algorithm applicable to different data. If l or r equals 0, the algorithm actually degenerates into a single random walk algorithm. This means that we only make predictions through lncRNA similarity or disease similarity alone instead of combining them. To test the performance of the algorithm under different l and r values, we grid search the combination of l and r in a certain range. The value of l and r were increased from 1 to 7 and increased 1 each time. Then we use BiwalkLDA to make prediction and calculate the AUC values by LOOCV. The experimental results are shown in Table 2. The experimental results show that when the values of l and r are relatively close, BiwalkLDA perform well. This shows that prediction through lncRNA similarity and disease similarity are equally important. Finally, we set l=6 and r=6 as default in three dataset.

**Comparison with other algorithms**

To test the performance of the BiwalkLDA, we compared BiWalkLDA with three the-state-of-art computational methods (LDAP, LRLSLDA, SIMCLDA) of lncRNA-disease association prediction in three datasets. The results of the algorithm are measured by AUC value and number of correctly retrieved association. Because limited code can be used, we also compare our algorithm with KATZHMDA which is be used to predict disease-microbe association. LRLSLDA used Laplacian normalization operation and construct cost function in lncRNA and disease space. Then making
prediction by minimize the cost function to obtain optimal classifier [15]. LDAP fused different data source and make prediction based on bagging SVM classifier [16]. SIMCLDA predicted lncRNA-disease association based on inductive matrix completion [12]. KATZHMDA integrated known microbes-disease associations and gaussian interaction profile kernel similarity for microbes and diseases and make prediction based on katz algorithm [21]. On dataset1, we can see that BiwalkLDA obtained an AUC of 0.8268 which is higher than others (LRLSLDA:AUC=0.7217, KATZHMDA:AUC=0.6510, LDAP:AUC=0.6987, SIMCLDA:AUC=0.7949) as shown in Fig. 3a. In addition to AUC, we also use the numbers of correctly retrieved association to measure the performance of the algorithm. If a predicted association in the first percent k of the candidate set, this association will be regard as a correctly retrieved association under given threshold k. So the numbers of correctly retrieved association can reflect the accuracy of the algorithm in top k%. AUC reflects the global performance of the algorithm. The experimental results are shown in Fig 3b. BiWalkLDA can predict more correctly retrieved association in Top10%. But it can also be seen that LRLSLDA performs better at lower thresholds in term of the numbers of correctly retrieved association. This result actually indicates that BiwalkLDA is more inclined to make global optimal predictions. This phenomenon can be explained as follows: (1) BiwalLDA processing samples with less information separately may significantly increase the AUC value, but it may also make incorrect predictions. (2) More comprehensive sequencing results can be obtained based on lncRNA similarity network and disease similarity network. This means that only samples that meet both of these inference criteria will be given a higher ranking. However, if one sample conforms to the similarity inference of lncRNA but does not conform to the other, it will not be given a high ranking. This problem can be solved by using non-linear algorithm. The results on the other two datasets are similar, so we will not discuss them one by one (Figs. 4 and 5). It can be concluded that BiWalkLDA also achieve the best result (AUC 0.8510 in dataset2 and AUC 0.8473 in dataset3) and BiwalkLDA is robust enough to different parameter selection.

**De novo lncRNA-disease prediction**

In section of disease similarity, we combine gauss similarity and gene ontology similarity. Fusion of multiple similarities network not only improves the performance of the algorithm, but also strengthen generalization ability of BiwalkLDA. To assess the performance of BiWalkLDA, we conduct de novo lncRNA-disease association prediction in dataset1. In the process of de novo prediction, each

| r | l=1 | l=2 | l=3 | l=4 | l=5 | l=6 | l=7 |
|---|-----|-----|-----|-----|-----|-----|-----|
| 1 | 0.7618 | 0.7231 | 0.6902 | 0.6714 | 0.6585 | 0.6448 | 0.6304 |
| 2 | 0.8124 | 0.7890 | 0.7292 | 0.6985 | 0.6802 | 0.6702 | 0.6564 |
| 3 | 0.8008 | 0.8214 | 0.8140 | 0.7295 | 0.7010 | 0.6838 | 0.6713 |
| 4 | 0.7910 | 0.8092 | 0.8230 | 0.8243 | 0.7285 | 0.7000 | 0.6850 |
| 5 | 0.7848 | 0.7989 | 0.8118 | 0.8238 | 0.8267 | 0.7269 | 0.6988 |
| 6 | 0.7778 | 0.7911 | 0.8006 | 0.8119 | 0.8236 | 0.8268 | 0.7255 |
| 7 | 0.7729 | 0.7834 | 0.7920 | 0.8007 | 0.8116 | 0.8233 | 0.8263 |

Fig. 3 Comparison of predicting methods on dataset1. a Receiver operating characteristic curve of all algorithm using LOOCV. (b) Number of correctly retrieved known lncRNA-disease association for given percentage.
queried disease (i) would be removed all known lncRNA-disease association of this disease. Different computational methods were used in the prediction problem. Notice that we still know the gene ontology information of the disease. The experimental results are shown in Fig 6. The performance of BiWalkLDA only has a slight drop (AUC: 0.8364) and is much higher than other algorithms. The result shows that BiWalkLDA can make good prediction even if there is absence in disease-lncRNA association information and combining heterogeneous data sources can deal with data missing situation. Note that AUC of LDAP only has 0.4762. This result is lower than random guess which AUC value is 0.5. This is because we are actually testing the performance of the algorithm in the absence of data. LDAP treats this problem as a classification problem and using a bagging SVM classifier to make prediction. If there is a serious lack of data, the features learned will be inaccurate and the effect of classification will be poor. There are two reasons for the good performance of BiwalkLDA. 1) Gene
ontology information was used as a supplementary data. 2) BiwalkLDA used the neighborhood information to predict connections for new unknown IncRNAs.

Case studies
We conduct case study for prostate cancer to test the performance of BiWalkLDA in predicting new IncRNA for a certain disease. Prostate cancer (PC) is the most commonly diagnosed malignancy and the third leading cause of cancer death among men in developed countries. Predicting prostate cancer related IncRNA will help us to understand the mechanism of prostate cancer and provide a high probability set of candidate IncRNA for scientist. We use dataset3 to make prediction and confirm the result by searching related paper. Prediction results of the BiwalkLDA are showed in Table 3. It can see that eight prostate cancer-related IncRNA (H19, MALAT1, HOTAIR, MEG3, PVT1, GAS5, NEAT1, UCA1) in the top ten candidates have been confirmed by previous studies. Long non-coding RNA H19 and H19-derived microRNA-675 (miR-675) were significantly down-regulated in the metastatic prostate cancer cell line M12 compared with the non-meta-static prostate epithelial cell line P69 [22]. MALAT1 was up-regulated in human prostate cancer tissues and cell line [23]. HOTAIR as an androgen-repressed IncRNA is markedly up-regulated following androgen deprivation therapies and in castration-resistant prostate cancer [4]. MEG3 decreased significantly in prostate cancer tissues relative to adjacent normal tissues [24]. Region surrounding rs378854 which is identified as a novel function prostate cancer-specific genetic variant interacts with the MYC and PVT1 promoters [25]. GAS5 promotes the apoptosis of prostate cell, and exonic sequence, i.e. GAS5 IncRNA, is sufficient to mediate this activity [26]. Nuclear enriched abundant transcript 1 (NEAT1) was identified as the most significantly over-expressed IncRNA in prostate cancer by using a combination of chromatin immunoprecipitation (ChIP) and RNA-sequencing data [27]. UCA1 was abnormally up-regulated in tumor tissues from prostate cancer patients and patients with high UCA1 levels had a significantly poorer prognosis [28]. Successful predictions for prostate cancer prove that BiWalkLDA can help us to find new relationships between IncRNA and disease base on historical data.

Table 3 Top ten reported IncRNAs for prostate cancer

| Rank | Name of IncRNA | PMID          |
|------|----------------|---------------|
| 1    | H19            | 24988946      |
| 2    | CDKN2B-AS1     | Unconfirmed   |
| 3    | MALAT1         | 23845456      |
| 4    | HOTAIR         | 26411689      |
| 5    | MEG3           | 26610246      |
| 6    | PVT1           | 21814516      |
| 7    | BCYRN1         | Unconfirmed   |
| 8    | GAS5           | 23676682      |
| 9    | NEAT1          | 25415230      |
| 10   | UCA1           | 26550172      |
Conclusion
Many recent studies suggest that IncRNAs are strongly associated with various complex human diseases and they play important roles in the gene expression regulation and post-transcription modification. Predicting IncRNA-disease association can help understand the biological mechanism of disease and reduce the cost of experimental verification. However, discovering the relationship between IncRNA and disease by means of computational model is still a very challenging problem. Therefore, the development of computational tools is much in demand. Although many computational models have been proposed. Their prediction accuracy still has a lot of room to improve. To improve the performance of existing algorithms, we present a novel algorithm, BiwalkLDA based on bi-random walks for the prediction of IncRNA-disease associations. It integrates gene ontology and interaction profile data together to calculate disease similarity, to solve the cold-start problem by using the local structure of IncRNAs neighbors information. Four state-of-art computational methods and BiwalkLDA are applied to predict IncRNA-disease associations on three different datasets. Results show that BiwalkLDA is superior to every other existing algorithms in terms of both accuracy and recall. There are still many problems to be dealt with. Existing models are based on small-scale datasets. Although algorithms can achieve high accuracy, their results are often repetitive. If the dataset is too large, the existing algorithms cannot be applied to large-scale data. In future work, we will consider to develop more effective algorithm to solve this problem.

Abbreviations
AR: Androgen; IMC: Inductive matrix completion; LOOCV: Leave-one-out cross validation; ROC: receiver-operating characteristics

Acknowledgments
Many thanks go to Dr. Bolin Chen and Dr. Jiajie Peng for discussion.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 20 Supplement 18: Selected articles from the Biological Ontologies and Knowledge bases workshop 2018. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-18.

Authors’ contributions
JH designed the computational framework, YG, JL, YZ, and JW performed all the analyses of the data and wrote the manuscript; XS is the major coordinator, who contributed a lot of time and efforts in the discussion of this project. All authors read and approved the final manuscript.

Funding
Publication costs were funded by the National Natural Science Foundation of China (Grant No. 61702420). This project has also been funded by the National Natural Science Foundation of China (Grant No. 61332014, 61702420 and 61772426); the China Postdoctoral Science Foundation (Grant No. 2017M613203); the Natural Science Foundation of Shaanxi Province (Grant No. 2017JQ6037); the Fundamental Research Funds for the Central Universities (Grant No. 3102018zy032); the Top International University Visiting Program for Outstanding Young Scholars of Northwestern Polytechnical University.

Availability of data and material
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
1 School of Computer Science, Northwestern Polytechnical University, 710072 Xi’an, China. 2 Centre for Multidisciplinary Convergence Computing, School of Computer Science, Northwestern Polytechnical University, 710129 Xi’an, China. 3 Ming De College, Northwestern Polytechnical University, 710124 Xi’an, China.

Published: 25 November 2019

References
1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):346.
2. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489(7414):101.
3. Gutman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding mas in mammals. Nature. 2009;458(7235):223.
4. Zhang A, Zhao J, Kim J, et al. Lncrna hoatn enhances the androgen-receptor-mediated transcriptional program and drives castration-resistant prostate cancer. Cell Rep. 2015;13(1):209–21.
5. Hu J, Gao Y, Zheng Y, Shang X. FK-finder: Identification of key factors from host-microbial networks in cervical cancer. BMC Syst Biol. 2018;12(5):54.
6. Hu J, Gao Y, He J, Zheng Y, Shang X. WebNetCoffee: a webbased application to identify functionally conserved proteins from Multiple PPI networks. BMC Bioinformatics. 2018;19(1):422.
7. Hu J, Zheng Y, Shang X. MiteFinder: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. BMC Med Genomics. 2018;11(55):101.
8. Hu J, Shang X. Detection of Network Motif Based on a Novel Graph Canonization Algorithm from Transcriptional Regulation Networks. Molecules. 2017;22(12):2194.
9. Hu J, Wang J, Li J, Lin J, Liu T, Zhong Y, Liu J, Zheng Y, Gao Y, He J, Shang X. MD-SVM: A novel SVM-based algorithm for the motif discovery of transcription factor binding sites. BMC Bioinformatics. 2019;20(57): https://doi.org/10.1186/s12859-019-2735-3.
10. Peng J, Guan J, Shang X. Predicting Parkinson’s disease genes based on node2vec and autoencoder. Front Genet. 2019;10: https://doi.org/10.3389/fgen.2019.00226.
11. Fu G, Wang J, Domeniconi C, Yu G. Matrix factorization based data fusion for the prediction of Lncrna-disease associations. Bioinformatics. 2017;34(9):1529–37.
12. Lu C, Yang M, Luo F, Wu FX, Li M, Pan Y, et al. Prediction of Lncrna-disease associations based on inductive matrix completion. Bioinformatics. 2018;34(19):3357–64. https://doi.org/10.1093/bioinformatics/bty327.
13. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, et al. Inferring novel Lncrna-disease associations based on a random walk model of a Lncrna functional similarity network. Mol Biosyst. 2014;10(8):2074–81.
14. Chen X, You ZH, Yan GY, Gong DW. Invidia: improved random walk with restart for Lncrna-disease association prediction. Oncotarget. 2016;7(36):57919–31.
15. Chen X, Yan GY. Novel human Lncrna-disease association inference based on Lncrna expression profiles. Bioinformatics. 2013;29(20):2617–24.
16. Lan W, Li M, Zhao K, Liu J, Wu FX, Pan Y, et al. Ldap: a web server for Lncrna-disease association prediction. Bioinformatics. 2017;33(3):458–60.
17. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. Lncrnadisease: a database for long-non-coding rna-associated diseases. Nucleic Acids Res. 2013;41(Database issue):D983–D986.
18. Chen X. Katzlda: katz measure for the lncrna-disease association prediction. Sci Rep. 2014;5:16840.
19. Aken BL, Ayling S, Barrett D, Clarke L, Curwen V, Fairley S, et al. The ensembl gene annotation system. Database J Biol Databases Curation. 2016;2016:baw093. https://doi.org/10.1093/database/baw093.
20. Bauer-Mehren A, Rautschka M, Sanz F, Furlong Li. Disgenet. Bioinformatics. 2010;26(22):2924–292.
21. Chen X, Huang YA, You ZH, Yan GY, Wang XS. A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2016;33(5):733–9.
22. Zhu M, Chen Q, Liu X, Sun Q, Zhao X, Deng R, et al. Lncrna h19/mir-675 axis represses prostate cancer metastasis by targeting tgfbi. Febs J. 2015;281(16):3766–75.
23. Ren S, Liu Y, Xu W, Sun Y, Lu J, Wang F, et al. Long non-coding rna malat-1 is a new potential therapeutic target for castration resistant prostate cancer. J Urol. 2013;190(6):2278–87.
24. Luo G, Wang M, Wu X, Tao D, Xiao X, Wang L, et al. Long non-coding rna meg3 inhibits cell proliferation and induces apoptosis in prostate cancer. Cell Physiol Biochem. 2015;37(6):2209.
25. Meyer KB, Mala AT, O’Reilly M, Ghousaini M, Prathalingam R, Portergill P, et al. A functional variant at a prostate cancer predisposition locus at 8q24 is associated with pvt1 expression. Plos Genet. 2011;7(7):e1002165.
26. Pickard MR, Mourtadamaarabouni M, Williams GT. Long non-coding rna gas5 regulates apoptosis in prostate cancer cell lines. Biochim Biophys Acta. 2013;1832(10):1613–23.
27. Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, et al. The oestrogen receptor alpha-regulated lncrna neat1 is a critical modulator of prostate cancer. Nat Commun. 2014;5:5383.
28. Na XY, Liu ZY, Ren PF, Yu R, Shang XS. Long non-coding rna uca1 contributes to the progression of prostate cancer and regulates proliferation through ifi4-krt6/13 signaling pathway. Int J Clin Exp Med. 2015;8(8):12609–16.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.