frontiers in Genetics

ORIGINAL RESEARCH
published: 28 May 2019
doi: 10.3389/fgene.2019.00476

LncRNA-Disease Association Prediction Using Two-Side Sparse Self-Representation

Le Ou-Yang[1,2†], Jiang Huang[3†], Xiao-Fei Zhang[4], Yan-Ran Li[3], Yiwen Sun[5], Shan He[6] and Zexuan Zhu[3]*

[1] Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, China, [2] FJKLMAA (Fujian Key Laborotary of Mathematical Analysis and Applications), Fujian Normal University, Fuzhou, China, [3] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, [4] School of Mathematics and Statistics and Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, China, [5] School of Medicine, Shenzhen University, Shenzhen, China, [6] School of Computer Science, University of Birmingham, Birmingham, United Kingdom

OPEN ACCESS

Edited by:
Philipp Kapranov, Huaqiao University, China

Reviewed by:
Remco Molenaar, University Medical Center Amsterdam, Netherlands
Shihua Zhang, Academy of Mathematics and Systems Science (CAS), China
Jie Zheng, ShanghaiTech University, China

***Correspondence:**
Zexuan Zhu
zhuzx@szu.edu.cn

†Joint first authors

Specialty section:
This article was submitted to RNA, a section of the journal Frontiers in Genetics

Received: *30 December 2018*
Accepted: *03 May 2019*
Published: *28 May 2019*

Citation:
Ou-Yang L, Huang J, Zhang X-F, Li Y-R, Sun Y, He S and Zhu Z (2019) LncRNA-Disease Association Prediction Using Two-Side Sparse Self-Representation. Front. Genet. 10:476. doi: 10.3389/fgene.2019.00476

Evidences increasingly indicate the involvement of long non-coding RNAs (lncRNAs) in various biological processes. As the mutations and abnormalities of lncRNAs are closely related to the progression of complex diseases, the identification of lncRNA-disease associations has become an important step toward the understanding and treatment of diseases. Since only a limited number of lncRNA-disease associations have been validated, an increasing number of computational approaches have been developed for predicting potential lncRNA-disease associations. However, how to predict potential associations precisely through computational approaches remains challenging. In this study, we propose a novel two-side sparse self-representation (TSSR) algorithm for lncRNA-disease association prediction. By learning the self-representations of lncRNAs and diseases from known lncRNA-disease associations adaptively, and leveraging the information provided by known lncRNA-disease associations and the intra-associations among lncRNAs and diseases derived from other existing databases, our model could effectively utilize the estimated representations of lncRNAs and diseases to predict potential lncRNA-disease associations. The experiment results on three real data sets demonstrate that our TSSR outperforms other competing methods significantly. Moreover, to further evaluate the effectiveness of TSSR in predicting potential lncRNAs-disease associations, case studies of Melanoma, Glioblastoma, and Glioma are carried out in this paper. The results demonstrate that TSSR can effectively identify some candidate lncRNAs associated with these three diseases.

Keywords: lncRNAs-disease associations prediction, computational approaches, sparse representation, lncRNA similarity, disease similarity

1. INTRODUCTION

Long non-coding RNAs (lncRNAs), which are a class of non-coding transcripts with the lengths longer than 200 nucleotides (Derrien et al., 2012; Harrow et al., 2012; Guttman et al., 2013; Chen et al., 2016b), have been proven to be involved in various biological processes (Chen et al., 2012, 2016b, 2018) and closely correlated with the development of complex diseases, such as cancers

and rheumatic diseases (Bussemakers et al., 1999; Managadze et al., 2011; Bhartiya et al., 2012; Schonrock et al., 2012; Li et al., 2013; Lu et al., 2013; Zhao et al., 2014; Chen et al., 2016b). For example, studies have revealed the roles of lncRNAs in regulating gene expression (Taft et al., 2010; Wapinski and Chang, 2011). As the development of complex diseases are closely related to the mutations and abnormalities of lncRNAs, to understand the pathogenesis of human diseases systematically, and identify the biomarkers of disease progression and prognosis, it is important to predict the potential associations between diseases and lncRNAs (Chen et al., 2016b; Yu et al., 2018). However, only a small number of lncRNA-disease associations have been validated. Therefore, efficient methods for predicting the associations between lncRNAs and diseases are emergent needed (Lu et al., 2018).

In recent years, identifying the associations between diseases and lncRNAs has attracted a lot of attentions (Chen and Yan, 2013; Lu et al., 2018). Prediction methods based on biological experiments or computational approaches are proposed to undertake this task. Due to the limitations of biological experiments such as time-consuming and expensive in cost, computational approaches provide an alternative for biological experiments and have been widely used to identify the associations between lncRNAs and diseases (Chen et al., 2016b). Existing computational approaches for association prediction can be roughly classified into three categories. The first category is based on machine learning approaches. These models predict the associations between diseases and lncRNAs based on known lncRNA-disease associations. For example, Chen et al. proposed a semi-supervised learning-based method named Laplacian Regularized Least Squares for LncRNA-disease Association (LRLSLDA) (Chen and Yan, 2013) to predict the associations between diseases and lncRNAs. Zheng et al. formulated the problem of association prediction as a matrix factorization problem and introduced a collaborative matrix factorization model (CMF) (Zheng et al., 2013) to predict the associations. However, the performance of machine learning-based methods depend on the choice of hyperparameters such as the dimensionality of the latent space in matrix factorization-based methods, and the suitable values for these hyperparameters are usually previously unknown and hard to determine.

The second category is based on random walk. These models identify potential lncRNA-disease associations by integrating known associations between diseases and lncRNAs and similarities among diseases and lncRNAs. For example, Zhou et al. predicted the associations between diseases and lncRNAs by implementing random walk with restart on the constructed similarity networks among lncRNAs and diseases (Zhou M. et al., 2015). The third category is based on data integration. These models focus on integrating multiple heterogeneous data sources. For example, Lu et al. (2018) developed a model named SIMCLDA for identifying the associations between diseases and lncRNAs based on disease-gene and gene-gene ontology associations. However, the above methods rely heavily on the similarity networks or external information (e.g., similarity networks among diseases and lncRNAs, and gene-gene associations) that are inferred based on predefined metrics. Moreover, the information extracted from other databases or data platforms may include some irrelevant or noise information that may mislead the prediction of associations.

To address the above problems, in this paper, we introduce a novel two-side sparse self-representation (TSSR) model for lncRNA-disease association prediction. Based on known lncRNA-disease associations, our model can adaptively learn two non-negative sparse self-representation matrices which capture the intra-similarities among lncRNAs and diseases respectively. Moreover, our model could also drawn support from the intra-associations among disease and lncRNAs that derived from external information of lncRNAs and diseases to generate more accurate estimation of the representation matrices. Experiment results on three real datasets demonstrate that compared with six state-of-the-art association prediction algorithms, our TSSR model could achieve more accurate prediction results. Furthermore, case studies on three cancers (i.e., Glioblastoma, Glioma, and Melanoma) also demonstrate the effectiveness of TSSR in predicting the associations between lncRNAs and diseases. The source code of TSSR is available at https://github.com/Oyl-CityU/TSSR.

The rest of this paper is organized as follows. In section 2, we formulate our two-side sparse self-representation model and introduce a relaxed Majorization-Minimization algorithm to solve the optimization problem. The experiment results and case studies are given in section 3. In section 4, we conclude our works.

2. METHODS

2.1. Notations and Problem Statement

In this paper, we use $D = \{d_i\}_{i=1}^{m}$ to represent the set of lncRNAs and $T = \{t_j\}_{j=1}^{n}$ to represent the set of diseases, where m and n denote the number of lncRNAs and the number of diseases, respectively. A binary matrix $Y = [Y_{ij}] \in \{0, 1\}^{m\times n}$ is introduced to represent the associations between lncRNAs and diseases, where $Y_{ij} = 1$ if there is an association between lncRNA d_i and disease t_j, and $Y_{ij} = 0$ otherwise. Note that there are two reasons that may lead to $Y_{ij} = 0$. The first reason is that it has been experimentally verified that there is no association between d_i and t_j. The second reason is that whether there is an association between d_i and t_j is still unknown. Therefore, we usually refer to the zero elements in Y as unknown pairs. The lncRNA-disease association prediction problem can be formulated as the problem of predicting the scores of unknown pairs in Y, which can be used for ranking the pairs. In this study, we first rank the unknown pairs in Y based on the predicted scores in descending order, and then select the top-ranked pairs as potential association pairs.

In particular, unlike matrix factorization methods that project lncRNAs and diseases into a shared latent space and predict lncRNA-disease associations based on the inner product of their latent vectors, we try to learn the intra-similarities among lncRNAs and diseases from the observed associations in Y, and utilize the learned similarity matrices to reconstruct Y and thus predict the scores of unknown pairs in Y. Here, instead of using predefined metrics to construct the similarity matrices of lncRNAs and diseases (which makes the predicted results sensitive to the selected metrics and input data), we

introduce a novel two-side sparse self-representation (TSSR) model to adaptively learn the intra-similarities among lncRNAs and diseases from the observed associations in Y, and effectively utilize external information of lncRNAs and diseases to enhance the prediction performance.

2.2. Two-Side Sparse Self-Representation Model

Sparse representation techniques which focus on finding a sparse representation of a sample in the form of a linear combination of basic elements (also called atoms) in a dictionary, have been widely used to numerous applications such as computer vision and machine learning (Zhang et al., 2015). In traditional sparse representation models, the objective is to solve the following problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad s.t. \quad \mathbf{y} = D\mathbf{x}. \tag{1}$$

where $\|\cdot\|_0$ denotes L_0 norm, $\mathbf{y} \in R^{m\times 1}$ is a sample vector, D is a $m \times l$ matrix which denotes the dictionary and $\mathbf{x} \in R^{l\times 1}$ is the sparse representation coefficient of $\mathbf{y}$. In practice, L_0 norm is usually replaced with L_1 norm to make the above problem (1) solvable in polynomial time. Since the above problem (1) needs to take extra time to construct the dictionary D and has not data-adaptiveness. Many approaches are proposed to employ the dataset itself as the dictionary, which results in the following sparse self-representation model

$$\min_{X} \|Y - YX\|_F^2 + \beta\|X\|_1. \tag{2}$$

where $\|.\|_F$ is the Frobenius norm, Y denotes the feature set of all samples (each row denotes a feature and each column represents the feature vector of a sample), X is the sparse self-representation coefficient matrix of the columns of Y (each column $X_{\cdot j}$ of X denotes the representation coefficient of j-th sample $Y_{\cdot j}$, with all samples in Y as dictionary) and β is a tuning parameter to control the trade off between the minimization error and the sparsity. By solving the above model (2), X can capture the most similar relationships among the columns of Y, based on the information provided in Y. In this study, $Y \in \{0, 1\}^{m\times n}$ describes the observed associations between lncRNAs and diseases and we would like to predict potential associations between lncRNAs and diseases based on their intra-similarities learned from Y. Thus, instead of just finding the representations of the columns of Y, we prefer to explore the representations of the rows and columns of Y simultaneously, which capture the intra-similarities within lncRNAs and diseases respectively. Based on the idea of sparse self-representation, we introduce a novel two-side sparse self-representation (TSSR) model to handle the task of lncRNA-disease association prediction. In particular, we formulate the framework of TSSR into the following optimization problem

$$\begin{aligned} &\min_{U,V} \|Y - UYV\|_F^2 + \beta(\|U\|_1 + \|V\|_1), \\ &s.t. \quad U \geq 0, V \geq 0, \sum_{z=1}^{m} U_{iz} = 1, \sum_{k=1}^{n} V_{kj} = 1. \end{aligned} \tag{3}$$

where $U = [U_{ii'}] \in \mathbb{R}_+^{m\times m}$ and $V = [V_{jj'}] \in \mathbb{R}_+^{n\times n}$ are two non-negative sparse matrices which represent the row and column representation coefficient matrices of Y, respectively, and β is a tuning parameter which controls the sparsity of U and V. Based on this definition, U denotes the coefficient matrix based on the dictionary YV, which captures the similarities among lncRNAs. For example, $U_{ii'}$ denotes the similarities between the i-th and i'-th lncRNAs, which correspond to the i-th and i'-th rows of Y. On the other hand, V denotes the coefficient matrix based on the dictionary UY, which captures the similarities among diseases. For example, $V_{jj'}$ denotes the similarities between the j-th and j'-th diseases, which correspond to the j-th and j'-th columns of Y. With the sparse regularization term, we can control the sparsity of the learned representation matrices U and V, and find the most similar relationships within lncRNAs and diseases. The constraints $\sum_{z=1}^{m} U_{iz} = 1$ and $\sum_{k=1}^{n} V_{kj} = 1$ are used to guarantee the probability properties of $U_{i\cdot}$ and $V_{\cdot j}$, respectively.

In the above objective function (3), the representation matrices are learned from the original data matrix Y, which means that they will be sensitive to the input data Y. If the input data only includes a small number of known associations, it may be hard to learn a comprehensive representation matrix. With the development of high-throughput experimental techniques and the accumulation of clinical information, we could also collect some functional annotations and phenotype information for lncRNAs and diseases respectively. Based on these prior information, we can infer the intra-associations among diseases and lncRNAs. To utilize these pairwise associations inferred from other databases to promote the estimation of two representation coefficient matrices U and V, two regularization terms are added to Equation (3). Moreover, we introduce a weight matrix W in a similar way to Zheng et al. (2013) to prevent unknown instances (for which association information is not available) from contributing to the determination of the row and column representations of Y (i.e., U and V). The final objective function of our TSSR model is as follows.

$$\begin{aligned} \min_{U,V} &\|W \odot (Y - UYV)\|_F^2 + \beta(\|U\|_1 + \|V\|_1) \\ &+ \lambda_d\|S_d - U\|_F^2 + \lambda_t\|S_t - V\|_F^2, \\ &s.t. \quad U \geq 0, V \geq 0, \sum_{z=1}^{m} U_{iz} = 1, \sum_{k=1}^{n} V_{kj} = 1. \end{aligned} \tag{4}$$

where λ_d and λ_t are two tuning parameters controlling the influences of prior intra-associations among lncRNAs and diseases, $S_d \in \mathbb{R}^{m\times m}$ and $S_t \in \mathbb{R}^{n\times n}$ denote the affinity matrices of lncRNA and disease respectively, where $(S_d)_{ii'}$ describes the association between lncRNAs d_i and $d_{i'}$, and $(S_t)_{jj'}$ describes the associations between diseases t_j and $t_{j'}$. $\odot$ denotes the element-wise product or Hadamard product of two matrices and $W \in \mathbb{R}^{m\times n}$ is a weight matrix where $W_{ij} = 0$ for unknown entries in Y and $W_{ij} = 1$ for known entries in Y. Consequently, unknown entries in Y do not contribute to the minimization of the first term of Equation (4).

2.3. Optimization Algorithm

Here, to handle the constraints in (4), we employ a relaxed Majorization-Minimization algorithm (Yang and Oja, 2011, 2012) to obtain the solution of objective function (4). For more details about this optimization method, please refer to Yang and Oja (2012). In particular, we denote ∇_U as the gradient of our objective function with respect to U.

$$\nabla_U = -2[W \odot (Y - UYV)]V^{\mathrm{T}}Y^{\mathrm{T}} - 2\lambda_d(S_d - U) + \beta. \quad (5)$$

Let $\nabla_U^+ = 2W \odot (UYV)V^{\mathrm{T}}Y^{\mathrm{T}} + 2\lambda_d U + \beta$ and $\nabla_U^- = 2(W \odot Y)V^{\mathrm{T}}Y^{\mathrm{T}} + 2\lambda_d S_d$ denote the positive and negative parts of ∇_U, respectively. Thus, we have $\nabla_U = \nabla_U^+ - \nabla_U^-$.

Due to the constraint $\sum_{z=1}^{m} U_{iz} = 1$ and $U_{iz} \geq 0$, we obtain the following updating rule for U_{iz}:

$$U_{iz}^{\text{new}} = U_{iz} \cdot \frac{a_i^U(\nabla_U^-)_{iz} + 1}{a_i^U(\nabla_U^+)_{iz} + b_i^U}. \quad (6)$$

where a_i^U and b_i^U can be obtained by Equations (7) and (8), respectively.

$$a_i^U = \sum_z \frac{U_{iz}}{(\nabla_U^+)_{iz}}, \quad (7)$$

$$b_i^U = \sum_z U_{iz}\frac{(\nabla_U^-)_{iz}}{(\nabla_U^+)_{iz}}. \quad (8)$$

Similarly, we denote ∇_V as the gradient of our objective function with respect to V.

$$\nabla_V = -2(Y^{\mathrm{T}}U^{\mathrm{T}})[W \odot (Y - UYV)] - 2\lambda_t(S_t - V) + \beta. \quad (9)$$

Let $\nabla_V^+ = 2Y^{\mathrm{T}}U^{\mathrm{T}}[W \odot (UYV)] + 2\lambda_t V + \beta$ and $\nabla_V^- = 2Y^{\mathrm{T}}U^{\mathrm{T}}(W \odot Y) + 2\lambda_t S_t$ denote the positive and negative parts of ∇_V, respectively, we have $\nabla_V = \nabla_V^+ - \nabla_V^-$.

Similarly, the updating rule for V_{kj} is as follows:

$$V_{kj}^{\text{new}} = V_{kj} \cdot \frac{a_j^V(\nabla_V^-)_{kj} + 1}{a_j^V(\nabla_V^+)_{kj} + b_j^V}. \quad (10)$$

where $a_j^V = \sum_k \frac{V_{kj}}{(\nabla_V^+)_{kj}}$ and $b_j^V = \sum_k V_{kj}\frac{(\nabla_V^-)_{kj}}{(\nabla_V^+)_{kj}}$.

The details of the optimization algorithm to the proposed TSSR model are described in Algorithm 1. U and V can be updated by Equations (6) and (10), respectively. In this study, we stop the iteration when the changes of U and V are less than 1e-6, measured by L_1 norm. Finally, the predicted label matrix $\hat{Y}$ can be returned by $\hat{Y} = UYV$ when algorithm arrives at the convergence conditions.

Algorithm 1: Algorithm for the TSSR model

- **Inputs:** Partial label matrix Y, lncRNA affinity matrix S_d, disease affinity matrix S_t, tuning parameter $\lambda_d, \lambda_t, \beta$, weight matrix W.
- **Output:** Predicted label matrix $\hat{Y}$.
- **Main algorithm:**

1. Initialize U and V;
2. **While** not converged **do**
3. Update U according to Equation (6)
$U_{iz}^{\text{new}} = U_{iz} \cdot \frac{a_i^U(\nabla_U^-)_{iz} + 1}{a_i^U(\nabla_U^+)_{iz} + b_i^U}$;
4. Update V according to Equation (10)
$V_{kj}^{\text{new}} = V_{kj} \cdot \frac{a_j^V(\nabla_V^-)_{kj} + 1}{a_j^V(\nabla_V^+)_{kj} + b_j^V}$;
5. Check the convergence conditions.
6. **End while**
7. **Return** $\hat{Y} = UYV$.

3. RESULTS

In this section, we demonstrate the performance of various algorithms on three real datasets. Furthermore, case studies of three cancer diseases (i.e., Melanoma, Glioblastoma, and Glioma) are performed to validate the effectiveness of our TSSR model. The materials, experimental settings, and parameter settings are described as follows.

3.1. Materials

3.1.1. LncRNA-Disease Associations

We collect three datasets to evaluate the performance of various prediction algorithms. The first dataset is downloaded from the supplementary data of a article (Lu et al., 2018), which contains 621 experimentally confirmed lncRNA-disease associations between 226 diseases and 285 lncRNAs from the LncRNADisease database[1] established in 2015. The second dataset involving 260 high-quality associations between 95 lncRNAs and 81 human disease is obtained from the supplementary files of the published article (Chen et al., 2015), which retrieved data from MNDR database[2] (Wang et al., 2013) in March 2015 . The third dataset is downloaded from the Lnc2Cancer database [3] in 2015. By getting rid of the duplicate lncRNA-disease associations for the same lncRNA-disease pair, we obtain 677 distinct associations, including 54 human cancers and 436 lncRNAs. The statistics of the three datasets are illustrated in **Table 1**.

3.1.2. Disease Similarities

As previous studies have discovered that diseases with similar phenotypes are usually related with similar dysfunctions of lncRNAs (Chen et al., 2015), incorporating the similarities among diseases estimated from other database may help to infer the

[1] http://www.cuilab.cn/lncrnadisease
[2] http://www.rna-society.org/mndr/
[3] http://www.bio-bigdata.com/lnc2cancer/

TABLE 1 | The statistics of three datasets.

Datasets	No.of lncRNA	No.of disease	No.of associations	Density
LncRNA Disease	285	226	621	0.01
MNDR	95	81	260	0.03
Lnc2Cancer	436	54	677	0.03

potential associations between diseases and lncRNAs based on known lncRNA-disease associations. Similar to previous studies (Wang et al., 2010; Chen et al., 2015), we construct the similarity matrix S_t of diseases by integrating the disease semantic similarity matrix inferred from the structure of directed acyclic graph that describes the relationships among diseases (Wang et al., 2010; Chen et al., 2015) and disease Gaussian interaction profile kernel similarity matrix inferred from known associations between diseases and lncRNAs (Chen and Yan, 2013; Chen et al., 2015). In particular, we obtain the similarity matrix S_t by averaging the disease similarity matrix and disease Gaussian interaction profile kernel similarity matrix (van Laarhoven et al., 2011; Chen and Yan, 2013; Chen et al., 2015, 2016a).

3.1.3. LncRNA Similarities

Since lncRNAs with similar functions tend to exhibit similar associations with diseases, calculating the similarities among lncRNAs will promotes the identification of potential associations between diseases and lncRNAs. In this study, we calculate the similarity matrix S_d of lncRNAs by integrating the functional similarity matrix calculated by the model of LNCSIM (Chen et al., 2015) and the lncRNA Gaussian interaction profile kernel similarity matrix estimated from known associations between lncRNAs and diseases (Chen and Yan, 2013). Similar to the disease similarity matrix S_t, we obtain the lncRNA similarity matrix S_d by averaging the lncRNA functional similarity matrix and Gaussian interaction profile kernel similarity matrix (van Laarhoven et al., 2011; Chen and Yan, 2013; Chen et al., 2015; Chen et al., 2016a).

3.2. Experimental Settings

To illustrate the effectiveness of our proposed TSSR model, we compare our method with other six state-of-the-art association prediction methods, namely NetlapRLS (Xia et al., 2010), BLM-NII (Mei et al., 2012), CMF (Zheng et al., 2013), PBMDA (You et al., 2017a), PRMDA (You et al., 2017b), and SIMCLDA (Lu et al., 2018). All these methods are designed for predicting the inter-associations between different types of biological entities and all of them can make use of the prior intra-associations among biological entities to improve their performance. Thus, all these algorithms are well suited for undertaking the task of lncRNA-disease association prediction. Moreover, our experiment results show that they are effective in inferring the associations between diseases and lncRNAs. Specifically, 15 repetitions of 10-fold cross validation (CV) are conducted for each model, with receiver operating characteristic (ROC) curve as the main metric to evaluate the performance. By stacking the columns of matrix Y, we obtained the vector, a $mn \times 1$ vector, denoted as vec(Y). In each repetition of 10-fold CV, we divide vec(Y) into ten disjoint folds randomly. Nine folds are treated as the training set while the remaining one fold is left out as the testing set. The AUC (Area Under Curve) score is calculated for each 10-fold CV repetition, and the final AUC score for each model are obtained by averaging over 15 such repetitions.

3.3. Parameter Settings

As each model has some hyperparameters that need to be predefined, we perform cross validation on the training set to determine the values of these hyperparameters. In particular, the parameter settings for various models are described as follows. For NetLapRLS (Xia et al., 2010), the hyperparameters satisfy $\frac{\gamma_{d2}}{\gamma_{d1}}=\frac{\gamma_{p2}}{\gamma_{p1}}$, $\beta_d = \beta_p$ with their values chosen from $\{10^{-6}, 10^{-5}, \ldots, 10^{2}\}$. For BLM-NII (Mei et al., 2012), the value of the linear combination weight α is chosen from {0, 0.1, 0.2, …, 1.0}. The max function is utilized to combine the interaction scores inferred from the disease and lncRNA sides. For the matrix factorization based methods, the dimensionality of the latent space K is selected from {50, 100} (Zheng et al., 2013). For CMF (Zheng et al., 2013), the regularization coefficient λ_1 is chosen from $\{2^{-2}, \ldots, 2^{1}\}$ (Zheng et al., 2013), while the values of λ_d and λ_t are chosen from $\{2^{-3}, 2^{-2}, \ldots, 2^{5}\}$. For PBMDA (You et al., 2017a), the maximum path length L is set to 3 and the weight threshold T is selected from $\{0.2, 0.3, \ldots, 0.8\}$ with the step size set to 0.1, while the decay factor α is set to 2.26. For SIMCLDA (Lu et al., 2018), we set the values of α_l and α_d from 0.1 to 1 with stepsize 0.1 and select the regularization parameter from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For TSSR, we choose the three parameters β and $\lambda_d = \lambda_t$ from $\{2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}\}$. Note that the most suitable hyper-parameters of a machine learning model on different datasets are usually different. Therefore, in this work, we adopt grid search (Bergstra and Bengio, 2012) to select the optimal hyperparameters for each model on each dataset.

3.4. Comparison With State-of-the-Art Methods

We conduct the experiments with 10-fold CV to shed light on the performance of TSSR in predicting potential lncRNA-disease associations, compared with other six state-of-the-art methods. Here, the AUC score is used to evaluate the predictive performance of various methods. The experiment results measured by AUC are shown in **Figures 1–3**. As shown in **Figure 1**, on LncRNADisease dataset, TSSR obtains an AUC score of 0.8736, which is higher than other methods (BLM-NII 0.8641, NetLapRLS 0.7837, CMF 0.7273, PBMDA 0.6885, PRMDA 0.7231, SIMCLDA 0.6067), indicating the superiority of our TSSR in predicting lncRNA-disease associations. We can find from **Figure 2** that on MNDR dataset, TSSR achieves the best AUC score (TSSR 0.8369, BLM-NII 0.7929, NetLapRLS 0.8210, CMF 0.8078, PBMDA 0.7722, PRMDA 0.6596, SIMCLDA 0.6187). On Lnc2Cancer dataset (the results are shown in **Figure 3**), TSSR still has competitive performance with other six methods with respect to AUC score (TSSR 0.9814, BLM-NII 0.9859, NetLapRLS 0.9392, CMF 0.9864, PBMDA 0.9680,

PRMDA 0.8179, SIMCLDA 0.6190). Note that on Lnc2Cancer, our TSSR achieves similar performance with BLM-NII and CMF. This may due to the parameter setting of TSSR. In this study, the values of the hyperparameters λ_d and λ_t (which control the influences of prior intra-similarities among lncRNAs and diseases) in our TSSR are set to same for simplicity, which is reasonable when the two data sets are balanced. However, the number of lncRNAs and diseases in Lnc2Cancer dataset are imbalanced. Thus, forcing λ_d and λ_t to be equal may limit the performance of TSSR. If the values of λ_d and λ_t are tuned separately, TSSR could achieve better performance. Moreover, to evaluate the effect of external information on the performance of TSSR, we remove the regularization terms related to the external information (i.e., setting $\lambda_d = \lambda_t = 0$) and show the results in **Figure 4**. As shown in this figure, the performance of TSSR and TSSR without external information (denoted by TSSR_original) is comparable (on LncRNADisease, TSSR 0.8736, TSSR_original 0.8735; on MNDR, TSSR 0.8369, TSSR_original 0.8367; on Lnc2Cancer, TSSR 0.9814, TSSR_original 0.9614), which means the improved performance of TSSR is mainly due to the self-representation learning. Thus, our TSSR does not depend heavily on the external information. All these results demonstrate the effectiveness of the proposed TSSR in predicting potential lncRNA-disease associations.

3.5. Effects of Parameters

The proposed TSSR involves three parameters, λ_d, λ_t, and β, where λ_d and λ_t control the influences of prior intra-associations among lncRNAs and diseases and β controls the sparsity of U and V. We will study how these parameters affect the performance of TSSR.

Figure 5 shows the prediction performance of TSSR on LncRNADisease dataset, MNDR dataset and Lnc2Cancer dataset, measured by AUC with respect to different values of λ_d and λ_t. As shown in **Figure 5**, the optimal value of $\lambda_d = \lambda_t$ for these three datasets is 2^{-10}, 2^0, and 2^2, respectively, while β is set to 2^1, 2^8, and 2^8, respectively. We find that TSSR usually performs well when the values of λ_d and λ_t are relatively small, which means the additional use of external information is not always helpful for performance improvement. On the contrary, if the external information contains noise, the performance of TSSR may decrease if we overemphasizing the effect of external information. These results demonstrate that our TSSR can effectively learn the representation matrices from known lncRNA-disease associations, and flexibly utilize external information to promote the prediction of potential lncRNA-disease associations.

In addition, we also study the impact of sparsity control parameter β. **Figure 6** illustrates the AUC scores obtained by TSSR in terms of different values of β. As shown in **Figure 6**, on these three datasets, TSSR achieves the best AUC score when the value of β is 2^1, 2^8, and 2^8, respectively, while $\lambda_d = \lambda_t$ is set to 2^{-10}, 2^0, and 2^2, respectively. We can also find from this figure that larger values of β can generally achieve better performance, which indicates the importance of controlling the sparsity of the representation matrices U and V.

3.6. Case Studies

To further validate the performance of our algorithm, based on the LncRNADisease dataset, we apply our TSSR model to identify the most possible lncRNAs that associated with three cancers (i.e., Melanoma, Glioma, and Glioblastoma). Here, all the known associations in the LncRNADisease dataset are used to train the model. Then we select the top 20 associated lncRNAs which get the highest predicted ranks for each cancer and verify these predictions based on MNDR and Lnc2Cancer databases. Moreover, the relevant literatures that support the prediction results are listed to indicate whether the predicted lncRNA-disease associations have been experimentally validated. Specially, MNDR database contains both experimental and prediction evidence (Ning et al., 2016; Ping et al., 2018). The results for the three cancers are shown in **Tables 2–4**, respectively. Note that we only show the predictions that are not included in the training set.

Melanoma is a deadly malignancy which develops from the pigment-containing cells with increasing incidence than that of any other types of cancer (Aladowicz et al., 2013). People with low level of skin pigment exposure in excess ultraviolet light (UV) have a high risk to be infected a melanoma (Kanavy and Gerstenblith, 2011). It has been estimated that by 2030, melanoma could overtake colorectal cancer as the fifth most common cancer (Rahib et al., 2014). Therefore, we apply our TSSR model to predict the potential melanoma-associated lncRNAs. According to the results shown in **Table 2** (the complete list of the top 20 identified lncRNAs is shown in **Supplementary Material**), 10 out of the top 20 identified lncRNAs have been verified. For example, Luan et al. (2016) discovered that MALAT1 could promote the cell proliferation, invasion and migration of melanoma. Li et al. observed that MEG3 was obviously decreased in melanoma cells (Li et al., 2018). They also found melanoma cell apoptosis was induced by up-regulation of MEG3, and consequently come to a conclusion that overexpression of MEG3 has a significant repression impact in melanoma cell migration and invasion ability.

Glioma is one of the most common primary malignant tumors originating in the brain, which comprises approximately 30% of all brain tumors (Goodenberger and Jenkins, 2012; Boele et al., 2015). Glioma can be graded from I to IV by World Health Organization (WHO) grading system according to their grade (Louis et al., 2016a,b). The exact causes of glioma are still unclear at the present (Kwiatkowska and Symons, 2013; Li et al., 2015). Studies have revealed the roles of lncRNAs in the development of human disease, including glioma (Zhou et al., 2018). Here, we utilize the TSSR to identify the potential lncRNAs that are more likely to related to glioma. Based on the experiment results, 9 out of the top 20 identified lncRNAs have been validated in the MNDR and Lnc2Cancer databases, and other relevant literatures. The results are shown in **Table 3** (the complete list of the top 20 identified lncRNAs is shown in **Supplementary Material**). For example, Ma et al. discovered that compared with paired

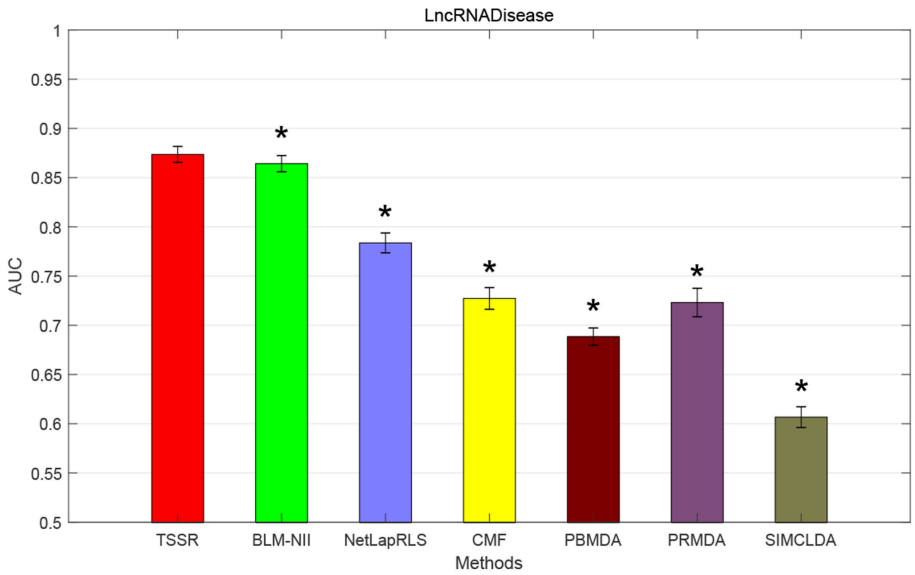

FIGURE 1 | AUC scores of various algorithms in LncRNADisease dataset (* indicates TSSR significantly outperforms the competitor with $p < 0.05$ using t-test, error bars denote 95% confidence intervals).

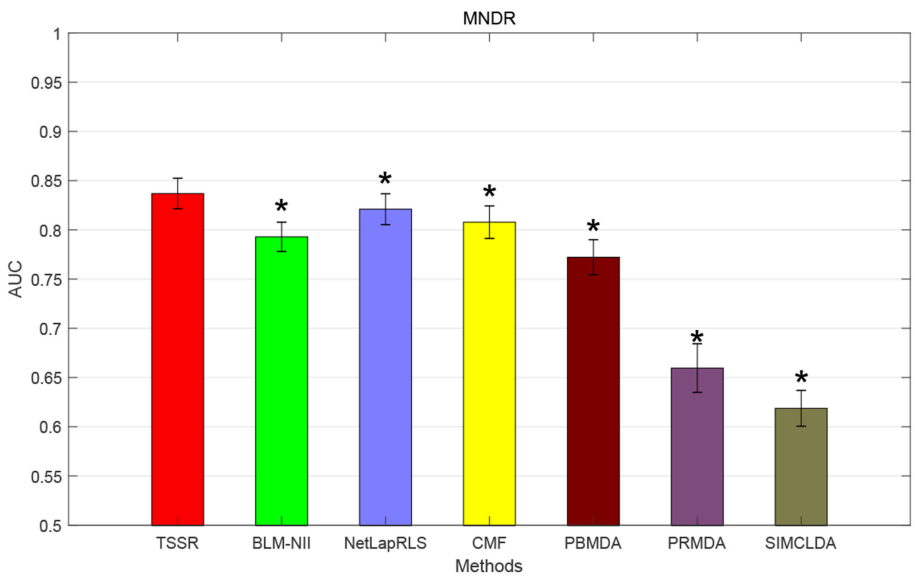

FIGURE 2 | AUC scores of various algorithms in MNDR dataset (* indicates TSSR significantly outperforms the competitor with $p < 0.05$ using t-test, error bars denote 95% confidence intervals).

normal tissues, the expression level of lncRNA MALAT1 was increased in glioma tissues, which means MALAT1 can be treated as a convictive marker for the prognosis of glioma patients (Ma et al., 2015). Zou et al. revealed that glioma patients with high PVT1 expression had low survival rate (Zou et al., 2017). Moreover, patients who received chemotherapy and radiotherapy could improve their survival by down-regulating PVT1. They also indicated that PVT1 could be served as potential target for the treatment of diffuse gliomas.

Glioblastoma, also known as glioblastoma multiform (GBM) (grade IV of Glioma), is the most common and aggressive form of primary brain tumors and kills nearly every patient in a median time of 15 months (Bleeker et al., 2012; Jovčevska et al., 2013). More importantly, there is still no clear way to prevent the disease (Gallego, 2015). Therefore, it is urgent to predict the potential glioblastoma-associated lncRNAs. In this study, we use our TSSR to undertake this task. As shown in **Table 4**, 8 out of the 20 lncRNAs have been verified in

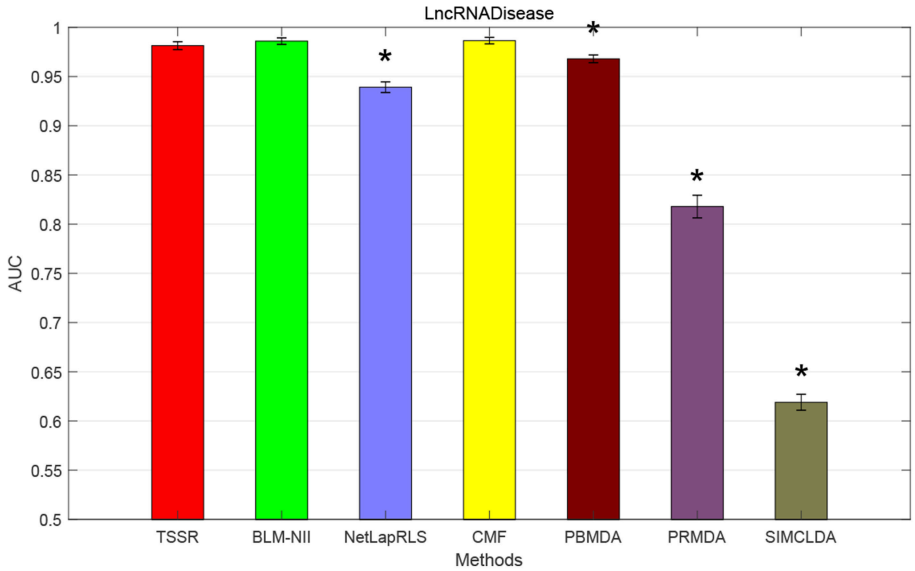

FIGURE 3 | AUC scores of various algorithms in Lnc2Cancer dataset (* indicates TSSR significantly outperforms the competitor with $p < 0.05$ using t-test, error bars denote 95% confidence intervals).

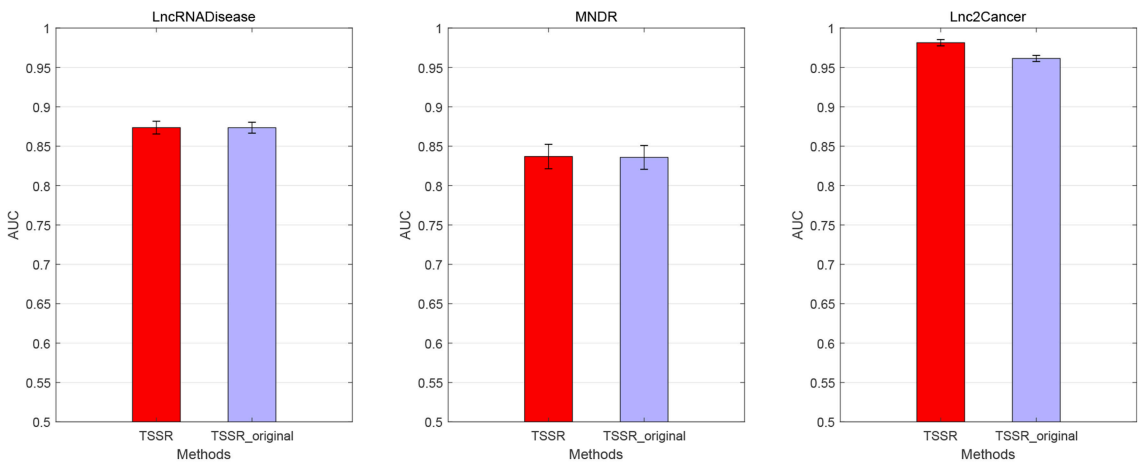

FIGURE 4 | Performance of TSSR with and without external information (denoted by TSSR and TSSR_original, respectively) on LncRNADisease, MNDR, and Lnc2Cancer datasets, measured by AUC (error bars denote 95% confidence intervals).

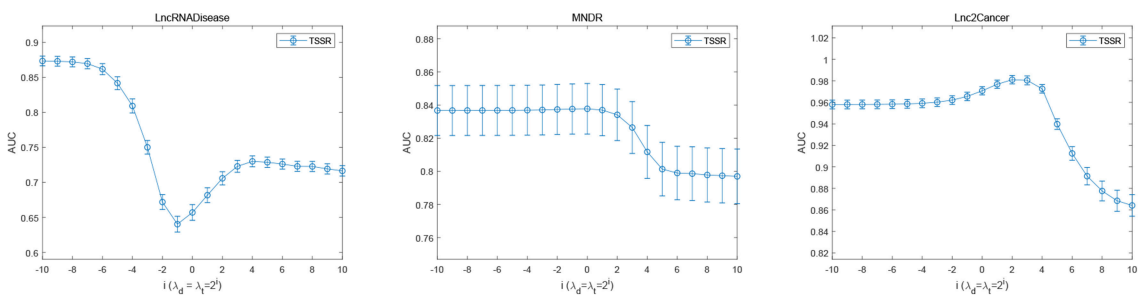

FIGURE 5 | Performance of TSSR on LncRNADisease, MNDR, and Lnc2Cancer datasets, measured by AUC with different values of λ_d and λ_t (error bars denote 95% confidence intervals).

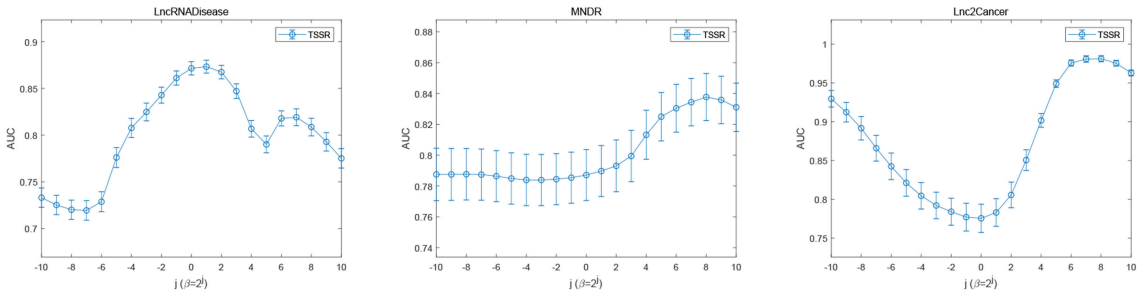

FIGURE 6 | Performance of TSSR on LncRNADisease, MNDR, and Lnc2Cancer datasets, measured by AUC with different values of β (error bars denote 95% confidence intervals).

TABLE 2 | The identified novel lncRNAs that have been verified to be associated with Melanoma.

Rank	lncRNA	Evidence(Database)	Evidence(PMID)	Expression pattern
1	CCAT2	MNDR	Prediction evidence	
2	TUSC7	MNDR	Prediction evidence	
9	GHET1	MNDR	Prediction evidence	
12	MEG3	MNDR/lnc2Cancer	29781534,29808164	Up-regulated, differential expression
13	HOTAIR	MNDR/lnc2Cancer	28067428,23862139	up-regulated
14	SOX2-OT	MNDR	Prediction evidence	
15	MALAT1	MNDR/lnc2Cancer	27725873, 27564100,27966454,24892958,19625619	Up-regulated,differential expression
17	SNHG5	MNDR/lnc2Cancer	26440365	Up-regulated
18	BCAR4	MNDR	Prediction evidence	
19	CCAT1	lnc2Cancer	28409554	Up-regulated

Prediction evidence denotes the prediction associations in MNDR database.

TABLE 3 | The identified novel lncRNAs that have been verified to be associated with Glioma.

Rank	lncRNA	Evidence(Database)	Evidence(PMID)	Expression pattern
2	HOTAIR	MNDR/lnc2Cancer	29323737,28083786 ,29218099, 27277755,24203894	Up-regulated, down-regulated
3	MALAT1	MNDR/lnc2Cancer	28551849,27134488,26649728,25613066,26619802	Up-regulated, down-regulated
4	GAS5	MNDR/lnc2Cancer	26370254,28666797	Up-regulated, down-regulated
7	PVT1	lnc2Cancer	28351322,29108264,29620147,29501773,29046366	Up-regulated, differential expression
11	SPRY4-IT1	MNDR/lnc2Cancer	29467908,27460732,26464658	Up-regulated
12	GHET1	MNDR	Prediction evidence	
15	IGF2-AS	MNDR	Prediction evidence	
18	LincRNA-p21	lnc2Cancer	28689810	Down-regulated
19	SNHG4	MNDR	Prediction evidence	

Prediction evidence denotes the prediction associations in MNDR database.

the MNDR and Lnc2Cancer databases, and other relevant literatures (the complete list of the top 20 identified lncRNAs is shown in **Supplementary Material**). For example, Zhou et al. described that HOTAIR has a significant increased expression in multiple human cancers including GBM and they found HOTAIR is necessary for GBM formation *in vivo* (Zhou X. et al., 2015). Thus, HOTAIR could be a potential therapeutic target in glioblastoma. Liu et al. found that NBAT1 has lower expressions in glioblastoma tissues compared with those in normal brain tissues and they also observed that up-regulated NBAT1 inhibits proliferation of T98 and U87 cells via regulating Akt, suggesting that NBAT1 may be related to prognosis of glioblastoma (Liu et al., 2018).

Based on the above case studies, we find that our TSSR is effective in identifying novel associations between lncRNAs and diseases based on known lncRNA-disease associations and intra-associations among lncRNAs and diseases.

4. CONCLUSION

Increasing evidences indicate the role of lncRNAs in biological processes, which motivates the development of computational

TABLE 4 | The identified novel lncRNAs that have been verified to be associated with Glioblastoma.

Rank	lncRNA	Evidence(Database)	Evidence(PMID)	Expression pattern
1	MEG3	MNDR/lnc2Cancer	27306825,28187000,22234798,25378224,26111795	Up-regulated
2	HOTAIR	MNDR/lnc2Cancer	27306825,25428914,25823657,26111795,26943771	Up-regulated
6	BCYRN1	MNDR	25561975	Differentially expressed
8	GAS5	MNDR/lnc2Cancer	27784795,23726844	Up-regulated, differentially expressed
10	NEAT1	lnc2Cancer	23046790	Up-regulated
11	HIF1A-AS2	MNDR/lnc2Cancer	27264189	Up-regulated
15	NBAT1	lnc2Cancer	29771423	Up-regulated
17	NDM29	MNDR	25561975	Differentially expressed

Prediction evidence denotes the prediction associations in MNDR database.

models to identify the potential associations between lncRNAs and diseases. Predicting the potential associations between lncRNAs and diseases based on known lncRNA-disease associations is equivalent to a recommendation problem with implicit feedback, where the task is to predict whether the unknown pairs in Y are potential associations or not. In this paper, we present a novel model, named two-side sparse self-representation (TSSR), to predict the scores of unknown pairs in Y. Based on these predicted scores, we could identify potential associations between lncRNAs and diseases. Unlike previous matrix factorization techniques that project lncRNAs and diseases into a shared latent space and predict lncRNA-disease associations based on the inner product of their latent vectors (where the dimension of latent space is previously unknown and hard to determine), our model directly learn the intra-similarities among lncRNAs and diseases from the observed associations in Y, and utilize the learned representation matrices to reconstruct Y by regarding original Y as a dictionary. As shown in Equation (4), our TSSR does not need to make many assumptions of the model in advance. Moreover, by forcing the representation matrices to be sparse, our TSSR could learn the most similar relationships among lncRNAs and diseases based on the observed associations in Y. Thus, our TSSR has data-adaptiveness and avoids the determination of some sensitive parameters such as the dimension of latent space and number of nearest neighbors. Unlike random walk-based or data integration-based methods that rely heavily on the similarity networks inferred from external information with predefined metrics, our model could adaptively learn the self-representations of lncRNAs and diseases according to their performance in reconstructing observed associations in Y. Moreover, in case the input data Y only includes a small number of known associations, our model could draw support from the intra-associations among lncRNAs and diseases derived from external information to enhance the learning of representation matrices. Therefore, our model could effectively predict potential lncRNA-disease associations by leveraging the information provided by known lncRNA-disease associations and external information of lncRNAs and diseases. Experiment results on three real data sets show that our TSSR could achieve better performance than other six state-of-the-art methods. The effectiveness of TSSR in predicting potential lncRNA-disease associations is also evaluated based on three case studies. As a link prediction algorithm, our TSSR model is flexible and could be used to handle other link prediction tasks in bipartite networks.

Furthermore, since external information of lncRNAs and diseases are utilized to enhance the performance of various methods, we also perform sensitivity analysis to assess the influences of noise information on the performances of various methods. In particular, we generate the similarity matrices S_d and S_t randomly (i.e., the elements in S_d and S_t are generated randomly) and test the performances of various methods. The detailed experiment results are shown in **Tables S4–S6**. As shown in these tables, although the performance of TSSR is affected by the noise information, it could still achieve the best performance, which means our TSSR could be used to undertake the lncRNA-disease prediction task even when the collected external information of lncRNAs and diseases contains a lot of noise.

With the development of high-throughput experimental techniques, an increasing number of data for lncRNAs and diseases are becoming available. We can calculate the similarities among lncRNAs (or diseases) based on different views of data and different metrics. How to efficiently seek the optimal combination of these similarities is an interesting future work. We will try to extend our model to handle this problem.

AUTHOR CONTRIBUTIONS

LO-Y and JH conceived and designed the study, performed the statistical analysis, and drafted the manuscript. ZZ conceived of the study, and participated in its design and coordination and helped to draft the manuscript. X-FZ and Y-RL participated in the design of the study, performed the statistical analysis, and helped to revise the manuscript. YS and SH participated in the design of the study and helped to revise the manuscript. All authors read and approved the final manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China under grants No. 61602309, 61871272, 61575125, 11871026, and 61402190, Shenzhen Fundamental Research Program, under grant JCYJ20170817095210760 and JCYJ20170302154328155, Natural Science Foundation of SZU [2017077], Guangdong Special Support Program of

Topnotch Young Professionals, under grants 2014TQ01X273, and 2015TQ01R453, Guangdong Foundation of Outstanding Young Teachers in Higher Education Institutions, under grant Yq2015141, Natural Science Foundation of Hubei province [ZRMS2018001337].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00476/full#supplementary-material

REFERENCES

Aladowicz, E., Ferro, L., Vitali, G. C., Venditti, E., Fornasari, L., and Lanfrancone, L. (2013). Molecular networks in melanoma invasion and metastasis. *Future Oncol.* 9, 713–726. doi: 10.2217/fon.1

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13, 281–305. doi: 10.1016/j.chemolab.2011.12.002

Bhartiya, D., Kapoor, S., Jalali, S., Sati, S., Kaushik, K., Sachidanandan, C., et al. (2012). Conceptual approaches for lncRNA drug discovery and future strategies. *Expert Opin. Drug Discov.* 7, 503–513. doi: 10.1517/17460441.2012.682055

Bleeker, F. E., Molenaar, R. J., and Leenstra, S. (2012). Recent advances in the molecular understanding of glioblastoma. *J. Neuro Oncol.* 108, 11–27. doi: 10.1007/s11060-011-0793-0

Boele, F. W., Rooney, A. G., Grant, R., and Klein, M. (2015). Psychiatric symptoms in glioma patients: from diagnosis to management. *Neuropsychiatr. Dis. Treat.* 11:1413. doi: 10.2147/NDT.S65874

Bussemakers, M. J., Van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F., Schalken, J. A., et al. (1999). Dd3:: A new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* 59, 5975–5979.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNAdisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi: 10.1093/nar/gks1099

Chen, L., Ma, D., Li, Y., Li, X., Zhao, L., Zhang, J., et al. (2018). Effect of long non-coding RNA PVT1 on cell proliferation and migration in melanoma. *Int. J. Mol. Med.* 41, 1275–1282. doi: 10.3892/ijmm.2017.3335

Chen, X., Huang, Y.-A., Wang, X.-S., You, Z.-H., and Chan, K. C. (2016a). FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. *Oncotarget* 7, 45948–45958. doi: 10.18632/oncotarget.10008

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5:11338. doi: 10.1038/srep11338

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016b). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinformatics* 18, 558–576. doi: 10.1093/bib/bbw060

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

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Gallego, O. (2015). Nonsurgical treatment of recurrent glioblastoma. *Curr. Oncol.* 22:e273. doi: 10.3747/co.22.2436

Goodenberger, M. L., and Jenkins, R. B. (2012). Genetics of adult glioma. *Cancer Genet.* 205, 613–621. doi: 10.1016/j.cancergen.2012.10.009

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240–251. doi: 10.1016/j.cell.2013.06.009

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Jovčevska, I., Kočevar, N., and Komel, R. (2013). Glioma and glioblastoma-how much do we (not) know? *Mol. Clin. Oncol.* 1, 935–941. doi: 10.3892/mco.2013.172

Kanavy, H. E., and Gerstenblith, M. R. (2011). Ultraviolet radiation and melanoma. *Semin. Cutan. Med. Surg.* 30, 222–228. doi: 10.1016/j.sder.2011.08.003

Kwiatkowska A., and Symons M. (2013). "Signaling determinants of glioma cell invasion," in *Glioma Signaling. Advances in Experimental Medicine and Biology,* Vol. 986, ed J. Barańska (Dordrecht: Springer), 121–141.

Li, J., Xuan, Z., and Liu, C. (2013). Long non-coding rnas and complex human diseases. *Int. J. Mol. Sci.* 14, 18790–18808. doi: 10.3390/ijms140918790

Li, J., Yuan, J., Yuan, X., Zhang, C., Li, H., Zhao, J., et al. (2015). Induction effect of microRNA-449a on glioma cell proliferation and inhibition on glioma cell apoptosis by promoting pkcα. *Eur. Rev. Med. Pharmacol. Sci.* 19, 3587–3592. Available online at: https://www.europeanreview.org/article/9593

Li, P., Gao, Y., Li, J., Zhou, Y., Yuan, J., Guan, H., et al. (2018). LncRNA meg3 repressed malignant melanoma progression via inactivating Wnt signaling pathway. *J. Cell. Biochem.* 119, 7498–7505. doi: 10.1002/jcb.27061

Liu, J., Wang, W., Zhang, X., Du, Q., Li, H., and Zhang, Y. (2018). Effect of downregulated lncrna nbat1 on the biological behavior of glioblastoma cells. *Eur. Rev. Med. Pharmacol. Sci.* 22, 2715–2722. doi: 10.26355/eurrev_201805_14968

Louis, D. N., Ohgaki, H., Wiestler, O. D., and Cavenee, W. K. (2016a). *WHO Classification of Tumours of the Central Nervous System*. International Agency for Research on Cancer.

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016b). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

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

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14:651. doi: 10.1186/1471-2164-14-651

Luan, W., Li, L., Shi, Y., Bu, X., Xia, Y., Wang, J., et al. (2016). Long non-coding RNA MALAT1 acts as a competing endogenous RNA to promote malignant melanoma growth and metastasis by sponging miR-22. *Oncotarget* 7, 63901–63912. doi: 10.18632/oncotarget.11564

Ma, K., Wang, H., Li, X, Li, T., Su, G., Yang, P., et al. (2015). Long noncoding RNA MALAT1 associates with the malignant status and poor prognosis in glioma. *Tumor Biol.* 36, 3355–3359. doi: 10.1007/s13277-014-2969-7

Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A., and Koonin, E. V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* 3, 1390–1404. doi: 10.1093/gbe/evr116

Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L., and Zheng, J. (2012). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 29, 238–245. doi: 10.1093/bioinformatics/bts670

Ning, S., Zhang, J., Peng, W., Hui, Z., Wang, J., Yue, L., et al. (2016). Lnc2cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* 44(Database issue), D980–D985. doi: 10.1093/nar/gkv1094

Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., and Pei, T. (2018). A novel method for lncRNA-disease association prediction based on an lncRNA-disease association network. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 688–693. doi: 10.1109/TCBB.2018.2827373

Rahib, L., Smith, B. D., Aizenberg, R., Rosenzweig, A. B., Fleshman, J. M., and Matrisian, L. M. (2014). Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* 74, 2913–2921. doi: 10.1158/0008-5472.CAN-14-0155

Schonrock, N., Harvey, R. P., and Mattick, J. S. (2012). Long noncoding RNAs in cardiac development and pathophysiology. *Circul. Res.* 111, 1349–1362. doi: 10.1161/CIRCRESAHA.112.268953

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: regulators of disease. *J. Pathol.* 220, 126–139. doi: 10.1002/path.2638

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

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

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

Wapinski, O., and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends Cell Biol.* 21, 354–361. doi: 10.1016/j.tcb.2011.04.001

Xia, Z., Wu, L.-Y., Zhou, X., and Wong, S. T. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* 4:S6. doi: 10.1186/1752-0509-4-S2-S6

Yang, Z., and Oja, E. (2011). Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Trans. Neural Netw.* 22, 1878–1891. doi: 10.1109/TNN.2011.2170094

Yang, Z., and Oja, E. (2012). "Clustering by low-rank doubly stochastic matrix decomposition," in *Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML'12* (Edinburgh: Omnipress), 707–714.

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., et al. (2017a). PBMDA: A novel and effective path-based computational model for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

You, Z.-H., Wang, L.-P., Chen, X., Zhang, S., Li, X.-F., Yan, G.-Y., et al. (2017b). PRMDA: personalized recommendation-based MiRNA-disease association prediction. *Oncotarget* 8, 85568–85583. doi: 10.18632/oncotarget.20996

Yu, J., Ping, P., Wang, L., Kuang, L., Li, X., and Wu, Z. (2018). A novel probability model for lncRNA–disease association prediction based on the naïve bayesian classifier. *Genes* 9:345. doi: 10.3390/genes9070345

Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. (2015). A survey of sparse representation: algorithms and applications. *IEEE Access* 3, 490–530. doi: 10.1109/ACCESS.2015.2430359

Zhao, W., Luo, J., and Jiao, S. (2014). Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci. Rep.* 4:6591. doi: 10.1038/srep06591

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, IL: ACM), 1025–1033. doi: 10.1145/2487575.2487670

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

Zhou, Q., Liu, J., Quan, J., Liu, W., Tan, H., and Li, W. (2018). lncRNAs as potential molecular biomarkers for the clinicopathology and prognosis of glioma: a systematic review and meta-analysis. *Gene* 668, 77–86. doi: 10.1016/j.gene.2018.05.054

Zhou, X., Ren, Y., Zhang, J., Zhang, C., Zhang, K., Han, L., et al. (2015). HOTAIR is a therapeutic target in glioblastoma. *Oncotarget* 6, 8353–8365.

Zou, H., Wu, L.-X., Yang, Y., Li, S., Mei, Y., Liu, Y.-B., et al. (2017). lncRNAs PVT1 and HAR1A are prognosis biomarkers and indicate therapy outcome for diffuse glioma patients. *Oncotarget* 8, 78767–78780.

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