frontiers in Genetics

METHODS
published: 23 January 2019
doi: 10.3389/fgene.2018.00685

Multiple Partial Regularized Nonnegative Matrix Factorization for Predicting Ontological Functions of lncRNAs

Jianbang Zhao[1]* *and Xiaoke Ma*[2]*

[1] College of Information Engineering, Northwest Agriculture & Forestry University, Xianyang, China, [2] School of Computer Science and Technology, Xidian University, Xi'an, China

Long non-coding RNAs (LncRNA) are critical regulators for biological processes, which are highly related to complex diseases. Even though the next generation sequence technology facilitates the discovery of a great number of lncRNAs, the knowledge about the functions of lncRNAs is limited. Thus, it is promising to predict the functions of lncRNAs, which shed light on revealing the mechanisms of complex diseases. The current algorithms predict the functions of lncRNA by using the features of protein-coding genes. Generally speaking, these algorithms fuse heterogeneous genomic data to construct lncRNA-gene associations via a linear combination, which cannot fully characterize the function-lncRNA relations. To overcome this issue, we present an nonnegative matrix factorization algorithm with multiple partial regularization (aka MPrNMF) to predict the functions of lncRNAs without fusing the heterogeneous genomic data. In details, for each type of genomic data, we construct the lncRNA-gene associations, resulting in multiple associations. The proposed method integrates separately them via regularization strategy, rather than fuse them into a single type of associations. The results demonstrate that the proposed algorithm outperforms state-of-the-art methods based network-analysis. The model and algorithm provide an effective way to explore the functions of lncRNAs.

Keywords: lncRNA, nonnegative matrix factorization, gene ontology, networks, regularization

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Qinghua Jiang,
Harbin Institute of Technology, China
Jianbo Pan,
Johns Hopkins Medicine, United States

***Correspondence:**
Jianbang Zhao
zhaojianbang@nwsuaf.edu.cn
Xiaoke Ma
xkma@xidian.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *29 October 2018*
Accepted: *10 December 2018*
Published: *23 January 2019*

Citation:
Zhao J and Ma X (2019) Multiple Partial Regularized Nonnegative Matrix Factorization for Predicting Ontological Functions of lncRNAs. Front. Genet. 9:685. doi: 10.3389/fgene.2018.00685

1. INTRODUCTION

Long non-coding RNAs (lncRNAs) are a type of non-coding RNAs with more than 200 nucleotides in length, which have very little or no potential to encode proteins (Mercer et al., 2009). In the past lncRNAs are categorized as "dark matter" and "junks." However, more and more evidence demonstrates that lncRNAs are critical regulators for biological processes, such as immune response, cell development and differentiation, as well as gene imprinting (Morris and Mattick., 2014; Turner et al., 2014; Ma et al., 2017). Furthermore, lncRNAs are highly related to diseases and cancers (Zou et al., 2015, 2016; Zhu et al., 2018). Largely due to the high-throughput biological techniques, particularly the next generation sequence (NGS), large numbers of lncRNAs have been identified (Iyer et al., 2015; Fang et al., 2018).

Compared to the protein-coding genes (genes for short), the functions of vast majority of lncRNAs are unknown. Thus, it is promising to predict the functions of lncRNAs, which

are critical for revealing the underlying mechanisms of gene regulation. The approaches for annotating the functions of lncRNAs are classified into two classes: the biological experiment and computational based methods. Currently, the functions of some lncRNAs are validated by the biological experiment based methods. For example, based on the RNA-sequencing data, the mechanistic analysis reveals that UCA1 physically interacts with PTBP1 and ALAS2, which stabilizes ALAS2 (Liu et al., 2018). Li et al. (2016) utilized the RT-PCR to detect the expression profiles of lncRNA TUG1 in glioma, and found that TUG1 is involved in the apoptosis and cell proliferation. Based on the cap analysis of gene expression (CAGE) data, FANTOME generated a comprehensive atlas of 27919 human lncRNA genes across 1829 samples from the major human primary cell types and tissues (Hon et al., 2017). Wang et al. (2018) identified the function of NEAT1 using the enhanced green fluorescent protein reporter in human cells.

Except the expression profiles, some lncRNAs execute their functions via interacting with other bio-molecules, such as DNAs, RNAs and proteins. Mercer and Mattick (2013) focused on the lncRNAs as epigenetic modulators via binding to chromatin-modifying proteins and recruiting their catalytic activity to specific sites in the genome. Efforts is devoted to investigate the lncRNA-DNA interactions, including the chromatin isolation by RNA purification (Chu et al., 2012; Nowak et al., 2014). Furthermore, Ferre et al. (2016) identified the protein-lncRNA interactions, offering essential clues for a better understanding of lncRNA cellular mechanisms and their disease-associated perturbations.

Even though the experiment based approaches for the functions of lncRNAs are reliable, they are criticized by the expensive cost and complicated operations. Thus, the computational algorithms for the prediction of lncRNA functions provide an alternative, which become more and more important. Based on the assumption that the molecules with the same or similar functions have the same or similar patterns. Some efforts explore the co-expression patterns (Lee et al., 2004; Necsulea et al., 2014). Furthermore, the gene set enrichment analysis (GSEA) based on the statistics is also adopted to identify the functions of lncRNAs (Guttman et al., 2009). To explore the knowledge from genes, (Liao et al., 2011) combined the expression profiles of lncRNAs and genes to construct a coding and non-coding gene co-expression network according to the expression profiles in the GEO database, then predicted the functions of more than 300 mouse lncRNAs based on the co-expression modules. In order to make use of the global information, Guo et al. (2013) constructed a bi-colored network via integrating the expression profiles of lncRNA and genes, then provided the lnc-GFP algorithm to predict the functions of lncRNAs. Jiang et al. (2015) employed the statistical test to annotate the functions of lncRNAs. Recently, Zhang et al. (2018) proposed the NeuralNetL2GO algorithm, which uses neural networks to annotate lncRNAs.

Actually, there are many different genomic data to link the lncRNA and genes, for example gene co-expression, connection to the diseases, protein binding sites. The current algorithms integrate multiple heterogeneous genomic data into a single network via weighted or unweighted linear functions, which are criticized for not fully characterizing the links between lncRNAs and genes. Evidence shows that the linear combination destroys the patterns in the integrated network (Ma and Dong, 2017; Ma et al., 2019). In fact, each type of genomic data provides a perspective of the links between lncRNAs and genes. The ultimate goal of this study is to provide a computational method to predict functions of lncRNAs by fusing heterogeneous data. As shown in **Figure 2**, we construct multiple bi-color networks for lncRNAs and genes. Then, the multiple partial regularized nonnegative matrix factorization (MPrNMF) algorithm is proposed to simultaneously factorize the multiple networks. In order to improve the accuracy, the regularization strategy is adopted, where the factorized feature matrix preserves the links between lncRNAs and genes. The results demonstrate that the proposed method outperforms these algorithms based on the single bio-colored network, implying the proposed method is promising.

The rest of this paper is organized as: section 2 briefly reviews the related works on the prediction of lncRNAs functions. Section 3 describes the procedure of the proposed method. Section 4 shows the experimental results. Finally, the conclusion is presented in section 5.

2. RELATED WORKS

In this section, we first introduce the mathematical notations that are widely used in the forthcoming sections. Then, we review state-of-the-art methods for the prediction of lncRNA functions.

2.1. Notations

The notations are summarized in **Table 1**. Let n be the number of entities in the networks. Generally speaking, let n_o be the number of ontological functions in Gene Ontology (GO), n_g be the number of proteins (genes) in the PPI network, n_l be the number of lncRNAs in the co-expression network. Let G_g, G_l be the PPI and lncRNA co-expression networks, respectively. The adjacency matrix for G_g, denoted by W_g, corresponds to a $n_g \times n_g$ matrix whose element $w_{ij}^{[g]}$ is the weight on edge (v_i, v_j) in G_g. The degree of vertex v_i in G_g is the sum of weights on edges connecting v_i, i.e., $d_i^{[g]} = \sum_j w_{ij}^{[g]}$. The degree matrix D_g is the diagonal matrix with degree sequence of G_g, i.e., $D_g = diag(d_1^{[g]}, d_2^{[g]}, \ldots, d_n^{[g]})$. The Laplacian matrix of G_g is defined as $L_g = I - D_g^{-1/2} W_g D_g^{-1/2}$. Analogously, the adjacent matrix of G_l is denoted by W_l. Let L_l be the Laplacian matrix for G_l. The associations between heterogeneous entities are denoted by matrix. Specifically, let X be the known lncRNA-ontology associations, Y be the known gene-lncRNA associations, and $Y_1(Y_2)$ be the known lncRNA-disease (gene-disease) associations, respectively.

2.2. Related Algorithms

The label propagation algorithm is successfully applied to predict phenotype-gene associations with various backgrounds (Li and Patra, 2010; Vanunu et al., 2010), where the principle of the label propagation algorithms is illustrated in **Figure 1A**. In details,

label propagation assumes that the well connected lncRNAs in G_l are very likely to be the same label, which leads to the following objective function

$$J_{LP} = \theta tr(\widehat{X}L_l\widehat{X}^{'}) + (1-\theta)\|\widehat{X} - X\|^2, \quad (1)$$

where $\widehat{X}$ is the predicted lncRNA-ontology associations, $\theta \in (0, 1)$ is the parameter controlling the contributions of two terms in Equation (1), $tr(A)$ is the trace of matrix A, i.e., $tr(A) = \sum_i a_{ii}$ and $\|A\|$ is the l_2 norm of matrix A. In Equation (1), the first item characterizes how the predicted lncRNA-ontology associations $\widehat{X}$ is consistent with the lncRNA co-expressed network, while the second one measures the good the predicted associations fit the initial labeling.

However, the number of predicted associations is largely determined by the sparsity of the known associations in X. When X is very sparse, the number of predicted associations is limited. Actually, X is very sparse since the GO functions of vast majority of lncRNAs are unknown. Fortunately, the GO functions of most proteins are known. Thus, the available algorithms overcome this limitation of the label propagation algorithm via integrating the proteins and lncRNAs as shown in **Figure 1B**. Specifically, given the known protein-GO associations X, PPI network G_g, lncRNA co-expression network G_l and lncRNA-gene associations Y, the ultimate goal is to predict the lncRNA-ontology associations via integrative analysis of heterogeneous data. The lnc-GFP algorithm (Guo et al., 2013) follows the label propagation method by using the bi-colored network, which is defined as

$$C = \begin{bmatrix} W_l & Y \\ Y^{'} & W_g \end{bmatrix}. \quad (2)$$

Thus, the objective function in Equation (1) is transformed into

$$J_{LP} = \theta tr(\widehat{X}L_C\widehat{X}^{'}) + (1-\theta)\|\widehat{X} - X\|^2, \quad (3)$$

where L_C is the Laplacian matrix of the bi-colored network C. The KATZLGO method (Zhang et al., 2017) predicts the GO functions of lncRNAs by using the KATZ score of the bi-colored network, which counts the paths with various lengths in the bi-colored networks.

The bi-colored based methods make use of lncRNA-gene associations to predict the functions of lncRNAs. To explore the knowledge in G_l and G_g, Petergrosso et al. (2017) proposed the dual label propagation (DLP) to predict the phenotome-genome associations. Specifically, the objective function in Equation(1) based on the DLP model can be re-written as

$$J_{DLP} = \|\widehat{X} - X\|^2 + \beta tr(\widehat{X}L_g\widehat{X}^{'}) + \gamma tr(\widehat{X}L_l\widehat{X}^{'}), \quad (4)$$

where $\beta \geq 0, \gamma \geq 0$ are tuning parameters. The first item measures the consistence between the predicted associations and the bi-colored network, and the last two ones measures the smoothness in the PPI and lncRNA networks.

Most of the available algorithms for the prediction of LncRNA functions are based on the bi-colored network model. In this study, we investigate the possibility to predict the functions of lncRNAs via integrating multiple networks, where each type of genomic data is used to construct the lncRNA-gene associations.

TABLE 1 | Notations and descriptions.

Symbol	Definition and description
n_o, n_g, n_l	Number of ontological functions, genes and lncRNAs
G	graph with vertex set V and edge set E
X	Known lincRNA-ontology associations
Y_1, Y_2	Known lncRNA-gene associations
G_g	Protein-Protein interaction (PPI) network
G_l	LncRNA co-expression network
$\overline{W}_g$	Normalized adjacent matrix of the PPI network $\overline{W}_g = D^{-1/2}W_gD^{-1/2}$
$\overline{W}_l$	Normalized adjacent matrix of lncRNA co-expression network $\overline{W}_g = D^{-1/2}W_lD^{-1/2}$
L_g	Normalized Laplacian matrix of G_g, i.e., $L_g = I - \overline{W}_g$
L_l	Normalized Laplacian matrix of G_l, i.e., $L_l = I - \overline{W}_l$

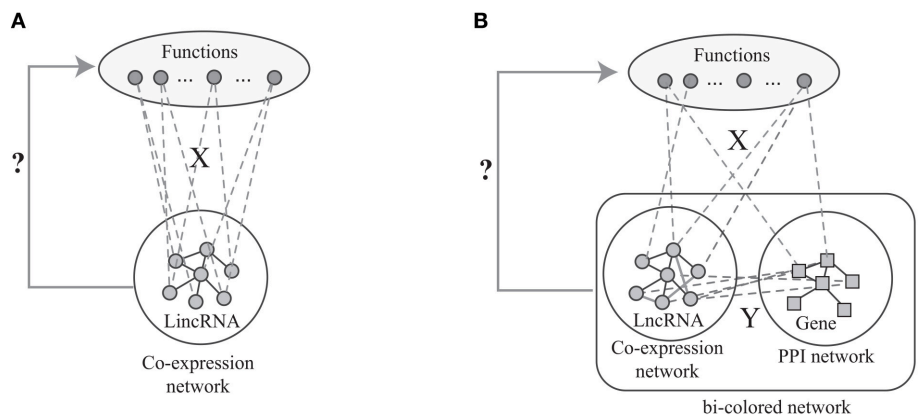

FIGURE 1 | The flowchart of the current algorithms based on network analysis: **(A)** label propagation method based on the lncRNA co-expression network, **(B)** label propagation method based on the bio-colored network.

3. METHODS

The procedure of MPrNMF is illustrated in **Figure 2**. In this section, we derive the objective function and optimizing rules of the proposed algorithm in turns.

3.1. Objective Function

All these bi-colored network based algorithms predict the lncRNA-ontology associations based on the single bi-colored network via integrating various genomic data. In this study, we construct two bi-colored networks, where each one corresponds to a view of the lncRNA-gene associations. In the first one, the lncRNA-gene associations are determined by the pearson correlation coefficient between the expression profiles of lncRNAs and genes. And, the second lncRNA-gene associations are determined by the diseases. In details, the lncRNA-gene association is the Jaccard index of the diseases related to lncRNAs and genes. The i-th view of the bi-colored network is denoted by

$$C_i = \begin{bmatrix} W_l & Y_i \\ Y_i^{'} & W_g \end{bmatrix}, \tag{5}$$

where $Y_i (i = 1, 2)$ is the lncRNA-gene associations in the i-th view.

Given the lncRNAs(genes)-ontology associations X, NMF aims at obtaining approximation of X via the product of two nonnegative matrices B_1 and F_t (Lee and Seung, 1999), i.e.,

$$J = \|X - BF\|^2, \quad s.t. \quad B \geq 0, F \geq 0, \tag{6}$$

where B is the basis matrix and F is the feature matrix. Furthermore, we also expect the feature matrix F also reflects the topological structure of multiple views of the bi-colored network, which is implemented via the regularization. To this end, the Equation (6) is reformulated as

$$J = \|X - BF\|^2 + \alpha \sum_{i=1}^{2} tr(FC_iF^{'}), \quad s.t. \quad B \geq 0, F \geq 0, \tag{7}$$

where parameter α controls the importance of the regularization items and $tr(A)$ is the trace of matrix A, i.e., $tr(A) = \sum_i a_{ii}$.

In the bi-colored network, the vertices consist of lncRNAs and genes. Thus, the feature matrix F is also re-written as $F = [F_l, F_g]$, where F_l denotes the part for the lncRNAs and F_g for genes. Thus, $tr(FC_iF^{'})$ is reformulated as

$$\begin{aligned} tr(FC_iF^{'}) &= tr\left([F_l, F_g]\begin{bmatrix} W_l & Y_i \\ Y_i^{'} & W_g \end{bmatrix}\begin{bmatrix} F_l^{'} \\ F_g^{'} \end{bmatrix}\right) \\ &= tr(F_gW_gF_g^{'} + F_lY_i^{'}F_g^{'} + F_lY_iF_g^{'} + F_lW_lF_l^{'}) \\ &= tr(F_gW_gF_g^{'}) + 2tr(F_lY_i^{'}F_g^{'}) + tr(F_lW_lF_l^{'}). \end{aligned}$$

The above equation indicates that the regularization item for the bi-colored network can be divided into three components: W_g, W_l and Y_i. In the two views, the only difference is the lncRNA-gene relations. Thus, we expect the regularization item can fully relect the lncRNA-gene relations Y_i. In this case, the objective function in Equation (7) is transformed into

$$\begin{aligned} \min J &= \|X - BF\|^2 + \alpha \sum_{i=1}^{2} tr(F_lY_iF_g^{'}) \\ s.t. \quad & B \geq 0, F \geq 0, F^{'}1_{n_l+n_g} = 1_{n_l+n_g} \end{aligned} \tag{8}$$

where 1_n is the column vector with all elements 1. The l_1-norm constraint on matrix F_t is adopted to obtain sparsity solutions.

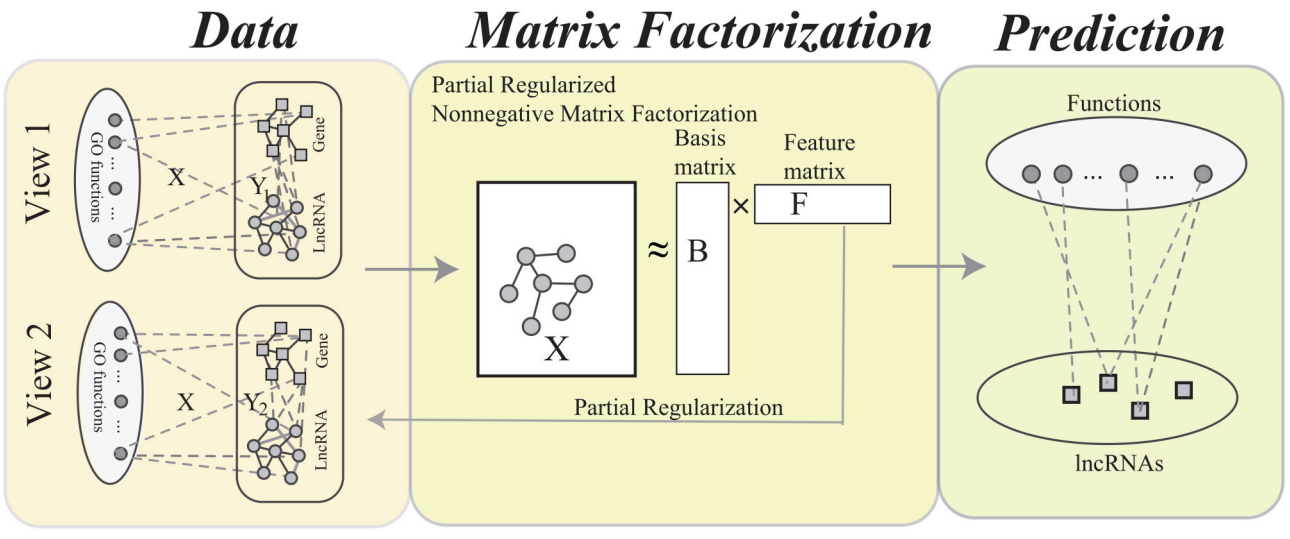

FIGURE 2 | The flowchart of the MPrNMF algorithm, which consists of three components: network construction, matrix factorization and function prediction. In the network construction, each type of heterogenous lncRNA-gene associations is used to construct a bi-colored network. The matrix factorization procedure obtains approximation of lncRNA(gene)-ontology associations X, where the feature matrix F reflects multiple lncRNA-gene associations. The function prediction procedure is based on the decomposed matrices.

3.2. Optimization Rules

To optimize the objective function in Equation (8), we derive the updating rules for matrix B and F. Since the objective function is non-convex, we update one matrix by fixing the other, which continues until the termination criterion is reached.

By integrating the sparsity constraint of matrix F, the Lagrange function for objective function is formulated as

$$\begin{aligned} L &= \|X - BF\|^2 + 2\alpha \sum_{i=1}^{2} tr(F_g Y_i F_l^{'}) + tr(\Lambda(F^{'} 1_{n_l+n_g} - 1_{n_l+n_g}) \\ &\quad (F^{'} 1_{n_l+n_g} - 1_{n_l+n_g})^{'}) \\ &= \|X - BF\|^2 + \alpha \sum_{i=1}^{2} tr(FC_i^* F^{'}) + tr(\Lambda(F^{'} 1_{n_l+n_g} - 1_{n_l+n_g}) \\ &\quad (F^{'} 1_{n_l+n_g} - 1_{n_l+n_g})^{'}), \end{aligned}$$

where matrix C_i^* is defined as

$$C_i^* = \begin{bmatrix} 0 & Y_i \\ Y_i^{'} & 0 \end{bmatrix}.$$

The derivative of L on B is calculated as

$$\frac{1}{2}\nabla_B L = XF^{'} - BFF^{'},$$

and the derivative of L on F is written as

$$\frac{1}{2}\nabla_F L = B^{'}X - B^{'}BF^{'} + \alpha \sum_{i=1}^{2} FC_i^* - 1_{n_l+n_g} 1_{n_l+n_g}^{'} \Lambda.$$

According to the Karush-Kuhn-Tucker condition, by setting $\frac{1}{2}\nabla_B L$=0, we obtain the updating rule for matrix B as

$$B = B \odot \sqrt{\frac{[BFF^{'}]}{[XF^{'}]}}, \tag{9}$$

where $\odot$ denotes element-wise product, $[\cdot]/[\cdot]$ denotes element-wise division and $\sqrt{\cdot}$ is the element-wise square root. Analogously, the updating rule for matrix F is derived as

$$F = F \odot \sqrt{\frac{[B^{'}BF^{'}]}{[B^{'}X + \alpha(FC_1^* + FC_2^*)]}}. \tag{10}$$

After obtaining matrices B and F, we divide the matrix $B = \begin{bmatrix} B_l \\ B_g \end{bmatrix}$. The prediction of lncRNA-ontology is obtained as $B_l F_l$. The procedure of the proposed algorithm is illustrated in Algorithm 1. Usually, the number of iterations is 100.

Algorithm 1 The MPrNMF algorithm

Input:

$Y_i (1 \geq i \geq n)$: The multiple views of lncRNA-gene associations;

X: The known lncRNA(gene)-ontology associations;

k: number of communities;

α: weight for multiple views;

Output:

$\widehat{X}_l$: the predicted lncRNA-ontology associations.

Part I: Matrix Decomposition

1: Initialing randomly B and F;
2: Fixed matrix F, update matrix B according to Equation (9);
3: Fixed matrix B, update matrix F according to Equation (10);
4: Continue Step 2 and 3 until the termination criterion is reached;

Part II: Predicting lncRNA-ontology associations

5: Predicting the lncRNA-ontology associations as $\widehat{X}_l = B_l F_l$;
6: **return** $\widehat{X}_l$.

4. RESULTS

4.1. Data

The PPI network is downloaded from the BioGrid database (https://thebiogrid.org/). We select the maximal connected subgraph in the PPI network for analysis. The lncRNAs are downloaded from the GENCODE database (https://www.gencodegenes.org/). The gene-disease associations are downloaded from the OMIM database (https://omim.org/), while the lncRNA-disease associations are downloaded from the LncRNADisease database (http://www.cuilab.cn/lncrnadisease). The expression profiles are downloaded from the COXPRESdb database Okamura et al. (2018) (http://coxpresdb.jp/), where the three preprocessed datasets, including Hsa.c4-1, Hsa2.c2-0, and Hsa3.c1-0, are used.

Since there is no available public database for the ontology of lncRNAs, Zhang and Ma (2018) manually curate a set of 55 lncRNAs with 129 GO terms by literature searching. We adopt this dataset as benchmark to test the performance of the proposed method.

4.2. Criterion

To predict the lncRNA-ontology associations, the output of the proposed algorithm is a real value in the interval [0,1]. Hence a threshold is need to determine the final prediction. Following the NeuraNetL2GO algorithm (Zhang and Ma, 2018), we use the Recall, Precision and Fmax to quantify the accuracy of algorithms. Specifically, let t be the threshold, and $P(t)$ be the set of predicted ontology, and T be the ontology in the benchmark dataset. For the i-th lncRNA, the true positives (TP), false positives (FP) and false negatives (FN) are defined as

$$TP_i = \sum_{o \in \mathcal{O}} I(f \in P_i(t) \wedge f \in T_i), \tag{11}$$

$$FP_i = \sum_{o \in \mathcal{O}} I(f \in P_i(t) \wedge f \notin T_i), \tag{12}$$

$$FN_i = \sum_{o \in \mathcal{O}} I(f \notin P_i(t) \wedge f \in T_i), \tag{13}$$

where o is an ontology, $\mathcal{O}$ denotes the set of all functions, and $I(x)$ is indicator function with value 1 if x is true, 0 otherwise. The recall, precision, and Fmax are defined as

$$Recall = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i}, \tag{14}$$

$$Precision = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i}, \tag{15}$$

$$Fmax = \max_t \frac{2Recall(t)Precision(t)}{Recall(t) + Precision(t)}. \tag{16}$$

4.3. Parameter Selection

There are two parameters involved in MPrNMF: parameter k is the number of features, and parameter α controls the relative importance of partial regularization items. On the parameter k, Wu et al. (2016) proposed the instability based NMF model for parameter selection. For each k, MPrNMF runs τ times with random initial solutions and obtains τ basis matrices, denoted by $B_1, \ldots, B_\tau$. Given two matrices B_1 and B_2, a $\tau \times \tau$ matrix H is defined where the element h_{ij} is the cross correlation between the i-th column of matrix B_1 and the j-th column of matrix B_2. The dissimilarity between B_1 and B_2 is defined as

$$diss(B_1, B_2) = \frac{1}{2k}(2k - \sum_j \max H_{\cdot j} - \sum_i \max H_{i\cdot}),$$

where $H_{\cdot j}$ denotes the j-th column of matrix Q. The instability is the discrepancy of all the basis matrices for k, which is defined as

$$\Upsilon(k) = \frac{2}{\tau(\tau - 1)} \sum_{1 \le i < j \le \tau} diss(B_i, B_j).$$

As shown in **Figure 3A**, the instability of MPrNMF changes as the number of features k ranges from 40 to 64 with gap 4. When k <52, the instability decreases, while it increases if k >52. The reason is that when k is small, the number of features cannot fully characterize topological structure of associations, while large k results in the redundance of features. It reaches minimum at $k = 52$. Thus, we set $k = 52$.

How the parameter α effects the performance of MPrNMF is illustrated in **Figure 3B**, where the Fmax changes as α increases from 0.1 to 2 with a gap 0.2. It is easy to assert that, when α increases from 0.1 to 1, the performance also improves. The accuracy of the proposed algorithm is robust when $\alpha > 1$. The reason is that when α is small, the objective function is dominated by the associations between lncRNA(gene)-ontology diseases. As α increases, the contribution of the regularization items for the multiple views of lncRNA-gene associations increases, improving the accuracy. Therefore, we set $\alpha = 1$ since it reaches a good balance between lncRNA(gene)-ontology associations and lncRNA-gene associations.

4.4. Performance

To fully validate the performance of MPrNMF, three algorithms are selected for a comparison, including lnc-GFP (Guo et al., 2013), Lnc2Function (Jiang et al., 2015) and NeuraNetL2GO (Zhang and Ma, 2018, because of their excellent performance. In this study, we only focus on the biological process of GO terms.

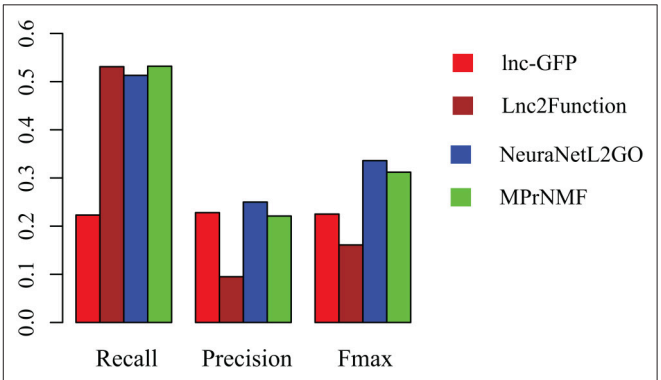

FIGURE 4 | The performance of various algorithms on the prediction of ontology of lncRNAs in terms of Recall, Precision and Fmax.

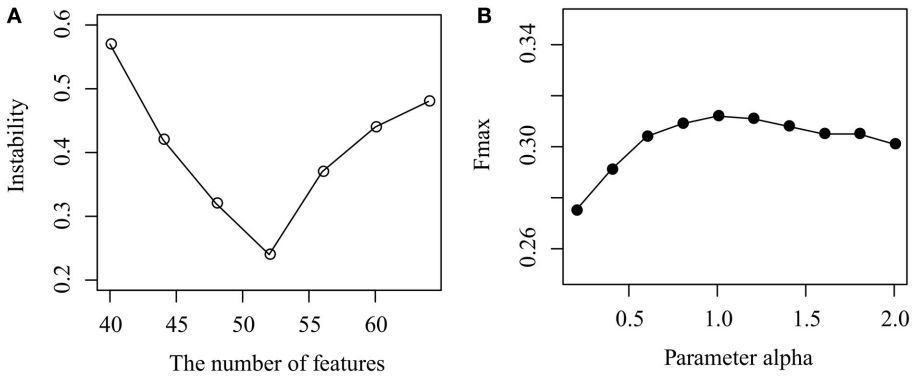

FIGURE 3 | Parameter selection: **(A)** how instability changes as the number of features k increases, and **(B)** how the parameter α effects the performance of the proposed algorithm.

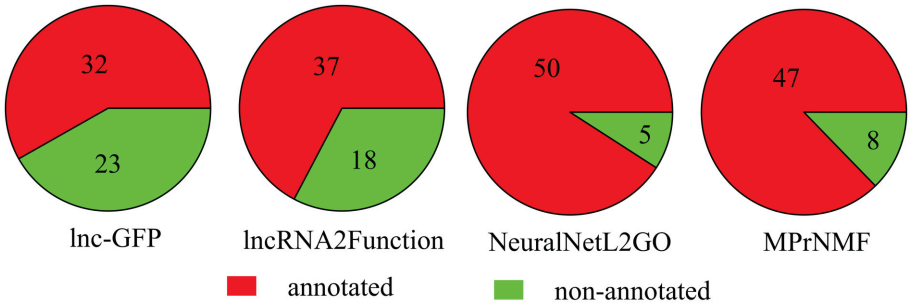

FIGURE 5 | The number of lncRNAs that are correctly annotated by various algorithms.

The accuracy of various algorithms is shown in **Figure 4**, where recall, precision and Fmax are adopted for measuring the performance. These result demonstrate that: (i) MPrNMF achieves the best performance on the recall; (ii) MPrNMF outperforms the lnc-GFP and Lnc2Function; (iii) MPrNMF is inferior to the NeuraNetL2GO. There two possible reasons why the proposed method is superior to lnc-GFP and Lnc2Function. First of all, MPrNMF integrates multiple heterogeneous genomic data via the matrix factorization, which is more accurate to characterize lncRNA-ontology associations. Second, the multiple heterogeneous genomic data are regularized separately, rather than fusing them via a linear function. However, the proposed algorithm is inferior to NeuraNetL2GO. In detail, the Fmax for MPrNMF is 0.309, while that of NeuraNetL2GO is 0.336. There also two possible reasons. First of all, the MPrNMF algorithm is also a network-based method, requiring the networks are connected, which excludes away many lncRNAs or genes for analysis. The second reasons is that MPrNMF does not fully explore the topological information of networks, while the NeuraNetL2GO makes use of graph embedding features from networks.

Furthermore, we also compare these algorithms in terms of the number of lncRNAs that are annotated with a least one biological process GO term. As shown in **Figure 5**, 47 lncRNAs are correctly annotated by the proposed method, which is significantly higher than lnc-GFP and Lnc2Function. Even though it is not as high as that of NeuraNetL2GO, the difference is not significant (p-value = 0.387, Fisher Exact Test).

In MNrNMF, multiple views of lncRNA-gene associations are used. Then, we investigate the performance of each view of the associations. The Fmax of the proposed algorithm based on co-expression lncRNA-gene associations is 0.242, while that based on the disease lncRNA-gene associations is 0.278. These results indicate that the effective integration of heterogeneous genomic data is promising on the prediction of lncRNA-ontology.

4.5. Case Study

In this subsection, we apply MPrNMF to lncRNA instance to show the application of the proposed algorithm. HOTAIRM1 is an intergenic lncRNA between HOXA1 and HOXA2. Evidence shows that HOTAIRM1 is a critical regulator for the expression level of HOXA1 and HOXA4 (Zhang et al., 2009, 2014), which is involved in cell growth in leukemia cells. We apply the MPrNMF algorithm to predict the functions of HOTAIRM1, and it discovers 5 ontology functions: biological regulation, cellular process and signal transduction. These functions have been validated by the previous studies, indicating that the proposed method is applicable to predict the ontological functions of lncRNAs.

5. CONCLUSION

More and more lncRNAs have been identified in the past few years. However, the functions of vast majority of lncRNAs are poorly characterized. In this study, we propose a novel algorithm to predict the functions of lncRNAs via integrating multiple types of genomic data. The results demonstrate that the proposed algorithm is superior to the network-analysis based methods. However, the proposed method has some limitations. First, only the expression and disease data are used to construct the lncRNA-gene associations, which cannot fully characterize the relations. However to construct more reliable lncRNA-gene associations is promising in predicting the functions of lncRNAs. Second, the proposed method cannot fully make use the topological information in the multiple networks, such as graph embedding features. In the further studying, we will investigate how to solve these two issues.

AUTHOR CONTRIBUTIONS

JZ and XM designed the method and JZ coded the algorithm. JZ and XM wrote the paper.

FUNDING

This work was supported by the NSFC (Grant No. 61772394), Scientific Research Foundation for the Returned Overseas Chinese Scholars of Shaanxi Province (Grant No. 2018003) and Fundamental Research Funding of Central Universities (Grant No. Z109021508, JB180304).

ACKNOWLEDGMENTS

The authors appreciate the reviewers for their suggestions.

REFERENCES

Chu, C., Quinn, J., and Chang, H. (2012). Chromatin isolation by rna purification (chirp). *J. Visual. Exper.* 61:3912. doi: 10.3791/3912

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). Noncodev5: a comprehensive annotation database for long non-coding RNAs. *Nuclear Acids Res.* 46, D308–D314. doi: 10.1093/nar/gkx1107

Ferrè, F., Colantoni, A., and Helmer-Citterich, M. (2016). Revealing protein-lncrna interaction. *Brief. Bioinformat.* 17, 106–116. doi: 10.1093/bib/bbv031

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., et al (2013). Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res.* 41:e35. doi: 10.1093/nar/gks967

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., and Feldser D., (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature07672

Hon, C. C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5 ends. *Nature* 543, 199–204. doi: 10.1038/nature21374

Iyer, M., Niknafs, Y., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genet.* 47, 199–208. doi: 10.1038/ng.3192

Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., et al. (2015). Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on RNA-seq data. *BMC Genomics* 16:S2. doi: 10.1186/1471-2164-16-S3-S2

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Res.* 14, 1085–1094. doi: 10.1101/gr.1910904

Li, J., Zhang, M., An, G., Ma, Q. (2016). Lncrna tug1 acts as a tumor suppressor in human glioma by promoting cell apoptosis. *Exper. Biol. Med.* 241, 644–649. doi: 10.1177/1535370215622708

Li, Y., and Patra, J. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogenous network. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btq108

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., et al. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.* 39, 3864–3878. doi: 10.1093/nar/gkq1348

Liu, J., Li, Y., Tong, J., Gao, J., Guo, Q., Zhang, L., et al. (2018). Long non-coding RNA-dependent mechanism to regulate heme biosynthesis and erythrocyte development. *Nature Commun.* 9:4386. doi: 10.1038/s41467-018-06883-x

Ma, X., and Dong, D. (2017). Evolutionary nonnegative matrix factorization algorithms for community detection in dynamic networks. *IEEE Trans. Knowl. Data Eng.* 29, 1045–1058. doi: 10.1109/TKDE.2017.2657752

Ma, X., Dong, D., and Wang, Q. (2019). Community detection in multi-layer networks using joint nonnegative matrix factorization. *IEEE Trans. Knowl. Data Eng.* 31, 273–286. doi: 10.1109/TKDE.2018.2832205

Ma, X., Yu, L., Wang, P., and Yang, X. (2017). Discovering dna methylation patterns for long non-coding RNAs associated with cancer subtypes. *Comput. Biol. Chem.* 69, 164–170. doi: 10.1016/j.compbiolchem.2017.03.014

Mercer, T. R., Dinger, M. E., and Mattick., J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* 10, 155–159. doi: 10.1038/nrg2521

Mercer, T. R., and Mattick, J. S. (2013). Structure and function of long noncoding rnas in epigenetic regulation. *Nature* 20, 300–307. doi: 10.1038/nsmb.2480

Morris, K. V., and Mattick., J. S. (2014). The rise of regulatory RNA. *Nature Rev. Genet.* 15, 423–437. doi: 10.1038/nrg3722

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., et al. (2014). The evolution of lncrna repertoires and expression patterns in tetrapods. *Nature* 505, 635–640. doi: 10.1038/nature12943

Nowak, E., Miller, J. T., Bona, M. K., Studnicka, J., Szczepanowski, R. H., Jurkowski, J., et al. (2014). Ty3 reverse transcriptase complexed with an RNA-DNA hybrid shows structural and functional asymmetry. *Nature Struc. Mol. Biol.* 21, 389–396. doi: 10.1038/nsmb.2785

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., et al. (2018). Coxpresdb in 2015: coexpression database for animal species by dna-microarray and rnaseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.* 43, D82–D86. doi: 10.1093/nar/gku1163

Petergrosso, R., Park, S., Hwang, T., and Kuang, R. (2017). Transfer learning across ontologies for phenome-genome association prediction. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btw649

Turner, M., Galloway, A., and Vigorito., E. (2014). Noncoding RNA and its associated proteins as regulatory elements of the immune system. *Nature Immunol.* 15, 484–491. doi: 10.1038/ni.2887

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* 6:e1000641. doi: 10.1371/journal.pcbi.1000641

Wang, Y., Hu, S. B., Wang, M. R., Yao, R. W., Wu, D., Yang, L., et al. (2018). Genome-wide screening of neat1 regulators reveals cross-regulation between paraspeckles and mitochondria. *Nature Cell Biol.* 20, 1145–1158. doi: 10.1038/s41556-018-0204-2

Wu, S., Joseph, A., Hammonds, A., Celniker, S. E., Yu, B., and Frise, E., (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4290–4295. doi: 10.1073/pnas.1521171113

Zhang, E., and Ma, X. (2018). Regularized multi-view subspace clustering for common modules across cancer stages. *Molecules* 23:1016. doi: 10.3390/molecules23051016

Zhang, J., Zhang, Z., Wang, Z., Liu, Y., and Deng, L. (2018). Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Boinformatics* 34, 1750–1757. doi: 10.1093/bioinformatics/btx833

Zhang, X., Lian, Z., Padden, C., Gerstein, M. B., Rozowsky, J., Snyder, M., et al. (2009). A myelopoiesis-associated regulatory intergenic noncoding rna transcript within the human hoxa cluster. *Blood* 113, 2526–2534. doi: 10.1182/blood-2008-06-162164

Zhang, X., Weissman, S., and Newburger, P. (2014). Long intergenic non-coding rna hotairm1 regulates cell cycle progression during myeloid maturation in nb4 human promyelocytic leukemia cells. *RNA Biol.* 11, 777–787. doi: 10.4161/rna.28828

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2017). Katzlgo: large-scale prediction of LncRNA functions by using the katz measure based on multiple networks. *IEEE/ACM Trans. Computat. Biol. Bioinformat.* doi: 10.1109/TCBB.2017.2704587. [Epub ahead of print].

Zhu, P., Wu, J., Wang, Y., Zhu, X., Lu, T., Liu, B., et al. (2018). Lncgata6 maintains stemness of intestinal stem cells and promotes intestinal tumorigenesis. *Nature Cell Biol.* 20, 1134–1144. doi: 10.1038/s41556-018-0194-0

Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi H., et al. (2015). Prediction of microrna-disease associations based on social network analysis methods. *Biomed. Res. Int.* 10:810514. doi: 10.1155/2015/810514

Zou, Q., Li, J., Song, L., Zeng, X., Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. *Brief. Funct. Genomics* 15, 55–61. doi: 10.1093/bfgp/elv024

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.