Bioinformatics, 34(9), 2018, 1529–1537
doi: 10.1093/bioinformatics/btx794
Advance Access Publication Date: 7 December 2017
Original Paper

Data and text mining

Matrix factorization-based data fusion for the prediction of lncRNA–disease associations

Guangyuan Fu[1], Jun Wang[1], Carlotta Domeniconi[2] and Guoxian Yu[1,*]

[1]College of Computer and Information Science, Southwest University, Chongqing 400715, China and [2]Department of Computer Science, George Mason University, Farifax, VA 22030, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on September 2, 2017; revised on December 1, 2017; editorial decision on December 4, 2017; accepted on December 5, 2017

Abstract

Motivation: Long non-coding RNAs (lncRNAs) play crucial roles in complex disease diagnosis, prognosis, prevention and treatment, but only a small portion of lncRNA–disease associations have been experimentally verified. Various computational models have been proposed to identify lncRNA–disease associations by integrating heterogeneous data sources. However, existing models generally ignore the intrinsic structure of data sources or treat them as equally relevant, while they may not be.

Results: To accurately identify lncRNA–disease associations, we propose a *M*atrix *F*actorization based *L*ncRNA–*D*isease *A*ssociation prediction model (MFLDA in short). MFLDA decomposes data matrices of heterogeneous data sources into low-rank matrices via matrix tri-factorization to explore and exploit their intrinsic and shared structure. MFLDA can select and integrate the data sources by assigning different weights to them. An iterative solution is further introduced to simultaneously optimize the weights and low-rank matrices. Next, MFLDA uses the optimized low-rank matrices to reconstruct the lncRNA–disease association matrix and thus to identify potential associations. In 5-fold cross validation experiments to identify verified lncRNA–disease associations, MFLDA achieves an area under the receiver operating characteristic curve (AUC) of 0.7408, at least 3% higher than those given by state-of-the-art data fusion based computational models. An empirical study on identifying masked lncRNA–disease associations again shows that MFLDA can identify potential associations more accurately than competing models. A case study on identifying lncRNAs associated with breast, lung and stomach cancers show that 38 out of 45 (84%) associations predicted by MFLDA are supported by recent biomedical literature and further proves the capability of MFLDA in identifying novel lncRNA–disease associations. MFLDA is a general data fusion framework, and as such it can be adopted to predict associations between other biological entities.

Availability and implementation: The source code for MFLDA is available at: http://mlda.swu.edu.cn/codes.php? name = MFLDA.

Contact: gxyu@swu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Long non-coding RNAs (lncRNAs) are drawing increasing attention in a variety of biological and medical research. This is because of their critical roles in biological, developmental and pathological processes (Core *et al.*, 2008; Esteller, 2011; Mercer *et al.*, 2009; Tsai *et al.*, 2010). LncRNAs have low expressions and modest sequence conservation with highly specific tissues, they participate in various disease processes and emerge as a class of important regulators

(Ponting *et al.*, 2009; Wang and Chang, 2011). For example, lncRNA 'HOTAIR' has 100 to ~2000 times expression levels in breast cancer metastases, and its expression level is correlated with metastasis and progression of other various cancers (i.e. colorectal cancer, gastric cancer, liver cancer and lung cancer) (Tsai *et al.*, 2010; Gupta *et al.*, 2010). 'HOTAIR' is considered as a potential biomarker in various types of cancers. Experimental studies also demonstrate that lncRNA 'BCAR4' is associated with breast cancer; it expresses in 27% of primary breast tumors. Specifically, in human ZR-75-1 and MCF7 breast cancer cells, the force expression of 'BCAR4' causes cell proliferation in the absence of estrogen and in the presence of various antiestrogens. This is an indication that 'BCAR4' can be considered as a proper target for the treatment of antiestrogen-resistant breast cancer (Godinho *et al.*, 2012). Given the fundamental roles of lncRNAs, identifying potential lncRNA–disease associations can not only help uncovering the disease mechanism at the lncRNA level, but also facilitate disease biomarker detection, analysis, and prevention.

Various efforts have been developed to automatically predict lncRNA–disease associations based on accumulated biological data, and to save the huge cost of wet-lab experiments (Chen *et al.*, 2017). Current efforts can be divided into three categories. The first category predicts potential associations based on known disease-related lncRNAs (Chen and Yan, 2013; Sun *et al.*, 2014). The related methods generally assume that similar diseases are often associated with functionally similar lncRNAs. Chen and Yan (2013) introduced a semi-supervised learning based model to effectively identify novel lncRNA–disease associations by using known associations and lncRNA expression profiles. Sun *et al.* (2014) developed a computational framework to detect potential lncRNA–disease associations by implementing a random walk with restart on a lncRNA–lncRNA functional similarity network. Most methods of this category cannot be applied to new diseases (lncRNAs) whose related lncRNAs (diseases) are completely unknown. The second category predicts novel lncRNA–disease associations based on known disease related genes or miRNAs (Chen, 2015a; Zhou *et al.*, 2015). Zhou *et al.* (2015) integrated three networks (miRNA–associated lncRNA–lncRNA crosstalk network, disease–disease similarity network and known lncRNA–disease association network) into a heterogeneous network and performed random walks on the heterogeneous network to predict lncRNA–disease associations. Chen (2015a) assumed lncRNA and disease significantly sharing common interacting miRNAs can be considered with a candidate association, and applied hypergeometric distribution test for each lncRNA–disease pair to identify lncRNA–disease associations.

Each data source provides a unique and incomplete view of the complex mechanism between biological molecules (i.e. genes, miRNAs and lncRNAs) and diseases, or disease related chemicals (i.e. drugs) (Barabasi and Oltvai, 2004; Barabasi *et al.*, 2011). Integrating multiple related data sources helps to reach a more comprehensive view of diseases and lncRNAs, and to neutralize (reduce) the impact of false positives of individual data sources (Gligorijevic and Przulj, 2015). The third category identifies disease-related lncRNAs by fusing multiple data sources (Chen *et al.*, 2016; Lan *et al.*, 2016; Zhang *et al.*, 2017). With the influx of heterogenous biological data, computational models in this category have become one of the most promising topics in lncRNA–disease association prediction. Chen *et al.* (2016) proposed a random walk with restart based model (IRWRLDA) to integrate multiple data sources for lncRNA–disease association prediction. Lan *et al.* (2016) used Karcher mean (Jeuris *et al.*, 2012) to integrate multiple data sources and to construct a composite disease similarity network and a composite lncRNA similarity network, respectively, and then applied bagged support vector machines to predict lncRNA–disease associations. Chen (2015b) introduced a KATZ measure based data fusion model to uncover potential lncRNA–disease associations. Moreover, some other models (i.e. ILNCSIM; Huang and Chen, 2015) focus on fusing lncRNA–lncRNA similarities derived from multiple data sources to obtain a composite lncRNA functional network, and then predict lncRNA–disease associations on the composite network. These data fusion-based methods achieve significantly improved results compared with those using individual data sources. However, they generally have to transform heterogeneous data sources into homogenous networks of lncRNAs and of diseases. This mandatory transformation may neglect the intrinsic structure and relevance of individual data sources towards association prediction.

Matrix factorization techniques recently have been extensively explored to fuse multiple heterogeneous data sources (Gligorijevic and Przulj, 2015; Zhang and Zhou, 2014). This type of techniques neither has to transform or map heterogeneous data sources onto homologous networks of lncRNAs or of diseases, nor has to develop separate models for individual data sources. They can explore and employ the intrinsic structure of different data sources. Wang *et al.* (2015) introduced a joint non-negative matrix factorization (NMF) model to simultaneously decompose multiple transcriptomics data matrices into one common sub-matrix along with multiple individual sub-matrices to identify differentially expressed genes. Wang *et al.* (2013) proposed a non-NMF (Lee and Seung, 2001) based matrix completion approach to predict new protein–protein interactions. Zitnik and Zupan (2015) developed a penalized matrix tri-factorization model to simultaneously factorize multiple data matrices for predicting gene functions and pharmacologic actions. These matrix factorization-based models show great potential in recovering latent associations between various biological molecules (or entities). However, similar to the aforementioned data fusion based lncRNA–disease association prediction methods, they implicitly assume each data source has *equal* relevance towards the target prediction task, and do *not* differentiate among the quality of different data sources. Therefore, their performance might be seriously compromised by noisy (irrelevant or low quality) data sources. How to selectively integrate multiple data matrices using a matrix factorization model is *rarely* studied and remains an open challenge (Gligorijevic and Przulj, 2015).

To account for the quality and relevance of different heterogeneous data sources, we introduce a Matrix Factorization based LncRNA–Disease Association prediction model (called MFLDA). MFLDA first encodes directly (or indirectly) relevant data sources related to lncRNAs or diseases in individual relational data matrices and presets weights for these matrices. Next, it simultaneously optimizes the weights and low-rank matrix tri-factorization of each relational data matrix. To obtain coherent and interpretable low-rank matrices, whose product pursues to approximate the respective data matrix, MFLDA regulates the low-rank matrix approximation process using constraints on the respective data sources and shared low-rank matrices across inter-related data sources. In a 5-fold cross validation experiment on verified lncRNA–disease associations, MFLDA achieves an AUC of 0.7408, and significantly outperforms related methods, including IRWRLDA (Chen *et al.*, 2016), KATZLDA (Chen, 2015b), ILNCSIM (Huang and Chen, 2015), LDAP (Lan *et al.*, 2016), S-NMTF (Wang *et al.*, 2013) and a variant of MFLDA, which equally treats all the data matrices. In the simulated experiments, designed by masking associations between lncRNAs and the most specific diseases, MFLDA also obtains higher AUC values than the competing methods. Case studies of stomach,

lung, and breast cancer further show that 38 out of 45 (84.44%) lncRNA–disease associations predicted by MFLDA are supported by the biomedical literature. In addition, our empirical study confirms that MFLDA can *selectively and differentially* fuse heterogenous data sources by assigning large weights to relevant data sources and small (or zero) weights to less relevant (or noisy) data sources. All these experimental studies corroborate the effectiveness and potential value of MFLDA in identifying novel lncRNA–disease associations.

2 Materials and methods

2.1 Problem formulation

Suppose there are m types of molecules directly or indirectly related to lncRNAs or diseases, and a collection of relational data sources $\mathcal{R}$, each of which relates a pair of object types. $\mathbf{R}_{ij} \in \mathcal{R}$ ($\mathbf{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$, $i, j \in \{1, 2, \ldots, m\}$) store the inter-relation between n_i objects of the ith type and n_j objects of the jth type. Note, $\mathbf{R}_{ij}$ can be asymmetric. A data source that describes the intra-relation between objects of the ith type is represented by a constraint matrix $\Theta_i \in \mathbb{R}^{n_i \times n_i}$. Matrix factorization based data fusion simultaneously decomposes $\mathbf{R} \in \mathbb{R}(\sum_i^m n_i) \times (\sum_i^m n_i)$ or its sub-matrices, constrained by sub-matrices of $\Theta \in \mathbb{R}(\sum_i^m n_i) \times (\sum_i^m n_i)$, into low-rank matrices to explore the latent relationship between objects of the same type and of different types. It then uses the low-rank matrices to reconstruct the target association matrix (i.e. $\mathbf{R}_{ij}$) to predict new associations between objects of the ith type (lncRNAs) and objects of the jth type (diseases).

2.2 Matrix factorization model for multi-relational and multi-type data

Matrix factorization-based data fusion has various variants (Wang *et al.*, 2013, 2015; Zitnik and Zupan, 2015). For a better problem analysis and presentation, we start with a recently proposed and representative framework suggested by Zitnik and Zupan (2015). The objective function of this framework is:

$$\min_{\mathbf{G}\geq 0} \mathcal{Z}(\mathbf{G}, \mathbf{S}) = \sum_{\mathbf{R}_{ij}\in\mathcal{R}} ||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||_F^2 + \sum_{t=1}^{\max_i t_i} tr(\mathbf{G}^T\Theta^{(t)}\mathbf{G})) \quad (1)$$

where $\mathbf{G}_i \in \mathbb{R}^{n_i \times k_i}$, $\mathbf{G}_j \in \mathbb{R}^{n_j \times k_j}$, $\mathbf{S}_{ij} \in \mathbb{R}^{k_i \times k_j}$ $(k_i \ll n_i, k_j \ll n_j)$, $\mathbf{G} = diag(\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_m)$, $tr(\cdot)$ and $||\cdot||_F^2$ are the matrix trace operator and Frobenius norm, respectively. Suppose the ith type of objects has t_i data sources, represented by t_i constraint matrices $\{\mathbf{\Theta}_i^t\}_{t=1}^{t_i}$. $\mathbf{\Theta}^{(t)}$ collectively stores all these block diagonal matrices: $\mathbf{\Theta}^{(t)} = diag(\mathbf{\Theta}_1^{(t)}, \mathbf{\Theta}_2^{(t)}, \ldots, \mathbf{\Theta}_m^{(t)})$, $\mathbf{\Theta}_i^{(t)}$ $(t \in \{1, 2, \ldots, \max_i t_i\}$, and the ith block matrix along the main diagonal of $\mathbf{\Theta}^{(t)}$ is zero if $t > t_i$. $\mathbf{S}_{ij}$ has much fewer vectors than $\mathbf{R}_{ij}$; it can be viewed as a compressed matrix which encodes latent inter-relations between objects of the ith type and objects of the jth type. $\mathbf{G}_i$ ($\mathbf{G}_j$) is the low-rank representation of objects of the ith (jth) type. Entries in the constraint matrices are positive for dissimilar objects, and negative for similar ones. The former are known as cannot-link constraints, since they force pairs of dissimilar objects to be far away from each other in the latent component space; and the latter are must-link constraints, as they force pairs of similar objects to be close in the latent component space. These constraints can reduce the value of the cost function during optimization. Suppose $\mathbf{G}_i$ is the low-rank representation of n_i lncRNAs, and $\mathbf{G}_j$ is the low-rank representation of n_j diseases. Then the potential lncRNA–disease associations can be predicted as $\mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T$.

Unlike kernel (network) integration and classifier ensemble based data fusion techniques, Equation (1) does not require to map other data sources $\mathbf{R}_{a,\ b}$ $(a \neq i$ or $b \neq j)$ onto $\mathbf{R}_{ij}$ or $\mathbf{\Theta}_i$, it directly works on multi-type objects with multi-relations, and thus it has the potential of exploring and employing the inter-structure between objects of different types and intra-structure of objects of the same type (Gligorijevic and Przulj, 2015; Zitnik and Zupan, 2015). In addition, it can take into account directly connected data sources ($\mathbf{R}_{ib}$, $b \neq j$ or $\mathbf{R}_{aj}$, $a \neq i$), and indirectly connected data sources ($\mathbf{R}_{a,b}$, $a \neq i$ and $b \neq j$) of the target association matrix, since $\mathbf{G}_i$ is not only optimized with respect to $\mathbf{R}_{ij}$, but also with respect to $\mathbf{\Theta}_i^{(t)}$ and other data sources that have inter-relations with objects of the ith type. Since low-rank matrix factorization can reduce the impact of noises (Lee and Seung, 2001; Meng and De La Torre, 2013), Equation (1) can reduce the impact of false positives of fusing multiple data sources to some extent.

However, we can clearly see that Equation (1) equally treats all the related matrices $\{\mathbf{R}_{ij}\}_{i,j=1}^m$. As such, its performance may be compromised by noisy (or irrelevant) data sources. Other matrix factorization based data fusion solutions suffer from the same issue (Gligorijevic and Przulj, 2015; Wang *et al.*, 2013, 2015).

2.3 Objective function of MFLDA

To selectively integrate heterogeneous relational data sources, we assign different weights to the data matrices and define the following objective function:

$$\min_{\mathbf{G}\geq 0} \mathcal{O}'(\mathbf{G}, \mathbf{S}, \mathbf{W}) = \sum_{\mathbf{R}_{ij}\in\mathcal{R}} \mathbf{W}_{ij}||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||_F^2 + \sum_{t=1}^{\max_i t_i} tr(\mathbf{G}^T\Theta^{(t)}\mathbf{G}) \quad (2)$$
$$s.t. \quad \mathbf{W} \geq 0,\ \textstyle\sum_{i,j=1}^m \mathbf{W}_{ij} = 1$$

where $\mathbf{W} \in \mathbb{R}^{m \times m}$ contains the weights assigned to $|\mathcal{R}|$ relational data sources. For $\mathbf{R}_{ab} \notin \mathcal{R}$, $\mathbf{W}_{ab} = 0$. Equation (2) aims at specifying different weights to different relational sources, and collaboratively factorizing $|\mathcal{R}|$ relational matrices into low-rank matrices. In this way, it can exploit the complimentary information spread across different types of objects and reduce (or remove) the misleading effect of noisy (or irrelevant) data sources.

However, Equation (2) may only assign $\mathbf{W}_{ij} = 1$ to $\mathbf{R}_{ij}$ if $\mathbf{R}_{ij}$ have the smallest *reconstruction loss* ($||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||_F^2$) among all the relational matrices. As a result, all the other relational data sources will be disregarded. In practice, complementary information is often embedded in more than one biological relational data source, and this trivial weight assignment should be avoided. To this end, we add an l_2-norm based regularization penalty on $\mathbf{W}$ and update the objective function as follows:

$$\min_{\mathbf{G}\geq 0} \mathcal{O}(\mathbf{G}, \mathbf{S}, \mathbf{W}) = \sum_{\mathbf{R}_{ij}\in\mathcal{R}} \mathbf{W}_{ij}||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||_F^2 + \sum_{t=1}^{\max_i t_i} tr(\mathbf{G}^T\Theta^{(t)}\mathbf{G}) + \alpha||vec(\mathbf{W})||_F^2 \quad (3)$$
$$s.t. \quad \mathbf{W} \geq 0, \sum \mathrm{vec}(\mathbf{W}) = 1$$

where $vec(\mathbf{W})$ is an operator that stacks the rows of $\mathbf{W}$, and $\alpha > 0$ is used to control the complexity of $vec(\mathbf{W})$. By adding the l_2-norm and minimizing Equation (3), a larger weight is assigned to the relational data matrix $\mathbf{R}_{ij}$ whose low-rank approximation is more accurate. In other words, the relational data source with a smaller $||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||_F^2$ will be assigned a larger weight. On the other

hand, the data source with a larger $||\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T||_F^2$ will be assigned a smaller (or zero) weight. We want to remark that MFLDA can automatically remove noisy relational data sources by assigning zero weights to them, and reduce the impact of relational data sources which are partially noisy by assigning smaller weights to them. In this way, MFLDA can further reduce the impact of false positives of individual data sources by selectively fusing them. These advantages of MFLDA are formally shown from the explicit solution of $\mathbf{W}$ in the Supplementary Material.

The objective function of MFLDA is non-convex in $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{W}$ altogether. Following the idea of auxiliary functions frequently used in the convergence proof of approximate matrix factorization algorithms (Lee and Seung, 2001; Zitnik and Zupan, 2015), we can show that $\mathcal{O}(\mathbf{G}, \mathbf{S}, \mathbf{W})$ is nonincreasing under the iterative updating rule for $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{W}$ that alternatively keeps fixed two of them and optimizes the third one. Due to page limitation, the details of the iterative optimization of $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{W}$ is provided in the Supplementary Material, along with the convergence proof. The operating principle of MFLDA is also described in Figure 1.

3 Results and discussion

3.1 Experimental setup

To investigate the performance of MFLDA, we consider six object types depicted in the left part of Figure 1: lncRNAs (Type 1), miRNAs (Type 2), genes (Type 3), Gene Ontology (Type 4), Disease Ontology (Type 5) and drugs (Type 6). We collected eleven (nine inter and two intra) relational data sources between these object types from public databases. Specifically, lncRNA–miRNA associations ($\mathbf{R}_{12}$) are collected from starBase v2.0 (Li *et al.*, 2013a); lncRNA–gene interactions ($\mathbf{R}_{13}$) from lncRNA2target (Jiang *et al.*, 2014); lncRNA–gene function associations ($\mathbf{R}_{14}$) from GeneRIF (Lu *et al.*, 2007) and pre-processed using Open Biomedical Annotator (Jonquet *et al.*, 2009); lncRNA–disease associations ($\mathbf{R}_{15}$) from LncRNADisease (Chen *et al.*, 2013), Lnc2Cancer (Ning *et al.*, 2016) and GeneRIF (Lu *et al.*, 2007); miRNA–gene interactions ($\mathbf{R}_{23}$) from miRTarBase (Hsu *et al.*, 2014); miRNA–disease associations ($\mathbf{R}_{25}$) from HMDD (Li *et al.*, 2013b); Gene Ontology annotations ($\mathbf{R}_{34}$) from Gene Ontology (Ashburner *et al.*, 2000); gene–disease associations ($\mathbf{R}_{35}$) from DisGeNet (Pinero *et al.*, 2015); and gene–drug associations ($\mathbf{R}_{36}$) from DrugBank (Law *et al.*, 2014). In addition, we collected gene–gene interactions ($\mathbf{\Theta}_3$) from BioGrid (Stark *et al.*, 2006), and drug–drug interactions ($\mathbf{\Theta}_6$) from DrugBank (Law *et al.*, 2014). We collected the latest version of these databases (access date: 31 May 2017) for the experiments. The details and statistics of these data sources are provided in Supplementary Tables S1 and S2.

We conduct 5-fold cross validation to measure and compare the performance of MFLDA against related methods, including KATZLDA (Chen, 2015b), ILNCSIM (Huang and Chen, 2015), IRWRLDA (Chen *et al.*, 2016), LDAP (Lan *et al.*, 2016), S-NMTF (Wang *et al.*, 2013) and a variant of MFLDA called MFLDA-nW, which assigns equal weights to all the relational data sources. For cross validation, we randomly partition known lncRNA–disease associations ($\mathbf{R}_{15}$) into five row-wise folds; the associations of 4-folds are used as training samples and the remaining associations of the fifth fold are used as testing samples for evaluation. We observe that the interactions between lncRNAs and other molecules (or entities) are kept fixed during cross validation. After each association has been tested in a single round of cross validation, we plot the receive operating characteristic (ROC) curve by varying the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1-specificity) at different rank cutoffs. Sensitivity measures the percentage of test associations that rank above a given cutoff, whereas specificity measures the percentage of candidate associations that rank below the given cutoff. The value of area under the ROC Curve (AUC) can be computed to quantify the overall performance. The larger the AUC value, the better the performance is, and a random guess corresponds to an AUC value of 0.5. In addition, we also report the Precision-Recall (PR) curve at different rank cutoffs, where precision is the percentage of correct associations among the predicted ones, while recall is the same as sensitivity in a ROC curve. We repeat the random partition and evaluation in 10 independent rounds, and report the average results.

3.2 LncRNA–disease association prediction with cross validation

To comparatively study the performance of MFLDA in predicting lncRNA–disease associations, we implement IRWRLDA (Chen *et al.*, 2016), KATZLDA (Chen, 2015 b), ILNCSIM (Huang and Chen, 2015), LDAP (Lan *et al.*, 2016), S-NMTF (Wang *et al.*, 2013) and a variant of MFLDA called MFLDA-nW, which equally treats

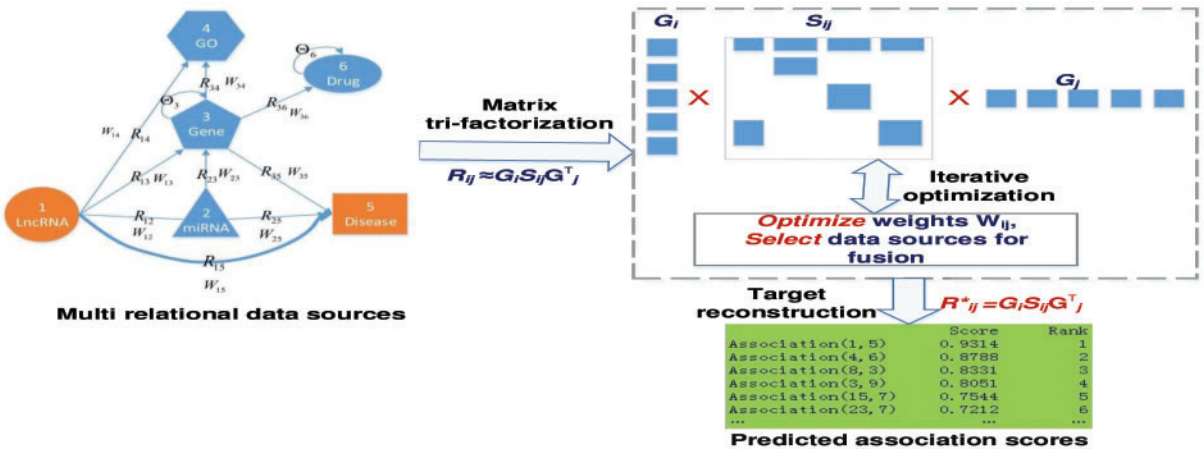


Fig. 1. The operating principle of MFLDA. MFLDA iteratively optimizes the low-rank matrices ($\mathbf{G}_i$) of multiple relational data matrices via matrix tri-factorization, and weights ($\mathbf{W}_{ij}$) assigned to these data matrices to selectively fuse them. It finally reconstructs the target association matrix based on the optimized low-rank matrices and weights.

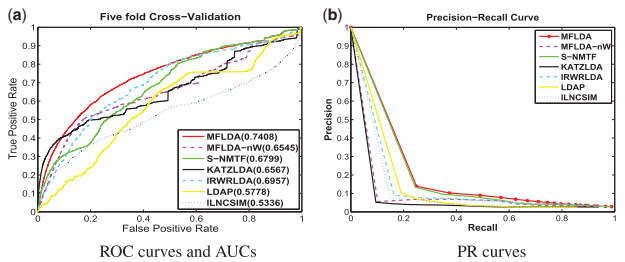
all the relational data matrices during the fusion process (see Equation 1). The input parameters of these methods are set as specified by the authors in their code, or optimized in the suggested ranges, and $\alpha = 10^5$ for MFLDA. The first four methods were introduced in the Section 1, S-NMTF is a nonnegative matrix tri-factorization based framework that can be adopted to fuse multiple relational data sources to predict lncRNA–disease associations. Figure 2 plots the ROC curves and PR curves of the comparing methods and reports their corresponding AUCs of 5-fold cross validation on experimentally verified lncRNA–disease associations. It is evident that MFLDA almost always has the highest TPRs under the same false negative rates, and achieves the highest AUC (0.7408) among the methods, whereas the AUCs of MFLDA-nW, KATZLDA, IRWRLDA, ILNCSIM, SNMTF and LDAP are 0.6545, 0.6567, 0.6957, 0.5336, 0.6799 and 0.5778, respectively. Figure 2b shows that, at first, both MFLDA and S-NMTF have the highest precision in correspondence of the same recall value; then, MFLDA achieves a higher precision with respect to all the other methods (including S-NMTF) for any given recall value. In addition, we also report the AUC values of PR curve and recalls at different top k cutoffs in Table 1. MFLDA again outperforms other comparing methods in terms of AUC and recalls.

MFLDA performs significantly better than MFLDA-nW, although also the latter can explore and exploit the intrinsic and shared structure of heterogenous data sources. The reason is that MFLDA-nW equally integrates all the relational data sources. S-NMTF also uses matrix tri-factorization to integrate heterogenous data sources; it obtains higher AUC than other methods, but still has lower AUC than MFLDA. The reason is that S-NMTF cannot differentiate the various data sources. In fact, it performs matrix factorization on a big $(\sum_{i=1}^{m} n_i) \times (\sum_{i=1}^{m} n_i)$ matrix, which includes all $\mathbf{R}_{ij} \in \mathcal{R}$. Furthermore, it requires the matrix to be symmetric.

Both IRWRLDA and KATZLDA integrate heterogeneous data sources to predict lncRNA–disease associations. Their AUCs are comparable to those of S-NMTF and MFLDA-nW, but lower than that of MFLDA. This is mainly because IRWRLDA and KATZLDA transform heterogeneous data sources onto lncRNA functional similarity and disease similarity networks, and then take advantage of these two networks and known lncRNA–disease associations to infer new lncRNA–disease associations. This mandatory transformation may enshroud the intrinsic structure of the data sources. Furthermore, IRWRLDA and KATZLDA are biased towards diseases (lncRNAs) with more known associated lncRNAs (diseases), and with more miRNAs or gene interaction partners. For the same reasons, LDAP has a low AUC, although it constructs a composite disease network by integrating many disease similarity networks. ILNCSIM cannot accurately estimate the similarity between lncRNAs, or between diseases when the test associations are left out for prediction. Furthermore, it suffers from combining results of two classifiers. As such, it has the lowest AUC among the compared methods.

In Figure 2a, MFLDA has a lower performance than both KATZLDA and S-NMTF for the top portion of the predicted list. This portion of predicted results is important as it identifies the potential lncRNA–disease associations predicted with high confidence. KATZLDA uses a graph-based computational method that transforms the problem of association prediction into the problem of calculating similarities between nodes in a heterogeneous network, which is constructed by integrating the lncRNA similarity network, the disease similarity network, and the lncRNA–disease association network. The adjacency matrix of this heterogeneous network is much smaller than that of MFLDA. KATZLDA uses only this small heterogeneous network; as such, KATZLDA can do better in the top portion of the predicted list. S-NMFT uses a large adjacency matrix as MFLDA does, and it globally factorizes the whole matrix for lncRNA–disease association prediction. In contrast, MFLDA jointly factorizes small relational data matrices. The global factorization can explore the underlying associations between lncRNAs and diseases based on heterogeneous data sources. However, since KATZLDA mandatorily maps heterogeneous data sources onto a homologous network of lncRNAs and diseases, and S-NMTF equally treats all the data sources, they both eventually have a performance lower than MFLDA.

In summary, these experimental results show that MFLDA cannot only explore and leverage the intrinsic and shared structure of heterogeneous data sources, as other matrix factorization based models do, but can also select and weigh the data sources for predicting lncRNA–disease associations.



ROC curves and AUCs PR curves

Fig. 2. Performance comparison between MFLDA, MFLDA-nW, KATZLDA, IRWRLDA, ILNCSIM, S-NMTF and LDAP in terms of ROC curve, AUC and PR curve. (**a**) ROC curves and AUCs. (**b**) PR curves

3.3 Predicting masked lncRNA–disease associations

By checking the timeline of discovered diseases associated with a certain lncRNA, we find that the discovered diseases become

Table 1. AUCs of MFLDA, MFLDA-nW, IRWRLDA, KATZLDA, ILNCLDA and LDAP on respective PR curves, and the recalls at different top k cutoffs

	AUC(PR)	Recall		
		$k = 100$	$k = 500$	$k = 1000$
MFLDA	**0.2045** ± 0.0023	**0.0326** ± 0.0033	**0.0768** ± 0.0052	**0.1049** ± 0.0022
MFLDA-nW	0.1064 ± 0.0056	0.0107 ± 0.0016	0.0237 ± 0.0035	0.0428 ± 0.0035
S-NMTF	0.1956 ± 0.0033	0.0247 ± 0.0036	0.0541 ± 0.0011	0.0845 ± 0.0053
IRWRLDA	0.1433 ± 0.0101	0.0087 ± 0.0013	0.0267 ± 0.0011	0.0567 ± 0.0043
KATZLDA	0.1356 ± 0.0034	0.0051 ± 0.0014	0.0178 ± 0.0004	0.0367 ± 0.0024
ILNCSIM	0.0904 ± 0.0019	0.0062 ± 0.0012	0.0153 ± 0.0053	0.0253 ± 0.0033
LDAP	0.1687 ± 0.0023	0.0103 ± 0.0022	0.0344 ± 0.0048	0.0589 ± 0.0026

Note: Results in boldface are significantly superior to other results in the same setting.

increasingly more specific. As shown in Figure 3, lncRNA 'ZFAS1' (ZNFX1-AS1) was associated with breast cancer (DO: 1612) in 2011, and it was also associated with ancestor diseases of DO: 1612 based on the Disease Ontology hierarchy (Schriml *et al.*, 2012). It was later found being associated with colon cancer (DO: 219) (Li *et al.*, 2015) in 2015, with hepatocellular carcinoma (DO: 684) (Thorenoor *et al.*, 2016) in 2016, and with gastric cancer (DO: 10534) (Nie *et al.*, 2017) in 2017. These newly discovered diseases correspond to subtypes of organ system cancer (DO: 0050686), which was already found being associated with the same lncRNA before 2011. Accurately identifying the association between lncRNAs and subtypes of a complex (or concrete) disease can provide more directional clues for precise treatment and biomarker discovery. Given this, we conduct simulated experiments to investigate the ability of MFLDA in predicting specific lncRNA–disease associations. In the simulated experiments, we assume the diseases associated with a particular lncRNA are complete, and then randomly mask q specific diseases corresponding to leaf nodes (i.e. DO: 1612, DO: 219, DO: 684 and DO: 10534) of the direct acyclic graph formed by diseases associated with the lncRNA. We observe that, if a leaf node is masked and its direct parent currently does not have any descendant associated with the lncRNA, then itself can be masked later. For example, if DO: 219 is masked for 'ZFAS1' in Figure 3, then its direct parent DO: 9256 becomes a leaf node and can be masker later. We want to remark that masking a non-leaf node is meaningless, since the non-leaf node's descendants are associated with that lncRNA and this masked association can be directly inferred from the descendants. We consider the masked lncRNA–disease associations as specific associations to be predicted, and then apply MFLDA and other methods to predict them. Note that, if q is larger than the number of diseases associated with a lncRNA, it means we are not masking all the diseases but keep at least one unmasked. To reduce the random effect, we repeat each comparing method under each $q \in \{1, 3, 5\}$ in 10 independent rounds and report the average results with standard deviations in Table 2. Similarly to the dataset used in 5-fold cross validation, 2697 associations between 240 lncRNAs and 412 diseases are used in the mask experiments. The average numbers of masked distinct diseases and associations under different q are also provided. Results in **boldface** are the best results in the same setting, with significance checked by pairwise t-test at 95% confidence level.

Fig. 3. Diseases associated with lncRNA 'ZAFS1'. Breast cancer (DO: 1612), colon cancer (DO: 219), hepatocellular carcinoma (DO: 684) and gastric cancer (DO: 10534) were recognized as being associated with 'ZAFS1' in 2011, 2015, 2016 and 2017, respectively

From Table 2, we can clearly see that MFLDA outperforms the competing methods across different input values of q. Both IRWRLDA and KATZLDA transform heterogenous data sources into a composite network to describe the similarity between lncRNAs, and into a composite network to encode the similarity between diseases. This transformation may enshroud the intrinsic structure of the data sources. As a consequence, they are outperformed by MFLDA, which directly works on the data matrices. For the same reason, LDAP and ILNCSIM are also outperformed by MFLDA. In fact, LDAP utilizes a bootstrap aggregating strategy to reconcile individual classifiers; it outperforms ILNCSIM and KATZLDA, but still loses to MFLDA when $q = 5$. Similarly to MFLDA, S-NMTF and MFLDA-nW also utilize non-negative tri-matrix factorization to reconstruct the target association matrix and to predict lncRNA–disease associations. Although they all can explore and leverage the intrinsic structure of the respective data sources, S-NMTF and MFLDA-nW are significantly outperformed by MFLDA. This is because they equally treat all heterogeneous data sources. The number of experimentally verified associations affects the prediction performance, since all these comparing methods show decreasing trend as q increase. The prediction performance is expected to be further improved as more associations available. These experimental results not only demonstrate the ability of MFLDA of identifying specific lncRNA–disease associations, but also prove the advantage of MFLDA in selectively integrating multiple heterogeneous data sources (Gligorijevic and Przulj, 2015).

3.4 Case study on breast, lung and stomach cancers

Case studies of three common and representative cancers, namely Breast, Stomach and Lung cancers, are investigated to further evaluate the capability of MFLDA in predicting novel lncRNA–disease associations. For this type of study, we consider all the known associations between the target cancer and lncRNAs, and use all these lncRNAs as test samples. The inter-connections of lncRNAs with other molecules are kept the same. We then select the top 15 plausible associations as the predicted lncRNA–disease associations for each cancer. After that, we check the predicted associations by referring to available

Table 2. AUC of MFLDA, MFLDA-nW, IRWRLDA, KATZLDA, ILNCLDA and LDAP on predicting (q) masked lncRNA–disease associations of each lncRNA

	$q = 1$	$q = 3$	$q = 5$
Masked diseases/associations	7.1/157.1	26.4/468.3	41.1/755.7
MFLDA	**0.9522** ± 0.0083	**0.9368** ± 0.0033	**0.9132** ± 0.0052
MFLDA-nW	0.9174 ± 0.0167	0.8947 ± 0.0060	0.8937 ± 0.0175
S-NMTF	0.9120 ± 0.0157	0.8895 ± 0.0036	0.8860 ± 0.0111
IRWRLDA	0.8973 ± 0.0019	0.8756 ± 0.0030	0.8110 ± 0.0011
KATZLDA	0.8836 ± 0.0026	0.8437 ± 0.0014	0.7751 ± 0.0004
ILNCSIM	0.8835 ± 0.0030	0.8262 ± 0.0012	0.7653 ± 0.0053
LDAP	0.8535 ± 0.0096	0.8344 ± 0.0042	0.7907 ± 0.0088

Note: Results in boldface are significantly superior to other results in the same setting.

associations in LncRNADisease (Chen *et al.*, 2013), Lnc2Cancer (Ning *et al.*, 2016) and GeneRIF (Lu *et al.*, 2007). For the predicted associations that cannot be supported by associations in these databases, we further manually check them on PubMed and list the supportive literature. The currently supported and un-supported associations are listed in Table 3 (for Breat cancer), Supplementary Table S3 (for Lung cancer), and Supplementary Table S4 (for Stomach cancer) of the Supplementary file. We highlight the associations supported by recent literature in PubMed but not included in these databases in **boldface**.

Breast cancer is the most common cancer in women, and it remains the most prevalent cancer in the world. In 2012, a total of 1.38 million new breast cancer cases were diagnosed, representing 23% of all cancers (Ferlay *et al.*, 2015). MFLDA is implemented to predict potential breast cancer-related lncRNAs. As a result, 13 out of 15 predicted lncRNAs are supported. For instance, Chang *et al.* (2011) report that five breast cancer cell lines (MCF-7, T-47 D, YCC-B1, YCC-B3 and YCC-B5) show synergism for the combinations of both SAHA/paclitaxel and SAHA/docetaxel, 'HCG9' (HLA Complex Group 9, non-protein coding) is down-regulated in the group of paclitaxel-resistant and combination synergistic cell lines as compared with paclitaxel-sensitive and combination antagonistic cells. Meanwhile, WT1 (Wilm's tumor suppressor gene1) plays a role in regulating the epithelial-mesenchymal balance of Breast cancer cells and WT1-expressing tumors are mainly associated with a mesenchymal phenotype (Artibani *et al.*, 2017). WT1-AS, as the antisense RNA to WT1, transcript includes both the first exon of the WT1 gene with the promoter positioned in intron 1 of the WT1 gene. Hypermethylation of the intron 1 region is observed in mesothelioma, renal cell carcinoma, and breast cancer (Kaneuchi *et al.*, 2005). The potential association to Breast cancer of 'WT1-AS' deserves more attention.

The case study analysis of Lung and Stomach cancers is included in the Supplementary Material. In summary, 38 (13 for Breast cancer, 13 for Lung cancer and 12 for Stomach cancer) out of the top 45 plausible lncRNA–disease associations are supported by currently available associations in the databases (Chen *et al.*, 2013; Lu *et al.*, 2007; Ning *et al.*, 2016) used for the experiments, and by manual biomedical text mining on PubMed. These case studies again confirm the potential of MFLDA in identifying novel lncRNA–disease associations with confidence. We want to remark that the 7 un-validated associations should not be viewed as incorrect associations. As more experimental evidence becomes available, some of them maybe further supported.

Table 3. MFLDA predicted lncRNAs associated with **Breast** cancer (top 15 in ranking list), and the corresponding evidence

lncRNA	Evidence(PMID)	Rank
HCG9	**20224928**	1
WT1-AS	**26046002**; **28345629**(WT1)	2
LINC00472	26564482	3
HOTAIR	24829860; 22996375; 21903344	4
CTBP1-AS	**26933806**(CTBP1)	5
MEG3	14602737	6
PSORS1C3	without evidence	7
XIST	17545591	8
MALAT1	22492512; 22996375; 24499465	9
UCA1	26439035; 26464647	10
GAS5	18836484; 20673990; 24789445	11
BCYRN1	9422992; 15240511	12
LINCMD1	without evidence	13
CDKN2B-AS1	17440112; 20956613; 20453838	14
SNHG12	**28337281**	15

3.5 Effects of weighting relational data matrices

From the explicit solution for $\mathbf{W}$ (see Supplementary Material), it is clear that once the value of α is specified, the weight assigned to $\mathbf{R}_{ij} \in \mathcal{R}$ can be computed based on the reconstruction loss of that matrix. To search for a feasible value of α, and to study the contribution of weighting the relational data matrices, we conduct 5-fold cross validation experiments to predict lncRNA–disease associations by varying α in $\{10^{-2}, 10^{-1}, \ldots, 10^{7}, 10^{8}\}$, and report the average AUC and standard deviation for each input value of α in Figure 4.

From Figure 4, we can observe that when $\alpha = 10^5$, MFLDA achieves the highest AUC. To further interpret this result, we also report the weights ($\mathbf{W}_{ij}$) assigned to the nine inter-relational data matrices when $\alpha = 10^5$, $\alpha = 10$, $\alpha = 10^3$ and $\alpha = 10^8$ in Figure 5. We can observe two extreme cases: when $\alpha = 10$, only the associations between lncRNAs and miRNAs are selected; when $\alpha = 10^8$, all nine inter-relational matrices are selected and assigned nearly equal weights. This behavior is expected from Equation (3), and the explicit solution for $\mathbf{W}_{ij}$ given in Equation (14) of the Supplementary Material. A (too) small α value does not have a sufficient regularization effect on the weights assigned to individual matrices, and thus only one data matrix is selected. On the other hand, a (too) large α value inflicts a strong regularization effect and forces similar weight assignments to all matrices. MFLDA with $\alpha = 10^8$ performs significantly better than with $\alpha = 10$; this is because complementary information is spread across multiple relational data sources. For the same reason, MFLDA with $\alpha = 10^3$ selects to fuse three data sources, and it also obtains better performance than MFLDA with $\alpha = 10$. Interestingly, there is a discrepancy between the AUC of MFLDA-nW (around 0.65) and that of MFLDA (around 0.72) with $\alpha = 10^8$. The reason is that MFLDA-nW assigns exactly equal weights to different data matrices, and MFLDA with $\alpha = 10^8$ assigns similar (but not equal) weights to different data matrices. MFLDA iteratively optimizes both the weights assigned to data matrices and the matrix factorization of these matrices, whereas MFLDA-nW only optimizes the matrix factorization of these matrices. When $\alpha = 10^5$, some inter-relational matrices [miRNA–gene associations ($\mathbf{R}_{23}$) and gene–GO associations ($\mathbf{R}_{34}$)] are completely removed from the fusion process. In addition, the selected data sources are given different weights. A possible explanation for this result is that the removed matrices contain numerous noisy associations, and the selected relational data matrices have different degrees of relevance towards the target prediction task. Overall, the experiments with different values of α show that MFLDA can selectively integrate different data sources during the fusion process. Based on these results, we set $\alpha = 10^5$ for the experiments.

The low-rank size ($k_i, i \in \{1, 2, \ldots, 6\}$) can also affect the performance of MFLDA and other low-rank matrix factorization based data fusion solutions (Wang *et al.*, 2013, 2015; Zitnik and Zupan, 2015). By referring to the size of the relational data matrices, and

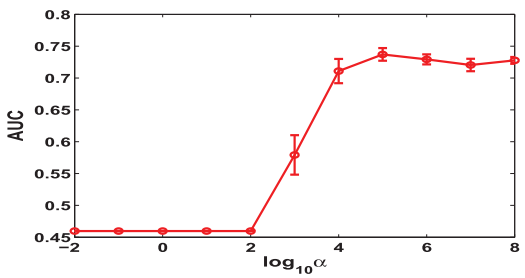


Fig. 4. AUC of MFLDA under different input values of α

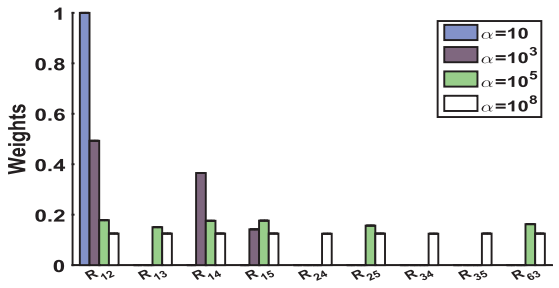


Fig. 5. Weights assigned to nine data sources with $\alpha = 10$, $\alpha = 10^3$, $\alpha = 10^5$ and $\alpha = 10^8$, respectively

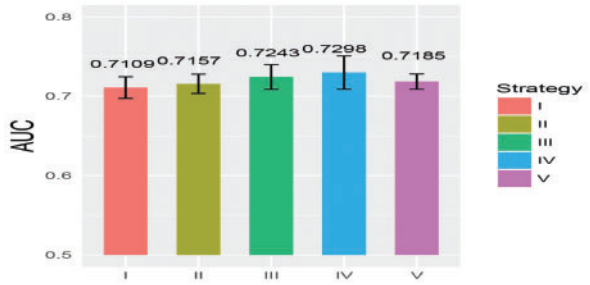


Fig. 6. Contribution of adding new data sources: (I) $\mathbf{R}_{15}$; (II) $\mathbf{R}_{15} + \mathbf{R}_{13}$; (III) $\mathbf{R}_{15} + \mathbf{R}_{13} + \mathbf{R}_{12}$; (IV) $\mathbf{R}_{15} + \mathbf{R}_{13} + \mathbf{R}_{12} + \mathbf{\Theta}_3$; (V) $\mathbf{R}_{15} + \mathbf{R}_{13} + \mathbf{R}_{12} + \mathbf{R}_{36} + \mathbf{\Theta}_3$

following the suggestion in (Zitnik and Zupan, 2015), we vary k_i from 10 to 210 to seek the optimal low-rank size using 5-fold cross validation, and show the AUC under each input value of k_i in Supplementary Figure S1. Since simultaneously optimizing all k_i is time-consuming and impractical, we independently seek the optimal k_i while fixing the other $k_j (j \neq i)$ to 50. Based on the results in the Figure, we set $k_i (i \in \{1, 2, \ldots, 6\})$ to 50, 110, 50, 70, 170 and 50, respectively. More discussion on the input values of k_i is provided in the Supplementary Material.

3.6 Contribution of multi-type data sources

To investigate the contribution of utilizing multi-type data sources, we conduct additional experiments by starting with only the target lncRNA–disease associations ($\mathbf{R}_{15}$), and then appending lncRNA–gene associations ($\mathbf{R}_{13}$), lncRNA–miRNA associations ($\mathbf{R}_{12}$), gene–gene interactions ($\mathbf{\Theta}_3$) and gene–drug associations ($\mathbf{R}_{36}$) one step at the time. Figure 6 reveals the AUC change as more data sources are added. We can observe that by adding data sources incrementally, AUC goes from 0.7112 to 0.7273. We used pairwise *t*-test to check the statistical difference between I and II, II and III and III and IV in Figure 6, and the *P*-values are 0.33, 0.02 and 0.09, respectively. Although the improvement is not significant in some cases, the trend still supports the contribution of multi-type data sources, and both the inter-associations between objects of different types and constraints between objects of the same type contribute to lncRNA–disease association prediction. However, AUC slightly downgrades to 0.7226 after adding gene–drug associations ($\mathbf{R}_{36}$). This behavior can be attributed to the sparsity of $\mathbf{R}_{36}$. In fact, the number of gene–drug associations in $\mathbf{R}_{36}$ is 3, 760, whereas there are 15, 527 genes and 8, 283 drugs. For this reason, $\mathbf{R}_{36}$ has a small reconstruction loss and is being assigned a large weight $\mathbf{W}_{ij}$. This observation suggests that the reconstruction error is not always the optimal criterion to evaluate the quality of a data source, and we will investigate alternative ones. Nevertheless, as shown in Figure 2, although MFLDA makes use of $\mathbf{R}_{36}$ (an indirect data source of lncRNAs and diseases), it still outperforms the competing methods. There is an apparent contradiction between MFLDA starting with a single data source in Figures 4 and 6. However, this is accountable. For the experiment in Figure 6, MFLDA with $\alpha = 10^5$ starts with $\mathbf{R}_{15}$ (lncRNA–disease associations), and incrementally adds other datasets. But for the experiments in Figure 4, MFLDA with $\alpha < 10^2$ only selects $\mathbf{R}_{12}$ (lncRNA–miRNA associations) for prediction, which is unable to reconstruct the target lncRNA–disease associations; as a result, it has a much lower AUC.

We further study the inner sequence similarity between lncRNAs computed by BLAST and the inner similarity between diseases measured by the Disease Ontology structure. The experimental results show that including them does not contribute to an improved AUC. Due to page limit, the details of the experimental setup and result analysis are provided in the Supplementary Table S5. In addition, we investigate the MFLDA ability of assigning zero (or smaller) weights to noisy data sources. For this study, we additionally include two noisy data sources ($\mathbf{R}_{16}$, relational data matrix for lncRNAs and drugs; and $\mathbf{R}_{24}$, relational data matrix for miRNAs and GO). The empirical study shows that MFLA can automatically assign zero weights to these two noisy data matrices. The details of the experimental setup and result analysis are also provided in the Supplementary Figure S2.

4 Conclusions and future work

In this article, we introduced a matrix factorization based data fusion model to identify lncRNA–disease associations. Unlike current matrix factorization based solutions, MFLDA can select and weigh heterogeneous data sources, thus achieving a superior performance. MFLDA also outperforms other data fusion techniques, which map heterogenous data sources onto homologous networks before the identification step, and thus may enshroud the intrinsic structure of the data sources. MFLDA can explore and exploit the intrinsic and shared structure of heterogenous data sources. Furthermore, MFLDA is a general matrix factorization based data fusion framework; as such, it can readily integrate various heterogenous data sources to predict associations between different types of entities, such as RNA–protein interactions, associations between genes and GO terms, and so on.

How to efficiently and jointly seek the optimal rank size of low-rank matrices is an interesting future work and may further improve the performance MFLDA. MFLDA uses the reconstruction loss criterion to evaluate the quality of different data matrices, and thus it may be biased towards data matrices with sparse entries, even if they have low quality or relevance. More fit criteria are interesting future pursues. MFLDA also depends on the quality of the initial low-rank approximations of each data matrix.

Funding

This work was supported by Natural Science Foundation of China [No. 61402378], Natural Science Foundation of CQ CSTC [No. cstc2016jcyj A0351], Fundamental Research Funds for the Central Universities of China [2362015XK07, XDJK2016B009 and XDJK2017D061] and Chongqing Graduate Student Research Innovation Project [No. CYS16070].

Conflict of Interest: none declared.

References

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Artibani,M. *et al*. (2017) WT1 expression in breast cancer disrupts the epithelial/mesenchymal balance of tumour cells and correlates with the metabolic response to docetaxel. *Sci. Rep*., **7**, 45255.

Barabasi,A. and Oltvai,Z. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*., **5**, 101–113.

Barabasi,A. *et al*. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*., **12**, 56–68.

Chang,H. *et al*. (2011) Identification of genes associated with chemosensitivity to SAHA/taxane combination treatment in taxane-resistant breast cancer cells. *Breast Cancer Res. Treat*, **125**, 55–63.

Chen,G. *et al*. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*., **41**, D983–D986.

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

Chen,X. (2015a) Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep*., **5**, 13186.

Chen,X. (2015b) KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep*., **5**, 16840.

Chen,X. *et al*. (2016) IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. *Oncotarget*, **7**, 57919–57931.

Chen,X. *et al*. (2017) Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform*., **18**, 558–576.

Core,L.J. *et al*. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.

Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet*., **12**, 861–874.

Ferlay,J. *et al*. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*, **136**, E359.

Gligorijevic,V. and Przulj,N. (2015) Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, **12**, 20150571.

Godinho,M. *et al*. (2012) BCAR4 induces antioestrogen resistance but sensitises breast cancer to lapatinib. *Br. J. Cancer*, **107**, 947–955.

Gupta,R.A. *et al*. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.

Huang,Y.A. and Chen,X. (2015) ILNCSIM: improved lncRNA functional similarity calculation model. *Oncotarget*, **7**, 25902–25914.

Hsu,S.D. *et al*. (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res*., **42**, D78–D85.

Jeuris,B. *et al*. (2012) A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electron. Trans. Numer. Anal*., **39**, 379–402.

Jiang,Q. *et al*. (2015) LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Res*., **43**, D193–D196.

Jonquet,C. *et al*. (2009) The open biomedical annotator. *Summit Transl. Bioinformatics*, **2009**, 56–60.

Kaneuchi,M. *et al*. (2005) WT1 and WT1-AS genes are inactivated by promoter methylation in ovarian clear cell adenocarcinoma. *Cancer*, **104**, 1924.

Lan,W. *et al*. (2016) LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics*, **33**, 458–460.

Law,V. *et al*. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*., **42**, (D1), D1091–D1097.

Lee,D.D. and Seung,H.S. (2001) Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Process. Syst*., **32**, 535–541.

Li,J.H. *et al*. (2013a) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*., **42**, D92–D97.

Li,T. *et al*. (2015) Amplification of long noncoding RNA ZFAS1 promotes metastasis in hepatocellular carcinoma. *Cancer Res*., **75**, 3181–3191.

Li,Y. *et al*. (2013b) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*., **42**, D1070–D1074.

Lu,Z. *et al*. (2007) GeneRIF quality assurance as summary revision. *Pacific Symposium on Biocomputing*, **12**, 269–280.

Meng,D. and De La Torre,F. (2013) Robust matrix factorization with unknown noise. In: *Proceedings of the IEEE International Conference on Computer Vision*, p 1337–1344. IEEE.

Mercer,T.R. *et al*. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet*., **10**, 155–159.

Nie,F. *et al*. (2017) Long noncoding RNA ZFAS1 promotes gastric cancer cells proliferation by epigenetically repressing KLF2 and NKD2 expression. *Oncotarget*, **8**, 38227.

Ning,S. *et al*. (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res*., **44**, D980–D985.

Pinero,J. *et al*. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028.

Ponting,C.P. *et al*. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Schriml,L.M. *et al*. (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*., **40**, D940–D946.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, D535–D539.

Sun,J. *et al*. (2014) Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. BioSyst*., **10**, 2074–2081.

Thorenoor,N. *et al*. (2016) Long non-coding RNA ZFAS1 interacts with CDK1 and is involved in p53-dependent cell cycle control and apoptosis in colorectal cancer. *Oncotarget*, **7**, 622–637.

Tsai,M.C. *et al*. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.

Wang,H. *et al*. (2013) Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *J. Comput. Biol*., **20**, 344–358.

Wang,H.Q. *et al*. (2015) jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. *Bioinformatics*, **31**, 572–580.

Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.

Zhang,J. *et al*. (2017) Integrating multiple heterogeneous networks for novel lncrna-disease association inference. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **99**, 1–12.

Zhang,S. and Zhou,X.J. (2014) Matrix factorization methods for integrative cancer genomics. *Cancer Genomics Proteomics Methods Protoc*., **1176**, 229–242.

Zhou,M. *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.

Zitnik,M. and Zupan,B. (2015) Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell*., **37**, 41–53.