Weighted matrix factorization based data fusion for predicting lncRNA-disease associations

Guoxian Yu[1], Yuehui Wang[1], Jun Wang[1,*], Guangyuan Fu[1], Maozu Guo[2], Carlotta Domeniconi[3]
[1]College of Computer and Information Sciences, Southwest University, China
[2]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, China
[3]Department of Computer Science, George Mason University, USA
Email: {gxyu, yuehuiwang, kingjun, fugy}@swu.edu.cn, maozuguo@bucea.edu.cn, carlotta@cs.gmu.edu

***Abstract*—Increasing biomedical studies have demonstrated important associations between lncRNAs and various human complex diseases. Developing data integrative models can boost the performance of lncRNA-disease association identification. However, existing models generally have to transform heterogenous data into homologous networks, and then sum up these networks into a composite network for integrative prediction. The transformation may conceal the intrinsic structure of the heterogeneous data, and the summation process may suffer from noisy networks. Both these issues compromise the performance. In this paper, we introduce a Weighted Matrix Factorization based data fusion solution to predict LncRNA-Disease Associations (WMFLDA). WMFLDA first directly encodes the inter-associations between different types of biological entities (such as genes, lncRNAs, and Disease Ontology terms) via a heterogeneous network, which also encodes multiple intra-association networks of entities of the same type. Next, it assigns weights to these inter-association and intra-association matrices, and performs collaborative low-rank matrix factorization to explore the latent relationships between entities. After that, it simultaneously optimizes these weights and low-rank matrices. In the end, it uses the optimized low-rank matrices and weights to reconstruct the lncRNA-disease association matrix and accomplish the prediction. WMFLDA achieves a larger area under the receiver operating curve (by at least 7.61%), and a larger area under the precision-recall curve (by at least 5.49%) than competitive data fusion approaches in different experimental scenarios. WMFLDA can not only maintain the intrinsic structure of the association matrices, but can also selectively and differentially combine them. The codes and datasets are available at http://mlda.swu.edu.cn/codes.php?name=WMFLDA**

***Index Terms*—LncRNA-disease associations, Matrix factorization, Data fusion, Heterogeneous networks**

I. Introduction

In the past few years, with the rapid development of both experimental technology and computational methods, an increasing number of long non-coding RNAs (lncRNAs) has been discovered in large-scale transcriptome analysis, and their importance has become increasingly evident [2]. LncRNAs have low expressions and modest sequence conservation with highly specific tissues, and they have been involved in a wide range of biological processes [11]. Increasing evidence has demonstrated critical associations between lncRNAs and a broad range of complex human diseases [9], [15]. For example, the decreased expression of lncRNA 'WT1-AS' is shown to promote cell proliferation and invasion in gastric cancer [6]. Various lncRNA (or disease) related biological data have been accumulated, however, only a few lncRNA-disease associations have been reported. The study of identifying novel lncRNA-disease associations has attracted increasing attention [4]. Computational approaches can identify the most probable associations for experimental validation [3], [5], thus avoiding expensive and time-consuming wet-lab experiments.

Developing effective computational models to predict potential lncRNA-disease associations in large scale has become one of the most important topics of bioinformatics [4]. Various computational methods have been developed, and they can be roughly divided into three categories. The first category is mainly based on the known lncRNA-disease associations [5], [14]; these methods generally assume that similar diseases are associated with functionally similar lncRNAs. Most of these methods cannot be applied to new diseases without known associated lncRNAs, and vice versa. The second category predicts novel lncRNA-disease associations using known disease related genes or miRNAs [12], [21]. Most methods of the second category transform related data into a heterogenous network and apply network based inference to accomplish the prediction. The third category fuses multiple data sources to identify lncRNA-disease associations [3], [19]. Each data source provides a partial view of the complex mechanism between diseases and lncRNAs, and combining diverse data sources can provide a more comprehensive view [8]. The latter data fusion methods typically obtain a better performance than the methods that use individual data sources alone. However, they often transform heterogenous data sources into homogeneous networks, and simply sum up these homologous networks. As such, they may neglect the intrinsic structure of the individual data sources, and may be impacted by noisy networks [7].

Matrix factorization techniques have been widely developed to fuse multiple heterogenous data sources [8]. They neither have to transform heterogenous data sources into homogeneous networks, nor develop separate models for each data source [22]. In addition, they can explore and employ the

* Corresponding author: kingjun@swu.edu.cn (Jun Wang)

intrinsic and shared structure for different types of data. Wang *et al.* [17] proposed a symmetric nonnegative matrix tri-factorization approach (S-NMTF) to simultaneously cluster multi-type relational data sources. Zitnik and Zupan [22] developed a penalized matrix tri-factorization based model (DFMF) to simultaneously factorize multiple data matrices for predicting gene functions and pharmacologic actions. Lu *et al.* [13] introduced an inductive matrix completion [10] based approach (SIMCLDA), which uses feature vectors extracted from Gaussian interaction profile kernel of lncRNAs and functional similarity of diseases to predict potential lncRNA-disease associations.

The aforementioned matrix factorization-based data fusion methods show great potential in recovering underlying associations between various types of biological data, but they implicitly assume that each data source has equal relevance towards the target prediction task, and ignore the different degrees of relevance each data source may have [7]. To overcome this problem, Fu *et al.* [7] introduced a matrix factorization based lncRNA-disease association prediction (MFLDA) model to assign weights to individual inter-association matrices between different types of biological entities, and simultaneously decompose these inter-association matrices into low-rank matrices. MFLDA then uses the optimized low-rank matrices and weights to reconstruct the target matrix to predict new associations between lncRNAs and diseases. However, MFLDA does not account for the different degrees of relevance of intra-association matrices of the same type of objects, and thus its performance may be compromised by the low-quality or irrelevant data sources.

To simultaneously account for the different degrees of relevance of inter-association matrices and of multiple intra-association matrices, we propose a Weighted Matrix Factorization based data fusion solution to predict LncRNA-Disease Associations (WMFLDA). WMFLDA first encodes inter-associations between different types of biological entities (such as genes, lncRNAs, and Disease Ontology terms) via a heterogeneous network, which also encodes multiple intra-association networks of the objects of the same type. Then, it presets weights for inter-association and intra-association matrices and performs collaborative low-rank matrix factorization. Next, it simultaneously optimizes the weights and the low-rank matrix approximations of the inter-association and intra-association data matrices. After that, WMFLDA completes the lncRNA-disease association matrix based on the product of optimized low-rank matrices and weights to predict potential disease-related lncRNAs. In five-fold cross validation experiments on experimentally confirmed lncRNA-disease associations, WMFLDA achieves an AUROC of 0.9037 and an AUPR of 0.3747. It significantly outperforms related comparing methods, including RWRlncD [14], KATZLDA [3], RWRHLDA [21], S-NMTF [17], SIMCLDA [13], and MFLDA [7]. In addition, the experiments confirm that WMFLDA can selectively and differentially combine inter-association and intra-association data matrices. We want to remark that WMFLDA can also be directly used to recommend links between other entities.

II. Methodology

A. Problem Formulation

Suppose there are m types of molecules directly or indirectly related to lncRNAs or diseases, and a collection of inter-association 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 associations between n_i objects of the i-th type and n_j objects of the j-th type. Note, $\mathbf{R}_{ij}$ can be asymmetric. Multiple intra-association data matrices for the i-th type of objects are denoted as $\mathbf{\Theta}_i^{(t)} \in \mathbb{R}^{n_i \times n_i}$, $t \in \{1, 2, \cdots, t_i\}$, where t_i is the number of intra-association matrices for the i-th object type. Matrix factorization based data fusion wants to collaboratively decompose $\mathbf{R}$ or its sub-matrices, constrained by sub-matrices of $\mathbf{\Theta}$, into low-rank matrices to explore the latent relationship between objects of different types. It then uses the low-rank matrices to reconstruct the target association matrix (i.e. $\mathbf{R}_{ij}$) and predict new associations between objects of the i-th type (lncRNAs) and objects of the j-th type (diseases).

B. Matrix Factorization based Data Fusion

The matrix factorization-based data fusion approach has various variants [16]–[18], [22]. To facilitate the discussion of our problem, we start with a recent and representative framework proposed by Zitnik and Zupan [22]. The objective function of this framework is:

$$\begin{aligned} \min_{G \geq 0} \mathcal{Z}(\mathbf{G}, \mathbf{S}) = & \sum_{\mathbf{R}_{ij} \in \mathcal{R}} \left\| \mathbf{R}_{ij} - \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T \right\|_F^2 \\ & + \sum_{t=1}^{\tau} tr(\mathbf{G}^T) \mathbf{\Theta}^{(t)} \mathbf{G} \end{aligned} \tag{1}$$

where $\mathbf{G}_i \in \mathbb{R}^{n_i \times k_i}$, $\mathbf{G}_j \in \mathbb{R}^{n_j \times k_j}$, $\tau = \max_i t_i$, $\mathbf{S}_{ij} \in \mathbb{R}^{k_i \times k_j}$ $(k_i \ll n_i, k_j \ll n_j)$, $\mathbf{G} = diag(\mathbf{G}_1, \mathbf{G}_2, \cdots, \mathbf{G}_m)$, $tr(\cdot)$ and $\|\cdot\|_F^2$ are the matrix trace operator and the Frobenius norm, respectively. $\mathbf{S}_{ij}$ has much fewer vectors than $\mathbf{R}_{ij}$ and it can be viewed as a compressed matrix, which encodes latent inter associations between objects of the i-th type and objects of the j-th type. $\mathbf{G}_i(\mathbf{G}_j)$ is the low-rank representation of objects of the i-th (j-th) type. $\mathbf{\Theta}^{(t)}$ collectively stores all the following block diagonal matrices: $\mathbf{\Theta}^{(t)} = diag(\mathbf{\Theta}_1^{(t)}, \mathbf{\Theta}_2^{(t)}, \cdots, \mathbf{\Theta}_m^{(t)})$, $t \in (1, 2, \cdots, \max_i t_i)$, and the i-th block matrix along the main diagonal of $\mathbf{\Theta}^{(t)}$ is zero if $t > t_i$. Entries in intra-association matrices are positive for dissimilar objects, and negative for similar ones. The positive entries are known as cannot-link constraints, because they force pairs of dissimilar objects to be far away from each other in the latent component space; and the negative entries are must-link constraints, since they force pairs of objects to be close in the latent component space. These intra-associations can guide the pursue of coherent low-rank matrices $\mathbf{G}_i$ and 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 $\mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j$.

Many network based data fusion methods [3], [20], [21] first map lncRNA (disease)-related data sources onto homologous networks of lncRNAs (diseases), and then fuse lncRNA (disease) similarity networks via network integration [19]. Eq. (1) directly works on multi-type objects with multi-relations rather than mapping them onto homologous networks, and thus it has the potential of exploring and employing the intrinsic structure of objects of the same type and of different types. In addition, Eq. (1) optimizes the target association matrix $\mathbf{R}_{ij}$ with respect to $\mathbf{G}_i$ and $\mathbf{G}_j$, and both $\mathbf{G}_i$ and $\mathbf{G}_j$ are also determined by other indirectly connected data sources of lncRNAs and diseases; therefore, it can also account for multiple indirect data sources. However, we can clearly see that Eq. (1) equally treats all the inter-association matrices $\mathbf{R}_{ij}$ and all the intra-association matrices $\mathbf{\Theta}_i^{(t)}, i = \{1, 2, \ldots, m\}, t \in \{1, 2, \cdots, \tau\}$. As such, its performance may be compromised by noisy (or irrelevant) inter-association networks and intra-association networks. To account for the different degrees of relevance of inter-association matrices towards the prediction task, Fu *et al.* [7] extended Eq. (1) by assigning different weights to $\mathbf{R}_{ij}$ $(i, j = 1, 2, \ldots, m)$. Empirical studies have shown that accounting for the different degrees of relevance can improve the prediction performance. However, the approach of Fu *et al.* [7] ignores the different relevance of the various intra-association data sources.

C. Objective Function of WMFLDA

From the above analysis we advocate to assign weights to different intra-association and inter-association matrices to selectively integrate the multi-relational data matrices. In this way, both the inter-association and the intra-association matrices can be selectively combined for integrative prediction, and the impact of noisy networks can be further reduced. To this end, the objective function of WMFLDA is defined as follows:

$$\begin{aligned}\min_{\mathbf{G}\geq 0} \mathcal{L}(\mathbf{G}, \mathbf{S}, \mathbf{W}^r, \mathbf{W}^h) = &\sum_{\mathbf{R}_{ij}\in\mathbb{R}} \mathbf{W}_{ij}^r \left\|\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T\right\|_F^2 \\ &+ \sum_{i=1}^{m}\sum_{t=1}^{\tau} \mathbf{W}_{it}^h tr(\mathbf{G}_i^T\mathbf{\Theta}_i^{(t)}\mathbf{G}_i) \\ s.t. \quad \mathbf{W}^r \geq 0, \mathbf{W}^h \geq 0, &\sum_{i,j=1}^{m} \mathbf{W}_{ij}^r = 1, \sum_{i=1}^{m}\sum_{t=1}^{\tau} \mathbf{W}_{it}^h = 1\end{aligned} \tag{2}$$

where $\mathbf{W}^r \in \mathbb{R}^{m\times m}$, $\mathbf{W}^h \in \mathbb{R}^{m\times\tau}$, $\mathbf{W}^r$ contains the weights assigned to $|\mathcal{R}|$ inter-association matrices and $\mathbf{W}_{it}^h$ is the weight of t-th intra-association matrix of the i-th object type. For $\mathbf{R}_{ij} \notin \mathcal{R}$, $\mathbf{W}_{ij}^r = 0$. For $\mathbf{\Theta}_i^t, t > \max_i t_i$, $\mathbf{W}_{it}^h = 0$. Unlike Eq. (1), Eq. (2) can explore the contribution of different intra-association and inter-association matrices by assigning weights to them. However, Eq. (2) may only set $\mathbf{W}_{ij}^r = 1$ to $\mathbf{R}_{ij}$ if $\mathbf{R}_{ij}$ has the smallest reconstruction loss ($\|\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j\|_F^2$) among all the inter-association matrices, and the other inter-association matrices will be disregarded. Eq. (2) may also assign $\mathbf{W}_{it}^h = 1$ to $\mathbf{\Theta}_i^{(t)}$, if $\mathbf{\Theta}_i^{(t)}$ has the fewest cannot-link constraints among all the intra-association matrices. In other words, the sparser the intra-association (inter-association) matrix is, the larger the weight assigned to it will be. As a result, the contribution of other intra-association matrices will be ignored.

Given the complementary information of different data sources, using only one inter-association matrix and one intra-association matrix may not produce reliable predictions. To avoid this trivial weight assignment, we add two l_2-norm based regularization terms on $\mathbf{W}^r$ and $\mathbf{W}^h$, and update the objective function as follows:

$$\begin{aligned}\min_{G\geq 0} \mathcal{L}(\mathbf{G}, \mathbf{S}, \mathbf{W}^r, \mathbf{W}^h) = &\sum_{R_{ij}\in\mathbb{R}} \mathbf{W}_{ij}^r \left\|\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j^T\right\|_F^2 \\ &+ \sum_{i=1}^{m}\sum_{t=1}^{\tau} \mathbf{W}_{it}^h tr(\mathbf{G}_i^T\mathbf{\Theta}_i^{(t)}\mathbf{G}_i) \\ &+ \alpha\left\|vec(\mathbf{W}^r)\right\|_F^2 + \beta\left\|vec(\mathbf{W}^h)\right\|_F^2 \\ s.t. \quad \mathbf{W}^r \geq 0, \mathbf{W}^h \geq 0, &\sum vec(\mathbf{W}^r) = 1, \sum vec(\mathbf{W}^h) = 1\end{aligned} \tag{3}$$

where $vec(\mathbf{W}^r)$ is the vectorization operator that stacks the rows of $\mathbf{W}^r$, $\alpha > 0$ and $\beta > 0$ are used to control the complexity of $vec(\mathbf{W}^r)$ and $vec(\mathbf{W}^h)$; α and β can also help to selectively integrate inter-association and intra-association data matrices.

The objective function of WMFLDA is non-convex in $\mathbf{G}$, $\mathbf{S}$, $\mathbf{W}^r$, and $\mathbf{W}^h$ altogether. We can optimize $\mathcal{L}(\mathbf{G}, \mathbf{S}, \mathbf{W}^r, \mathbf{W}^h)$ according to the idea of auxiliary functions frequently used in the convergence proof of approximate matrix factorization algorithms [22]. For $\mathbf{G}$, $\mathbf{S}$, $\mathbf{W}^r$, and $\mathbf{W}^h$, we alternatively consider three of them as constants and optimize the other one. We obtain the optimal $\mathbf{G}$ and $\mathbf{S}$ by taking the partial derivative of Eq. (3) with respect to $\mathbf{G}$ and $\mathbf{S}$ with importing the Lagrangian multipliers and Karush-Kuhn-Tucker complementary condition [1]. The explicit solution of $\mathbf{W}^r, \mathbf{W}^h$ are shown as follows:

$$\mathbf{W}_{ij}^r = \begin{cases} \frac{\gamma - \mathbf{L}_{ij}}{2\alpha} & \text{if } \gamma > \mathbf{L}_{ij} \text{ and } \mathbf{R}_{ij} \in \mathcal{R} \\ 0 & \text{if } \gamma \leq \mathbf{L}_{ij} \text{ or } \mathbf{R}_{ij} \notin \mathcal{R} \end{cases}, \tag{4}$$

where $\mathbf{L}_{ij} = \|\mathbf{R}_{ij} - \mathbf{G}_i\mathbf{S}_{ij}\mathbf{G}_j\|_F^2$ be the reconstruction loss for $\mathbf{R}_{ij}$, $\gamma = \frac{2\alpha + \sum_{p'=1}^{p}\mathbf{v}_\mathbf{L}(p')}{p}$. Let $\mathbf{v}_L \in \mathbb{R}^{|\mathcal{R}|}$ store the entries of vector $vec(\mathbf{L})$ in ascending order with entries corresponding to $\mathbf{R}_{ij} \notin \mathcal{R}$ removed. Let $p \in \{1, 2, \ldots, |\mathcal{R}|\}$, there exists a appropriate p', satisfying $\mathbf{v}^r(p') = \frac{\gamma - \mathbf{v}_L(p')}{2\alpha}$ when $p' \leq p$, $\mathbf{v}^r(p') = 0$ when $p' > p$. $\mathbf{v}^r \in \mathbb{R}^{|\mathcal{R}|}$ stores the corresponding entries of $vec(\mathbf{W}^r)$ with entries corresponding to $\mathbf{R}_{ij} \notin \mathcal{R}$ removed.

$$\mathbf{W}_{it}^h = \begin{cases} \frac{\mu - \mathbf{K}_i^{(t)}}{2\beta} & \text{if } \mu > \mathbf{K}_i^{(t)} \text{ and } t \leq \max_i t_i \\ 0 & \text{if } \mu \leq \mathbf{K}_i^{(t)} \text{ and } t > \max_i t_i \end{cases}, \tag{5}$$

where $\mathbf{K}_i^{(t)} = tr(\mathbf{G}_i^T \mathbf{\Theta}_i^{(t)} \mathbf{G}_i)$, $\mu = \frac{2\beta + \sum_{q'=1}^{q} \mathbf{v}_\mathbf{K}(q')}{q}$. Let $\mathbf{v}_K$ store the entries of vector $vec(\mathbf{K})$ in ascending order with entries corresponding to $\mathbf{\Theta}_i^{(t)}, t > \max_i t_i$ removed. Let $q \in \{1, 2, \ldots, |\mathbf{\Theta}|\}$, $|\mathbf{\Theta}|$ is the number of intra-association matrices for all the types, there exists an appropriate q', satisfying $\mathbf{v}^h(q') = \frac{\mu - \mathbf{v}_\mathbf{K}(q')}{2\beta}$ when $q' \leq q$, $\mathbf{v}^h(q') = 0$ when $q' > q$. $\mathbf{v}^h$ stores the corresponding entries of $vec(\mathbf{W}^h)$ with entries corresponding to $\mathbf{\Theta}_i^{(t)}, t > \max_i t_i$ removed.

From the explicit solution of $\mathbf{W}^r$ and $\mathbf{W}^h$, we can easily see that the inter-association matrix $\mathbf{R}_{ij}$ that has the smallest reconstruction loss will be assigned a larger weight. Similarly, the intra-association matrix $\mathbf{\Theta}_i^t$ that has fewer cannot-link constraints will be assigned a larger weight. On the other hand, a small (or zero) weight will be assigned to the inter-association matrix with a larger reconstruction loss, and a small (or zero) weight will be assigned to the intra-association matrix with more cannot-link constraints. A larger reconstruction loss may often be caused by the noisy entries of the respective association matrices. We want to highlight that WMFLDA has the potential to automatically remove the noisy inter-association and intra-association matrices, simply by assigning zero weights to the respective data matrices. In addition, we want to remark that low-rank matrix factorization is also robust to noisy associations. As such, WMFLDA can be more robust to noisy data sources, and provide a more prominent performance than the matrix factorization based data fusion solutions [7], [13], [17], [22].

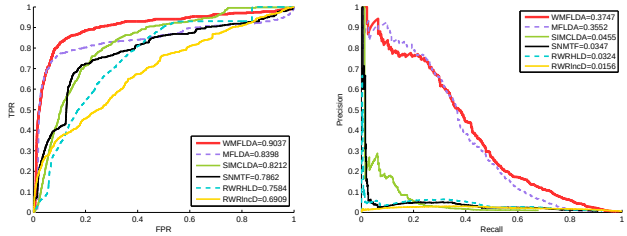
III. Results and Discussion

A. Experimental setup

To investigate the performance of WMFLDA, we consider 6 object types: lncRNAs (Type 1), miRNAs (Type 2), genes (Type 3), Gene Ontology (Type 4), Disease Ontology (Type 5), and drugs (Type 6). We collected 25 matrices (9 inter-association matrices, 16 intra-association matrices) between these objects from public databases. The details of the data sources are listed in Table I. We used the latest version of the databases (access date: 15 June 2018) for the experiments.

To comparatively study the effectiveness of the proposed WMFLDA, we compare it against five representative and competitive solutions for lncRNA-disease association prediction, which include KATZLDA [3], RWRlncD [14], RWRHLD [21], S-NMTF [17], SIMCLDA [13], and MFLDA [7]. The first four methods equally treat all the association matrices during the fusion process, and MFLDA only assigns weights to different inter-association matrices. The input parameters of these methods are set as specified by the authors in their code, or optimized in the suggested ranges. $\alpha = 10^7$ and $\beta = 10^6$ are adopted for WMFLDA. The low-rank size $k_i, i \in \{1, 2, \ldots, 6\}$ are separately specified to 220, 220, 160, 20, 50 and 50 for experiments.

Five fold cross-validation is adopted to investigate the performance. After each association has been tested in a single round of cross validation, we plot the receiver operating characteristic (ROC) curve by varying the true positive rate (TPR, sensitivity) against the false positive rate (FPR, 1-specificity) at different rank cutoffs. The value of the area under the ROC Curve (AUROC) can be computed to quantify the overall performance. The larger the AUROC value, the better the performance is, and a random guess corresponds to an AUROC value of 0.5. In addition, we also use the Precision-Recall (PR) curve to measure the performance of the methods and the area under the PR curve (AUPRC) to quantify the overall performance. The AUROC and AUPRC values can quantify the performance from different aspects, and they can provide a more comprehensive evaluation of the proposed solution.



(a) ROC curves and AUROCs (b) PR curves and AUPRCs

Fig. 1. Performance comparison between WMFLDA, MFLDA, SIMCLDA, S-NMTF, RWRHLD, and RWRlncD. (a) ROC curve and AUROCs. (b) PR curve and AUPRCs.

B. LncRNA-Disease Association Prediction with Cross Validation

In this section, we perform five-fold cross validation on experimentally confirmed lncRNA-disease associations to study the performance of WMFLDA. Particularly, we randomly divide known lncRNA-disease associations ($\mathbf{R}_{15}$) into five folds; the associations in four folds are used as training samples and the remaining associations of the other fold are alternatively used as testing samples for evaluation. Fig. 1(a) plots the ROC curves of the comparing methods and reports their corresponding AUROCs of 5-fold cross validation. It is evident that WMFLDA always has the highest TPRs under the same FPRs, and achieves the highest AUROC (0.9037) among the methods; the AUROCs of MFLDA, SIMCLDA, S-NMTF, RWRHLDA, and RWRlncD are 0.8398, 0.8212, 0.7862, 0.7584, and 0.6909, respectively. WMFLDA improves the AUROC by at least 7.61% with respect to the comparing methods. As for the PR curves and AUPRC in Fig. 1(b), WMFLDA again consistently outperforms the other methods, and it improves the AUPRC by at least 5.49%. These comparisons demonstrate the effectiveness of WMFLDA in selectively combining multiple inter-association and intra-association data matrices for accurate lncRNA-disease association prediction.

WMFLDA performs significantly better than MFLDA, although MFLDA also optimizes the weights assigned to different inter-association matrices. This is because MFLDA ignores the different relevance levels of multiple intra-association matrices towards the target prediction task. SIMCLDA uses principle component analysis to extract informative feature

TABLE I
DETAILS ON THE COLLECTED INTER-ASSOCIATION AND INTRA-CONSTRAINT MATRICES FROM DIFFERENT DATA SOURCES

Type	Source	Mapped Samples	Mapped Associations		Website
LncRNA-Disease	LncRNADisease Lnc2Cancer	240×412	2697	$\mathbf{R}_{15}$	http://www.cuilab.cn/lncrnadisease/ http://www.bio-bigdata.net/lnc2cancer/
LncRNA-miRNA	StarBase v2.0	240×495	1002	$\mathbf{R}_{12}$	http://starbase.sysu.edu.cn/mirLncRNA.php/
LncRNA-Gene	LncRNA2Target	240×15527	6186	$\mathbf{R}_{13}$	http://www.lncrna2target.org/
LncRNA-GO	GeneRIF	240×6428	3094	$\mathbf{R}_{14}$	ftp://ftp.ncbi.nih.gov/gene/GeneRIF/
miRNA-Disease	HMDD	495×412	13562	$\mathbf{R}_{25}$	http://www.cuilab.cn/hmdd/
miRNA-Gene	miRTarBase	495×15527	135852	$\mathbf{R}_{23}$	http://mirtarbase.mbc.nctu.edu.tw/
Gene-GO	GO Annotation	15527×6428	1191503	$\mathbf{R}_{34}$	http://geneontology.org/
Gene-Disease	DisGeNET	15527×412	115317	$\mathbf{R}_{35}$	http://www.disgenet.org/
Gene-Drug	DrugBank	15527×8283	3760	$\mathbf{R}_{36}$	http://www.drugbank.ca/
Gene-Gene	DIP	2719×2719	4551	$\mathbf{\Theta}_3^{(1)}$	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
	HPRD	7898×7898	32097	$\mathbf{\Theta}_3^{(2)}$	http://hprd.org/index_html
	I2D	13106×13106	283306	$\mathbf{\Theta}_3^{(3)}$	http://ophid.utoronto.ca/ophidv2.204/index.jsp
	IntAct	11778×11778	113973	$\mathbf{\Theta}_3^{(4)}$	http://www.ebi.ac.uk/intact/
	MINT	7898×7898	32097	$\mathbf{\Theta}_3^{(5)}$	http://mint.bio.uniroma2.it/
	BioGrid	13086×13086	289961	$\mathbf{\Theta}_3^{(6)}$	http://thebiogrid.org/
Drug-Drug	CredibleMeds	63×63	83	$\mathbf{\Theta}_6^{(1)}$	http://www.crediblemeds.org/
	DDIcorpus2011	314×314	582	$\mathbf{\Theta}_6^{(2)}$	http://labda.inf.uc3m.es/doku.php?id=en:labda_ddicorpus
	DDIcorpus2013	495×495	1133	$\mathbf{\Theta}_6^{(3)}$	http://labda.inf.uc3m.es/doku.php?id=en:labda_ddicorpus
	ANSM	832×832	27159	$\mathbf{\Theta}_6^{(4)}$	http://ansm.sante.fr/Dossiers/Interactions-medicamenteuses/Interactions-medicamenteuses/(offset)/0
	NLMcorpus	189×189	408	$\mathbf{\Theta}_6^{(5)}$	http://github.com/dbmi-pitt/public-PDDI-analysis/tree/master/PDDI-Datasets/NLM-Corpus
	ONC	295×295	4029	$\mathbf{\Theta}_6^{(6)}$	http://github.com/dbmi-pitt/public-PDDI-analysis/tree/master/PDDI-Datasets/ONC-High-Priority
	OSCAR	225×225	8585	$\mathbf{\Theta}_6^{(7)}$	http://sites.google.com/site/oscarusermanual/-oscar-emr/3-0-clinical-functions/3-7-1
	PKcorpus	164×164	416	$\mathbf{\Theta}_6^{(8)}$	http://dbmi-icode-01.dbmi.pitt.edu/dikb-evidence/package-insert-DDI-NLP-corpus.html
	SemMedDB	571×571	4762	$\mathbf{\Theta}_6^{(9)}$	http://skr3.nlm.nih.gov/SemMed/
	DrugBank	2419×2419	453436	$\mathbf{\Theta}_6^{(10)}$	http://www.drugbank.ca/

vectors based on multiple inter-association matrices and then completes the lncRNA-disease association matrix using the primary feature vectors. However, it does not utilize the intra-association information. For this reason, it always loses to WMFLDA. S-NMTF integrates multiple inter-association data sources using matrix tri-factorization; it still has lower AUROC and AUPRC than WMFLDA. The reason is that S-NMTF cannot selectively combine different 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 inter-association matrices and requires the matrices to be symmetric. This is the cause of the high running time of S-NMTF. RWRHLD applies a random walk with restart on a heterogenous network to predict lncRNA-disease associations; its AUROC and AUPRC are marginally higher than RWRlncD. This is mainly because RWRHLD transforms heterogeneous data sources onto lncRNA functional similarity and disease similarity, and then takes advantage of these two networks and known lncRNA-disease associations to infer new associations. This transformation may conceal the intrinsic structure of the data sources. RWRlncD has the lowest AUROC and AUPRC among the comparing methods, since it only utilizes known lncRNA-disease associations to infer additional lncRNA-disease associations. These results show that data fusion methods for lncRNA-disease association prediction achieve a better performance than computational methods that use individual data sources alone.

C. Effects of weighting intra-association data matrices

From the explicit solution of $\mathbf{W}^r$ in Eq. (4), we can clearly see that once the value of α is specified, the weight $\mathbf{W}_{ij}^r$ assigned to $\mathbf{R}_{ij} \in \mathcal{R}$ can be computed based on the reconstruction loss of that matrix. In addition, from Eq. (5), we can easily see that once the value of β is specified, the weight assigned to $\mathbf{\Theta}_i^{(t)} \in \mathbf{\Theta}^{(t)}$ can be determined based on the number of cannot-link constraints for intra-association matrices $tr(\mathbf{G}_i^T \mathbf{\Theta}_i^{(t)} \mathbf{G}_i)$. Given this, both α and β play important roles in determining the performance of WMFLDA. To search for a feasible value of α and β, following the experimental settings in Section III-B, we conduct five-fold cross validation to predict lncRNA-disease associations by varying α and β in $10^{-2}, 10^{-1}, \ldots, 10^{9}, 10^{10}$, and report the average AUPRC under each combination of α and β in Fig. 2.

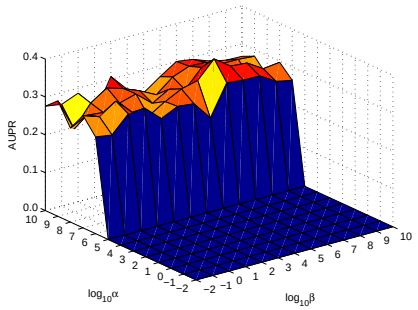


Fig. 2. AUPRC of WMFLDA under different input values of α and β.

From Fig. 2, we observe that when $\alpha = 10^7$ and $\beta = 10^6$, WMFLDA achieves the highest AUPRC. The input value of α significantly affects the performance; the AUPRC value increases as α increases, and reaches a plateau when $\alpha > 10^4$. This is because a too small α value assigns little emphasis to the inter-association matrices, and the target inter-association matrix is underrated as a result. The input value of β also affects the performance; the AUPRC value significantly increases as β gets larger, and then it slightly decreases when $\beta > 10^6$. This observation shows that both the input values of α and β have an impact on the performance of WMFLDA.

To further investigate the capability of WMFLDA in selectively combining intra-association matrices, we report in Fig. 3 the weights ($\mathbf{W}_{it}^h$) assigned to 10 intra-association matrices of drugs under different input values of β with $\alpha = 10^7$. We can

observe that when $\beta = 10^5$, only the intra-association matrix $\mathbf{\Theta}_6^{(1)}$ for drugs is selected. $\mathbf{\Theta}_6^{(1)}$ has the fewest cannot-link constraints (83 cannot-link constraints) among all the intra-association matrices for drugs in Table I. When $\beta \geq 10^9$, all the ten intra-association matrices $\mathbf{\Theta}_6^{(t)}, t = (1, 2, \ldots, 10)$ are selected and assigned nearly equal weights. This behavior is expected from Eq. (5). A (too) small β value does not have a sufficient regularization effect on the weights assigned to different intra-association matrices. On the other hand, a (too) large β value results in a strong regularization effect, and forces similar weight assignments to all matrices. When $\beta = 10^6$, 10^7 or 10^8, some intra-association matrices are selected for fusion, and WMFLDA has the highest AUROC and AUPRC when $\beta = 10^6$. The exclusion of other intra-association matrices is possible because these matrices may contain more noisy intra-associations than the selected sources, and the selected intra-association matrices have more reliable intra-associations to achieve an accurate lncRNA-disease associations prediction. In summary, the experiments with different values of β confirm that WMFLDA indeed can selectively integrate different intra-association matrices, and selectively combining these intra-association data matrices contributes to an improved performance.

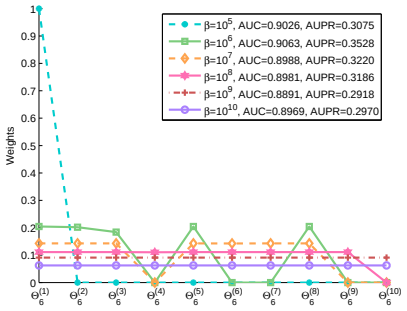


Fig. 3. Weights ($\mathbf{W}_{6t}^h$, $t = 1, 2, \cdots, 10$) assigned to ten intra-association matrices $\mathbf{\Theta}_6^{(t)}$ of drugs when $\beta = 10^5$, $\beta = 10^6$, $\beta = 10^7$, $\beta = 10^8$, $\beta = 10^9$, $\beta = 10^{10}$ and $\alpha = 10^7$.

IV. Conclusions

In this paper, we introduce a Weighted Matrix Tri-factorization based data fusion solution to predict LncRNA-Disease Associations (WMFLDA). Unlike other computational methods, WMFLDA can selectively integrate inter-associations and intra-associations of multi-relational data sources, and it can also explore and exploit the intrinsic and shared structure of heterogeneous data sources. The effectiveness of WMFLDA in predicting novel lncRNA-disease associations is confirmed by various experiments. WMFLDA can also be directly applied to predict links between different types of objects.

Acknowledgment

This work is supported by Natural Science Foundation of China (61741217, 61872300, 61873214, 61571163 and 61532014), the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228 and cstc2016jcyjA0351), and the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2017A05).

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[2] M. N. Cabili *et al.*, "Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses," *Genes & Development*, vol. 25, no. 18, pp. 1915–1927, 2011.
[3] X. Chen, "Katzlda: Katz measure for the lncrna-disease association prediction," *Scientific Reports*, vol. 5, p. 16840, 2015.
[4] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding rnas and complex diseases: from experimental results to computational models," *Briefings in bioinformatics*, vol. 18, no. 4, pp. 558–576, 2016.
[5] X. Chen and G. Yan, "Novel human lncrna-disease association inference based on lncrna expression profiles," *Bioinformatics*, vol. 20, no. 29, pp. 2617–2624, 2013.
[6] T. Du *et al.*, "Decreased expression of long non-coding rna wt1-as promotes cell proliferation and invasion in gastric cancer," *BBA Molecular Basis of Disease*, vol. 1862, no. 1, pp. 12–19, 2016.
[7] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix factorization-based data fusion for the prediction of lncrna–disease associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529–1537, 2017.
[8] V. Gligorijevic and N. Przulj, "Methods for biological data integration: perspectives and challenges," *Journal of The Royal Society Interface*, vol. 12, no. 112, 2015.
[9] R. A. Gupta *et al.*, "Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis," *Nature*, vol. 464, pp. 1071–1076, 2010.
[10] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *CoRR*, vol. abs/1306.0626, 2013.
[11] A. E. Kornienko *et al.*, "Gene regulation by the act of long non-coding rna transcription," *BMC Biology*, vol. 11, no. 1, p. 59, 2013.
[12] Y. Liu *et al.*, "Construction of a lncrnacpcg bipartite network and identification of cancer-related lncrnas: a case study in prostate cancer," *Molecular BioSystems*, vol. 11, pp. 384–393, 2015.
[13] C. Lu *et al.*, "Prediction of lncrnacdisease associations based on inductive matrix completion," *Bioinformatics*, p. bty327, 2018.
[14] J. Sun *et al.*, "Inferring novel lncrna-disease associations based on a random walk model of a lncrna functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.
[15] M.-C. Tsai *et al.*, "Long noncoding rna as modular scaffold of histone modification complexes," *Science*, vol. 329, no. 5992, pp. 689–693, 2010.
[16] H. Wang *et al.*, "jnmfma: a joint non-negative matrix factorization meta-analysis of transcriptomics data," *Bioinformatics*, vol. 31, no. 4, pp. 572–580, 2015.
[17] H. Wang *et al.*, "Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization," *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 279–284, 2011.
[18] H. Wang *et al.*, "Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization," *Journal of Computational Biology*, vol. 20, pp. 344–358, 2012.
[19] G. Yu, G. Fu, C. Lu, Y. Ren, and J. Wang, "Brwlda: bi-random walks for predicting lncrna-disease associations," *Oncotarget*, vol. 8, no. 36, pp. 60 429–60 446, 2017.
[20] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction using multi-label ensemble classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1045–1057, 2013.
[21] M. Zhou *et al.*, "Prioritizing candidate disease-related long non-coding rnas by walking on the heterogeneous lncrna and disease network," *Molecular BioSystems*, vol. 11, no. 3, pp. 760–769, 2015.
[22] M. Zitnik and B. Zupan, "Data fusion by matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 41–53, 2015.