Bioinformatics, 34(19), 2018, 3357–3364
doi: 10.1093/bioinformatics/bty327
Advance Access Publication Date: 26 April 2018
Original Paper

Systems biology

Prediction of lncRNA–disease associations based on inductive matrix completion

Chengqian Lu[1], Mengyun Yang[1], Feng Luo[2], Fang-Xiang Wu[3], Min Li[1], Yi Pan[4], Yaohang Li[5] and Jianxin Wang[1,*]

[1]School of Information Science and Engineering, Central South University, Changsha 410083, People's Republic of China, [2]School of Computing, Clemson University, Clemson, SC 29634, USA, [3]Division of Biomedical Engineering, University of Saskatchewan, Saskatchewan S7N 5A9, Canada, [4]Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994, USA and [5]Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on December 8, 2017; revised on April 4, 2018; editorial decision on April 18, 2018; accepted on April 25, 2018

Abstract

Motivation: Accumulating evidences indicate that long non-coding RNAs (lncRNAs) play pivotal roles in various biological processes. Mutations and dysregulations of lncRNAs are implicated in miscellaneous human diseases. Predicting lncRNA–disease associations is beneficial to disease diagnosis as well as treatment. Although many computational methods have been developed, precisely identifying lncRNA–disease associations, especially for novel lncRNAs, remains challenging.
Results: In this study, we propose a method (named SIMCLDA) for predicting potential lncRNA–disease associations based on inductive matrix completion. We compute Gaussian interaction profile kernel of lncRNAs from known lncRNA–disease interactions and functional similarity of diseases based on disease–gene and gene–gene onotology associations. Then, we extract primary feature vectors from Gaussian interaction profile kernel of lncRNAs and functional similarity of diseases by principal component analysis, respectively. For a new lncRNA, we calculate the interaction profile according to the interaction profiles of its neighbors. At last, we complete the association matrix based on the inductive matrix completion framework using the primary feature vectors from the constructed feature matrices. Computational results show that SIMCLDA can effectively predict lncRNA–disease associations with higher accuracy compared with previous methods. Furthermore, case studies show that SIMCLDA can effectively predict candidate lncRNAs for renal cancer, gastric cancer and prostate cancer.
Availability and implementation: https://github.com//bioinfomaticsCSU/SIMCLDA
Contact: jxwang@mail.csu.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Traditional central dogma of molecular biology postulates that genetic information is stored in the protein-coding genes (Yanofsky, 2007). However, accumulating evidences indicate that non-coding genes are not just plentiful but also play important roles. With the completion of ENCODE project, researchers found that ~74.7% of human genome are transcribed (Djebali *et al.*, 2012), in contrast to only ~1.5% encoding proteins (Lander *et al.*, 2001). Furthermore, accumulating evidences show that non-coding RNAs (ncRNAs) participate in a wide-repertoire of biological processes. In particular, long non-coding RNAs (lncRNAs) with length $\geq$ 200 nt (nucleotides) consist of the largest subclass of ncRNAs. Although lncRNAs

have no potential to encode, facts show that they perform various biological functions such as translational and post-translational regulation, cell differentiation, proliferation and apoptosis, epigenetic regulation and so on (Guttman *et al.*, 2009). Consequently, mutation and dysregulation of lncRNAs can cause miscellaneous human diseases, including breast cancer (Vincent-Salomon *et al.*, 2007), lung cancer (Chen *et al.*, 1997), Alzheimer's disease (Faghihi *et al.*, 2008; Liu *et al.*, 2018) and others. The details of mechanisms of lncRNAs are still a conundrum, but lncRNAs are considered as molecules for disease diagnosis and therapy. Predicting the relationships between lncRNAs and diseases has attracted more and more attentions. It is not only beneficial to disease diagnosis, but also helpful in understanding biological processes. Furthermore, computational methods provide a possible way to decrease the time and cost of experiments with limited known lncRNA–disease associations.

Computational methods proposed to predict potential lncRNA–disease associations could be classified into the following three categories. Methods in the first category identify lncRNA–disease associations by utilizing biological information of lncRNA, such as genome location, expression profile, tissue specificity and so on. Chen *et al.* (2013) predicted lncRNA–disease associations by virtue of neighbourship between lncRNAs and genes in genome location, based on known gene–disease associations. Since location and identification of lncRNAs is still an impediment, this model just works for small parts of lncRNAs. Liu *et al.* (2014) identified potential associations by combing lncRNA's tissue specificity and gene–lncRNA co-expression. Chen (2015) used the KATZ measure to find potential associations, integrating lncRNA expression profiles, lncRNA functional similarity, known lncRNA–disease associations, disease semantic similarity and Gaussian interaction profile kernel. However, tissue-specific expression and low expression level of lncRNAs limit these methods predicting for all lncRNAs. Methods in the second category uses machine learning models to identify potential associations. Chen and Yan (2013) proposed a semi-supervised learning method LRLSLDA to identify possible associations between lncRNAs and diseases by using Laplacian regularized least squares. However, this method suffered from combing two classifiers reasonably. Lan *et al.* (2016) fused different data sources and used a bagging SVM classifier to predict latent interactions between lncRNAs and diseases. Nevertheless, effectively fusing different kernels of lncRNAs is still a big problem. Based on the commonly accepted assumption that phenotypically similar diseases are prone to be associated with functionally similar lncRNAs and vice versa Wu *et al.* (2008), methods in the third category takes use of random walk. Sun *et al.* (2014) constructed a lncRNA–lncRNA functional similarity network through diseases semantic similarity, and used the random walk with restart to predict lncRNA–disease associations. Chen *et al.* (2016) took use of random walk on the lncRNA similarity network with an initial probability vector, specified by lncRNA expression similarity and disease semantic similarity. Zhang *et al.* (2017a) applied a flow propagation algorithm on a constructed network, incorporating information of lncRNAs, proteins and diseases. Yao *et al.* (2017) designed a heterogeneous random walk on the constructed multi-level composite network, integrating lncRNA, gene and phenotype. However, these models predict potential associations just utilizing the most important feature vector, corresponding to the maximum eigenvector of known information. Moreover, there is still a challenge to achieve significant performance for prediction.

In this study, we formulate lncRNA–disease association prediction as a recommendation system problem. We propose the use of an inductive matrix completion (IMC) method (Jain and Dhillon, 2013) for predicting lncRNA–disease associations (named SIMCLDA) by using informative feature vectors corresponding to the top singular vectors of the lncRNA and disease feature matrices. Generally, a recommendation system is an information filtering system that seeks to predict the preference that user would give to a certain item, given only partial known preference information. Recently, recommendation system methods have been applied to association prediction in a variety of bioinformatics problems. For example, Zheng *et al.* (2013) proposed a collaborative matrix factorization method to predict drug-target interactions; Luo *et al.* (2018) used a matrix completion method for drug repositioning. Analogously, the lncRNA–disease association prediction task takes the set of lncRNAs, the set of diseases and the set of partially known associations between lncRNAs and diseases, then recommends lncRNAs for a given disease, using prior information about lncRNAs and diseases. Similar to the assumption in the user-item recommendation system that users with similar behavior share similar preferences towards items, the lncRNA–disease prediction assumes that functionally similar lncRNAs exhibit similar interaction patterns with diseases. Here, we model the lncRNA–disease association prediction problem as a recommendation task and solve it with speedup IMC (Xu *et al.*, 2013). SIMCLDA uses principle components analysis (PCA) (Jolliffe, 1986) to extract informative feature vectors, based on disease–gene, gene–gene ontology and known lncRNA–disease associations. For a new lncRNA, SIMCLDA calculates the interaction profile according to its sequence-similar neighbors. Then, SIMCLDA completes the association matrix using the primary feature vectors from the constructed feature matrices. SIMCLDA outperforms the other methods in leave-one-out experiments. Moreover, case studies show that SIMCLDA is capable of inferring potential lncRNAs for renal cancer, gastric cancer and prostate cancer. In summary, incorporating effective features extracted from interaction profiles with neighboring information into fast IMC computation leads to SIMCLDA performance enhancement.

2 Materials and methods

Given a lncRNA–disease interaction matrix $A \in \mathbb{R}^{m \times n}$ from known lncRNA–disease associations, where m and n are the number of lncRNAs and diseases, respectively, and each row corresponds to a lncRNA while each column represents a disease. A_{ij} is 1 if lncRNA i is linked to disease j, and A_{ij} is 0 if their relationship is unknown. Predicting potential associations between lncRNAs and diseases is deemed as completing A, which is a problem of matrix completion.

2.1 Methods overview

To infer potential lncRNAs associated with diseases under consideration, we propose a method SIMCLDA based on IMC (Jain and Dhillon, 2013). SIMCLDA consists of five steps shown in Figure 1. Step 1, SIMCLDA computes Gaussian interaction profile kernel of lncRNAs (*Gkl*) from known lncRNA–disease interactions. Step 2, SIMCLDA uses Jaccard similarity coefficient to calculate functional similarity of diseases (*Dis*). Step 3, SIMCLDA extracts primary feature vectors from *Gkl* and *Dis* by PCA, respectively. Step 4, SIMCLDA calculates the interaction profile for a new lncRNA with the interaction profiles of its neighbors. Step 5, SIMCLDA completes the association matrix with IMC using primary feature vectors and constructed interaction profiles.

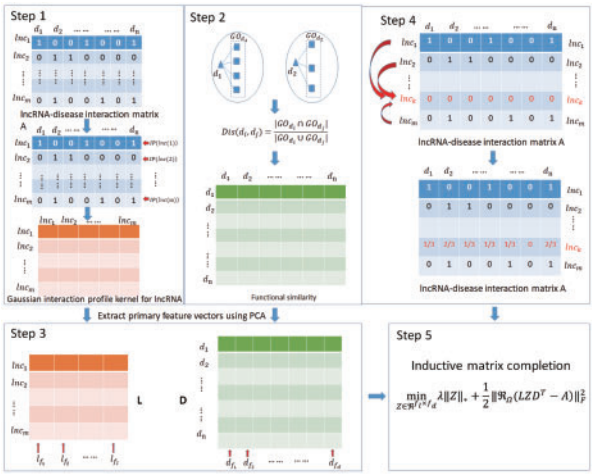

Fig. 1. Scheme of SIMCLDA. Step 1: computing *Gkl*. Step 2: calculating functional similarity of disease *Dis* based on disease–gene and gene–gene ontology associations. Step 3: extracting primary feature vectors from *Gkl* and *Dis* by PCA, respectively. Step 4: calculating interaction profile for a new lncRNA according to the interaction profiles of its neighbors. Step 5: completing the association matrix with IMC

2.2 Gaussian interaction profile kernel of lncRNAs

Based on the assumption that functionally similar lncRNAs exhibit similar interaction patterns with diseases, we use $Gkl \in \mathbb{R}^{m\times m}$ to define the lncRNA latent feature space including feature matrix. The interaction profile of lncRNA i ($IP(lnc_i)$) is the ith row vector of association matrix $A \in \mathbb{R}^{m\times n}$. It is a binary vector indicating the presence or absence of its associations with diseases. Then, the distance between any two row vectors is computed as Gaussian interaction profile kernel of their corresponding lncRNAs,

$$Gkl(lnc_i, lnc_j) = \exp\left(-\gamma_l \|IP(lnc_i) - IP(lnc_j)\|^2\right) \tag{1}$$

Here γ_l determines the kernel bandwidth, which is normalized by the average number of associations with diseases per lncRNA and is computed as follows,

$$\gamma_l = \frac{1}{m}\sum_{i=1}^{m} \|IP(lnc_i)\|^2 \tag{2}$$

2.3 Disease similarity

Recent studies discover that phenotypically similar diseases are related with similar dysfunctions of genes (Schlicker *et al.*, 2010). Gene ontology annotations provide a way to measure semantic function of genes. We calculate functional similarity of diseases ($Dis \in \mathbb{R}^{n\times n}$) using Jaccard similarity coefficient (Jaccard, 1908) based on disease–gene and gene–gene ontology associations,

$$Dis(d_i, d_j) = \frac{|GO_{d_i} \cap GO_{d_j}|}{|GO_{d_i} \cup GO_{d_j}|}, \tag{3}$$

where GO_{d_i} denotes the gene ontology terms associated with disease i.

2.4 Extracting primary feature vectors

IMC model assumes that the feature vectors of lncRNAs and disease interacting in the latent space determine the associations. Due to the low quality of the original data, we use PCA to extract the primary feature vectors from $Gkl \in \mathbb{R}^{m\times m}$ and $Dis \in \mathbb{R}^{n\times n}$, respectively. SIMCLDA employs singular value decomposition (SVD) to perform PCA. Because both *Gkl* and *Dis* are symmetric, then $\boldsymbol{Gkl} = U_l S_l U_l^T$ for lncRNAs, where U_l is an $m\times m$ unitary matrix and S_l is an $m\times m$ diagonal matrix with singular values deposited on the diagonal in descending order, and $\boldsymbol{Dis} = U_d S_d U_d^T$ for diseases, where U_d means an $n\times n$ unitary matrix and S_d is an $n\times n$ diagonal matrix with singular values on the diagonal in descending order. Based on the dominating energy strategy (Ji *et al.*, 2016), we find appropriate parameters f_l and f_d with specified α_l and α_d, which are the percentage of the sum of singular values in the diagonal matrix S_l and S_d such that

$$f_l = \arg\min_x \left\{ \frac{\sum_{i=1}^{x} (S_l)_{ii}}{\sum_{j=1}^{m} (S_l)_{jj}} \geq \alpha_l \right\} \tag{4}$$

and

$$f_l = \arg\min_x \left\{ \frac{\sum_{i=1}^{x} (S_d)_{ii}}{\sum_{j=1}^{n} (S_d)_{jj}} \geq \alpha_d \right\} \tag{5}$$

Then, we set $L = \left(V_{l_1}, V_{l_2}, \ldots, V_{l_{f_l}}\right)$ and $D = \left(V_{d_1}, V_{d_2}, \ldots, V_{d_{f_d}}\right)$. The primary feature vectors in the latent spaces are top singular vectors corresponding to the top singular values, which contain most intrinsic information.

2.5 Calculating interaction profile for a new lncRNA

During the prediction process, a new lncRNA without any known lncRNA–disease associations results in the cold-start problem (details in Supplementary 3.1), i.e. for a new lncRNA i, all elements of its interaction profile $\boldsymbol{IP}(lnc_i)$ are 0, indicating that no prior association knowledge could be used for prediction. Inspired by the solutions to the cold-start problem in collaborative filtering, we calculate the interaction profile for a new lncRNA using the mean of its neighbors' interaction profiles based on the assumption that the similar lncRNAs interact with the similar diseases. By virtue of the computed interaction profile, we are able to incorporate prior interaction patterns of the neighbors of this new lncRNA and extract effective feature vectors. For example, we compute a new interaction profile $\boldsymbol{IP}'(lnc_i)$ for lncRNA i. If similarities between other lncRNAs and the lncRNA i were larger than the mean of the sequence similarity, these lncRNAs can be defined as the neighbors of lncRNA i. Then, we calculate the interaction profile for lncRNA i by using the mean of its neighbors' interaction profiles. Moreover, we replace the corresponding interaction profiles with the newly formed one such that

$$\boldsymbol{IP}'(lnc_i) = \frac{\sum_{j\in N(lnc_i)} \boldsymbol{IP}(lnc_j)}{|N(lnc_i)|}, \tag{6}$$

where $N(lnc_i)$ is the set of the neighbors of lncRNA i and $|\cdot|$ denotes the cardinality of a set.

2.6 Inductive matrix completion

2.6.1 Standard matrix completion

In the standard model, matrix completion is to recovery the missing values in an $m\times n$ matrix $\boldsymbol{M}$ given partially observed entry sets Ω. Let $\Re_\Omega(M) : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^{m\times n}$ be a linear projection operator such that

$$\Re_\Omega(M)_{ij} = \begin{cases} A_{ij}, & if(i,j) \in \Omega \\ 0, & if(i,j) \notin \Omega \end{cases} \tag{7}$$

Under the common assumption that the entry values of $\boldsymbol{M}$ are determined by only a few latent factors, matrix completion can be

modeled as a constraint satisfaction problem by minimizing the rank of matrix $\boldsymbol{M}$, i.e.

$$\begin{aligned} &\text{minimize} \quad rank(M) \\ &s.t. \quad M_{ij} = A_{ij},\ (i,j) \in \Omega \end{aligned}, \tag{8}$$

Unfortunately, it is an NP-hard problem. By replacing the objective function in (Fazel, 2002) with the nuclear norm, the matrix completion can be reformulated as the following convex optimization problem,

$$\begin{aligned} &\text{minimize} \quad \|M\|_* \\ &s.t. \quad M_{ij} = A_{ij},\ (i,j) \in \Omega \end{aligned}, \tag{9}$$

where $\|\cdot\|_*$ is the nuclear norm defined as the sum of the singular values.

2.6.2 SIMCLDA

We use the kernel matrix of lncRNAs and diseases to represent the latent space by including feature vectors of side information based on kernel extension in Yu *et al.* (2014). We complete A based on the low-rank assumption in Xu *et al.* (2013), the column vectors in A lie in the subspace spanned by the column vectors in L, and the row vectors in A lie in the subspace spanned by the column vectors in D. Then, the problem can be defined as:

$$\begin{aligned} &\min_{Z \in \mathbb{R}^{f_l \times f_d}} \|Z\|_* \\ &s.t. \Re_\Omega\left(LZD^{\mathrm{T}}\right) = \Re_\Omega(A), \end{aligned} \tag{10}$$

where Z is the objective matrix to complete A.

Relaxing the constraint of $\Re_\Omega\left(LZD^{\mathrm{T}}\right) = \Re_\Omega(A)$ to $f(Z) = \frac{1}{2}\|\Re_\Omega\left(LZD^{\mathrm{T}} - A\right)\|_F^2$ and denoting $p(Z) = \|Z\|_*$, the above optimization problem (10) becomes

$$\min_{Z \in \mathbb{R}^{f_l \times f_d}} \lambda p(Z) + f(Z), \tag{11}$$

where λ is the regularization parameter controlling the extent of the nuclear norm.

For any given $Y \in \mathbb{R}^{f_l \times f_d}$, $f(Z)$ can be approximated by the following quadratic approximation

$$f(Z) \approx \tilde{f}(Z,Y) = f(Y) + \langle \nabla f(Y), Z - Y \rangle + \frac{s}{2}\|Z - Y\|_F^2 \tag{12}$$

$$= \frac{s}{2}\|Z - (Y - \frac{1}{s}\nabla f(Y))\|_F^2 + f(Y) - \frac{1}{2s}\|\nabla f(Y)\|_F^2, \tag{13}$$

where $\nabla f(Y) = L^{\mathrm{T}}\Re_\Omega\left(LYD^{\mathrm{T}} - A\right)D$ is the gradient of $f(Z)$ at Y, $\langle . \rangle$ denotes matrix inner product, and s is a proximal parameter for estimating the second-order gradient $\nabla^2 f(Y)$. Accordingly, the minimization model (11) becomes

$$\min_{Z \in \mathbb{R}^{f_l \times f_d}} \lambda\|Z\|_* + \frac{s}{2}\left\|Z - \left(Y - \frac{1}{s}\nabla f(Y)\right)\right\|_F^2 \tag{14}$$

We then generate an accelerated gradient descent (APG) (Toh and Yun, 2010) style iterative scheme for (14)

$$Y_k \leftarrow Z_k + \theta_k\left(\theta_{k-1}^{-1} - 1\right)(Z_k - Z_{k-1}) \tag{15a}$$

$$Z_{k+1} \leftarrow \arg\min_Z \lambda\|Z\|_* + \frac{s}{2}\left\|Z - \left(Y_k - \frac{1}{s}\nabla f(Y_k)\right)\right\|_F^2 \tag{15b}$$

$$\theta_{k+1} \leftarrow \left(\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2\right)/2 \tag{15c}$$

Particularly, (15b) can be obtained by recasting the linearized Bregman iterations as a special form of Uzawa's algorithm (Cai *et al.*, 2010) such that

$$Z_{k+1} \leftarrow D_{\frac{\lambda}{s}}\left(Y_k - \frac{1}{s}L^{\mathrm{T}}\Re_\Omega\left(LY_kD^{\mathrm{T}} - A\right)D\right) \tag{16}$$

Here $D_{\frac{\lambda}{s}}(.)$ denotes the matrix shrinkage operator based on SVD on the operand matrix with respect to threshold $\frac{\lambda}{s}$ such that

$$D_{\frac{\lambda}{s}}(X) = \sum_i^{\sigma_i \geq \frac{\lambda}{s}} \left(\sigma_i - \frac{\lambda}{s}\right) u_i v_i^{\mathrm{T}}, \tag{17}$$

where u_i and v_i are the left and right singular vectors of X corresponding to σ_i, respectively.

The prediction procedure is summarized in Algorithm 1. Based on inductive matrix completion, SIMCLDA iteratively updates the approximation using linearized Bregman iteration until convergence is reached.

3 Results and discussion

3.1 Data collection

We retrieve known lncRNA–disease associations from the gold standard dataset in LncRNADisease database. In order to evaluate the performance of our proposed SIMCLDA, we use three datasets. The first dataset is downloaded from LncRNADisease established in October 2012, which contains 293 experimentally validated lncRNA–disease associations. The second dataset is obtained from LncRNADisease established in 2014, which contains 351 lncRNA–disease associations with 156 lncRNAs and 189 diseases. The third dataset is retrieved from LncRNADisease established in 2015, which contains 685 lncRNA–disease associations with 256 lncRNAs and 189 diseases. After correcting names of diseases (according to UMlS,

Algorithm 1 SIMCLDA Algorithm

Input: Sequence similarity of lncRNAs, disease–gene associations, gene–GO associations and the incomplete lncRNA–disease association matrix $\boldsymbol{A}$

Output: Predicted association matrix $\boldsymbol{M}$

1: Calculate ***GKL*** based on known lncRNA–disease associations
2: Calculate disease similarity matrix ***Dis*** from disease–gene and gene–GO, using Jaccard similarity coefficient
3: Extract primary feature vectors from ***GKL*** and ***Dis*** using PCA
4: Construct new interaction profile for a new lncRNA
5: Initialize threshold $\epsilon, \theta_1 = \theta_2 \in (0,1], Z_1 = Z_2, s, \gamma > 1$
6: $k = 2$
7: **do**
8: $\quad Y_k = Z_k + \theta_k(\theta_{k-1}^{-1} - 1)(Z_k - Z_{k-1})$
9: $\quad Z_{k+1} = D_{\frac{\lambda}{s}}(Y_k - \frac{1}{s}L^{\mathrm{T}}\Re_\Omega(LY_kD^{\mathrm{T}} - A)D)$
10: $\quad$ **while** $\tilde{f}(Z_{k+1}) \geq \tilde{f}(Z_{k+1}, Y_k)$ **do** ▷ adjust s if overestimated
11: $\quad\quad s = s * \gamma$
12: $\quad\quad Z_{k+1} = D_{\frac{\lambda}{s}}(Y_k - \frac{1}{s}L^{\mathrm{T}}\Re_\Omega(LY_kD^{\mathrm{T}} - A)D)$
13: $\quad$ **end while**
14: $\quad \theta_{k+1} = (\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2)/2$
15: $\quad k = k + 1$
16: **while** $|f(Z_{k+1}) - f(Z_k)| \geq \epsilon$

Table 1. Details of final datasets

Datasets	No. of lncRNAs	No. of diseases	No. of interactions
Dataset1	112	150	276
Dataset2	131	169	319
Dataset3	285	226	621

Mesh and NCBI) and lncRNAs (according to HGNC, NCBI, Lncipedia and lncrnadb), we remove all repeating records with the same lncRNA and disease, and all the wrong entries that do not belong to human beings. The statistics of the final datasets are shown in Table 1, and the overlaps between the datasets are shown in Supplementary Figures S1, S2 and S3. We use BioMart to download associations between genes and gene ontology terms of human beings from Ensemble database (Aken *et al.*, 2016). Disease–gene associations are derived from DisGeNET database (Piñero *et al.*, 2017).

3.2 Leave-one-out cross validation

In order to evaluate the performance of SIMCLDA in predicting potential lncRNA–disease associations, we perform leave-one-out cross-validation (LOOCV) on known experimentally verified lncRNA–disease associations and prioritize candidates of disease-associated lncRNAs. For a given disease d_i, each known lncRNA associated to d_i is left out in turn as the test sample, and the other known experimentally verified lncRNAs associated with d_i are considered as training samples. All the lncRNAs without known associations with d_i make up the d_i-associated candidate samples. In the candidate samples, the test sample is deemed as a positive sample, and the others are negative samples. After performing prediction, the probabilities of associations between candidate samples and d_i are calculated. Meanwhile, all candidate lncRNAs are ranked by the predicted probabilities. The higher the candidate lncRNA is ranked, the better SIMCLDA performs. After all known associations have been tested, we calculate both true positive rate (TPR) and false positive rate (FPR) as follows:

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

where TP is the number of positive samples, whose rank is higher than a given rank cutoff and FN is the number of negative samples that are identified incorrectly.

$$FPR = \frac{FP}{FP + TN} \tag{19}$$

where FP is the number of negative samples, whose rank is lower than a given rank cutoff and TN is the number of negative samples that are identified correctly.

TPR indicates the percentage of the test sample rank is higher than a given rank cutoff. FPR means that the percentage of candidate lncRNAs ranked lower than a given rank cutoff. Then the receiver operating characteristic (ROC) curve is utilized to evaluate the performance, which plots TPR versus FPR with respect to various cutoffs. The latent feature space *Gkl* depends on the known lncRNA–disease associations. Since each known lncRNA–disease association is taken as a test sample in turn during LOOCV, we need to recalculate *Gkl* corresponding to different test samples.

3.3 The effects of parameters

3.3.1 The effects of α_l and α_d

Primary feature vectors of lncRNA and disease determine the possibility of interactions. Thus, the numbers of lncRNAs' and diseases'

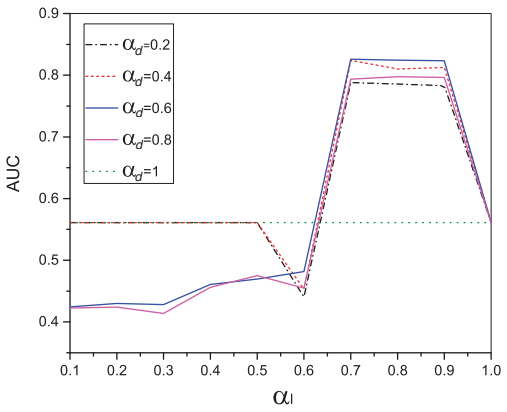

Fig. 2. Effects of parameters α_l and α_d on Dataset1

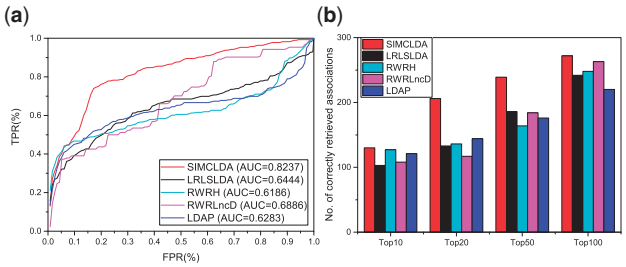

Fig. 3. Comparison of predicting methods on Dataset1. (**a**) Performance of all methods in terms of ROC curve using LOOCV. (**b**) Number of correctly retrieved known lncRNA–disease associations for specified rank thresholds

primary feature vectors f_l and f_d play important roles. SIMCLDA sets α_l and α_d to find proper f_l and f_d by using the dominant energy strategy on three datasets. We test effects of α_l, ranged from 0.1 to 0.9, and α_d, ranged from 0.2 to 1. As shown in Figure 2, we can see that the AUC values of SIMCLDA based on Dataset1 vary slightly for all α_d except $\alpha_d = 1$, when $0.1 \leq \alpha_l \leq 0.6$. It suggests that the primary feature vectors of lncRNAs are informative but certain important information is missing. The performance becomes much better by increasing α_l from 0.6 to 0.8, indicating important information is gradually integrated. Meanwhile, AUC values fall rapidly when $0.9 \leq \alpha_l \leq 1$, indicating SIMCLDA incorporates irrelevant information. SIMCLDA performs the best when α_d equals 0.6 for all datasets. The results of SIMCLDA on Dataset2 and Dataset3 are shown in Supplementary Figures S4 and S5, respectively. In summary, we set $\alpha_l = 0.8$ and $\alpha_d = 0.6$ as the default.

3.3.2 The effect of λ

In Equation (11), the regularization parameter is used to balance the nuclear norm and approximation error. Intuitively, if the magnitude of λ is too large, the nuclear norm brings in a large penalty which may lead to large approximation error. On the other hand, if the magnitude of λ is too small, the low-rank property of the solution matrix Z may not be ensured. Typically, the proper value of λ is determined by cross-validation. Here, we randomly divide the dataset into ten folds and use nine folds as training set and one fold as test set. Table 2 shows the mean (Mean) and the standard deviation (SD) of AUC values on the test sets in the ten-fold cross-validation with respect to λ ranging from 0.001 to 1000 increasing in the power of 10. Through the training process, SIMCLDA minimizes the objective function (11) with respect to the specified value of λ and estimates the performance on the test set. The results for the three datasets are

Table 2. The effect of λ

λ	Dataset1		Dataset2		Dataset3	
	Mean	SD	Mean	SD	Mean	SD
0.001	0.6435	0.0072	0.6944	0.0083	0.7332	0.0042
0.01	0.6490	0.0080	0.6989	0.0079	0.7397	0.0049
0.1	0.6704	0.0158	0.7045	0.0309	0.7791	0.0049
1	**0.8008**	**0.0086**	**0.8267**	**0.0051**	**0.8340**	**0.0038**
10	0.5810	0.0001	0.5613	0.0007	0.5421	0.0001
100	0.5731	0.0009	0.5523	0.0005	0.5328	0.0003
1000	0.5610	0.0002	0.5473	0.0004	0.5197	0.0001

Results are marked in bold on three datasets with $\lambda = 1$.

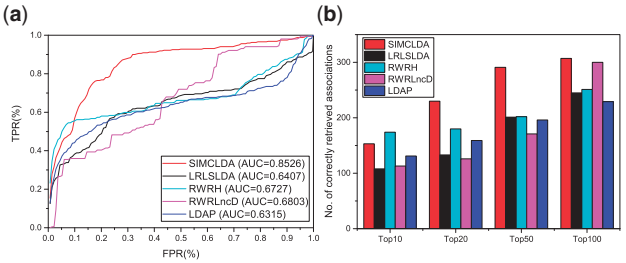

Fig. 4. Comparison of predicting methods on Dataset2. (**a**) Performance of all methods in terms of ROC curve using LOOCV. (**b**) Number of correctly retrieved known lncRNA–disease associations for specified rank thresholds

shown in the Table 2. One can find that SIMCLDA yields the best performance when λ adopts the magnitude of 1, which is consistent across the three datasets. As a result, we set $\lambda = 1$ as the default for SIMCLDA.

3.4 Comparison with other methods

We compare SIMCLDA with other four the-state-of-art computational methods (LRLSLDA, RWRlncD, RWRH and LDAP) in terms of AUC and number of correctly retrieved associations on the same three datasets. LRLSLDA (Chen *et al.*, 2013) used Laplacian regularized least squares, a semi-supervised learning method, to identify the possible associations between lncRNAs and diseases by incorporating lncRNA expression profiles. RWRlncD (Sun *et al.*, 2014) was a random walk with restart method for each disease to predict lncRNA–disease associations, based on constructed lncRNA–lncRNA functional similarity. RWRH (Li and Patra, 2010) was a random walk with restart on a heterogeneous network. LDAP (Lan *et al.*, 2016) fused different data sources and used a bagging SVM classifier to predict latent associations between lncRNAs and diseases. On Dataset1, we can see that SIMCLDA obtained an AUC of 0.8237, which is significantly higher than AUCs of others (LRLSLDA 0.6444, RWRH 0.6186, RWRLncD 0.6886 and LDAP 0.6283), suggesting that our method exhibits greatly improved accuracy compared to these prediction methods (Fig. 3a). Furthermore, the numbers of correctly retrieved lncRNA–disease associations are shown in Figure 3b. If a predicted association is ranked higher than the specified rank threshold, then it is regarded as a correctly retrieved association. SIMCLDA outperforms the other methods by predicting more true associations. On Dataset2, SIMCLDA also performs better with the AUC of 0.8526 than the others (LRLSLDA 0.6407, RWRH 0.6727, RWRLncD 0.6803 and LDAP 0.6315), whose results are shown in Figure 4a. As shown in Figure 4b, SIMCLDA can retrieve more correct associations. On Dataset3, SIMCLDA is much better with AUC of 0.8578 than the others (LRLSLDA 0.73, RWRH 0.7078, RWRLncD 0.6121 and LDAP 0.6393), as shown in Figure 5a. SIMCLDA can find more correct associations on the whole shown in Figure 5b. Overall, our method performs is more effective than the other existing methods.

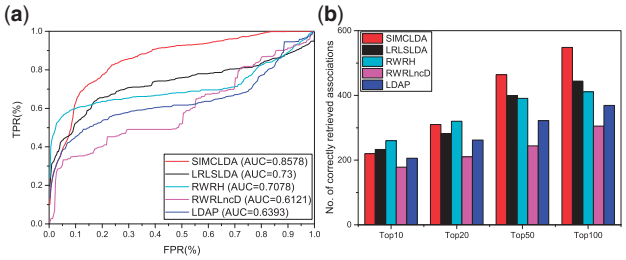

Fig. 5. Comparison of predicting methods on Dataset3. (**a**) Performance of all methods in terms of ROC curve using LOOCV. (**b**) Number of correctly retrieved known lncRNA–disease associations for specified rank thresholds

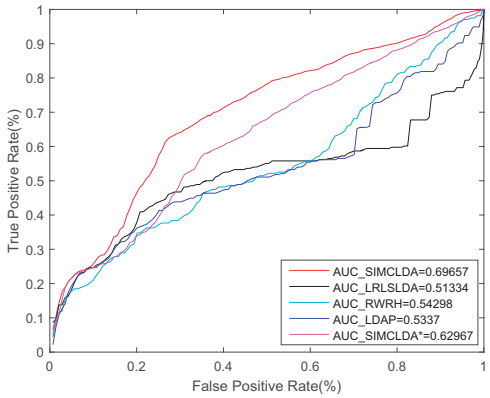

Fig. 6. Comparison of predicting methods in de novo prediction test on Dataset1

3.5 *De novo* lncRNA–disease prediction

To assess the performance of SIMCLDA in predicting potential associations for new lncRNAs, we conduct a *denovo* lncRNA–disease prediction test. After removing all known lncRNA–disease associations for each queried lncRNA *i*, computational methods are used to predict its associations. In order to evaluate the effectiveness of calculating interaction profiles, SIMCLDA* that does not calculate the interaction profile for a new lncRNA is also compared. Then, we compare SIMCLDA, LRLSLDA, RWRH, LDAP and SIMCLDA* on three datasets, except RWRLncD. Because similarity among lncRNAs is computed by lncRNA–disease associations, RWRLncD cannot be used in the*de novo* test. As shown in Figure 6, SIMCLDA achieves an AUC value of o.69657, which is much higher than those of other prediction methods. Moreover, we can find that calculating interaction profiles is an effective way for prediction. Comparison with other methods on Dataset2 and Dataset3 are reported in Supplementary Figures S6 and S7. In the *de novo* experiment, SIMCLDA derives interaction patterns for a new lncRNA based on its neighbors' interaction profiles, extracts feature vectors using PCA to reduce noise, and predicts lncRNA–disease associations using the speedup IMC model. In summary, all of the above factors contribute to the accuracy improvement in SIMCLDA.

3.6 Case studies

To demonstrate the capability of SIMCLDA in predicting new lncRNAs related to a queried disease, we conduct case studies for

Table 3. Top ten candidate lncRNAs for renal cancer

Rank	Name of lncRNA	References
1	H19	Wang *et al.* (2015a)
2	MALAT1	Hirata *et al.* (2015)
3	GAS5	Seles *et al.* (2016)
4	MEG3	Wang *et al.* (2015b)
5	XIST	Zhang *et al.* (2017b)
6	UCA1	Li *et al.* (2016)
7	DRAIC	Sakurai *et al.* (2015)
8	PCAT29	Unknown
9	NEAT1	Liu *et al.* (2017)
10	SRA1	Unknown

renal cancer, gastric cancer and prostate cancer on Dataset3, and confirm the new predicted lncRNA–disease associations in the top-10 by manually mining literatures. Case study for renal cancer is described as follows.

Renal cancer is one of the 10 most common cancers, with more than 250 000 new cases diagnosed each year worldwide (Zhou *et al.*, 2014). It is important to find the associations between progression of renal cancer and dysregulations of some lncRNAs. SIMCLDA has inferred associations between all the candidate lncRNAs for renal cancer, and finds that 8 renal cancer-associated lncRNAs (H19 1st, MALAT1 2nd, GAS5 3rd, MEG3 4th, XIST 5th, UCA1 6th, DRAIC 7th and NEAT1 9th) are in the top-10 rank of prediction (Table 3). The expression level of H19 is significantly higher in clear cell renal carcinoma compared with the normal renal tissues (Wang *et al.*, 2015a). MALAT1 expresses higher in human renal cell carcinoma, and MALAT1 silencing decreases renal cell carcinoma proliferation and invasion and increases apoptosis (Hirata *et al.*, 2015). Compared with non-tumorous renal tissue, GAS5 expression level is significantly lower in renal cell carcinoma samples *in vitro* and *in vivo* (Seles *et al.*, 2016). In renal cancer cell, MEG3 is significantly down-regulated in comparison to normal renal tissue *in vivo* and in cultured cells (Wang *et al.*, 2015b). The lncRNA XIST regulates the tumorigenicity of renal cell carcinoma cells via the miR-302c/SDC1 axis (Zhang *et al.*, 2017b). UCA1 expression levels are significantly increased in renal cell carcinoma tissues and cells (Li *et al.*, 2016). DRAIC overexpression indicates a favorable prognosis in many kinds of malignancies including renal cell carcinoma (Sakurai *et al.*, 2015). NEAT1 knockdown suppresses renal cell–carcinoma cell proliferation by inhibiting cell cycle progression, and inhibits renal cell–carcinoma cell migration and invasion by reversing the epithelial-to-mesenchymal transition phenotype (Liu *et al.*, 2017).

Case studies for gastric cancer and prostate cancer on Dataset3 are described in the Supplementary Tables S1 and S2. Furthermore, we use Dataset1 and Dataset2 to predict experimentally verified lncRNA–disease associations in Dataset3 for renal cancer, gastric cancer and prostate cancer, e.g. testing associations between renal cancer and the watched lncRNAs. The watched lncRNAs are existing in Dataset1 and Dataset2, but without relation to the watched disease. For renal cancer, six watched lncRNAs in the top-10 predicted lncRNAs are verified in Dataset3 (Supplementary Table S3), and 6 of 7 watched lncRNAs are verified in Dataset3 (Supplementary Table S4). Case studies for gastric cancer and prostate cancer on Dataset1 and Dataset2 are shown in Supplementary Tables S5–S8. These successful predictions demonstrate that SIMCLDA has the potential to infer novel lncRNAs for diseases.

4 Conclusions

Predicting lncRNA–disease associations is not only helpful in understanding critical roles of lncRNAs in biological progresses, but also beneficial to disease diagnosis, treatment, prognosis and prevention. In this study, we have proposed a computational method SIMCLDA to predict lncRNA–disease associations from known data using IMC, based on the assumption that functionally similar lncRNAs tend to interact with phenotypically similar diseases. Compared to other methods, our methods performs better in terms of AUC values on three datasets. SIMCLDA surpasses others for most conditions according to the number of correctly retrieved associations. In a *de novo* prediction test, SIMCLDA also outperforms other methods. In addition, we have conducted case studies on renal cancer and found that 80% predicted candidate lncRNAs could be confirmed for each disease of interest by literature mining. Our study has two major contributions in predicting lncRNA–disease associations. First, we can find more precise primary feature vectors from *Gkl* and *Dis* to improve accuracy. Second, we can predict lncRNA–disease associations for new lncRNAs. SIMCLDA could be a useful tool for studying lncRNA–disease relationship.

The fundamental idea of using inductive matrix completion for LncRNA–disease association prediction is to find a low-rank matrix that can integrate prior knowledge about lncRNA and disease to complete the lncRNA–disease association matrix. By factorizing a matrix to low-rank matrices (Ramlatchan *et al.*, 2018), matrix factorization provides a framework (Xi *et al.*, 2017a) for dimension reduction and matrix completion, which can also be applied to lncRNA–disease association prediction. For example, the optimization problem (10) can be modeled as

$$\min_{M,N} \|A - LMND^T\|_F + \lambda_1\|M\|_2 + +\lambda_2\|N\|_2, \qquad (20)$$

when the matrix factorization framework is applied (Jain and Dhillon, 2013). Different regularization techniques such as $L0$ regularization, $L1$ regularization and network regularization (Xi *et al.*, 2017b) may be incorporated. Various matrix factorization forms such as SVD, non-negative matrix factorization (Xi and Li, 2016) and non-negative orthogonal matrix factorization may also be adopted. Alternatively, matrix factorization can be applied to complete the association matrix of a heterogeneous network (Luo *et al.*, 2018) integrating the lncRNA-similarity network, disease-similarity network and lncRNA–disease network. Nevertheless, the inductive matrix completion method for lncRNA–disease association prediction described in this article has the flexibility to incorporate feature vectors from multiple sources. Moreover, the nuclear norm regularization adopted in (10) ensures convex optimization.

LncRNAs interact with RNA-binding proteins to regulate gene expressions (Wang *et al.*, 2008). Hence, mutations and dysfunction of lncRNAs lead to dysfunction of biological functionalities of proteins, and then result in human diseases. Computational methods have been developed to discover the interactions between lncRNAs and related proteins (Ge *et al.*, 2016). The accurate prediction of lncRNA–disease relationship by SIMCLDA can help reduce false positive in lncRNA–protein interaction and identify disease related lncRNA–protein interaction. Furthermore, combining the predication of lncRNA-protein interaction with lncRNA–disease relationship can help understand biological molecular mechanisms of diseases, and then provide prevention and treatment for human disease in the future.

Funding

This work was supported by the National Natural Science Foundation of China under Grant No. 61728211, No. 61622213, No. 61420106009, and No. 61602156.

Conflict of Interest: none declared.

References

Aken,B.L. *et al*. (2016) The ensembl gene annotation system. *Database*, **2016**, baw093. 2016.

Cai,J.-F. *et al*. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim*., **20**, 1956–1982.

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

Chen,W. *et al*. (1997) Expression of neural bc200 rna in human tumours. *J. Pathol*., **183**, 345–351.

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

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

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

Djebali,S. *et al*. (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Faghihi,M.A. *et al*. (2008) Expression of a noncoding rna is elevated in Alzheimers disease and drives rapid feed-forward regulation of β-secretase expression. *Nat. Med*., **14**, 723.

Fazel,M. (2002) *Matrix rank minimization with applications*. PhD thesis, Stanford University.

Ge,M. *et al*. (2016) A bipartite network-based method for prediction of long non-coding RNA–protein interactions. *Genomics Proteomics Bioinf*., **14**, 62–71.

Guttman,M. *et al*. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, **458**, 223–227.

Hirata,H. *et al*. (2015) Long noncoding rna malat1 promotes aggressive renal cell carcinoma through ezh2 and interacts with mir-205. *Cancer Res*., **75**, 1322–1331.

Jaccard,P. (1908) Nouvelles recheres sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat*., **44**, 223–270.

Jain,P. and Dhillon,I.S. (2013) Provable inductive matrix completion. *arXiv preprint arXiv: 1306.0626*.

Ji,H. *et al*. (2016) A rank revealing randomized singular value decomposition (r3svd) algorithm for low-rank matrix approximations. *arXiv preprint arXiv: 1605.08134*.

Jolliffe,T. *et al*. (1986) Principal Component Analysis and Factor Analysis. In: *Principal Component Analysis*. Springer, pp. 115-128.

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

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

Li,Y. *et al*. (2016) Identification of long-non coding rna uca1 as an oncogene in renal cell carcinoma. *Mol. Med. Rep*., **13**, 3326–3334.

Liu,F. *et al*. (2017) The long non-coding rna neat1 enhances epithelial-to-mesenchymal transition and chemoresistance via the mir-34a/c-met axis in renal cell carcinoma. *Oncotarget*, **8**, 62927.

Liu,J. *et al*. (2018) Applications of deep learning to mri images: a survey. *Big Data Min. Anal*., **1**, 1–18.

Liu,M.-X. *et al*. (2014) A computational framework to infer human disease-associated long noncoding RNAs. *PLoS One*, **9**, e84408.

Luo,H. *et al*. (2018) Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, **34**, 1904–1912.

Piñero,J. *et al*. (2017) Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*., **45**, D833–D839.

Ramlatchan,A. *et al*. (2018) A survey of matrix completion methods for recommendation systems. *Big Data Min. Anal*., in press.

Sakurai,K. *et al*. (2015) The lncRNA draic/pcat29 locus constitutes a tumor suppressive nexus. *Mol. Cancer Res*., **13**, 828–838.

Schlicker,A. *et al*. (2010) Improving disease gene prioritization using the semantic similarity of gene ontology terms. *Bioinformatics*, **26**, i561–i567.

Seles,M. *et al*. (2016) Current insights into long non-coding RNAs in renal cell carcinoma. *Int. J. Mol. Sci*., **17**, 573.

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

Toh,K.-C. and Yun,S. (2010) An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pac. J. Optim*., **6**, 15.

Vincent-Salomon,A. *et al*. (2007) X inactive–specific transcript RNA coating and genetic instability of the x chromosome in brca1 breast tumors. *Cancer Res*., **67**, 5134–5140.

Wang,L. *et al*. (2015a) Down-regulated long non-coding RNA h19 inhibits carcinogenesis of renal cell carcinoma. *Neoplasma*, **62**, 412–418.

Wang,M. *et al*. (2015b) Long non-coding rna meg3 induces renal cell carcinoma cells apoptosis by activating the mitochondrial pathway. *J. Huazhong Univ. Sci. Technol. [Med. Sci.]*, **35**, 541–545.

Wang,X. *et al*. (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, **454**, 126.

Wu,X. *et al*. (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol*., **4**, 1–189.

Xi,J. and Li,A. (2016) Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition. *IEEE/ACM Trans. Comput. Biol. Bioinf*., **13**, 656–668.

Xi,J. *et al*. (2017a) Discovering potential driver genes through an integrated model of somatic mutation profiles and gene functional information. *Mol. BioSystems*, **13**, 2135–2144.

Xi,J. *et al*. (2017b) A novel network regularized matrix decomposition method to detect mutated cancer genes in tumour samples with inter-patient heterogeneity. *Sci. Rep*., **7**, 2855.

Xu,M. *et al*. (2013). Speedup matrix completion with side information: application to multi-label learning. In: *Advances in Neural Information Processing Systems*, pp. 2301–2309.

Yanofsky,C. (2007) Establishing the triplet nature of the genetic code. *Cell*, **128**, 815–818.

Yao,Q. *et al*. (2017) Global prioritizing disease candidate lncrnas via a multi-level composite network. *Sci. Rep*., **7**, 39516.

Yu,H.-F. *et al*. (2014). Large-scale multi-label learning with missing labels. In: International *Conference on Machine Learning*, pp. 593–601.

Zhang,J. *et al*. (2017a) Integrating multiple heterogeneous networks for novel lncRNA–disease association inference. *IEEE/ACM Trans. comput. Biol. Bioinf*, doi:10.1109/TCBB.2017.2701379.

Zhang,J. *et al*. (2017b) The lncRNA xist regulates the tumorigenicity of renal cell carcinoma cells via the mir-302c/sdc1 axis. *Int. J. Clin. Exp. Pathol*., **10**, 7481–7491.

Zheng,X. *et al*. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1025–1033.

Zhou,S. *et al*. (2014) An emerging understanding of long noncoding RNAs in kidney cancer. *J. Cancer Res. Clin. Oncol*., **140**, 1989–1995.