Bioinformatics, 34(11), 2018, 1904–1912
doi: 10.1093/bioinformatics/bty013
Advance Access Publication Date: 22 January 2018
Original Paper

Systems biology

Computational drug repositioning using low-rank matrix approximation and randomized algorithms

Huimin Luo[1,2], Min Li[1], Shaokai Wang[1], Quan Liu[1], Yaohang Li[3,*] and Jianxin Wang[1,*]

[1]School of Information Science and Engineering, Central South University, ChangSha 410083, China, [2]School of Computer and Information Engineering, Henan University, KaiFeng 475001, China and [3]Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on May 20, 2017; revised on September 22, 2017; editorial decision on January 7, 2018; accepted on January 18, 2018

Abstract

Motivation: Computational drug repositioning is an important and efficient approach towards identifying novel treatments for diseases in drug discovery. The emergence of large-scale, heterogeneous biological and biomedical datasets has provided an unprecedented opportunity for developing computational drug repositioning methods. The drug repositioning problem can be modeled as a recommendation system that recommends novel treatments based on known drug–disease associations. The formulation under this recommendation system is matrix completion, assuming that the hidden factors contributing to drug–disease associations are highly correlated and thus the corresponding data matrix is low-rank. Under this assumption, the matrix completion algorithm fills out the unknown entries in the drug–disease matrix by constructing a low-rank matrix approximation, where new drug–disease associations having not been validated can be screened.

Results: In this work, we propose a drug repositioning recommendation system (DRRS) to predict novel drug indications by integrating related data sources and validated information of drugs and diseases. Firstly, we construct a heterogeneous drug–disease interaction network by integrating drug–drug, disease–disease and drug–disease networks. The heterogeneous network is represented by a large drug–disease adjacency matrix, whose entries include drug pairs, disease pairs, known drug–disease interaction pairs and unknown drug–disease pairs. Then, we adopt a fast Singular Value Thresholding (SVT) algorithm to complete the drug–disease adjacency matrix with predicted scores for unknown drug–disease pairs. The comprehensive experimental results show that DRRS improves the prediction accuracy compared with the other state-of-the-art approaches. In addition, case studies for several selected drugs further demonstrate the practical usefulness of the proposed method.

Availability and implementation: http://bioinformatics.csu.edu.cn/resources/softs/DrugRepositioning/DRRS/index.html

Contact: yaohang@cs.odu.edu or jxwang@mail.csu.edu.cn

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

1 Introduction

New drug discovery is a risky, time-consuming and quite expensive process (Chong and Sullivan, 2007). Though the last decades have witnessed tremendous investments in drug development, the number of new drugs approved by the US Food and Drug Administration (FDA) is unsatisfactory. The total worldwide cost of drug research and development has risen up to 141 billion USD; however, the number of drug approvals per year remains low (Schuhmacher *et al.*, 2016). In light of these challenges, drug repositioning, which concerns the identification and development of new therapeutic uses for existing drugs, has emerged as a noteworthy alternative to the traditional drug development. Currently, some repositioned drugs have been identified successfully through serendipitous or rational observations. However, these manual investigations show low efficiency for the search space having massive interaction information (for example, drug–target or drug–disease interactions) (Mullen *et al.*, 2016). To address this issue, systematic computational repositioning methods have attracted increasing attention given its efficiency and scalability.

Generally, the aim of computational drug repositioning is to find novel indications for existing drugs and apply the newly identified drugs to the treatment of diseases other than the drugs' originally intended ones (Shim and Liu, 2014). Recently, the usage of computational drug repositioning in drug discovery has become a common practice, where an increasing number of machine learning, network analysis, text mining and semantic inference methods have been proposed (Li *et al.*, 2016). For instance, Gottlieb *et al.* (2011) proposed a computational method PREDICT to identify potential drug indications by integrating various drug–drug and disease–disease similarities used to construct classification features. Based on these features, a logistic regression classifier was learned to score the novel drug–disease associations. Napolitano *et al.* (2013) combined drug related data to calculate drug similarities and trained a multi-class SVM (Support Vector Machine) classifier to predict novel alternative therapeutic indications. Moreover, some existing methods applying matrix factorization have been used to drug repositioning. Dai *et al.* (2015) proposed a matrix factorization model, which has incorporated the topological information of gene interaction network to detect novel drug indications. Based on the known drug–disease associations, the model is learned to predict novel drug indications. Yang *et al.* (2014) constructed causal networks connecting drug-target-pathway-gene-disease to compute drug–disease association scores and learned a PMF (Probabilistic matrix factorization) model based on known drug–disease associations to classify drug–disease associations. The computed association scores and association types are used to predict drug–disease associations. However, these computational methods based on matrix factorization models cannot be applied to predict novel drug indications when drug–target and disease–gene associations are unavailable.

Moreover, network-based strategy has attracted much interest due to the large-scale generation of high-throughput biological data, which has enabled the construction of complex biological interaction networks. Wang *et al.* (2014) have proposed a computational framework, TL_HGBI, to infer novel treatments for diseases based on a heterogeneous network integrating similarity and association data about diseases, drugs and drug targets. Martínez *et al.* (2015) have developed a network-based prioritization method named DrugNet to predict new therapeutic indications for drugs and novel treatments for diseases. This method identified novel drug–disease associations by propagating information in a heterogeneous network which is constructed by using information about diseases, drugs and targets. Luo *et al.* (2016) proposed a computational method to find novel indications for existing drugs by applying comprehensive similarity measures and Bi-Random Walk algorithm. According to previous studies, the computational drug repositioning method, which applies random walk on the heterogeneous network integrating various biological data, has demonstrated certain success in computational drug repositioning.

Repositioning computational drugs can also be thought of constructing a recommendation system that recommends the top-ranked diseases for given drugs. Typically, a recommendation system is a class of applications that involve predicting users responses based on their preferences. The recommendation system approach narrows what could become a complex, difficult decision to just a few recommendations, which has attracted a lot of attention in many applications. The most well-known example is how the Google search algorithm recommends the best related websites to view when a few keywords (features) are supplied. Another well-known example is the recommendation of the most likely purchase goods from Amazon based on unique customer behaviors. Recently, recommendation system technologies have been applied to drug–target interaction prediction or drug repositioning. For example, Wang *et al.* (2015) presented a recommendation system based drug repositioning approach to infer novel drug indications and side effects simultaneously. Inspired by the recent success of the recommendation system approach, we hereby design a drug–disease recommendation system to predict the most likely indications for given drugs.

From mathematical point of view, the process of random walk on the heterogeneous network is equivalent to that of approximating the eigenvector associated with the largest eigenvalue of its transition matrix. Nevertheless, unless the largest eigenvalue is absolutely dominating, the other dominant eigenvectors also play non-negligible roles. The formulation of drug–disease recommendation system is based on matrix completion, which is designed to fill out the unknown entries in the association matrix according to its eigenspace with respect to all dominant eigenvalues. Assuming that the underlying factors contributing to drug–disease associations are highly correlated and thus the number of underlying independent factors is much smaller than the existing number of diseases or drugs, the fundamental idea of matrix completion is to construct a low-rank matrix to approximate the drug–disease association matrix. In fact, matrix completion methods have recently started to attract interest in bioinformatics and biomedical applications (Kapur *et al.*, 2016; Natarajan and Dhillon, 2014).

In this work, we construct a heterogeneous network by integrating individual networks based on drug similarities, disease similarities and verified drug–disease associations. Then, we design a drug repositioning recommendation system (DRRS) based on the Singular Value Thresholding (SVT) algorithm (Cai *et al.*, 2010) to complete the association matrix of the heterogeneous network. A recycling rank-revealing randomized singular value decomposition algorithm (R^4SVD) (Li and Yu, 2017) is employed to fast and adaptively approximate the dominant singular values and their associated singular vectors so that the recommendation system is scalable to handle large adjacency matrices generated from large-scale drug–disease networks. DRRS is compared with several state-of-the-art repositioning methods with respect to prediction performance. The experiment results demonstrate the effectiveness of DRRS on discovering novel drug indications, including the drugs without previously known associations. In case studies, the top-ranked diseases for several selected drugs are examined. Many top-ranked, novel drug–disease associations are strongly supported by the public databases, which further confirm the effectiveness of

our approach. The main contribution of this paper involves: (i) our proposed method performed matrix factorization on a matrix generated from the heterogeneous network containing drug similarity, disease similarity and drug–disease association information. The advantage of this approach is justified in theory by comparing with the popularly used random walk methods. The effectiveness of this approach has also been shown in our results on different datasets; (ii) we propose an automatic scheme that can be used to determine the appropriate rank for the complete matrix.

2 Materials and methods

In this study, we propose a novel drug repositioning recommendation system approach DRRS to infer potential drug indications. First, we give a brief description of the used datasets and construct a heterogeneous network by integrating multi-source data. Then, based on the known drug similarity, disease similarity and drug–disease association data, matrix completion algorithm is utilized to recover the missing associations in heterogeneous network.

2.1 Datasets

The gold standard datasets used for inferring novel drug indications is obtained from Gottlieb *et al.* (2011), and is collected from multiple data sources. This dataset includes 593 drugs, 313 diseases and 1933 validated drug–disease associations totally. Drugs are extracted from DrugBank database (Wishart *et al.*, 2006) which is a comprehensive database containing extensive information about drugs and their targets. Diseases are collected from human phenotypes defined in the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2002) which is a public resource providing information about human genes and diseases. The numbers of drugs, diseases, targets and all the interactions used in this study are shown in Table 1.

Here, the similarities between drugs are calculated by the Chemical Development Kit (CDK) (Steinbeck *et al.*, 2003) based on SMILES (Weininger, 1988) chemical structures and the pairwise drug similarity is represented as the Tanimoto score of their 2D chemical fingerprints. The similarities between diseases are obtained from MimMiner (Van Driel *et al.*, 2006), which measures the degree of pairwise disease similarity by the text mining analysis of their medical descriptions information in the OMIM database.

2.2 Construction of the heterogeneous network

Currently, most matrix completion methods used in prediction applications conduct prediction by simply completing the known association matrix. Previous studies have shown that the integration of heterogeneous, multi-source data can successfully lead to prediction accuracy improvements. Hence, we integrate drug similarity, disease similarity and drug–disease association information to construct a heterogeneous network by incorporating heterogeneous data and then predict novel drug indications by completing its adjacency matrix.

Correspondingly, we construct a heterogeneous network composed of three sub-networks, namely drug–drug network, disease–disease network and drug–disease network. For drug–drug network, let $R = \{dr_1, dr_2, \ldots, dr_m\}$ denotes m drugs and each edge connecting two drugs is weighted by the pairwise chemical structures similarity value. Similarly, for disease–disease network, let $D = \{ds_1, ds_2, \ldots, ds_n\}$ denotes n diseases and each edge connecting two diseases is weighted by the pairwise phenotype similarity value. The drug–disease network is modeled as a bipartite graph $G(R, D, E)$, where $E(G) \subseteq R \times D$, $E(G) = \{e_{ij}\}$ contains edges between drug dr_i and disease ds_j. The weight of e_{ij} is initially set to 1 if there exist a known association between drug dr_i and disease ds_j, otherwise 0.

Finally, the heterogeneous network is constructed by connecting drug–drug network and disease–disease network via drug–disease interaction network, as shown in Figure 1. The adjacency matrix $\mathbf{A}$ of the heterogeneous network is defined by (1).

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{RR} & \mathbf{A}_{RD} \\ \mathbf{A}_{RD}^{T} & \mathbf{A}_{DD} \end{bmatrix}. \tag{1}$$

In matrix $\mathbf{A}$, the diagonal submatrices $\mathbf{A}_{RR}$ and $\mathbf{A}_{DD}$ are the adjacency matrices of drug network and disease network, respectively. Both $\mathbf{A}_{RR}$ and $\mathbf{A}_{DD}$ are dense. The off-diagonal submatrix $\mathbf{A}_{RD}$ is the adjacency matrix of drug–disease network and $\mathbf{A}_{DR} = \mathbf{A}_{RD}^{T}$, where $\mathbf{A}_{RD}^{T}$ denotes the transpose of $\mathbf{A}_{RD}$. Due to the fact that the connectivities in each individual biological networks are bidirectional and their weights are positive, the adjacency matrix $\mathbf{A}$ of the heterogeneous network is symmetric and semi-positive definite. Hence, the eigenvalues of the adjacency matrix are real, positive and are equal to the singular values. Moreover, the left singular vectors of $\mathbf{A}$ are the same as the right singular vectors, which are also equal to $\mathbf{A}$'s eigenvectors. The unknown entries are only presented in the off-diagonal submatrices $\mathbf{A}_{RD}$ and $\mathbf{A}_{DR}$, representing the unknown associations to be predicted. After all, the goal of the drug–disease association prediction problem is cast as filling out the missing entries in the adjacency matrix $\mathbf{A}$.

2.3 Prediction using low-rank matrix completion

Based on the premise that similar drugs tend to treat similar disease, the hidden factors that govern the drug–disease associations are highly correlated, which results in an also highly correlated data matrix. Our drug–disease recommendation system model is based on constructing an r-rank matrix $\mathbf{A}^*$ to approximate the $(m+n) \times (m+n)$ adjacency matrix $\mathbf{A}$ of the drug–disease heterogeneous network described in Section 2.2, where $r \ll m+n$. We denote Ω as a set of indices of all known elements in $\mathbf{A}$. Clearly, Ω contains the indices of all elements in $\mathbf{A}_{RR}$ and $\mathbf{A}_{DD}$, including 0 s, as well as the known associations in $\mathbf{A}_{RD}$ and $\mathbf{A}_{DR}$. The construction of $\mathbf{A}^*$ tries to minimize the rank of $\mathbf{A}^*$, i.e.

$$\begin{aligned} &\min\ rank(\mathbf{A}^*) \\ &s.t. P_{\Omega}(\mathbf{A}^*) = P_{\Omega}(\mathbf{A}). \end{aligned} \tag{2}$$

It is important to notice that if $\mathbf{A}$ is composed by only $\mathbf{A}_{RD}$ instead of the one from the heterogeneous network, matrix completion by rank minimization will not lead to meaningful results. This is due to the fact that all of the known drug–disease association samples are positive in drug-repositioning. Filling out $\mathbf{A}_{RD}$ will lead to an optimal solution for the rank minimization problem, i.e. a matrix of all one's with rank 1.

Unfortunately, this rank minimization problem is known to be NP-hard (Natarajan, 1995) and thus it is impractical for drug–disease association prediction problems involving large number of diseases and drugs. Here we adopt a relaxation form proposed by Candès and Recht (2009) via minimizing the sum of the singular values of $\mathbf{A}^*$, which is known as the nuclear norm of $\mathbf{A}^*$. Correspondingly, the matrix completion problem is remodeled as a proximal optimization problem (Cai *et al.*, 2010) such as

Table 1. Statistics of the gold standard dataset used in this study

Dataset	Drugs	Diseases	Interactions	Sparsity[a]
	593	313	1 933	1.041^{-2}

[a]The sparsity is defined as the ratio of the number of known interactions to the number of all possible interactions.

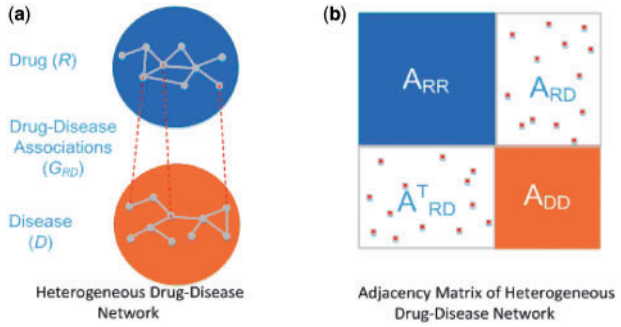

Fig. 1. The heterogeneous drug–disease network and its adjacency matrix

$$\begin{aligned} &\min \tau ||\mathbf{A}^*||_* + \frac{1}{2}||\mathbf{A}^*||_F^2 \\ &s.t. P_\Omega(\mathbf{A}^*) = P_\Omega(\mathbf{A}), \end{aligned} \tag{3}$$

where $\|.\|_*$ denotes the nuclear norm and τ is a singular value threshold parameter. The solution of the proximal problem is a good approximation to that of the nuclear norm minimization when τ is large. Nevertheless, if τ is too large, convergence of the optimization process will become very slow. In DRRS, we set $\tau = \|P_\Omega(\mathbf{A})\|_F(m+n)/\sqrt{|\Omega|}$ to balance the approximation accuracy and convergence speed. Candès and Recht (2009) have also showed that under certain conditions, the solution obtained by optimizing the nuclear norm is equivalent to the one by rank minimization. The matrix completion program by optimizing the nuclear norm can be addressed by using the singular value thresholding (SVT) algorithm (Cai *et al.*, 2010).

Starting from $\mathbf{Y}^{(0)} = \lceil \tau/(\delta||P_\Omega(\mathbf{A})||)\rceil \delta P_\Omega(\mathbf{A})$, SVT reformulates the Uzawa's algorithm (Arrow *et al.*, 1958) or linearized Bregman iteration (Yin *et al.*, 2008) by generating a series of matrices $(\mathbf{X}^{(i+1)}, \mathbf{Y}^{(i+1)})$ via

$$\begin{cases} \mathbf{X}^{(i+1)} = D_\tau(\mathbf{Y}^{(i)}) \\ \mathbf{Y}^{(i+1)} = \mathbf{Y}^{(i)} + \delta P_\Omega(\mathbf{A} - \mathbf{X}^{(i+1)}), \end{cases} \tag{4}$$

where δ is the iteration step size set to $(m+n)/\sqrt{|\Omega|}$ as suggested by Li and Yu (2017) and the SVT operator $D_\tau(\cdot)$ is a soft thresholding operator such that

$$D_\tau(\mathbf{Y}^{(i)}) = \sum_{j=1}^{\sigma_j^{(i)} \geq \tau} (\sigma_j^{(i)} - \tau) u_j^{(i)} v_j^{(i)T}, \tag{5}$$

where σ_j's include singular values that are larger than τ, while u_j and v_j are the corresponding left and right singular vectors, respectively.

During the SVT process of matrix completion, estimating the singular values of $\mathbf{Y}^{(i)}$ greater than τ to compute $D_\tau(\cdot)$ is required at each iteration step. This can be straightforwardly obtained by computing full Singular Value Decompostion (SVD) on $\mathbf{Y}^{(i)}$ and then shrinking it by selecting the singular values greater than τ and their corresponding singular vectors. However, numerically computing the full SVD of an adjacency matrix from a large heterogeneous network is often computationally costly as well as memory intensive. In fact, during SVT iterations, $D_\tau(\cdot)$ only concerns the singular values in $\mathbf{Y}^{(i)}$ that are greater than τ. This allows us to apply fast SVD algorithms that focus on approximating the dominant singular values of interest to enhance the computational efficiency of the matrix completion algorithm. A rank-revealing randomized SVD algorithm (R^3SVD) (Ji *et al.*, 2016) has been proposed to fast approximate the SVT operator by projecting $\mathbf{Y}^{(i)}$ onto a small Gaussian matrix and applying power iterations. R^3SVD builds up a low-rank QB decomposition incrementally based on orthogonal Gaussian projection and then derives the low-rank SVD. Extending R^3SVD to a recycling rank-revealing randomized SVD (R^4SVD) algorithm (Li and Yu, 2017) by taking advantage of the singular vectors obtained from the previous iterations further improves the computational efficiency of the SVT process. Here, R^4SVD algorithm is incorporated in DRRS for fast computation of $D_\tau(\cdot)$. The fast implementation of SVT algorithm using R^4SVD, so-called SVT-R^4SVD, performs matrix completion in our DRRS method.

2.4 Determining the optimal rank

Our recommendation system for drug repositioning includes two phases. The first phase is to determine the appropriate rank value r in the completed matrix that yields the optimal prediction performance. Rank r of the completed matrix is a critical parameter in matrix completion—an underestimated r will lower prediction accuracy while an overestimated r may lead to overfitting as well as unnecessary computational costs. However, the appropriate value of r is unknown beforehand in the drug repositioning problem. Here, we use an approach based on validation to determine the appropriate value of r. First of all, we randomly designate 10% of the known drug–disease associations as a validation set. Then, we monitor the Area Under Curve (AUC, see Section 3.1) value for the validation set during the SVT-R^4SVD process where r is gradually increasing in general. The r value is gradually increasing, although not always strictly, because SVT is a convex optimization method—during the SVT process, the approximation error $||P_\Omega(\mathbf{A}^*) - P_\Omega(\mathbf{A})||$ is decreasing gradually and therefore the rank r of the completed matrix $\mathbf{A}^*$ generally has to increase accordingly to satisfy the approximation error. The rank r with respect to the best AUC value is regarded as the optimal rank *bestr* and will be selected as the target rank for the next phase. In the second phase, we rerun the SVT-R^4SVD program on the whole known drug–disease association dataset until the target rank is reached. The corresponding completed drug–disease association matrix $\mathbf{A}^*_{RD}$ is then retrieved and the recommendations are made by sorting the predicted entry values.

Our two-phase recommendation system is illustrated in Algorithm 1. Function SVT-R^4SVD(·) carries out one SVT iteration, whose implementation details can be found in (Li and Yu, 2017). Function AUC(·) computes the AUC value for the validation set generated in the first phase. The drug–disease association information is considered to be more important in prediction, and therefore similarity matrix is multiplied by a weight coefficient less than 1. We set the coefficient to 0.8 in this study.

2.5 Comparison with random walk algorithms

The random walk method (Berger *et al.*, 2010; Köhler *et al.*, 2008; Li and Patra, 2010) has been popularly used as a prediction model for inferencing and ranking associations among the biological networks. It simulates a random walker starting from a set of randomly selected seed nodes and then calculates the ranking scores for all the nodes as the stationary distribution, representing the probability of being reached by the random walker when equilibrium is reached. A random walk iteration is typically described as

$$p_i = (1-\gamma)\mathbf{P}^T p_{i-1} + \gamma p_0, \quad (6)$$

where p_0 is the initial probability vector, γ is the random walk restart probability, and $\mathbf{P}$ is the transition matrix transformed from the affinity matrix of the heterogeneous network.

The mathematical foundation of the random walk model is the power iteration, i.e. applying the high power of the transition matrix $\mathbf{P}$ on the initial vector. It is well known that the dominant eigenvalue of the transition matrix λ_1 is 1 while the magnitudes of the rest eigenvalues are less than 1. During power iterations, the high power of the transition matrix allows the dominant eigenvalue to remain 1 while the rest decade to 0. Eventually, the stationary distribution vector at equilibrium is equivalent to the dominant eigenvector of the transition matrix. More precisely, the random walk model can be thought as completing a rank-1 matrix while the unknown associations are predicted to fit the top dominant eigenvector of the transition matrix. In contrast, the matrix completion model takes all dominant eigenvalues into account and is cast to fit the unknown associations with respect to all eigenvectors corresponding to the dominant eigenvalues. Hence, in theory, the matrix completion model should yield better accuracy than random walk but with the tradeoff of more computational time. Nevertheless, in drug repositioning, the prediction accuracy is of most importance while the computational cost problem can be addressed by advance of algorithms and computer architectures.

Algorithm 1: Completing Matrix using DRRS

Input: drug similarity matrix $\mathbf{A}_{\mathrm{RR}}$ and its indices set Ω_{RR}, disease similarity matrix $\mathbf{A}_{\mathrm{DD}}$ and its indices set Ω_{RR}, drug-disease association matrix $\mathbf{A}_{\mathrm{RD}}$ and its indices set Ω_{RD}.

Output: Completed drug-disease association matrix $\mathbf{A}^*_{\mathrm{RD}}$.

`/* Phase I: determine optimal rank */`

randomly designate 10% of indices of Ω_{RD} as validation set $\Omega^{va}_{\mathrm{RD}}$ such that $\mathbf{A}_{\mathrm{RD}} = \mathbf{A}^{'}_{\mathrm{RD}} + \mathbf{A}^{va}_{\mathrm{RD}}$ and $\Omega_{\mathrm{RD}} = \Omega^{'}_{\mathrm{RD}} \cup \Omega^{va}_{\mathrm{RD}}$;

$\mathbf{A}^* \leftarrow \begin{bmatrix} \mathbf{A}_{\mathrm{RR}} & \mathbf{A}^{'}_{\mathrm{RD}} \\ \mathbf{A}^{'}_{\mathrm{RD}}{}^T & \mathbf{A}_{\mathrm{DD}} \end{bmatrix}$;

$\Omega \leftarrow \Omega_{\mathrm{RR}} \cup \Omega_{\mathrm{DD}} \cup \Omega^{'}_{\mathrm{RD}} \cup \Omega^{'}_{\mathrm{DR}}$;

$\tau \leftarrow \|P_\Omega(\mathbf{A}^*)\|_F (m+n)/\sqrt{|\Omega|}$;

$\delta \leftarrow (m+n)/\sqrt{|\Omega|}$;

$bestr \leftarrow 0; r \leftarrow 0; maxauc \leftarrow 0$;

while *True* **do**

  $\left(\begin{bmatrix} \mathbf{A}^*_{\mathrm{RR}} & \mathbf{A}^*_{\mathrm{RD}} \\ \mathbf{A}^*_{\mathrm{RD}}{}^T & \mathbf{A}^*_{\mathrm{DD}} \end{bmatrix}, r\right) \leftarrow$ SVT-R^{4}SVD($\mathbf{A}^*, \Omega, \tau, \delta$);

  $auc \leftarrow$ AUC($\mathbf{A}^*_{\mathrm{RD}}, \mathbf{A}^{va}_{\mathrm{RD}}, \Omega^{va}_{\mathrm{RD}}$);

  if $auc > maxauc$ **then**

    $maxauc \leftarrow auc; bestr \leftarrow r;$ `/* optimal rank */`

  end

  if $r \geq \min(m,n)$ **then** break;

  $\mathbf{A}^* \leftarrow \begin{bmatrix} \mathbf{A}^*_{\mathrm{RR}} & \mathbf{A}^*_{\mathrm{RD}} \\ \mathbf{A}^*_{\mathrm{RD}}{}^T & \mathbf{A}^*_{\mathrm{DD}} \end{bmatrix}$;

end

`/* Phase II: matrix completion */`

$\mathbf{A}^* \leftarrow \begin{bmatrix} \mathbf{A}_{\mathrm{RR}} & \mathbf{A}_{\mathrm{RD}} \\ \mathbf{A}_{\mathrm{RD}}{}^T & \mathbf{A}_{\mathrm{DD}} \end{bmatrix}$;

$\Omega \leftarrow \Omega_{\mathrm{RR}} \cup \Omega_{\mathrm{DD}} \cup \Omega_{\mathrm{RD}} \cup \Omega_{\mathrm{DR}}$;

$\tau \leftarrow \|P_\Omega(\mathbf{A}^*)\|_F (m+n)/\sqrt{|\Omega|}$;

$\delta \leftarrow (m+n)/\sqrt{|\Omega|}$;

while *True* **do**

  $\left(\begin{bmatrix} \mathbf{A}^*_{\mathrm{RR}} & \mathbf{A}^*_{\mathrm{RD}} \\ \mathbf{A}^*_{\mathrm{RD}}{}^T & \mathbf{A}^*_{\mathrm{DD}} \end{bmatrix}, r\right) \leftarrow$ SVT-R^{4}SVD($\mathbf{A}^*, \Omega, \tau, \delta$);

  `/* A*: input parameter of SVT-R4SVD */`

  if $r \geq bestr$ **then** break;

  $\mathbf{A}^* \leftarrow \begin{bmatrix} \mathbf{A}^*_{\mathrm{RR}} & \mathbf{A}^*_{\mathrm{RD}} \\ \mathbf{A}^*_{\mathrm{RD}}{}^T & \mathbf{A}^*_{\mathrm{DD}} \end{bmatrix}$;

end

return $\mathbf{A}^*_{\mathrm{RD}}$;

3 Experiments and results

In this section, we systematically evaluate the performance of DRRS using the golden standard datasets. First, the evaluation metrics used in this study are introduced. Then, we compare DRRS with several state-of-the-art algorithms in terms of prioritizing candidate diseases for a given drug of interest. Next, case studies are conducted to further illustrate the practical usefulness of DRRS. Finally, we perform prediction on the other two collected datasets to further verify the effectiveness of DRRS.

3.1 Evaluation metrics

To systematically evaluate the ability of DRRS in identifying candidate diseases for a specific drug, ten-fold cross-validation experiments are conducted. In the golden datasets, there are 1933 known drug–disease associations and the others having not been verified are considered as candidate associations. All known associations are randomly divided into ten partitions that are roughly equal in size. Each partition is taken in turn as the test set, while the remaining nine partitions serve as the training set.

After performing matrix completion based on the training set, for each drug, the test associations are ranked together with the candidate associations and are sorted in descending order according to their predicted values assigned by the matrix completion method. For each specific ranking threshold, true positive (TP), false negative (FN), false positive (FP) and true negative (TN) are calculated based on the ranking results. A test association is considered as a correctly identified positive sample if it has higher rank than the given threshold. A candidate association is considered as a correctly identified negative sample if it has lower rank than the given threshold. Here, TP and TN represent the number of positive samples and negative samples identified correctly, respectively. FP and FN denote the number of positive samples and negative samples identified incorrectly, respectively. By varying the rank threshold, True Positive Rate (TPR), False Positive Rate (FPR) and Precision can be calculated to construct the Receiver Operating Characteristic (ROC) curve and the Precision-Recall curve. For the ROC curve, FPR and TPR are plotted on the x- and y-axes, respectively. For PR curve, Recall is plotted on the x-axis and Precision is plotted on the y-axis (Davis and Goadrich, 2006). The area under ROC curve (AUC) value and precision are utilized to evaluate the overall performance of the prediction methods. To obtain convincing results, ten-fold cross validation is repeated ten times and the average value is reported as the final result. Strictly, the ROC curve and PR are not measuring exactly the precisions or recalls of the predictions, instead of ranking the known associations on top of the unknowns. Nevertheless, due to the fact that the true associations in reality is scarce compared to the total number of unknowns, measuring the properties of the ROC and PR while treating the unknown as the true negatives is still meaningful.

Moreover, comprehensive prediction experiments using all known associations as training set are conducted to evaluate DRRS. Here, each unknown drug–disease association is assigned a predicted score according to the completed matrix by DRRS. Then,

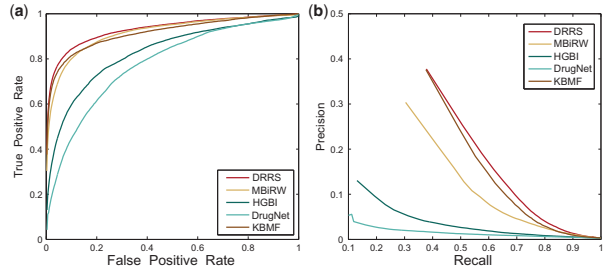

Fig. 2. Prediction results of different methods in identifying potential diseases for drugs. (**a**) ROC curves of prediction results obtained by applying DRRS and other competitive methods. (**b**) PR curves of identifying candidate diseases for drugs

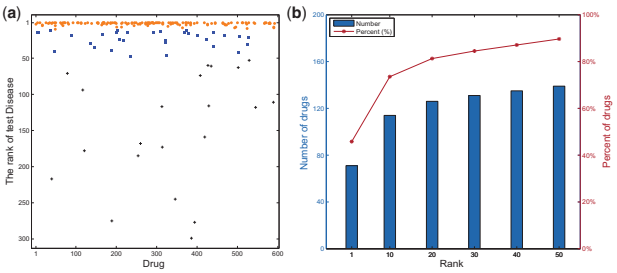

Fig. 3. (**a**) Prediction results in one fold of ten-fold cross-validation. The rank results of test associations based on their predicted scores by DRRS. 141 associations (orange) ranked in top 10, 32 (blue) ranked in top 10–50 and 21 (black) ranked over 50. (**b**) The number and the percentage of drugs correctly predicted with respect to different rank cutoffs (Color version of this figure is available at *Bioinformatics* online.)

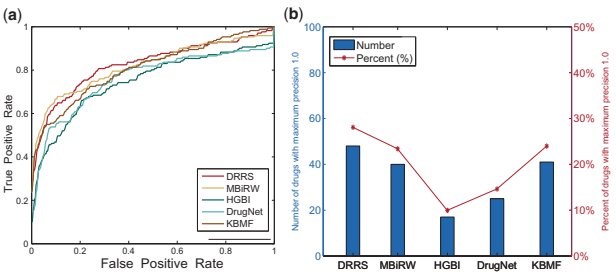

Fig. 4. (**a**) ROC curves of prediction results obtained by applying DRRS and other methods. (**b**) The number and the percentage of drugs with maximum precision value 1.0 predicted by all methods

selecting several drugs as the examples, we analyze the top-ranked candidate diseases for each selected drug by searching evidences from public databases.

3.2 Comparison with other methods

To assess the performance of DRRS, we compare it with four state-of-the-art methods: MBiRW (Luo *et al.*, 2016), DrugNet (Martínez *et al.*, 2015), HGBI (Wang *et al.*, 2013) and KBMF (Gönen *et al.*, 2013). MBiRW utilizes comprehensive similarity measures and Bi-Random Walk algorithm to identify potential novel indications for a given drug. DrugNet is a generic network-based drug repositioning method, which propagates information between networks and can be utilized to perform both drug–disease and disease-drug prioritization on drug–disease network or drug–target -disease network. HGBI is introduced based on the guilt-by-association principle and an intuitive interpretation of information flow on the heterogeneous graph. Although HGBI is originally developed for drug–target association prediction, it can be applied in predicting candidate diseases for drugs. The parameters used in MBiRW, HGBI and DrugNet are determined according to their literatures. KBMF is a kernelized Bayesian matrix factorization method, which can work with multiple data side information sources and can be applied in recommendation systems, and the subspace dimensionality parameter R used in KBMF is set to 40, which performs best in the cross validation test.

The overall performance of all methods is evaluated by applying ten-fold cross-validation specified in Section 3.1. The experiment results in terms of ROC curves and PR curves are depicted in Figure 2.

As shown in the experiment results, our proposed DRRS method outperforms the other competitive methods in terms of AUC and best precision values. More Specifically, DRRS achieves AUC value of 0.93, while MBiRW, HGBI, DrugNet and KBMF obtain inferior results of 0.917, 0.829, 0.778 and 0.915, respectively. The PR curves show that DRRS achieves the best precision with 0.378, indicating that it can successfully prioritize 37.8% true drug–disease associations as the ones with the highest rank. KBMF has similar performance with the best precision 0.376.

In one fold of the ten-fold cross validation, the prediction results of test set are analyzed. There are 194 drug–disease associations in the test set, and these associations involve 155 drugs. After matrix completion, for each drug, all of its test associations and candidate associations are ranked in descending order, whose results are shown in Figure 3. DRRS has prioritized 72.7% test associations in top-10 rankings. Furthermore, the test associations of 71 (71/155 ≈ 45.8%) drugs have been ranked at top 1 by DRRS.

Moreover, the running times of all methods on the golden datasets in one ten-fold cross validation run are compared in Supplementary Table S1. The results show that DRRS is faster than matrix completion method KBMF and takes less than 3 min.

3.3 Predicting indications for new drugs

DRRS can also be used for drugs without previously known disease associations. We analyze the performance of all methods for drugs which have only one known disease association in the golden datasets. In this case, for one given drug, the known associated disease is removed from the datasets, and the drug will have no association information during the process of prediction. Therefore, the test on these drugs is used to evaluate the ability of the method to predict associations for new drugs that have no known association with any disease.

There are 171 drugs which have only one known associated disease in the golden datasets. After prediction, the results in terms of ROC curves as well as the measures of the number and the percentage of drugs with maximum precision value 1.0 are reported in Figure 4a and b, respectively. The maximum precision value 1.0 means that the test disease is ranked successfully as number one out of all candidate diseases associated with the specific drug. One can find that DRRS has achieved superior performance over the other methods. For example, DRRS achieves AUC value of 0.824, while MBiRW, HGBI, DrugNet and KBMF obtain inferior AUC values of 0.818, 0.746, 0.759 and 0.806, respectively. Moreover, 48 (48/171 ≈ 28.07%) drugs are predicted with maximum precision value 1.0 in DRRS. In comparison, for MBiRW, HGBI, DrugNet and KBMF, there are 40 (23.39%), 17 (9.94%), 25 (14.62%) and 41 (23.98%) drugs predicted with maximum precision value 1.0, respectively.

Table 2. The number of verified novel drug–disease associations as top-5 and top-20 for the 593 drugs

	DRRS	MBiRW	DrugNet	HGBI	KBMF
Top 5	155	153	23	76	145
Top 20	397	364	203	231	354

3.4 Comprehensive prediction for novel drug–disease associations

After confirming the prediction ability of DRRS by cross validation experiments, we conducted a comprehensive prediction of novel associations between all drugs and diseases. In the inference process, all known drug–disease associations in the gold standard dataset are used as the training set and the remaining drug–disease pairs are regarded as the set of candidate drug–disease associations. DRRS can predict the potential disease associations for all drugs simultaneously. By applying DRRS, all candidate diseases for a specific drug are ranked according to their predicted values assigned by DRRS.

We have conducted case studies to verify whether the predicted top-ranked diseases are true or not according to two public biological databases: KEGG (Kanehisa *et al.*, 2014) and CTD (Davis *et al.*, 2013). KEGG and CTD have been constantly updated to include newly verified drug–disease associations and provide a foundation for our validation. We examined the most potential indications for each of the 593 drugs. The predicted results by all methods are summarized in Table 2. One can observe that 155 of top-5 and 397 of top-20 predicted novel drug–disease associations by DRRS have been annotated in KEGG and CTD, respectively, which are more than the other prediction methods.

We choose several drugs as examples and list the verified information of the top-5 candidate diseases for each selected drug in Supplementary Tables S2–S5. We find several novel drug–disease associations of the top-ranked predictions that have been annotated in KEGG or CTD database. Zoledronic acid has been predicted to treat for Breast cancer, Osteoporosis, RCC (Renal cell carcinoma, nonpapillary) and Prostate cancer, as confirmed in public database. Furthermore, DRRS predicts other novel treatments including: Risperidone for OCD (Obsessive-compulsive disorder) and PAND1 (Panic disorder 1); Prednisolone for Autoimmune Disease, Asthma; Paclitaxel for Prostate Cancer and Multiple Myeloma. These novel treatments have also been confirmed in CTD or KEGG database.

As a result, the confirmations of top-ranked predictions in CTD and KEGG databases support the practical application of DRRS on discovering novel indications for drugs. More importantly, the other top-ranked predictions that are not yet reported may also exist and deserve further scientific exploration by related experiments.

3.5 Validation on the other datasets

The robustness of DRRS is further validated to perform prediction on two other datasets: Cdataset and DNdataset, which have been used in our previous research (Luo *et al.*, 2016). Cdataset includes 663 drugs registered in DrugBank, 409 diseases listed in OMIM database, and 2353 verified drug–disease associations. DNdataset contains 4516 diseases annotated by Disease Ontology (DO) terms, 1490 drugs registered in DrugBank and 1008 known drug–disease associations derived from DrugBank. The numbers of drugs, diseases and interactions included in the two datasets are shown in Table 3.

Firstly, we conduct ten times ten-fold cross-validation to validate the prediction accuracy of our proposed method on Cdataset and DNdataset. The results of applying different methods in terms of ROC curves and PR curves are depicted in Figures 5 and 6, respectively. DRRS yields the best prediction accuracy in comparison with the other methods on Cdataset. DRRS achieves an AUC value of 0.947 while MBiRW, HGBI, DrugNet and KBMF obtain inferior results with 0.933, 0.858, 0.804 and 0.928, respectively. Moreover, the maximum precision achieved by DRRS is 0.402, which is similar to that of KBMF and higher than other methods. On DNdatasets, the AUC value obtained by DRRS is 0.935, which is lower than that obtained by MBiRW and DrugNet. This may be due to the fact that the known drug–disease associations in DNdataset is much sparser than gold standard dataset and Cdataset. However, in terms of precision measure of the PR curves that is more important in practice, DRRS obtains the maximum precision of 0.348, which is higher than that obtained by the second best method MBiRW (0.321). In summary, DRRS demonstrates high prediction accuracy on different datasets. Moreover, the running times of all methods on Cdatasets and DNdataset are recorded in Supplementary Table S1, which shows DRRS is faster than KBMF, particularly for handling large datasets. It should be noted that, in one ten-fold cross validation, KBMF on DNdatasets is so slow to perform ten times ten-fold cross-validation in acceptable time, and then the explicit results on DNdatasets have not been obtained and reported in our study.

Table 3. Statistics of Cdataset and DNdataset

	Drugs	Diseases	Interactions	Sparsity
Cdataset	409	663	2532	9.337^{-3}
DNdataset	1490	4516	1008	1.498^{-4}

Moreover, the experiments of predicting indications for new drugs are also conducted on the two datasets, and results are reported in Figures 7 and 8, respectively. There are 177 drugs and 347 drugs with only one associated disease in Cdatasets and DNdatasets, respectively. The test on these drugs is used to evaluate the performance of all methods in predicting candidate indications for new drugs. From the experiment results, we can clearly see that DRRS outperforms other competing methods. On Cdatasets, DRRS achieves AUC value of 0.824, while MBiRW, HGBI, DrugNet and KBMF obtain inferior results with 0.804, 0.732, 0.754 and 0.798, respectively. In addition, 45 (45/177 $\approx$ 25.42%) drugs have achieved maximum precision value 1.0. For MBiRW, HGBI, DrugNet and KBMF, there are 41 (23.16%), 19(10.73%), 23(12.99%) and 33(18.64%) drugs with maximum precision value 1.0, respectively. On DNdatasets, DRRS achieves AUC value of 0.943, which is slightly worse than MBiRW and DrugNet. The number of drugs with maximum precision value 1.0 is 139, which is equal to that of applying MBiRW. For HGBI and DrugNet, there are 80 (23.05%) and 84 (24.21%) drugs with maximum precision value 1.0, respectively. All these results further demonstrate the effectiveness of our proposed method in predicting indications for new drugs.

4 Conclusion

In this study, we propose using recommendation system approach to address the problem of drug repositioning. A novel computational method for drug repositioning, so-called DRRS, is developed to identify novel disease indications for given drugs. In DRRS, a heterogeneous drug–disease network is constructed by integrating drug–drug network, disease–disease network and drug–disease association network. The proposed method formulates the prediction of

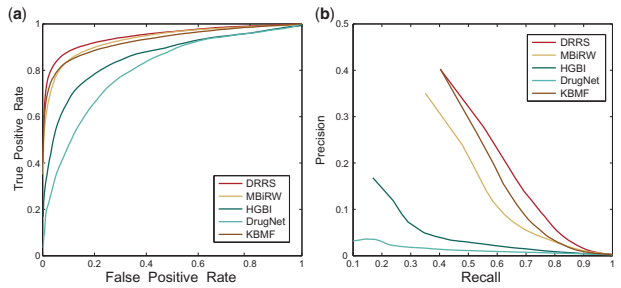

Fig. 5. Prediction results of different methods on Cdataset. (**a**) ROC curves of prediction results obtained by applying DRRS and other competitive methods. (**b**) PR curves of identifying candidate diseases for drugs

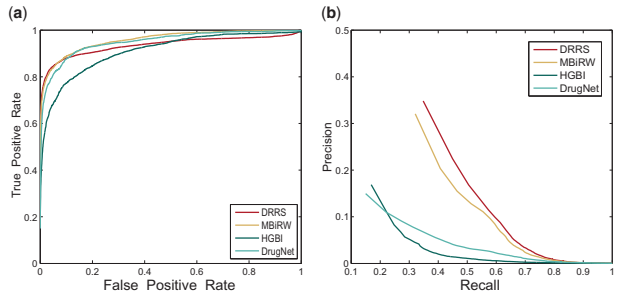

Fig. 6. Prediction results of different methods on DNdataset. (**a**) ROC curves of prediction results obtained by applying DRRS and other competitive methods. (**b**) PR curves of identifying candidate diseases for drugs

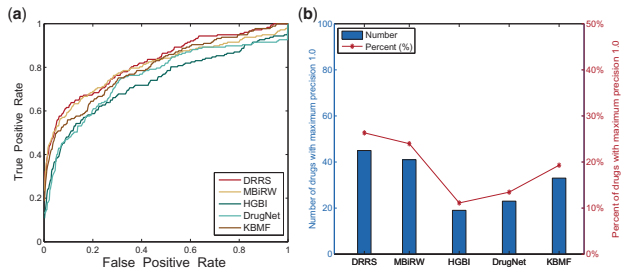

Fig. 7. (**a**) ROC curves of prediction results obtained by applying DRRS and other competitive methods on Cdatasets. (**b**) The number and percent of drugs with maximum precision value 1.0 obtained by all methods

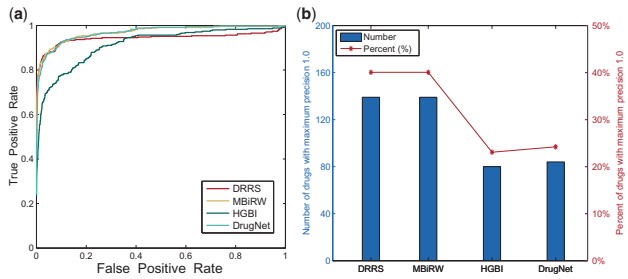

Fig. 8. (**a**) ROC curves of prediction results obtained by applying DRRS and other competitive methods on DNdatasets. (**b**) The number and percent of drugs with maximum precision value 1.0 obtained by all methods

potential drug–disease association as a matrix completion problem based on the association matrix of the heterogeneous drug–disease network. Then, the SVT-R^{4}SVD algorithm is adopted to identify novel drug–disease associations having not been validated yet by filling out the unknown entries in the drug–disease association matrix.

Theoretically, DRRS is superior to the existing drug repositioning methods based on random walk algorithm because all dominant eigenvalues and their associated eigenvectors of the adjacency matrix are taken into account. Moreover, DRRS has the ability of handling large datasets by using recycling rank-revealing randomized SVD algorithms to fast approximate SVD operations in SVT iterations. Comprehensive experiments including ten-fold cross-validation and case studies have been conducted to validate the performance of the proposed method on identifying novel indications for existing drugs. Promising results in the experiments demonstrate the effectiveness of DRRS, which is consistent with our theoretical analysis. However, the experimental results based on the three datasets indicate that the prediction capability of DRRS may be affected by sparsity of the datasets and similarity measures, as shown in DNdataset. In future studies, collecting and incorporating more relevant association data from more databases and literature may expand the application scope of our approach. In addition, the performance of DRRS can be further enhanced by improving the matrix completion algorithm itself or incorporating more effective drug and disease features.

Funding

This work was supported in part by the Natural Science Foundation of China under Grant No. 61420106009, No. 61622213, No. 61732009, No. 61728211 and National Science Foundation under Grant No. 1066471.

Conflict of Interest: none declared.

References

Arrow,K.J. *et al.* (1958) *Studies in Linear and Non-Linear Programming*, Stanford Univeristy Press, Stanford, CA.

Berger,S.I. *et al.* (2010) Systems pharmacology of arrhythmias. *Sci. Signal.*, **3**, ra30.

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

Candès,E.J. and Recht,B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

Chong,C.R. and Sullivan,D.J. (2007) New uses for old drugs. *Nature*, **448**, 645–646.

Dai,W. *et al.* (2015) Matrix factorization-based prediction of novel drug indications by integrating genomic space. *Comput. Math. Methods Med.*, **2015**, 1.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 233–240.

Davis,A.P. *et al.* (2013) The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Gönen,M. *et al.* (2013) Kernelized Bayesian matrix factorization. In: International Conference on Machine Learning, pp. 864–872.

Hamosh,A. *et al.* (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Ji,H. *et al.* (2016) A Rank Revealing Randomized Singular Value Decomposition (R3SVD) Algorithm for Low-rank Matrix Approximations, *arXiv: 1605.08134*.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

Kapur,A. *et al.* (2016) Gene expression prediction using low-rank matrix completion. *BMC Bioinformatics*, **17**, 243.

Köhler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

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

Li,J. *et al*. (2016) A survey of current trends in computational drug repositioning. *Brief. Bioinf.*, **17**, 2–12.

Li,Y. and Yu,W. (2017) A Fast Implementation of Singular Value Thresholding Algorithm using Recycling Rank Revealing Randomized Singular Value Decomposition, *arXiv: 1704.05528*.

Luo,H. *et al*. (2016) Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics*, **32**, 2664–2671.

Martínez,V. *et al*. (2015) DrugNet: network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med.*, **63**, 41–49.

Mullen,J. *et al*. (2016) Mining integrated semantic networks for drug repositioning opportunities. *PeerJ*, **4**, e1558.

Napolitano,F. *et al*. (2013) Drug repositioning: a machine-learning approach through data integration. *J. Cheminf.*, **5**, 30.

Natarajan,B.K. (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.*, **24**, 227–234.

Natarajan,N. and Dhillon,I.S. (2014) Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, **30**, i60–i68.

Schuhmacher,A. *et al*. (2016) Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.*, **14**, 1.

Shim,J.S. and Liu,J.O. (2014) Recent advances in drug repositioning for the discovery of new anticancer drugs. *Int. J. Biol. Sci.*, **10**, 654–663.

Steinbeck,C. *et al*. (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo-and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Wang,H. *et al*. (2015) Mining drug–disease relationships as a complement to medical genetics-based drug repositioning: where a recommendation system meets genome-wide association studies. *Clin. Pharmacol. Therap.*, **97**, 451–454.

Wang,W. *et al*. (2013) Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput.*, **18**, 53–64.

Wang,W. *et al*. (2014) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30**, 2923–2930.

Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.

Wishart,D.S. *et al*. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Yang,J. *et al*. (2014) drug–disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization. *J. Chem. Inf. Model.*, **54**, 2562–2569.

Yin,W. *et al*. (2008) Bregman iterative algorithms for compressed sensing and related problems. *SIAM J. Imag. Sci.*, **1**, 143–168.