Expert Systems With Applications 84 (2017) 281–289

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Computational drug repositioning using collaborative filtering via multi-source fusion

Jia Zhang[a,b], Candong Li[c], Yaojin Lin[d], Youwei Shao[a,b], Shaozi Li[a,b,*]

[a] Department of Cognitive Science, Xiamen University, Xiamen, 361005, PR China
[b] Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen University, Xiamen, 361005, PR China
[c] College of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou, 350122, PR China
[d] School of Computer Science, Minnan Normal University, Zhangzhou, 363000, PR China

ARTICLE INFO

Article history:
Received 26 January 2017
Revised 12 April 2017
Accepted 3 May 2017
Available online 4 May 2017

Keywords:
Drug repositioning
Multiple data sources
Collaborative filtering
Similarity
Optimization objective function

ABSTRACT

Drug repositioning contributes to a remarkable reduction in time and cost in traditional *de novo* drug discovery. In this study, we propose a multi-source-based drug repositioning method by using collaborative filtering to discover new indications of drugs. First, we integrate multiple data sources which are drug chemical structures, drug target proteins, and drug-disease associations to extract similarity matrices of drugs and diseases, respectively. Based on different similarity matrices, collaborative filtering is utilized to predict the drug-disease incidence matrix. Then an optimization objective function is designed to adjust the weight of each data source, and informative sources are noticed with the larger weights. Finally, experimental results on benchmark data sets reveal that the proposed algorithm is helpful to improve the prediction performance, by taking Alzheimer's disease and stroke as two examples, it is confirmed that the proposed algorithm can produce credible repositioning drugs in the treatment for these two diseases.

1. Introduction

De novo drug discovery is used as a strategy to discover new uses of old drugs, namely drug repositioning, and the term "*de novo*" is a latin word meaning "anew" or "from the beginning" (Adams & Brantne, 2006; Ashburn & Thor, 2004; DiMasi, Hansen, & Grabowski, 2003; Gilbert, Henske, & Singh, 2003; Kuang et al., 2016). However, traditional *de novo* drug discovery is a risky, expensive, and inefficient process. For launching a new drug from idea to the mature market, it is well known that plenty of time over a decade is required, and the estimated cost is more than 800 million dollars (Adams & Brantne, 2006; DiMasi et al., 2003). Furthermore, the overall success rate is low in less than 10% (Ashburn & Thor, 2004; Gilbert et al., 2003). In consideration of the challenge of the traditional way, computational drug repositioning is rising (Cano et al., 2017; Devi, Sathya, & Coumar, 2015; Martnez, Navarro, Cano, Fajardo, & Blanco, 2015; Prez-Snchez, Cano, & Garca Rodrguez, 2014), which has attracted increasing interests from the research community and pharmaceutical industry (Hurle et al., 2004; Shameer, Readhead, & Dudley, 2015).

As we know, drug repositioning is a task for identifying the potential of drug therapy, which can lower cost, and speed up the drug development cycle as short as 3–12 years (Dudley, Deshpande, & Butte, 2011). There are numerous successful examples of what can be taken to come true with drug repositioning. Sildenafil was originally marketed as a drug to treat cardiovascular disease, and discovered to treat erectile dysfunction in a clinical trail accidentally (Booth & Zemmel, 2003). It is later shown that low doses of Sildenafil also can be used in the treatment for pulmonary hypertension (Sardana et al., 2011). Thalidomide was originally indicated for morning sickness of pregnant women, and abandoned soon for the reason that was correlated with severe birth defects. Until 1998, Thalidomide was repurposed by the US Food and Drug Administration (FDA) to treat cutaneous manifestations of erythema nodosum leprosum in leprosy (Ashburn & Thor, 2004). In addition to Sildenafil and Thalidomide, the success of drug repositioning abounds in new uses of old drugs, such as Captopril, Oseltamivi, and Zanamivir, etc (Clark, 2006; Talele, Khedkar, & Rigby, 2010).

Inspired by the difficulty of the traditional way and huge commercial opportunities, many researchers are devoted to designing effective drug repositioning approaches. Hu and Agarwal (2009) analyzed genomic expression profiles of drugs and diseases, and proposed a method based on the Connectivity Map (CMap) data. By constructing a causal network between drugs and dis-

* Corresponding author.
E-mail addresses: zhangjia_gl@163.com (J. Zhang), fjzylcd@126.com (C. Li), yjlin@mnnu.edu.cn (Y. Lin), 392074149@qq.com (Y. Shao), szlig@xmu.edu.cn (S. Li).

http://dx.doi.org/10.1016/j.eswa.2017.05.004
0957-4174/

eases, (Li & Lu (2013)) introduced a pathway-based method using causal inference. Dakshanamurthy et al. (2012) presented a molecular docking method to match drug indications based on structural features of target proteins/compounds. Chiang and Butte (2009) predicted unknown drug-disease pairs by measuring the relationship between diseases according to the guilt-by-association (GBA) principle. Besides, phenotypic profiles, such as gene expression (Sirota et al., 2011; Wang, Sun et al., 2013) and side effects (Campillos, Kuhn, Gavin, Jensen, & Bork, 2008; Yang & Agarwal, 2011), are also widely used to achieve drug repositioning. However, the aforementioned methods only focus on different related information of drugs or diseases, which are sensitive to the noise in data and easily lead to predictive error (Li et al., 2016b; Lin, Hu, & Wu, 2014a, b). To handle such tasks, Gottlieb, Stein, Ruppin, and Sharan (2011) first constructed classification features based on a variety of similarity measures of drugs and diseases, and then applied logistic regression classifier to generating prediction. It is noteworthy that this method allowed to integrate additional similarity measures. Li and Lu (2012) presented a bipartite-graph-based method via considering drug chemical structures, drug target proteins and their interactions. Napolitano et al. (2013) merged drug-related features (e.g. target protein similarity and chemical structure similarity) into a single similarity matrix for the support vector machine (SVM) classification. Zhang, Wang, and Hu (2014) designed an optimization framework to analyze drug-disease associations based on similarity matrices of drugs and diseases, and the proposed method showed high efficiency. It is conceivable that these methods can help to improve the prediction performance by fusing multiple related sources. Nevertheless, previous studies seldom utilize drug-disease association information to improve drug/disease measures. In light of this point, Luo et al. (2016) introduced a Bi-Random walk method to achieve drug repositioning based on comprehensive similarity measures, whereas this method is inflexible in multi-source fusion.

In this paper, multiple data sources, including drug-disease association information, drug chemical structures, and drug target proteins, are involved in the similarity calculation, and collaborative filtering is introduced to build the prediction model. Essentially, collaborative filtering is a kind of recommendation technology, which recommends items/users for the active user/target item according to available entities as the neighbors (Adomavicius & Tuzhilin, 2005; Liu & Lee, 2010; Shi, Larson, & Hanjalic, 2014; Zhang, Lin, Lin, & Liu, 2016). For simplicity, we compare a drug to a user, and a disease is compared to an item, while Top-k similar drugs/diseases are selected as the neighbors for suggesting a target disease/active drug. Based on this hypothesis, we propose an effective drug repositioning method using collaborative filtering via multi-source fusion. First, similarity measures are utilized to compute drug/disease similarities based on different data sources as previous works (Liu et al., 2015; Luo et al., 2016). Then, collaborative filtering is applied to generating the prediction result of each drug candidate on corresponding target disease according to the similarity constructed by relevant data source. Next, we design an optimization objective function to analyze the weights of all data sources and adjust weak predicted results which are not informative. By conducting extensive experiments, we can conclude that the proposed algorithm is not only superior on various evaluation indexes, but also has the advantage to discover the potential of drug therapy. Finally, major contributions of our work are summarized as follows:

- A general optimization framework is constructed to model the drug repositioning task, which can deal with multiple related sources with less time, and lays the foundation for heterogeneous data fusion.
- An optimization objective function is designed for the weight assignment of different sources, and a linear fusion method is proposed to obtain the final prediction.
- Inspired by the recommended system, collaborative filtering is modified to search for drug candidates in the treatment of specific diseases.
- Extensive experiments demonstrate that our proposed algorithm can improve the prediction performance, and has superior capability with application to drug repositioning.

The rest of this paper is organized as follows. Section 2 provides an overview of related work in computational drug repositioning. In Section 3, we introduce our proposed algorithm in detail, and then demonstrate and explain our experimental results in Section 4. Finally in Section 5, we conclude the work and discuss several issues for future work.

2. Related work

Computational drug repositioning is becoming a rising topic in many areas (Shameer et al., 2015). With the drug-related and disease-based data growth, many computational repositioning strategies have emerged with available data from various sources (Li et al., 2016a). Here, drug chemical structures and drug target proteins are discussed as specific examples.

Drug chemical structures are evidently beneficial to improve the repositioning opportunity, which is based on the principle: similar chemical structures often affect biological systems in similar ways (Dudley et al., 2011; Liu et al., 2015). Therefore, chemical structure-based approaches to associate a drug with the others are on the strength of the similarity of the extracted features (Dudley et al., 2011; Eckert & Bajorath, 2007). Different with the drug chemical structures, the underlying assumption of target protein-based approaches is according to a key insight that the therapeutic effect of drugs on diseases is through their binding to biological targets, which are relevant to the diseases themselves (Wang, Yang, Zhang, & Li, 2014). In a general way, incorporating target protein information with other data sources directly is proved to be effective for rational drug discovery (Gottlieb et al., 2011; Wang et al., 2014). In addition to the drug chemical structures and drug target proteins, typical computational drug repositioning approaches are also designed on the basis of the other common data sources, such as molecular activity, shared molecular pathology, genetic, and side effect data (Li et al., 2016a; Zhang et al., 2014).

In recent years, the related data sources turn to be integrated for supporting the discovery of new indications of approved drugs. Compared with the methods which only focus on single data source, making fully use of multiple aspects of drug-related and disease-based data is a more efficient manner for constructing a robust and high-performance repositioning system (Li & Lu, 2012; Luo et al., 2016; Martnez et al., 2015; Napolitano et al., 2013; Zhang et al., 2014). Both of the literatures (Napolitano et al., 2013; Wang, Chen, Deng, & Wang, 2013a) focused on the drug-centered approach, and leveraged the data to train prediction models to correlate drugs with diseases. By using drug-related features, the authors constructed the drug similarity matrix as a kernel to train the SVM model for classification, and the experimental results revealed the prediction performance can be in a significant improvement via data integration. Zhang, Agarwal, and Obradovic (2013) presented a computational repositioning algorithm that incorporated the data sources of drug chemical structures, protein targets, and side-effects. Specifically, each data source was scored by a k-nearest neighbor classifier, then multiple predicted scores were integrated via large margin method. Through the Network analysis, Martnez et al. (2015) presented a network-based repositioning prioritization method called DrugNet that integrated drugs,

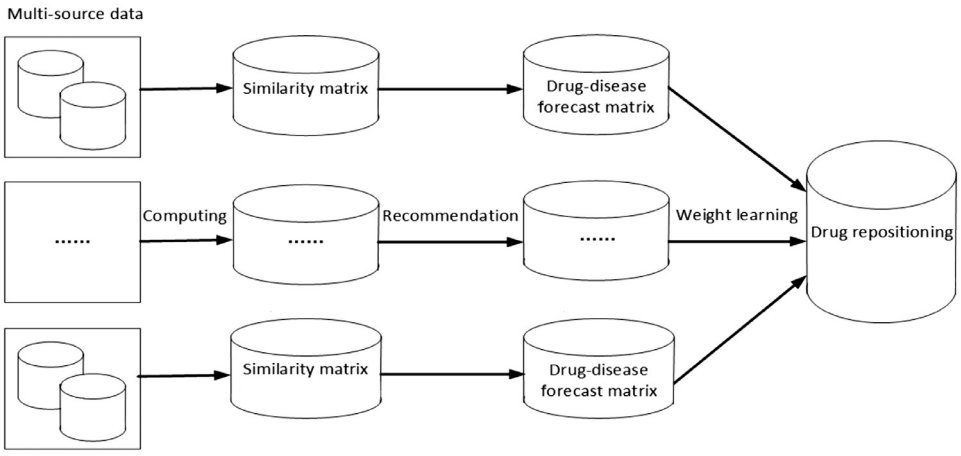


Fig. 1. Algorithm framework.

disease, and protein targets to perform disease-drug and drug-disease prioritization. In addition, Li and Lu (2012,2013) brought forward a bipartite network model and a causal network to identity new therapeutic uses of drugs successively.

3. The proposed algorithm

In this section, a novel multi-source-based drug repositioning method using collaborative filtering is described in detail, and the framework of the proposed algorithm is shown in Fig. 1. In which, similarity measures are conducted to generate the relationship of drugs and diseases based on different data sources respectively, and then collaborative filtering is applied to obtaining multiple prediction scores. Finally, multiple prediction scores integrate into an optimization result via weight learning method for the purpose of the drug repositioning task.

3.1. Collaborative filtering-based prediction algorithm

As one of popular and effective recommendation technologies, collaborative filtering is based on a key assumption that similar drugs/diseases may be share common indications/drug candidates, which predicts a drug score on target disease by aggregating the scores that similar drugs have previously known to the disease, or searches for similar diseases to guide the forecast of a drug candidate on target disease. Thus, we identify similar drugs/diseases based on multi-source data in advance.

3.1.1. Similarity calculation

For calculating the disease similarity, the data of drug-disease association information is collected from the Unified Medical Language System (UMLS) (Bodenreider, 2004), in which either a drug and a disease are treatment relationship (the drug score on the disease is denoted by 1), or the therapeutic use of a drug is unconfirmed on a disease (the drug score on the disease is denoted by 0). Then the Tanimoto coefficient is applicable to the similarity calculation, as follow:

$$sim_{ee'} = \frac{|I_{ee'}|}{|I_e| + |I_{e'}| - |I_{ee'}|}, \tag{1}$$

In Eq. (1), $sim_{ee'}$ represents the similarity between diseases e and e'. I_e and $I_{e'}$ denote the set of drugs rated on diseases e and e' respectively, and $|I_{ee'}|$ denotes the number of drugs shared by the two diseases.

The drug similarity is calculated based on drug chemical structures. The Chemical Development Kit (CDK) (Steinbeck et al., 2003) is utilized to encode each chemical component into an 881-dimensional chemical substructure defined in PubChem database (Wang et al., 2009). If a drug has a chemical substructure, the score on this chemical substructure is set to 1, and otherwise to 0. For simplicity, the Tanimoto coefficient is reused to calculate the pairwise similarity of 2D chemical fingerprints as the literature (Gottlieb et al., 2011) suggested. It is noteworthy that the Tanimoto coefficient is applied for both computation based on different data sources, and satisfies the requirements for the similarity calculation of drugs and diseases effectively. Given two drugs d and d', the computational formula is defined as follow:

$$sim_{dd'} = \frac{|I_{dd'}|}{|I_d| + |I_{d'}| - |I_{dd'}|}, \tag{2}$$

Except for the data source of drug chemical structures, the activity of a target protein in human body is modified by a drug for producing a desirable therapeutic effect. Therefore, we collect the data of drug target proteins from DrugBank (Wishart et al., 2008), and the presence or absence of each target protein is denoted by 1 or 0, respectively. Then the Cosine correlation coefficient (Liu et al., 2015) is adapted to calculate the pairwise target profile similarity between drugs, as follow:

$$sim_{dd'} = \frac{\sum_{i \in I_{dd'}} s_{di} \times s_{d'i}}{\sqrt{\sum_{i \in I_{dd'}} s_{di}^2} \times \sqrt{\sum_{i \in I_{dd'}} s_{d'i}^2}}, \tag{3}$$

In Eq. (3), s_{di} and $s_{d'i}$ are scores of drugs d and d' on target protein i respectively, and $I_{dd'}$ is the set of target proteins shared by these two drugs.

3.1.2. Prediction method

The similarity calculation is an important step for generating recommendation. After that, recommendation step is executed to predict the final score of a drug candidate on target disease. In order to calculate the final prediction score, we propose the multi-source-based fusion method, as follow:

$$p_{aq}^{*} = \sum_{k=1}^{K} \omega_k \times p_{aq}^{k}, \tag{4}$$

Where p_{aq}^{*} is the final predicted result of drug candidate a on target disease q. K is the number of data sources, and ω_k is the weight of data source k. Besides, p_{aq}^{k} is the predicted result based on data source k.

For calculating p_{aq}^{k}, the similarity between drug candidate a and other drugs is first calculated, and then Top-k similar drugs are selected as the neighbors in form of neighbor set NN of a. By em-

ploying the neighbors in NN, we can obtain the predicted result, as follows (Kaleli, 2014):

$$p_{aq}^{k} = \bar{s}_a + \frac{\sum_{d \in NN} sim_{ad}^{k} \times (s_{dq} - \bar{s}_d)}{\sum_{d \in NN} sim_{ad}^{k}}, \tag{5}$$

Where sim_{ad}^{k} is the similarity between drugs a and d based on data source k, and $\bar{s}_a$ and $\bar{s}_d$ are the average scores of drugs a and d respectively.

The above method is only suitable to the scenario that the drug similarity has been identified. Thus, for achieving the disease-based collaborative filtering, the prediction formula is modified based on the neighbors of target disease, as follow (Sarwar, Karypis, Konstan, & Riedl, 2001):

$$p_{aq}^{k} = \frac{\sum_{i \in NN'} s_{ai} \times sim_{qi}^{k}}{\sum_{i \in NN'} sim_{qi}^{k}}, \tag{6}$$

Where sim_{qi}^{k} is the similarity between diseases q and i based on data source k, and NN' is the neighbor set of target disease q. Generally speaking, $|NN|$ and $|NN'|$ are set as the same value for reducing the complexity of the final prediction.

In short, the final score is estimated through combining drug-based and disease-based collaborative filtering methods, and multi-source data is integrated to obtain the more accurate and reasonable predicted result other than single data source. However, owing to the sparsity of known drug-disease pairs, all neighbors of a drug candidate/target disease possibly have no scores. In this case, the drug candidate is regarded as a invalid positioning to fight the corresponding disease, hence the predicted result is assigned score of zero.

3.2. Weight learning for multi-source fusion

Multi-source fusion is to learn the source weights so that reliable data sources can play a crucial role in deriving the final optimization result. Based on this principle, a weight learning method is presented by constructing the following optimization objective function:

$$\begin{aligned} \arg\min_{\mathcal{W}} \quad & f(\mathcal{W}) = \sum_{k=1}^{K} w_k \sum_{\{(d,i)|s_{di}=1\}} \frac{(s_{di} - p_{di}^{k})^2}{std(p_{di}^{1}, \ldots, p_{di}^{K})} \\ \text{subject to:} \quad & \delta(\mathcal{W}) = 1. \end{aligned} \tag{7}$$

In Eq. (7), d and i denote a drug and a disease respectively, which satisfy $s_{di} = 1$. In fact, we wanna search for each known drug-disease pair (d, i) for training. By minimizing optimization objective function $f(\mathcal{W})$, the weights of all data sources are in assignment. In addition, $\delta(\mathcal{W})$ is a regularization function, which is defined as follow:

$$\delta(\mathcal{W}) = \sum_{k=1}^{K} exp(-w_k), \tag{8}$$

Where $\delta(\mathcal{W})$ regularizes w_k by constraining the sum of $exp(-w_k)$, which is conducive to expanding the difference of source weights. Reliable data sources will assign the higher weights, and vice versa.

Theorem 1. *Given optimization objective function Eq. (7) with its regularization function Eq. (8), the global optimal solution of w_k can be learned, as follow:*

$$w_k = -log\left(\frac{\sum_{\{(d,i)|s_{di}=1\}} (s_{di} - p_{di}^{k})^2}{\sum_{k'=1}^{K} \sum_{\{(d,i)|s_{di}=1\}} (s_{di} - p_{di}^{k'})^2}\right), \tag{9}$$

Proof. We define a variable v_k, which satisfies $v_k = exp(-w_k)$. As in the following, optimization function Eq. (7) can be transformed into a constrained optimization problem of v_k:

$$\begin{aligned} \arg\min_{\{v_k\}_{k=1}^{K}} \quad & f(v_k) = \sum_{k=1}^{K} -log(v_k) \sum_{\{(d,i)|s_{di}=1\}} \frac{(s_{di} - p_{di}^{k})^2}{std(p_{di}^{1}, \ldots, p_{di}^{K})} \\ \text{subject to} \quad & \sum_{k=1}^{K} v_k = 1. \end{aligned} \tag{10}$$

Where $f(v_k)$ is a linear combination of negative logarithm functions, which is convex in nature. Meanwhile, the limiting condition is linear in terms of $f(v_k)$. Thus, the aforementioned optimization objective function is convex (Boyd & Vandenberghe, 2004).

To solve this optimization problem, we use the method of Lagrange multipliers and define the Lagrangian of Eq. (10), as follow:

$$\begin{aligned} L(\{v_k\}_{k=1}^{K}, \lambda) = & \sum_{k=1}^{K} -log(v_k) \sum_{\{(d,i)|s_{di}=1\}} \frac{(s_{di} - p_{di}^{k})^2}{std(p_{di}^{1}, \ldots, p_{di}^{K})} \\ & + \lambda \left(\sum_{k=1}^{K} v_k - 1 \right), \end{aligned} \tag{11}$$

Where λ is the Lagrange multiplier. Let the partial derivative of Lagrangian is equal to zero in regard to v_k, and we can derive the following formula:

$$\sum_{\{(d,i)|s_{di}=1\}} \frac{(s_{di} - p_{di}^{k})^2}{std(p_{di}^{1}, \ldots, p_{di}^{K})} = \lambda v_k, \tag{12}$$

Combining Eq. (12) with $\sum_{k=1}^{K} v_k = 1$, we can obtain that

$$\lambda = \sum_{k'=1}^{K} \sum_{\{(d,i)|s_{di}=1\}} \frac{(s_{di} - p_{di}^{k'})^2}{std(p_{di}^{1}, \ldots, p_{di}^{K})}, \tag{13}$$

Plugging Eq. (13) into Eq. (12), we get the calculation formula of v_k. Owing to $v_k = exp(-w_k)$, then the weight w_k can be computed, as shown in Eq. (9). □

The theorem is proved.

3.3. Algorithm

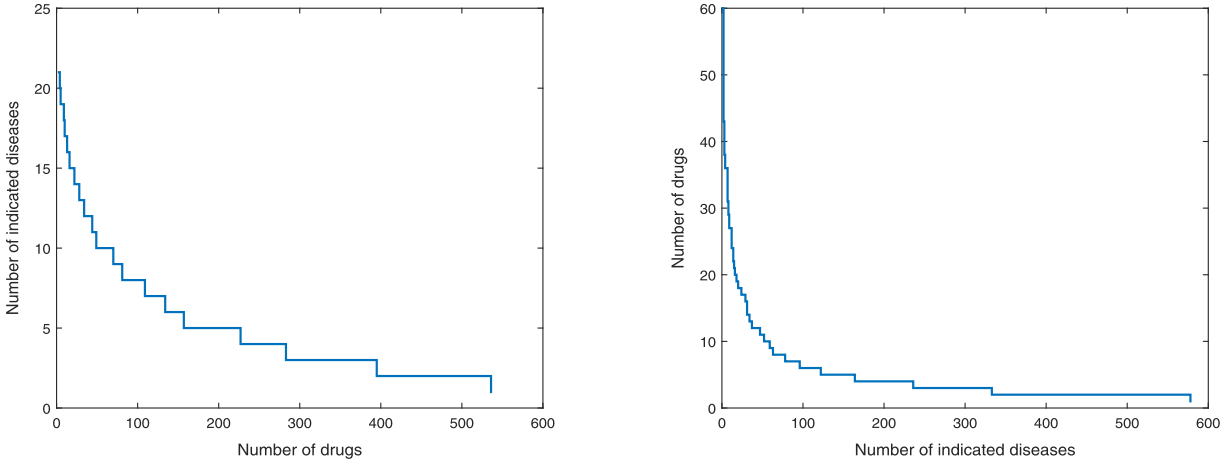
As stated in the above section, the pseudo code of our proposed algorithm is outlined in Algorithm 1.

Algorithm 1 Computational drug repositioning using collaborative filtering by fusing multi-source data (DRCFFS).

Input: Drug-disease incidence matrix $R(m, n)$; Data sets of drug chemical structures and drug target proteins; Parameter ϑ

```
Calculate the disease similarity by using Eq. (1), and construct corresponding disease similarity matrix;
Calculate the drug similarity by using Eqs. (2) and (3) respectively, and construct multiple drug similarity matrices;
for k = 1 to K do
    for a = 1 to m do
        for q = 1 to n do
            Predict p_aq^k according to ϑ neighbors of drug candidate a or target disease q by using Eq. (5) or (6);
        end for
    end for
    Return prediction matrix PR^k;
end for
Learn the weight distribution W by minimizing Eq. (7);
Predict drug-disease incidence matrix PR* by using Eq. (4).
```

Output: PR^* and $\mathcal{W} = \{w_1, \ldots, w_K\}$



(a) Number of indicated diseases per drug (b) Number of drugs per indicated disease

Fig. 2. Statistics of drug-disease incidence matrix.

In lines 1 and 2, the proposed algorithm allocates memory to preserve multiple similarity matrices, and the time complexity of the similarity calculation of drugs and diseases is $O(m^2)$ and $O(n^2)$ respectively, where m is the number of drugs, and n is the number of diseases. In lines $3 \sim 10$, source weights with their corresponding predicted results are estimated, and both of them require the time cost of $O(K \times m \times n)$, where K is the number of data sources. Finally, the proposed algorithm predicts unknown drug-disease pairs for the drug repositioning task.

4. Experiments

To verify the feasibility and validity of our proposed algorithm, extensive experiments are conducted in this section, which is divided into six subsections, including data description, an introduction of evaluation metrics, method comparison, data source comparison, case studies, and discussion. Details are as follows.

4.1. Data sets

Therapeutic uses of 536 approved drugs on 578 diseases are collected from the National Drug File-Reference Terminology, which is a part of the Unified Medical Language System (UMLS) (Bodenreider, 2004). According to all known drug-disease pairs, a drug-disease incidence matrix is constructed as a benchmark data set, in which the treatment relationship between a drug and a disease is represented by 1, and unconfirmed drug-disease pairs are represented by 0. In order to display this data set, the statistics of this data set are plotted as shown in Fig. 2. We can see from Fig. 2(a) that approximately 75% of drugs have indications in less than 5 diseases, while only 9% of drugs have indications over 10 diseases. In Fig. 2(b), it is observed most of diseases (about 83%) have drugs in less than 5, and only 9% of diseases can be prevented over 10 available drugs.

Furthermore, we collect drug related information considering drug chemical structures and drug target proteins. Towards the data source of drug chemical structures, chemical structures of this 536 drugs are extracted from PubChem (Wang et al., 2009). Each drug is denoted by an 881-dimensional chemical substructure, and the presence or absence of each substructure is denoted by 1 or 0, respectively. Moreover, the information of relationships between the 536 drugs and 775 target proteins is also collected from DrugBank (Wishart et al., 2008) to measure the drug similarity.

Table 1
Confusion matrix.

	Actual value	
Predicted value	True positive (TP) False negative (FN)	False positive (FP) True negative (TN)

4.2. Evaluation metrics

We compare the proposed algorithm with other methods by using receiver operating characteristic (ROC) curve and precision-recall curve. Besides, precision, recall, F-score, and the area under ROC curve (AUC) are also applied as evaluation metrics (Wang, Chen, Deng, & Wang, 2013b; Zhang et al., 2013). In order to give the clear definitions of evaluation metrics, we first define the confusion table which is designed by comparing actual ratings with predicted values, and two classes are denoted by positive and negative respectively, as shown in Table 1.

Then we can define these four evaluation metrics. Precision is the proportion of known drug-disease associations that appear among of the ranked list according to a specific threshold, as follows:

$$Precision = TP/(TP + FP), \tag{14}$$

Recall is the proportion of known associations that are correctly predicted, whose calculation formula is as:

$$Recall = TP/(TP + FN), \tag{15}$$

According to the computed results of precision and recall, the formula of F-score is defined as follow:

$$F - score = 2 \times Precision \times Recall/(Precision + Recall), \tag{16}$$

Since a specific threshold needs to be set for obtaining precision, recall, and F-score, as in experiments, we select the specific threshold which makes F-score maximum. Besides, the AUC score is measured by the area under ROC curve, which is plotted according to the true positive rate (TPR) and false positive rate (FPR).

$$TPR = TP/(TP + FN), \tag{17}$$

$$FPR = FP/(FP + TN), \tag{18}$$

In addition, 10-fold cross validation is adopted to evaluate the prediction performance systematically. In detail, all known drug-disease pairs in drug-disease incidence matrix are randomly divided into 10 equal subsets, each subset is held-out in turn for

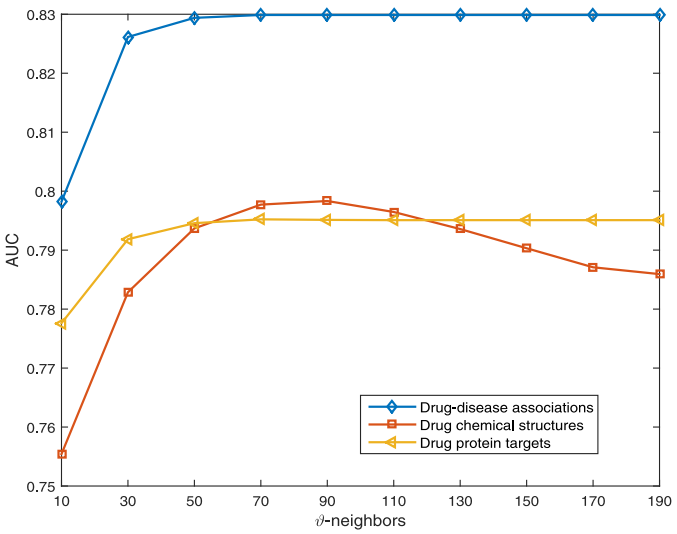


Fig. 3. The change of the AUC score with different ϑ-neighbors.

testing, while the remaining data is merged to train the prediction model. As the validation is iterated ten times, the averaged metric values out of ten runs are calculated for the algorithm.

4.3. Method comparison

As a preliminary endeavor, we try to search for an optimal parameter to maximize the prediction performance of each data source. Thus, predicted results are first generated based on drug-disease associations, drug chemical structures, and drug target proteins respectively. According to the predicted results, we can obtain change curves of the AUC score with different ϑ-neighbors, as shown in Fig. 3. We can see from Fig. 3 that variation trends are similar based on drug-disease associations and drug protein targets, and their AUC scores increase with the increasing of parameter ϑ when $\vartheta \leq 70$, and then remain unchanged. Different with the former, based on the data source of drug chemical structures, the AUC score is up to maximum when $\vartheta = 90$, and begins to decrease when $\vartheta > 90$. Therefore, for the sake of computational cost, all the results of the proposed algorithm shown in later experiments are obtained with parameter ϑ set to be 90.

After determining parameter ϑ uniquely, we compare our proposed algorithm with some other methods. MBiRW (Luo et al., 2016) is one of state-of-the-art drug repositioning algorithms selected as a compared method, which calculates the drug similarity based on drug chemical structures, while the disease similarity is estimated based on drug-disease association information. Based on the constructed similarity matrices of drugs and diseases, the heterogeneous network is designed via the bipartite graph for *de novo* drug-disease prediction. In addition, the parameters in MBiRW have little effects on test data in terms of the prediction performance, hence we set parameters $\alpha = 0.3$, and l, r to 2 as the literature (Luo et al., 2016) suggested. Furthermore, a collaborative filtering-based drug repositioning method based on single data source of chemical structures (DRCFCS) is designed to participate in this method comparison, which calculates the similarity based on drug chemical structures, and applies collaborative filtering to generating prediction. All of them, including our proposed algorithm, generate the final predicted scores in the drug-disease incidence matrix.

Fig. 4 demonstrates ROC curves and precision-recall curves of the three methods on the same experimental conditions. In which the three ROC curves are plotted as shown in Fig. 4(a). Explanatorily, since some predicted results are possible to be assigned scores

Table 2
Method comparison according to AUC, precision, recall, and F-score.

Method	AUC	Precision	Recall	F-score
DRCFFS	0.8566	0.8461	0.7141	0.7731
MBiRW	0.8097	0.7769	0.6756	0.7223
DRCFCS	0.7983	0.7245	0.6716	0.6959

Table 3
Data source comparison according to AUC, precision, recall, and F-score.

Data source	AUC	Precision	Recall	F-score
DDAS	0.8299	0.8628	0.6844	0.7629
CHST	0.7983	0.7245	0.6716	0.6959
TAPR	0.7951	0.8975	0.6104	0.7259
DDAS+CHST	0.8566	0.8461	0.7141	0.7731
DDAS+TAPR	0.8681	0.8241	0.7761	0.7991
CHST+TAPR	0.8285	0.8644	0.6543	0.7437
All sources	0.8696	0.8700	0.7561	0.8085

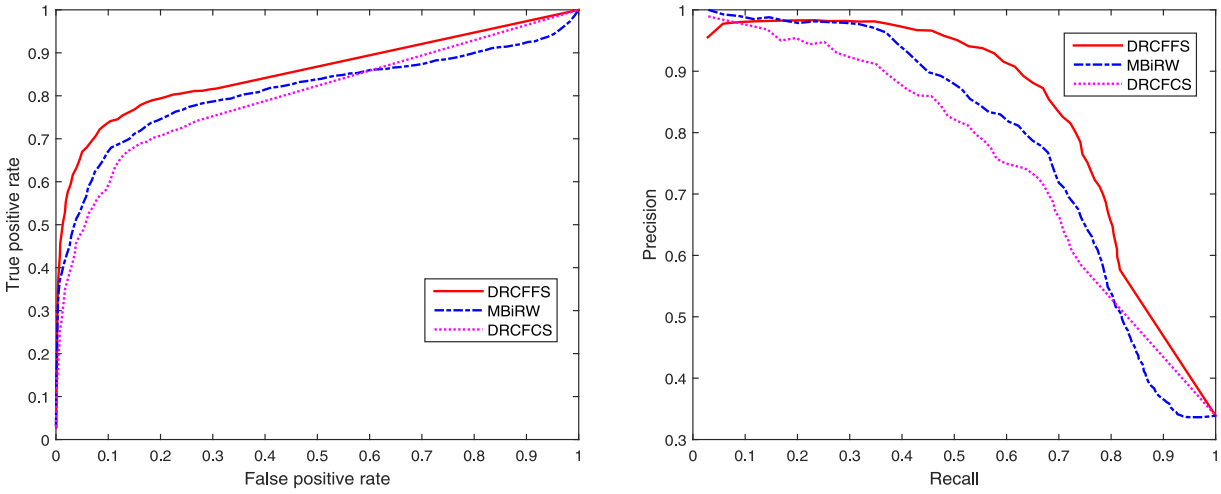
of zero, they are useless on the drug repositioning task in fact. To handle this situation, they are considered to be identified as positive examples, which causes the straight line segment of the ROC curves of DRCFFS and DRCFCS at the liberal region (Kuang et al., 2016). From Fig. 4(a), we can see that DRCFFS obtains the highest AUC score (AUC = 0.8566), while the AUC score of MBiRW is 0.8097 and DRCFCS gets the lowest AUC score of 0.7983. In Fig. 4(b), it is observed that DRCFFS outperforms the other methods according to the precision-recall curves. In addition, precision, recall, and F-score are recorded in Table 2. The results demonstrate that DRCFFS also produces the best results in terms of precision, recall, and F-score. In conclusion, our proposed algorithm can obtain the better performance compared with the other methods, which is helpful to make an accurate and reasonable predicted result.

4.4. Data source comparison

In this subsection, multiple data sources are involved in data source comparison, which are DDAS (drug-disease associations) and CHST (chemical structures). Besides, the data source of TAPR (target proteins) is added to further verify the effectiveness.

Table 3 summarizes the evaluation result. We can conclude from Table 3 that data set DDAS appears to be the most significant among the three data sources. Specifically, compared with CHST and TAPR, the predicted result based on DDAS has the higher AUC score, recall, and F-score, while the best precision is obtained based on TAPR. Nevertheless, the predicted result based on TAPR has the lowest recall. In addition, we can see from Table 3 that the combination of different data sources is helpful to improve the performance. For example, the combination of CHST (AUC = 0.7983, F-score = 0.6959) and TAPR (AUC = 0.7951, F-score = 0.7259) can obtain the higher AUC score 0.8285 and F-score 0.7437, and the combination of DDAS and TAPR has a significant performance improvement with respect to the AUC score and F-score. Secondly, the phenomenon, considering the variation of precision and recall, suggests that fusing data sources can provide the more balanced and meaningful result. Finally, the highest AUC score and F-score are obtained by combining all sources, and we get evaluation results of precision and recall which are 0.8773 and 0.7561 respectively. Thus, we may reasonably arrive at the conclusion that the proposed algorithm is feasible and effective via multi-source fusion.

Concerning the normalized weight distribution of all data sources, we reckon that DDAS has the highest weight of 0.68



(a) ROC curves (b) Precision-recall curves

Fig. 4. Method comparison according to ROC and precision-recall curves.

Table 4
Method comparison of Simple Average and the proposed algorithm.

Method	AUC	Precision	Recall	F-score
Simple Average	0.8690	0.8773	0.7446	0.8049
DRCFFS	0.8696	0.8700	0.7561	0.8085

among the three data sources, TAPR gets the weight of 0.31, while CHST is the lowest. In this case, we assume that the defined regularization function results in the huge gap of weights, hence most of informative sources can be fully utilized out of all data sources. Compared with the Simple Average method, which assigns data sources with the same weights, the proposed algorithm has the better result, as shown in Table 4. Meanwhile, it is observed from Table 4 that the Simple Average method obtains an outstanding prediction performance unexpectedly. By the in-depth analysis, we can conclude that simply combining predicted results based on collaborative filtering with equally importance does help to integrate into an optimization result. With the increasing of low-quality data sources involved in the prediction task, however, the proposed algorithm can be adapted to this scenario whereas the Simple Average method is less effective or even harmful.

4.5. Case studies

New indications of approved drugs can be obtained by our proposed method. Particularly, drug-disease associations in the drug-disease incidence matrix with drug-related information, such as chemical structures and target proteins, are used as the training set, and all unknown drug-disease pairs form the candidate set.

As an example, Top-10 predicted drug candidates for Alzheimer's disease are analyzed, as shown in Table 5. In which Valproic Acid and Selegiline are known drugs for Alzheimer's disease appeared in the drug-disease incidence matrix, and the remaining 8 drugs could be considered as the repositioning result. For these 8 drug candidates, some of them are supported by clinical trails on web ClinicalTrials.gov. Isocarboxazid, one of effective drugs to treat depressive disorder, is used as a potential therapy for Alzheimer's disease according to NCT02323217. Risperidone, originally indicated for autistic disorder and dementia, has been tested to treat Alzheimer's disease in multiple clinical trails (NCT00034762, NCT00417482, and NCT01119638). In addition to Isocarboxazid and Risperidone, clinical trails (NCT00015548 and NCT00245206) have been conducted to evaluate Olanzapine in the treatment for Alzheimer's disease.

Table 5
Top-10 drugs for Alzheimer's disease predicted by the proposed algorithm.

Drug name	Examples of original use(s)	Clinical trail
Valproic Acid	Alzheimer's Disease, Bipolar Disorder	–
Pyridostigmine	Myasthenia Gravis	No
Carbamazepine	Post-Traumatic, Trigeminal Neuralgia	No
Selegiline	Alzheimer's Disease, Parkinson Disease	–
Isocarboxazid	Depressive Disorder	Yes
Ethosuximide	Epilepsy (Absence)	No
Phenelzine	Depressive Disorder	No
Risperidone	Autistic Disorder, Dementia	Yes
Clozapine	Bipolar Disorder, Schizophrenia	No
Olanzapine	Bipolar Disorder, Schizophrenia	Yes

Table 6
Top-10 drugs for stroke predicted by the proposed algorithm.

Drug name	Examples of original use(s)	Clinical trail
Pentoxifylline	Diabetic Angiopathies, Stroke	–
Warfarin	Stroke, Thrombophlebitis	–
Sildenafil	Hypertension, Impotence (Vasculogenic)	Yes
Tadalafil	Hypertension	Yes
Iloprost	Hypertension	No
Caffeine	Apnea	No
Theophylline	Bronchial Spasm, Pulmonary Emphysema	No
Adenosine	Wolff-Parkinson-White Syndrome	No
Etomidate	Head Injuries (Closed), Brain Ischemia	Yes
Tolazoline	Raynaud Disease, Spasm	No

One other example is analyzed, as shown in Table 6, we list Top-10 drug candidates for stroke. For the Top-10 predicted drugs, Pentoxifylline and Warfarin are known treatments to stroke, and Sildenafil, Tadalafil and Etomidate are predicted and supported by clinical trails from ClinicalTrials.gov. In which the study of Sildenafil and stroke recovery is currently recruiting participants (NCT02628847). Tadalafil, originally indicated for Hypertension, is in practice as a possible treatment for stroke (NCT02801032), and clinical trails (NCT02453373 and NCT02822144) have been proposed to test Etomidate as a potential therapy for stroke.

4.6. Discussion

Drug repositioning is a promising and high-efficiency approach to discover new indications of old drugs on diseases. With the

development of the information technology, computational repositioning has a huge potential achieving precision medicine in the era of big data. For example, The explosive growth of multiple related data sources, such as drug chemical structures and target proteins data, as well as the data of drug-disease association information, creates a favorable and friendly atmosphere for *de nevo* drug discovery. For remedying the insufficient in multi-source learning of previous works mentioned in Section 1, we proposed an effective multi-source-based drug repositioning method using collaborative filtering. From the experimental results, we can conclude that our proposed algorithm can obtain the better prediction performance than some other methods. Moreover, we have considered known drug-disease pairs to improve similarity measures, and the proposed multi-source method is proved to be feasible to promote the predictive quality. Finally, two kinds of special diseases, namely Alzheimer's disease and stroke, are selected as treatment objects, and we can find relevant proofs to testify the effectiveness of our proposed algorithm.

Our proposed algorithm is manipulable to integrate additional related sources, such as the data of drug side effects. Side effects indicate unintended consequences of the drug action, and provide a pathway to associate drugs with diseases. This is because they result in the physiological consequence in human body. Furthermore, the phenotypic expression of a side effect is also helpful to measure the similarity of diseases. Together with the property of drugs and diseases, these data sets are important resources to construct computational drug repositioning mechanism.

In fact, our proposed algorithm is also applicable in many other fields, such as the identification of health-state in traditional Chinese medicine (TCM). As we know, the TCM syndrome diagnosis is based on a summary of comprehensive signals, which mainly include inspection, listening and smelling, inquiry, and palpation, and the proposed method can work efficiently to discover the valuable clinical knowledge from the multi-source data or multi-model data, and achieves the diagnosis of disease location and nature of patients. In brief, our proposed algorithm is appropriate in these cases to make well-founded decisions.

5. Conclusions

This work proposed a novel drug repositioning method, which is based primarily on two aspects. First, in light of an important assumption that similar drugs/diseases have similar indications/drug candidates, the collaborative filtering method of both drug-based and disease-based were introduced to generate the repositioning result. The second is according to the principle that fusing multiple related sources is conducive to building a robust prediction model with high efficiency. As we know, the prediction model based on one-sided related information is easily affected by the data noise. Therefore, multi-source data, including drug-disease associations, was used to improve drug/disease measures in this study. For achieving the multi-source fusion, an optimization objective function was constructed to learn the weight distribution of data sources. With this, we utilized multiple related sources with different weights to calculate drug and disease similarities, and obtained a sophisticated result in the end. In experiments, the results showed that the proposed method is feasible and effective. Compared with some other methods, the proposed method had the advantage with respect to the prediction performance, and it is conformed that the performance of our proposed method had a significant improvement via multi-source fusion. Finally, we have discovered several drug candidates to treat Alzheimer's disease and stroke. Isocarboxazid, Risperidone, and Olanzapine were predicted and supported by clinical trails as possible treatments for Alzheimer's disease, and Sildenafil, Tadalafil and Etomidate were proposed to treat stroke. The other repositioning result can provide the theoretical foundation in clinical trials.

In future, we will consider to integrate additional data sources into the optimization objective function (e.g. gene expression and side effects), and the construction of novel fusion method on the knowledge level is our pursuit to deal with such optimization problem. Furthermore, we will pay close attention to improve the designed multi-source-based model to other research fields, such as the identification of health-state in traditional Chinese medicine.

Acknowledgments

The authors would like to thank the anonymous reviewers and the editor for their constructive and valuable comments. This work is supported by grants from the Nature Science Foundation of China (Nos. 61572409, 61402386, 81230087, 61571188, 61672272, and 61303131), Fujian Province 2011 Collaborative Innovation Center of TCM Health Management and Collaborative Innovation Center of Chinese Oolong Tea Industry Collaborative Innovation Center (2011) of Fujian Province.

References

Adams, C. P., & Brantne, V. V. (2006). Estimating the cost of new drug development: Is it really 802 million dollars? *Health Affairs, 25*, 420–428.
Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*, 734–749.
Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery, 3*, 673–683.
Bodenreider, O. (2004). The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research, 32*, D267–D270.
Booth, B., & Zemmel, R. (2003). Quest for the best. *Nature Reviews Drug Discovery, 2*, 838–841.
Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.
Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science, 321*, 263–266.
Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., & Thapa, A. (2017). Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications, 72*, 151–159.
Chiang, A. P., & Butte, A. J. (2009). Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical Pharmacology and Therapeutics, 85*, 507–510.
Clark, D. E. (2006). What has computer-aided molecular design ever done for drug discovery? *Expert Opinion on Drug Discovery, 1*, 103–110.
Dakshanamurthy, S., Issa, N. T., Assefnia, S., Seshasayee, A., Peters, O. J., & Madhavan, S. (2012). Predicting new indications for approved drugs using a proteochemometric method. *Journal of Medicinal Chemistry, 55*, 6832–6848.
Devi, R. V., Sathya, S. S., & Coumar, M. S. (2015). Evolutionary algorithms for *denovo*, drug design-a survey. *Applied Soft Computing, 27*, 543–552.
DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics, 22*, 151–185.
Dudley, J. T., Deshpande, T., & Butte, A. J. (2011). Exploiting drug-disease relationships for computational drug repositioning. *Briefings in Bioinformatics, 12*, 303–311.
Eckert, H., & Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today, 12*, 225–233.
Gilbert, J., Henske, P., & Singh, A. (2003). Rebuilding big pharma's business model. *In Vivo, 21*, 73–80.
Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). Predict: A method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology, 7*, 496.
Hu, G., & Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *Plos One, 4*, e6536.
Hurle, M. R., Yang, L., Xie, Q., Rajpal, D. K., Sanseau, P., & Agarwal, P. (2004). Drug repositioning: Identifying and developing new uses for existing drugs. *Clinical Pharmacology and Therapeutics, 93*, 335–341.
Kaleli, C. (2014). An entropy-based neighbor selection approach for collaborative filtering. *Knowledge-Based Systems, 56*, 273–280.
Kuang, Z., Thomson, J., Caldwell, M., Peissig, P., Stewart, R., & Page, D. (2016). Computational drug repositioning using continuous self-controlled case series. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, CA, USA* (pp. 491–500).
Li, J., & Lu, Z. (2012). A new method for computational drug repositioning using drug pairwise similarity. In *Proceedings of the IEEE international conference on bioinformatics and biomedicine, philadelphia, PA, USA* (pp. 1–4).

Li, J., & Lu, Z. (2013). Pathway-based drug repositioning using causal inference. *BMC Bioinformatics, 14*, S3.
Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., & Lu, Z. (2016a). A survey of current trends in computational drug repositioning. *Briefings in Bioinformatics, 17*, 2–12.
Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., & Fan, W. (2016b). Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery. *IEEE Transactions on Knowledge and Data Engineering, 28*, 1986–1999.
Lin, Y., Hu, X., & Wu, X. (2014a). Ensemble learning from multiple information sources via label propagation and consensus. *Applied Intelligence, 41*, 30–41.
Lin, Y., Hu, X., & Wu, X. (2014b). Quality of information-based source assessment and selection. *Neurocomputing, 133*, 95–102.
Liu, F., & Lee, H. J. (2010). Use of social network information to enhance collaborative filtering performance. *Expert Systems with Applications, 37*, 4772–4778.
Liu, Z., Guo, F., Gu, J., Wang, Y., Li, Y., & Wang, D. (2015). Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. *Bioinformatics, 31*, 1788–1795.
Luo, H., Wang, J., Li, M., Luo, J., Peng, X., & Wu, F. X. (2016). Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics, 32*, 2664–2671.
Martnez, V., Navarro, C., Cano, C., Fajardo, W., & Blanco, A. (2015). Drugnet: Network-based drug-disease prioritization by integrating heterogeneous data. *Artificial Intelligence in Medicine, 63*, 41–49.
Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., & D'Amato, M. (2013). Drug repositioning: A machine-learning approach through data integration. *Journal of Cheminformatics, 5*, 30.
Prez-Snchez, H., Cano, G., & Garca Rodrguez, J. (2014). Improving drug discovery using hybrid softcomputing methods. *Applied Soft Computing, 20*, 119–126.
Sardana, D., Zhu, C., Zhang, M., Gudivada, R. C., Yang, L., & Jegga, A. (2011). Drug repositioning for orphan diseases. *Briefings in Bioinformatics, 12*, 346–356.
Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on world wide web, hong kong, china* (pp. 285–295).
Shameer, K., Readhead, B., & Dudley, J. T. (2015). Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Current Topics in Medicinal Chemistry, 15*, 5–20.
Shi, Y., Larson, M., & Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges. *ACM Computing Surveys, 47*, 3:1–3:45.
Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., & Sweet–Cordero, A. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science Translational Medicine, 3*, 96ra77.
Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The chemistry development kit (cdk): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences, 43*, 493–500.
Talele, T. T., Khedkar, S. A., & Rigby, A. C. (2010). Successful applications of computer aided drug discovery: Moving drugs from concept to the clinic. *Current Topics in Medicinal Chemistry, 10*, 127–141.
Wang, K., Sun, J., Zhou, S., Wan, C., Qin, S., & Li, C. (2013). Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. *Plos Computational Biology, 9*, e1003315.
Wang, W., Yang, S., Zhang, X., & Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics, 30*, 2923–2930.
Wang, Y., Chen, S., Deng, N., & Wang, Y. (2013a). Drug repositioning by kernel-based integration of molecular structure, molecular activity. *Plos One, 8*, e78518.
Wang, Y., Chen, S., Deng, N., & Wang, Y. (2013b). Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics, 29*, 1317–1324.
Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). Pubchem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research, 37*, W623–W633.
Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., & Tzur, D. (2008). Drugbank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research, 36*, D901–D906.
Yang, L., & Agarwal, P. (2011). Systematic drug repositioning based on clinical side–effects. *Plos One, 6*, e28025.
Zhang, J., Lin, Y., Lin, M., & Liu, J. (2016). An effective collaborative filtering algorithm based on user preference clustering. *Applied Intelligence, 45*, 230–240.
Zhang, P., Agarwal, P., & Obradovic, Z. (2013). Computational drug repositioning by ranking and integrating multiple data sources. In *Proceedings of the joint european conference on machine learning and knowledge discovery in databases* (pp. 579–594).
Zhang, P., Wang, F., & Hu, J. (2014). Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity. *AMIA 2014, american medical informatics association annual symposium, washington, DC, USA*.