An Optimized Drug Similarity Framework for Side-effect Prediction

Yi Zheng, Shameek Ghosh and Jinyan Li

Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

Abstract

Drug side-effects are crucial issues in both the pre-market drug developing process and post-market drug clinical applications. They contribute to one-third of drug failures and cause significant fatality and severe morbidity. Thus the early identification of potential drug side-effects is of great interests. Most existing methods essentially rely on leveraging few drug similarities directly for side-effect predictions, ignoring the performance improvement by drug similarity integration and optimization. In this study, we proposed an optimized drug similarity framework (ODSF) to improve the performance of side-effect predictions. First, this framework integrates four different drug similarities into a comprehensive similarity. Next, the comprehensive similarity is optimized via clustering and then enhanced by indirect drug similarity. Finally, the optimized drug similarity is employed for side-effect predictions. The performance of ODSF was evaluated on simulative side-effect predictions of 917 drugs from the DrugBank. Extensive comparison experiments demonstrate that ODSF is competent to capture drug features from diverse perspectives and the prediction performance is significantly improved owing to the optimized drug similarity.

1. Introduction

Drug side-effects are effects which are secondary to the intended effects [1]. They have drawn attention of the society because they cause a large number of morbidity and fatality every year. Therefore, the early identification of potential side-effects to avoid serious harms and financial loss is of great importance. For this purpose, experimental approaches which test compounds with *in vitro* biochemical and cellular assays were proposed. However, drug side-effect prediction remains challenging because of the expensive and long-term process of experimental approaches. In recent years, several computational methods have been proposed for side-effect predictions. Most of these methods are based on the hypothesis that similar drugs are more likely to share the same side-effects. According to the similarity types they adopt, these methods can be classified as follows:

(1) Target-protein similarity based methods, which utilize target-protein similarity directly or pathway similarity which involves target-proteins indirectly to measure the similarity between drugs. Two previous studies demonstrated that side-effects and target proteins have strong relations [2] [3]. Huang et al. developed a novel computational framework by combining clinical observation data with drug target data to predict side-effects of trial drugs [4]. Their results showed that the prediction performance improved significantly owing to incorporating prior knowledge including the drug target data. Pathways are series of actions among molecules in a cell. They can trigger the assembly of new molecules, including proteins. Previous enrichment analyses using KEGG and Gene Ontology reveal that the correlated sets were significantly enriched with proteins involved in the same biological pathways [5]. Side-effects can be seen as phenotypic outcomes by drugs targeting proteins in the same correlated set. Thus there is some-how relationship between side-effects and pathways. Fukuzaki et al. developed an efficient algorithm method named “CoopeRative Pathway Enumerator” to identify co-operative pathways which share common active conditions [6]. Finally, these identified cooperative pathways were leveraged to predict drug side-effects and achieved satisfactory results.

(2) Chemical structure similarity based methods, which measure the drug similarity by their chemical structures. Atias et al. conducted the canonical correlation analysis (CCA) between drug chemical structures and side-effects to predict new side-effects [7]. It is one of the pioneering work on predicting multiple side-effects at a time. Yoshihiro et al. tried to improve side-effect prediction by integrating drug chemical structures and target proteins [8]. Simulative prediction of side-effects from the DrugBank demonstrated that the prediction performance was improved significantly by integration of the two information sources.

Most existing computational prediction methods focus on prediction from one or few drug similarity sources. Too few similarity sources may not capture enough drug features. Moreover, these methods utilize drug similarities directly for predictions, ignoring the value of optimization

ISSN: 2325-887X DOI:10.22489/CinC.2017.128-068

from such similarities and indirect similarities that could improve the prediction performance. In this study, we proposed an optimized drug similarity framework (ODSF) to improve the side-effect prediction performance. This framework integrates four different drug similarities into a comprehensive similarity first. Next the comprehensive similarity is further optimized via clustering and then enhanced by indirect similarity. Finally, the optimized drug similarity is used for side-effect predictions.

2. Materials

2.1. Drug Side-effect Profiles

The side-effect data set was downloaded from SIDER [9]. We focus on side-effects of drugs which are grouped as "Small Molecules" in DrugBank. Our basic idea lies in predicting side-effects by drug similarities. Therefore, those drugs whose similarity information are not available were removed. Finally, we obtained a dataset constituted by 917 drugs, 500 side-effects and 78,855 drug side-effect associations.

2.2. Drug Similarity Data

Four types of drug similarity will be integrated as the original comprehensive drug similarity (OCDS) using the following formula.

$$\begin{aligned} Simcom(d_j, d_k) = [&S_{chem}(d_j, d_k) + S_{pro}(d_j, d_k) \\ &+S_{sub}(d_j, d_k) + S_{thera}(d_j, d_k)]/4 \end{aligned} \quad (1)$$

A. Chemical Structure Similarity The chemical-structure fingerprints of drugs were retrieved using CDK from their SMILES files downloaded from the DrugBank. For a drug d, it can be represented by its fingerprint $f^d(f_i^d \in \{0,1\}, i \in \{1...1024\})$. Then the chemical similarity score between drug d_j and drug d_k is given by:

$$S_{chem}(d_j, d_k) = \frac{\sum_{l=1}^{1024}(f_l^j \wedge f_l^k)}{\sum_{l=1}^{1024}(f_l^j \vee f_l^k)} \quad (2)$$

where $\wedge$ and $\vee$ are bitwise "and" and "or" operators respectively; f_l^j and f_l^k are the l^{th} bit of fingerprints of drug d_j and drug d_k respectively.

B. Drug Target Protein Similarity The similarity between two proteins is calculated based on the overlapping rate of their associated Gene Ontology (GO) terms. Suppose GO^m and GO^n are the GO term sets for protein p_m and protein p_n respectively, the similarity score between p_m and p_n would be

$$S_{go}(p_m, p_n) = \frac{GO^m \cap GO^n}{GO^m \cup GO^n} \quad (3)$$

where $\cap$ and $\cup$ are intersection and union operators respectively. The GO terms of target proteins were downloaded from the EMBL-EBI website. Then the drug target protein similarity between each pair of drugs was calculated by integrating protein similarities of their target proteins.

$$S_{pro}(d_j, d_k) = \frac{\sum_{m=1}^{N_j}\sum_{n=1}^{N_k} S_{go}(p_m, p_n)}{N_j * N_k} \quad (4)$$

where N_j and N_k are the total number of proteins in the interacted protein sets of drug d_j and drug d_k respectively.

C. Drug Substituent Similarity The drug substituent similarity between drug d_j and drug d_k is calculated via Jaccard score which is defined as follows:

$$S_{sub}(d_j, d_k) = \frac{SUB_j \cap SUB_k}{SUB_j \cup SUB_k} \quad (5)$$

where SUB_j and SUB_k are the substituent sets of drug d_j and d_k respectively.

D. Drug Therapeutic Similarity

The Anatomical Therapeutic Chemical (ATC) codes used in this study were extracted from the DrugBank. There are 5 levels in the ATC code. Consequently, we calculated the drug therapeutic similarity at each level separately first. The l^{th} level drug therapeutic similarity (S_l) between the drug d_j and d_k is defined as follows:

$$S_l(d_j, d_k) = \frac{ATC_l(d_j) \cap ATC_l(d_k)}{ATC_l(d_j) \cup ATC_l(d_k)} \quad (6)$$

where $ATC_l(d_j)$ denotes the l^{th} level ATC code for drug d_j. The average value of the five-level similarity scores is used as the therapeutic similarity of a drug pair:

$$S_{thera}(d_j, d_k) = \frac{\sum_{l=1}^{n} S_l(d_j, d_k)}{n} \quad (7)$$

where $n = 5$, is the total number of ATC code levels.

3. Methods

The overall framework of ODSF is illustrated in Figure 1. At the beginning, drugs and side-effects which don't satisfy the requirements are removed. Then four different drug similarities are integrated as the original comprehensive drug similarity. Next the original comprehensive similarity is further optimized via ClusterONE clustering and indirect similarity optimization. After that, the two types of drug similarities are integrated as the optimized similarity. Later, a balanced drug training set is built for each side-effect using the proposed strategy. Afterwards, drugs are vectorized according to their optimized similarities with each drug in the training set. Finally, a classifier is built and trained for each side-effect. Corresponding classifier is employed to predict potential drugs which could cause the side-effect.

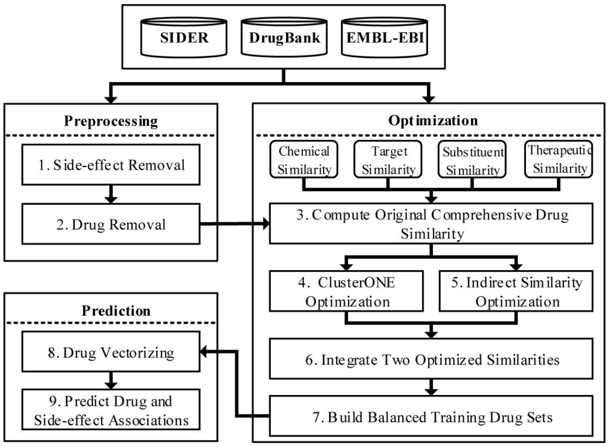
SIDER
DrugBank
EMBL-EBI
Preprocessing
1. Side-effect Removal
2. Drug Removal
Optimization
Chemical Similarity
Target Similarity
Substituent Similarity
Therapeutic Similarity
3. Compute Original Comprehensive Drug Similarity
4. ClusterONE Optimization
5. Indirect Similarity Optimization
6. Integrate Two Optimized Similarities
7. Build Balanced Training Drug Sets
Prediction
8. Drug Vectorizing
9. Predict Drug and Side-effect Associations

Figure 1. Flow diagram of drug side-effect prediction based on ODSF.

3.1. Drug Similarity Optimization

A. Optimization based on Clustering

Based on the hypothesis that similar drugs are more likely to share the same side-effects, we proposed to optimize drug similarity via clustering. We first built a weighted drug sharing network based on the known drug-side-effect associations, called DrugNetwork. In DrugNetwork, the vertexes $V = \{d_1, d_2, \ldots, d_n\}$ denote the set of n drugs, the edges represent the drug-drug associations, and the edge weights denote the number of common side-effects shared by corresponding drug pairs. Then we leveraged a graph clustering method “ClusterONE” to identify potential drug clusters in DrugNetwork. According to ClusterONE, the cohesiveness of a cluster C_i is defined as follows:

$$Coh(C_i) = \frac{W_{in}(C_i)}{W_{in}(C_i) + W_{bound}(C_i)} \tag{8}$$

where $W_{in}(C_i)$ is the sum of the edge weights within the cluster C_i, $W_{bound}(C_i)$ denotes total weights of edges which connect vertexes from cluster C_i to the rest of the DrugNetwork. For drug d_j and d_k belong to the same cluster C_i, their similarity value will be optimized as $Simcluster(d_j, d_k) = (1 + Coh(C_i)) * Simcom(d_j, d_k)$, where $Coh(C_i)$ is the cohesiveness of cluster C_i and $Simcom(d_j, d_k)$ is the OCDS between drug d_j and drug d_k. Note for the cluster-based optimized similarity between two different drugs which is equal or great than 1, we normalized it as 0.9999.

B. Optimization by Indirect Drug Similarity Indirect drug similarity refers to similarity values which are not measured by drug properties directly but by existing similarity. In this study, we developed a propagation framework to compute the indirect similarity between each drug pair. Here we leverage a drug pair, i.e., d_j and d_k, to illustrate the process of computing indirect drug similarity. Suppose d_j and d_k both belong to the drug set $D = \{d_1, d_2, \ldots, d_n\}$. Then their indirect drug similarity will be as follows:

$$Simids(d_j, d_k) = \frac{\sum Simcom(d_j, d_i) * Simcom(d_i, d_k)}{n - 2} \tag{9}$$

where $Simcom(d_j, d_i)$ and $Simcom(d_i, d_k)$ are the OCDS between drug d_j and d_i, and d_i and d_k respectively $(1 \leq i \leq n, i \neq j$ and $i \neq k)$.

C. Integration of Optimized Drug Similarity We adopted the following formula to integrate the two different types of drug similarity.

$$Simop(d_j, d_k) = \frac{Simcluster(d_j, d_k) + Simids(d_j, d_k)}{2} \tag{10}$$

where $Simcluster(d_j, d_k)$ and $Simids(d_j, d_k)$ are the cluster-based optimized similarity and indirect similarity.

3.2. Optimization via Building Balanced Drug Training Sets

After careful analysing, we found that the number of labeled drugs for different side-effects is different. Therefore, the training set will be unbalanced if we directly take the labeled and unlabeled drugs as positive and negative samples respectively. However, unbalanced training sets would largely degrade the prediction performance. Thus, to improve the prediction performance, we proposed the following steps to build a balanced drug training set for each side-effect. (a) Obtain the smaller number n_s from the labeled drug number and the unlabeled drug number; (b) Select n_s labeled drugs and n_s unlabeled drugs to form the positive and negative sample set respectively.

4. Results

F1-Score and macro-averaging F1-Score are leveraged to evaluate the prediction performance in this study. To demonstrate the improvement of our method, we evaluated the performance of side-effect prediction based on OCDS and our optimization framework over the 5-folds experiment. Four different classifiers namely KNN (K-Nearest Neighbors), SVM (Support Vector Machine), ELM (Extreme Learning Machine) and RBF (Radial Basis Function) network were employed in the experiment. Related results are illustrated in Figure 2 and Table 1. Figure 2 shows the scatter plots of F1-Scores using the two methods, where x-axis denotes F1-Scores based on OCDS and y-axis denotes F1-Scores based on ODSF. To better visualize the comparison results, we added a reference line “$y = x$” on which F1-Scores are equal to each sub-figure.

It can be seen clear from Figure2 that most dots distribute on the top-left area. It means the prediction performance based on our ODSF outperformed that based on OCDS for most side-effects. To investigate how much improvement is made by our method, we further calculated the macro-averaging F1-Scores of the top 50, top 100, top 200 and all side-effects as listed in Table 1. Clearly, the predictions based on our optimization framework achieved signicantly higher performance than that base on OCDS. For example, for the four classifiers from KNN to ELM, the macro-averaging F1-Score improvement of top-100 side-effects is 6.9%, 18.4%, 19.1% and 9.3%.

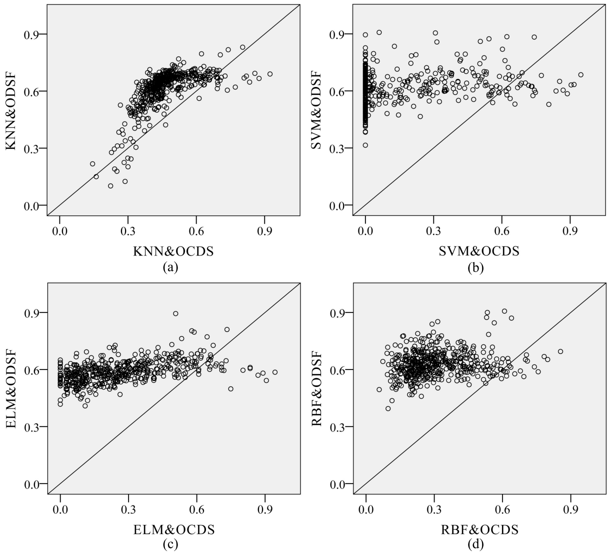

Figure 2. Scatter plots of F1-Scores from predictions based on OCDS and ODSF.

Table 1. Macro-averaging F1-Scores of predictions based on OCDS and ODSF respectively.

Measure	Top 50	Top 100	Top 200	All
KNN & OCDS	0.694	0.634	0.565	0.464
KNN & ODSF	**0.719**	**0.703**	**0.687**	**0.614**
SVM & OCDS	0.672	0.548	0.355	0.305
SVM & ODSF	**0.773**	**0.732**	**0.692**	**0.613**
RBF & OCDS	0.608	0.536	0.447	0.309
RBF & ODSF	**0.761**	**0.727**	**0.692**	**0.629**
ELM & OCDS	0.663	0.579	0.464	0.264
ELM & ODSF	**0.699**	**0.672**	**0.643**	**0.588**

5. Conclusion

In this study, we proposed a method to optimize drug similarity for side-effect prediction. First, four different types of similarities which could measure the similarity between drugs from different perspectives were fused together as OCDS. Then a clustering-based method was adopted to optimize OCDS. Next, indirect drug similarity which could reinforce the direct drug similarity was computed. Finally, the clustering-optimized similarity and indirect similarity were integrated into a unified framework for drug side-effect prediction. Extensive comparison experiments on drugs from DrugBank demonstrate that our optimized similarity based prediction method achieved much better performance than that based on OCDS.

References

[1] Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C. Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys CSUR 2015;47(4):56.

[2] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321(5886):263–266.

[3] Yamanishi Y, Kotera M, Moriya Y, Sawada R, Kanehisa M, Goto S. Dinies: drug–target interaction network inference engine based on supervised analysis. Nucleic acids research 2014;42(W1):W39–W45.

[4] Huang LC, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC genomics 2011;12(5):S11.

[5] Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y. Relating drug–protein interaction network with drug side effects. Bioinformatics 2012;28(18):i522–i528.

[6] Fukuzaki M, Seki M, Kashima H, Sese J. Side effect prediction using cooperative pathways. In IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2009; 142–147.

[7] Atias N, Sharan R. An algorithmic framework for predicting side effects of drugs. Journal of Computational Biology 2011;18(3):207–218.

[8] Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. Journal of chemical information and modeling 2012; 52(12):3284–3292.

[9] Kuhn M, Letunic I, Jensen LJ, Bork P. The sider database of drugs and side effects. Nucleic acids research 2015;gkv1075.

Address for correspondence:

Jinyan Li
15 Broadway Ultimo NSW 2007, Australia
Jinyan.Li@uts.edu.au