A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity

Jiao Li and Zhiyong Lu
National Center for Biotechnology Information (NCBI),
National Institutes of Health (NIH)
Bethesda, USA
lij10@ncbi.nlm.nih.gov, zhiyong.lu@nih.gov

***Abstract*—The traditional *de novo* drug discovery is known as a high cost and high risk process. In response, recently there is an increasing interest in discovering new indications for known drugs—a process known as drug repositioning—using computational methods. In this study, we present a new systematic approach for identifying potential new indications of an existing drug through its relation to similar drugs. Different from the previous similarity-based methods, we adapted a novel bipartite-graph based method when considering common drug targets and their interaction information. Furthermore, we added drug structure information into the calculation of drug pairwise similarity. In cross-validation experiments, our method achieved a sensitivity of 0.77 and specificity of 0.92 (AUC = 0.888) and compared favorably to the state of the art. When compared with a control group of drug uses, our drug repositioning results were found to be significantly enriched in both the biomedical literature and clinical trials. Our results indicate that combining chemical structure and drug target information results in better prediction performance and that the proposed approach successfully captures the implicit information between drug targets.**

Keywords-drug repositioning; bipartite graph; target similarity; chemical similarity; target interaction

I. Introduction

In response to the high cost and risk in traditional *de novo* drug discovery [1], discovering potential new uses for existing drugs, also known as drug repositioning, has attracted increasing interests from both the pharmaceutical industry and research community. Nowadays, the advances in molecular measurements laid a foundation for a surging domain—computational drug repositioning [2]. For instance, with the availability of the Connectivity Map (CMap) [3], a comprehensive reference collection of ranked gene expression profiles produced by different drug candidates, several approaches have been developed to leverage such drug molecular information. Iorio *et al*. used gene expression profiles of drugs in the CMap to compute drug pairwise similarity [4] and the resulting drug-drug network to explore repositioning opportunities for known drugs. Hu *et al.* compared the gene expression profiles of drugs with those of diseases and identified the correlation/anti-correlation between drugs and diseases [5]. They further showed that the anti-correlation relationships in the resulting disease-drug network can suggest new therapeutic uses for existing drugs. In addition to the genomic data, other drug-related information has also been investigated in similarity-based approaches, which assume that similar drugs are indicated for similar diseases. For instance, Campillos *et al.* used drug adverse effects to identify novel drug-target relationships (off-target interactions) which further connected drugs to new uses [6]. Li *et al.* integrated disease, gene/protein and drug connectivity information based on protein interaction networks and literature mining [7]. More recently, Chiang *et al*. presented a 'Guilt by Association' (GBA) approach to predict novel drug uses based on the known treatment relationships between drugs and diseases [8]. In this study, we implemented the GBA method and used it for comparison.

In this study, we proposed a new systematic method to identify a drug's potential new uses through its similar drugs found. Different from other similarity-based methods, we adapted a novel bipartite-graph based method when considering common drug target proteins and their interaction information. By applying it to our data, we were able to boost target similarity by making use of their corresponding interaction information and to obtain target similarity scores for drug pairs in cases where no common targets can be found.

II. Methods

In this study, we identify a target drug d_x*'s* potential new indications through its similar drugs (*e.g.*, d_y) as follows:

> *If two drugs d_x and d_y are found to be similar, and d_y is used for treating disease s, then d_x is a repositioning candidate for disease s treatment.*

When computing pairwise similarity between a drug pair d_x and d_y, we combine the similarities of their chemical structures $SIM_{chem}(d_x, d_y)$ and target profiles $SIM_{target}(d_x, d_y)$.

A. Computing Similarity of Drug Chemical Structures

For each drug pair, we compute the chemical structure similarity $SIM_{chem}(d_x, d_y)$ as the Tanimoto coefficient of their 2D chemical fingerprints $f(d_x)$ and $f(d_y)$:

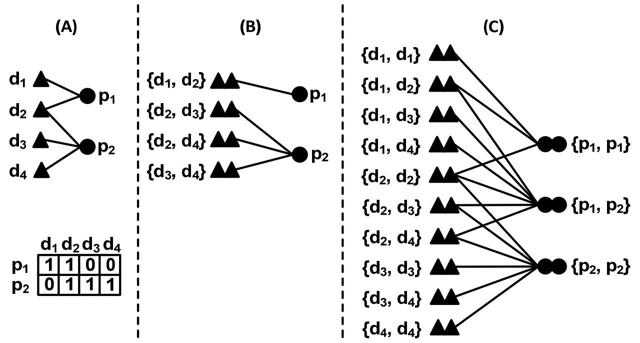
$$SIM_{chem}(d_x, d_y) = \frac{f(d_x) \bullet f(d_y)}{|f(d_x)| + |f(d_y)| - f(d_x) \bullet f(d_y)} \qquad (1)$$

Where, $|f(d_x)|$ and $|f(d_y)|$ are the count of structure fragments drugs d_x and d_y respectively. The dot product $f(d_x)\bullet f(d_y)$ represents the number of structure fragments shared by two drugs.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

B. *Computing Similarity of Drug Target Profiles*

We represent the relationships between drugs and their target proteins as a bipartite graph *G(V, E)* for computing $SIM_{target}(d_x, d_y)$. The node set, *V(G)={D, P}*, consists of two types of object (*i.e.*, the drug set *D* and protein set *P*). The edge set, $E(G)\subseteq D\times P$, consists of relationships between drugs and their target proteins. Fig 1(A) shows an example bipartite graph, where there are four drugs $D=\{d_1, d_2, d_3, d_4\}$, two proteins $P=\{p_1, p_2\}$, and five links (proteins p_1 and p_2 are the targets of drugs $\{d_1, d_2\}$ and $\{d_2, d_3, d_4\}$ respectively). Given a drug *d*, we represent its target protein set as *P(d)*. In this example, $P(d_1)=\{p_1\}$, $P(d_2)=\{p_1, p_2\}$, $P(d_3)=\{p_2\}$, and $P(d_4)=\{p_2\}$. Likewise, we represent a protein's linked drug set as *D(p)*. For instance, $D(p_1)=\{d_1, d_2\}$. Based on this bipartite graph, many methods can be applied to compute $SIM_{target}(d_x, d_y)$. Perhaps the most straightforward approach is to simply count the number of common proteins shared by two drugs *i.e.*, $P(d_x, d_y)=P(d_x)\cap P(d_y)$. As shown in Fig 1(B), drug pairs are only connected if they share common target proteins. This is not ideal because no target protein stands alone in biological systems. With an aim to capture the interactions between target proteins, we derived a graph model G^2 [9] from the bipartite graph *G(V, E)*, as shown in Fig 1(C). Where, the nodes in G^2 are all the possible combinations of drug pairs and protein pairs $V^2=\{D^2,P^2\}=\{D\times D, P\times P\}$. Let $R(d_x, d_y)$ and $R(p_a, p_b)$ denote similarity of drug pairs and protein pair respectively. For self-pairs such as $\{d_1, d_1\}$ and $\{p_1, p_1\}$, their similarity scores are set to be 1. The edges between drug and protein pairs in G^2 are built based on the drug-protein connections in the original bipartite graph *G*. For instant, an edge is established in G^2 between a drug pair $\{d_1, d_2\}$ and protein pair $\{p_1, p_2\}$ because there exist edges $<d_1, p_1>$ and $<d_2, p_2>$ in *G*.



	d1	d2	d3	d4
p1	1	1	0	0
p2	0	1	1	1

Figure 1. Models for computing drug pairwise similarity.

Given the G^2 graph model, we can iteratively compute the pairwise similarity of drug pairs $R_{2k+1}(d_x, d_y)$ and protein pairs $R_{2k+2}(p_a, p_b)$ as follows:

$$\begin{cases} R_{2k+1}(d_x,d_y)=\dfrac{1}{|P(d_x)||P(d_y)|}\displaystyle\sum_{i=1}^{|P(dx)|}\sum_{j=1}^{|P(dy)|}R_{2k}(P_i(d_x),P_j(d_y)) \\ R_{2k+2}(p_a,p_b)=\dfrac{1}{|D(p_a)||D(p_b)|}\displaystyle\sum_{i=1}^{|D(p_a)|}\sum_{j=1}^{|D(p_b)|}R_{2k+1}(D_i(p_a),D_j(p_b)) \end{cases} \quad (2)$$

As can be seen in equation (2), the drug pairwise similarity $R_{2k+1}(d_x, d_y)$ is the average similarity of protein pairs they connected to in the G^2 graph. In turn, the protein pairwise similarity $R_{2k+2}(p_a , p_b)$ also depends on the drug pairwise similarities. The iterative calculation is initialized with the protein pairwise similarity $R_0(p_a , p_b)$ as follows:

$$R_0=\begin{cases} 1 & \text{if a = b} \\ 0.5 & \text{if } p_a \text{ interacts with } p_b \text{ when a} \neq \text{b} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

In theory, the similarity of drug target profiles should be calculated as:

$$SIM_{target}(d_x,d_y)=\lim_{k\to\infty}(R_{2k+1}(d_x,d_y)) \quad (4)$$

It has been reported that the similarity score is rapidly converged, with relative rankings stabilizing within a fixed number of iterations to perform [9]. We have the same observation on our large-scale real-world data (see the Result section for details).

C. *Computing Drug Pairwise Similarity*

The final drug pairwise similarity $SIM(d_x, d_y)$ score is derived by summing up the weighted chemical similarity and target similarity as shown in equation (5), which readily integrates drug chemical structure, drug target and target interaction in one score ranging from 0 to 1.

$$SIM(d_x,d_y)=(1-\lambda)*SIM_{chem}(d_x,d_y)+\lambda*SIM_{target}(d_x,d_y) \quad (5)$$

Where, λ ($0<\lambda<1$) is a predefined constant for weighting the target similarity.

III. Results and Discussion

A. *Experimental Data*

1) Approved Drug List and Target Protein Information: From DrugBank [10], a widely used public database of drug data, we collected 1007 approved small-molecule drugs with their corresponding target protein information.

2) Drug-Disease Treatment Relationships: From the National Drug File - Reference Terminology (NDF-RT) [11], we extracted therapeutic uses for 799 drugs out of the 1007 drugs, which constructed a gold standard set of 3250 treatment relationships between 799 drugs and 719 diseases.

3) Protein-protein Interactions: From the Human Protein Reference Database (HPRD) [12], we collected 39,240 binary interactions between 9673 human proteins.

B. *Test of Our Method Assumption*

In this study, we built our method on the basis that similar drugs are indicated for similar diseases and conditions. To confirm this, we compared the computed pairwise similarities of 4066 drug pairs involved in treating cardiovascular diseases (e.g., both '*Doxazosin*' and

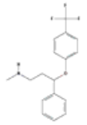
'*Terazosin'* are known to treat hypertension) against 4,000 randomly selected drug pairs. As expected, the drug pairs with similar therapeutic uses have significantly higher chemical and target similarities (t-test P value < 2.2×10^{-16}).

C. *Leave-One-Out Cross Validation*

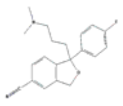
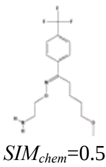
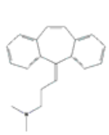
To assess our method in predicting novel indications, we conducted cross-validation experiments in which we used the known treatment relationships between drugs and diseases as the gold standard. Specifically, for each target drug, we removed its known uses and attempted to recover them through its top N similar drugs found. For instance, '*Fluoxetine*' is a drug known to treat 4 different diseases 'Bulimia', 'Depressive Disorder', 'Obsessive-Compulsive Disorder', and 'Panic Disorder' in our gold standard. Table I shows its top 3 most similar drugs found by our method. To measure our prediction performance, we report sensitivity, specificity and positive predictive value (PPV) in this work. In the '*Fluoxetine*' example in Table I (B), the corresponding values for the three metrics are 0.5, 0.99 and 0.33 respectively when only considering the first returned drug, and the performance increase to 0.75, 0.99 and 0.43 when the top 2 drugs are considered.

TABLE I. TOP 3 DRUGS SIMIALR WITH 'FLUOXETINE'

(A)

Given drug	*Chemical Structure*	*Target*
Fluoxetine		**P28223** (5-hydroxytryptamine 2A receptor) **P31645** (Sodium-dependent serotonin transporter)

(B)

	Similar Drug (SIM)	*Drug Chemical Structure (SIM_{chem})*	*Drug Targe (SIM_{target})*	*Original Use*
1	Citalopram *SIM*=0.556	 SIM_{chem}=0.66	P31645 SIM_{targe}=0.53	• Alcoholism • Depressive Disorder • Diabetic Neuropathies • Obsessive-Compulsive Disorder • Tobacco Use Disorder • Dementia
2	Fluvoxamine *SIM*=0.542	 SIM_{chem}=0.59	P31645 SIM_{targe}=0.53	• Depressive Disorder • Obsessive-Compulsive Disorder • Panic Disorder
3	Cyclobenzaprine *SIM*=0.532	 SIM_{chem}=0.54	P28223 SIM_{targe}=0.53	• Myositis • Muscle Rigidity • Pain • Spasm • Muscle Cramp • Muscle Spasticity • Tetanus

To show the performance over the entire dataset of 799 drugs, we calculated overall sensitivity and specificity tradeoffs for different drug pair-wise similarity calculation by varying N—the number of similar drugs—from 1 to 798. The area under the ROC curve (AUC) score was used as the evaluation metric.

1) Comparison of Three Different Ways of Computing Target Similarity: (a) the number of overlapping target proteins *(*$|P(d_x, d_y)|$); (b) Pearson's correlation of drug targets (Pearson); and (c) drug target similarity using the our G^2 method with iterations varying from R_1 to R_9 (see Fig 2). As can be seen, our G^2 method achieved stable performance after 3-5 iterations as the result of the rapid convergence of drug pairwise similarity and stabilized relative ranking. Hereafter, we set SIM_{target} to be $R_5(d_x, d_y)$.

Using target similarity alone, our G^2 method achieved higher AUC score (0.876) than using Pearson's correlation (0.842) or simply counting the overlap (0.838). This indicates that our method is able to better capture interactions between target proteins through iteratively propagating similarities from protein pairs to drug pairs, and vice versa.

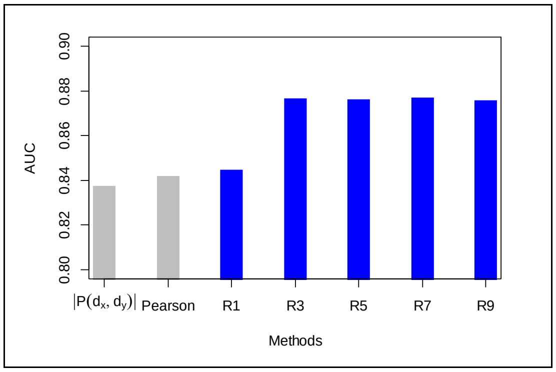


Figure 2. Comparison of different target similarity calculation methods.

2) Combination of Chemical Similarity and Target Similarity: We assessed the performance of combining target similarity ($R_5(d_x, d_y)$) with chemical similarity. By experimenting with different values (from 0 to 1) of the weight parameter λ in equation 5, we observed the highest performance (AUC=0.888) when λ = 0.8. This confirmed our hypothesis that the two similarities can complement each other in identifying similar drugs. We show in Fig 3 the overall performance of our method with respect to the number of top-ranked similar drugs returned in a ROC curve. As highlighted in Fig 3, when N (the number of most similar drugs returned) was 20, our method achieved a specificity of 0.92 and sensitivity of 0.77.

3) Comparison with the state of the art: We implemented the guilt-by-association (GBA) approach [8] and evaluated it on our data. As shown in Fig 3, the GBA approach yielded a specificity of 0.85 and sensitivity of 0.74, which is below the ROC curve of our method. Not only does our method outperform the GBA approach, it is also able to rank its prediction results (the GBA approach cannot), an important feature for prioritizing drug repositioning candidates in practice.

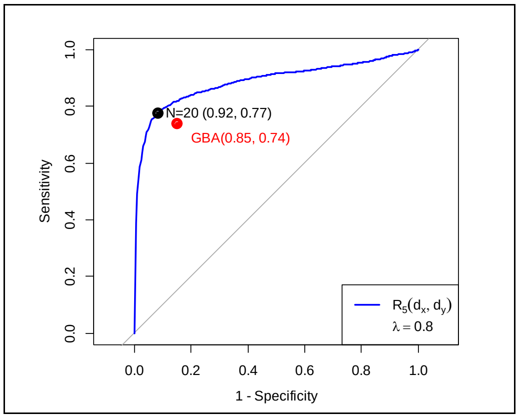


Figure 3. ROC curve of our method with combining SIM_{chem} and SIM_{target} with respect to different N (iteration $R_5(d_x, d_y)$, weight λ = 0.8)

D. *Analysis of Novel Predictions in Clinical Trials and the Biomedical Literature*

In addition to cross validation, we further evaluated the validity of our novel drug use prediction by searching the predicted drug-disease pairs against the trials in ClinicalTrials.gov and scientific abstracts in PubMed. Take the drug '*Fluoxetine*' for example (see Table I). As stated above, our method would predict 6 indications based on its most similar drug '*Citalopram*'. Two of the predicted uses are known uses, thus leaving the other 4 as novel predictions. When searching for their evidence, we found that the 'Alcoholism' use is indicated in a clinical trial (NCT00027378) which was conducted to study *Fluoxetine* in treatment adolescents with alcohol use disorder and major depression and that the other three uses have been investigated with study results published in the literature [13-15].

When setting λ = 0.8 and N = 20 (best performance obtained in cross-validation experiments), our method predicted 30,872 novel indications for the 1,007 drugs. 1,340 of these predictions can be found in clinical trials. As a matter of fact, it is 5 times more likely for our predicted uses to be found in a trial than those drug uses not predicted by our method (Chi^2 test P value < 2.2×10^{-16}). In addition, 8,564 (~30%) of the predicted novel uses can be found in the literature. Hence, we conclude that the novel uses predicted by our method are significantly enriched in both scientific literature and clinical trials.

IV. Conclusions and Future Work

Computational drug repositioning offers promise for discovering new uses of existing drugs, as drug related molecular, chemical, and clinical information has increased over the past decade and become broadly accessible. In this study, we developed a systematic method for mining potential new drug indications by exploring both chemical and molecular features in similar drugs. The proposed bipartite graph model successfully boosted target similarity by iteratively integrating explicit evidence (common target proteins shared by drugs) and implicit evidence (common drugs shared by target proteins).

Our method has some limitations. First, our method relies on existing knowledge of drugs, targets, protein interactions. The incompleteness of such information would limit our prediction power. Second, our method would fail to identify any reusable drugs for a disease if no current treatment is available for that disease. This is because our predicted indications are based on the known uses of other drugs. Lastly, in this work we limit our method to only the approved small molecules with known target proteins. This would exclude some drugs which is not a small molecular (e.g., *Rituximab*) or whose protein targets are not known yet (e.g., *Mannitol*). We plan to investigate these issues in future.

Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The authors would like to thank Profs. Xiaoyan Zhu and Jake Chen for their valuable discussion at the beginning of this work, and thank Dr. W. John Wilbur for his helpful comments and proofreading.

References

[1] T.T. Ashburn and K.B. Thor, Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov, 2004. **3**(8): p. 673-83.

[2] A.F. Shaughnessy, Old drugs, new tricks. BMJ, 2011. **342**: p. d741.

[3] J. Lamb, E.D. Crawford, D. Peck, et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 2006. **313**(5795): p. 1929-1935.

[4] F. Iorio, R. Bosotti, E. Scacheri, et al., Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A, 2010. **107**(33): p. 14621-6.

[5] G. Hu and P. Agarwal, Human disease-drug network based on genomic expression profiles. PLoS One, 2009. **4**(8): p. e6536.

[6] M. Campillos, M. Kuhn, A.C. Gavin, et al., Drug target identification using side-effect similarity. Science, 2008. **321**(5886): p. 263-6.

[7] J. Li, X. Zhu, and J.Y. Chen, Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol, 2009. **5**(7): p. e1000450.

[8] A.P. Chiang and A.J. Butte, Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther, 2009. **86**(5): p. 507-10.

[9] G. Jeh and J. Widom, SimRank: a measure of structural-context similarity. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002: p. 538-543.

[10] D.S. Wishart, C. Knox, A.C. Guo, et al., DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.

[11] National Drug File - Reference Terminology (NDF-RT). http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/.

[12] T.S. Keshava Prasad, R. Goel, K. Kandasamy, et al., Human Protein Reference Database--2009 update. Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.

[13] M.B. Max, S.A. Lynch, J. Muir, et al., Effects of desipramine, amitriptyline, and fluoxetine on pain in diabetic neuropathy. N Engl J Med, 1992. **326**(19): p. 1250-6.

[14] K.K. Saules, L.M. Schuh, C.L. Arfken, et al., Double-blind placebo-controlled trial of fluoxetine in smoking cessation treatment including nicotine patch and cognitive-behavioral group therapy. Am J Addict, 2004. **13**(5): p. 438-46.

[15] A. Mowla, M. Mosavinasab, H. Haghshenas, et al., Does serotonin augmentation have any effect on cognition and activities of daily living in Alzheimer's dementia? A double-blind, placebo-controlled clinical trial. J Clin Psychopharmacol, 2007. **27**(5): p. 484-7.