DATA SCIENCE JOURNAL

Li, J and Lu, Z 2015 An Integrative Approach for Discovery of New Uses of Existing Drugs. *Data Science Journal,* 14: 9, pp. 1–11, DOI: http://dx.doi.org/10.5334/dsj-2015-009

PROCEEDINGS PAPER

An Integrative Approach for Discovery of New Uses of Existing Drugs

Jiao Li[1] and Zhiyong Lu[2]

[1] Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China
li.jiao@imicams.ac.cn

[2] National Center for Biotechnology Information (NCBI), National Institutes of Health (NIH), Bethesda, MD 20894, USA
zhiyong.lu@nih.gov

The discovery of new uses of existing drugs offers the possibility to reduce time and risk because the approved drugs have passed several phases along the drug development pipeline. In this study, we present a computational method for novel drug use prediction based on the idea that similar drugs are indicated for similar diseases. When computing drug pairwise similarities, we considered both chemical structure and drug target similarities. In validation, our new drug use predictions were found to be significantly enriched in both the biomedical literature and clinical trials. These results indicate that our method is able to successfully integrate both biomedical scientific data and literature for drug discovery.

Keywords: Biomedical scientific data; Biomedical literature; Integrative mining; Target similarity; Chemical similarity; Target interaction

1 Introduction

In today's drug discovery, the number of new drugs discovered each year has not kept pace with the enormously increasing investment in pharmaceutical R&D. A recent report shows that the number of new drugs approved per billion US dollars has halved approximately every 9 years since 1950 (Scannell et al., 2012). In response to this, finding new uses for existing drugs (known as drug repositioning) has been proposed, which offers opportunities for faster development times and reduced risk (Ashburn & Thor, 2004). This is because the repositioning candidates should already have passed through development stages and efficacy/toxicity tests for their original indications. Many repositioning success stories offer great promise for the feasibility and effectiveness of the drug repositioning strategy. For example, GlaxoSmithKline received approval to market *bupropion hydrochloride* branded as WELLBUTRIN® for depression in 1985 and as ZYBAN® for smoking cessation in 1997 (The U.S. FDA, 2013). Although repositioning existing drugs for alternative indications is not new, it is only recently that large-scale computational methods are being developed and used (Shaughnessy, 2011).

Computational drug repositioning has become a new frontier (Dudley et al., 2011; Hurle et al., 2013; Lu et al., 2013) in today's drug discovery research. Recent methods focus on systematically exploring novel drug-disease therapeutic relationships from large-scale molecular data, such as transcriptomics, genome-wide association study (GWAS), and target screening data. For instance, with the availability of the Connectivity Map (CMap) (Lamb et al., 2006), which is 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 and the resulting drug-drug network to explore repositioning opportunities for known drugs (Iorio et al., 2010). Hu and Agarwal (2009) compared the gene expression profiles of drugs with those of diseases and identified the correlation/anti-correlation between drugs and diseases. 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. (2008) used drug adverse effects to identify novel drug-target relationships (off-target interactions), which further connected drugs to new uses. Li et al. (2009) integrated disease, gene/protein and drug connectivity information based on protein interaction networks and literature mining. Chiang and Butte (2009) presented a 'Guilt by Association' (GBA) approach to predict novel drug uses based on the known treatment relationships between drugs and diseases. Gottlieb et al. (2011) developed a computational method called PREDICT where the drug pairwise similarity was measured by similarities of chemical structures, side effects, and drug targets. These computed similarities were then used as features of a logistic regression classifier for predicting the novel associations between drugs and diseases. Li et al. (2013) built a causal network (CauseNet)–a layered drug-target-pathway-gene-disease causal inference network–to identify new therapeutic uses of existing drugs.

In this paper, we describe our previous approach in more detail (Li & Lu, 2012) for identifying new uses of an existing drug through its relationship to similar drugs (see **Figure 1**), along with additional experimental results. More specifically, we represent the relationships between drugs and their target proteins as a bipartite graph. As shown in **Figure 1**, drug d_1 is known for treating disease s_1 and d_2 for s_2. If the drug pair (d_1, d_2) obtains a high similarity score, we predict that they can be repositioned into each other's therapeutic area. That is, drug d_1 is predicted for disease s_2 treatment and d_2 for s_1. In order to validate our predictions, we perform a cross-validation experiment by comparing the predicted drug-disease pairs against known drug uses. In addition, we search evidence from both published biomedical literature and current clinical trials to support our predictions.

Our method is most related to the GBA and PREDICT approach mentioned above in that they all identify a drug's potential new uses through its similarity to existing drugs. Different from the GBA approach that

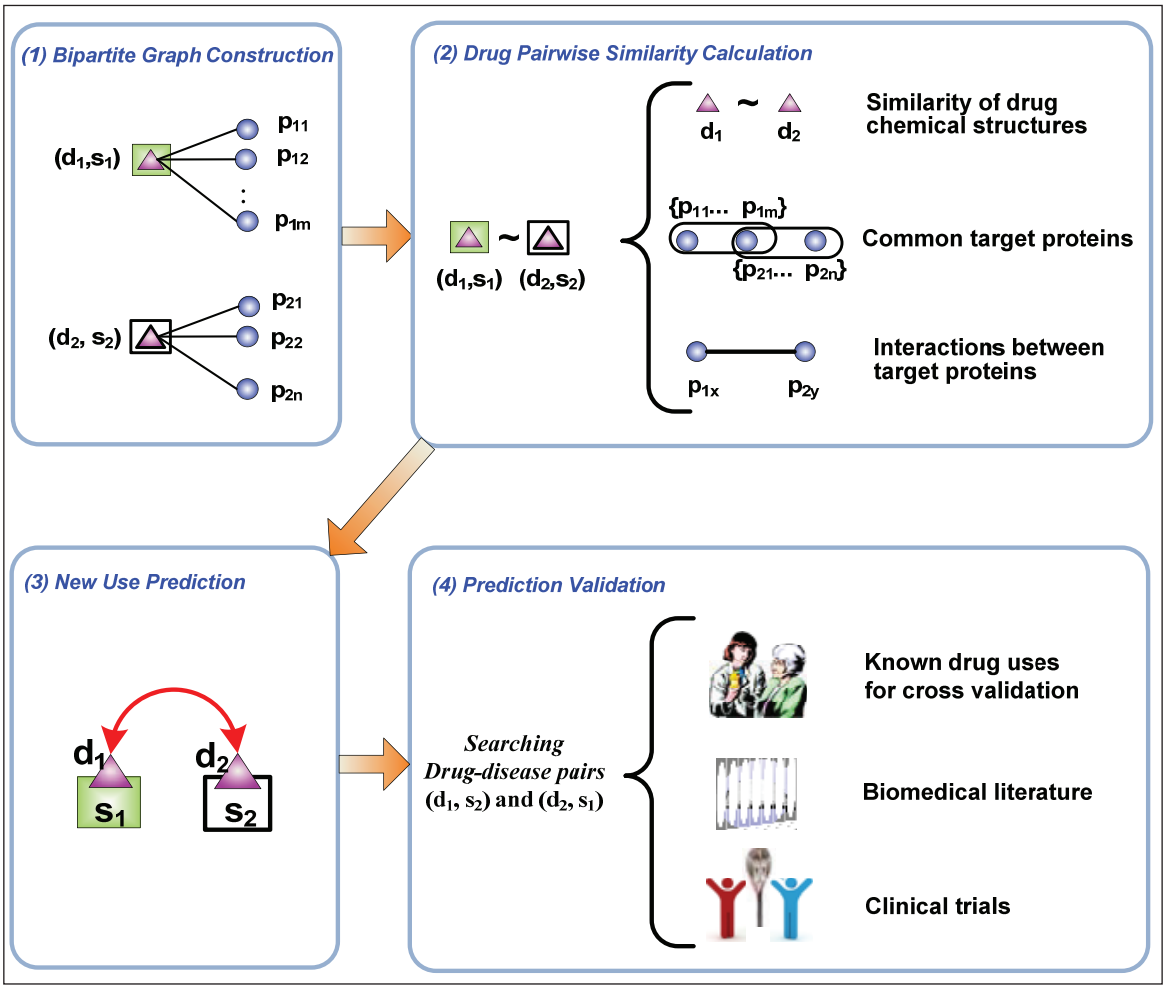


Figure 1: Overview of the bipartite graph-based method for drug repositioning.

relied on a single measure, both PREDICT and our method use multiple features (e.g., drug chemical structure and target profile) in computing drug pairwise similarity. More importantly, unlike PREDICT and other similarity-based methods, we adopted a novel bipartite-graph based method when considering common drug target proteins and their interaction information. This method assumes that two objects are similar if they are related to similar objects. By applying it to our data, we are 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. In other words, this approach empowers us to take advantage of such information as indirect protein interaction that is implicitly embedded in complex biological systems. Lastly, we differ from both approaches with respect to the dataset assembled. The drug-disease connections in this study were obtained only from public sources (GBA used a private source) and include various kinds of diseases (only OMIM diseases were included in PREDICT).

2 Methods

In our method, a drug's potential new indications are identified via its similar drugs. For example, two drugs d_x and d_y are found to be similar, and d_x is known to be used for treating disease *s*; thus d_y is potentially useful for disease *s* treatment. In drug discovery, a drug's chemical structure and its target profile are two important features evidently associated with its therapeutic use. Hence, 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 $SIM_{target}(d_x, d_y)$.

2.1 Similarity of Drug Chemical Structures

Our method for calculating the drug chemical structure similarity $SIM_{chem}(d_x, d_y)$ is based on the 2D chemical fingerprint descriptor of each drug's chemical structure in PubChem (Wang et al., 2009; Li et al., 2010). That is, each drug is represented by a binary fingerprint $f(d_x)$ in which each bit indicates the presence of a predefined chemical structure fragment. The pairwise chemical similarity between two drugs d_x and d_y is computed as the Tanimoto coefficient of their fingerprints:

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

where $|f(d_x)|$ and $|f(dy)|$ are the number of structure fragments drugs d_x and d_y respectively $f(d_x)$. $f(d_y)$, the dot product of fingerprints, is the number of structure fragments shared by two drugs.

2.2 Similarity of drug target profiles

Our method for calculating the drug target similarity $SIM_{target}(d_x, d_y)$ is based on both common target proteins and interactions between target proteins. The relationships between drugs and their target proteins can be represented as a bipartite graph G(V, E):

The node set of graph G, $V(G) = \{D, P\}$, consists of two types of object (i.e., the drug set *D* and protein set *P*).

The edge set of graph G, $E(G) \subseteq D \times P$, consists of relationships between drugs and their target proteins.

Given a drug *d*, its target protein set is noted as *P(d).* Likewise, a protein's linked drug set is noted as *D(p).*

Figure 2(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). In this example, $P(d_1) = \{p_1\}$, $P(d_2) = \{p_1, p_2\}$, $P(d_3) = \{p_2\}$, and $P(d_4) = \{p_2\}$; while $D(p_1) = \{d_1, d_2\}$ and $D(p_2) = \{d_2, d_3, d_4\}$.

Based on the bipartite graph, drug target similarity $SIM_{target}(d_x, d_y)$. can be computed by counting the number of common proteins shared by two drugs *i.e.*, $P(d_x, d_y) = P(d_x) \cap P(d_y)$. **Figure 2(B)** shows a bipartite graph *G'* where drug pairs are only connected if they share common target proteins. This is not ideal because no target protein stands alone in biological systems.

For better capturing the interactions between target proteins, we derived a bipartite graph model $G^2(V^2, E^2)$:

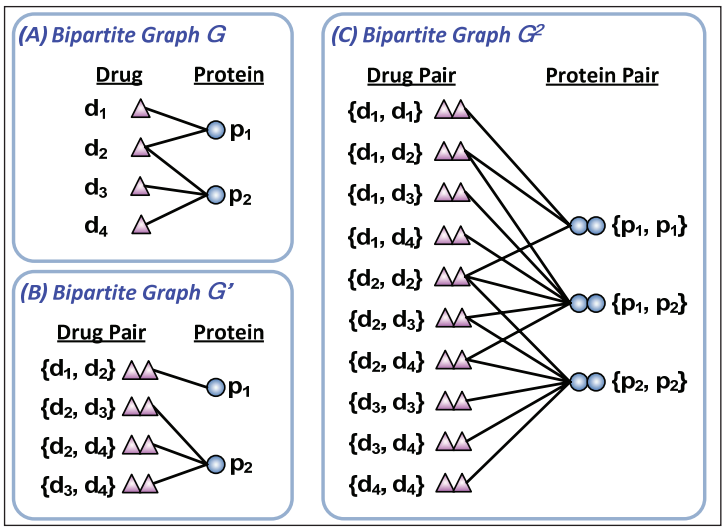


Figure 2: Bipartite graph models for computing drug target similarity.

The node set of graph G^2, $V^2 = \{D^2, P^2\} = \{D \times D, P \times P\}$, consists of two types of object (i.e., the drug pair and protein pair). Let $R(d_x, d_y)$ and $R(p_a, p_b)$ represent the similarity of drug pair and protein pair, respectively.

The edge set of graph G^2, $E^2 \subseteq D^2 \times P^2$, represents connections between drug pairs and drug pairs, which are derived from the edges in the original bipartite graph G.

Figure 2(C) shows the bipartite graph G^2 derived from G. The drug pair set contains all possible combinations of any two drugs including self-pairs (e.g., $\{d_1, d_1\}$ and $\{d_2, d_2\}$). Similarly, the protein pair set contains all possible protein combinations. An edge exists in G^2 between a drug pair $\{d_1, d_2\}$ and protein pair $\{p_1, p_2\}$ if and only if their respective edges $<d_1, p_1>$ and $<d_2, p_2>$ exist in G. Such a G^2 graph can capture a common target via edges between non-self-drug pairs and self-protein pairs (e.g., the edge between $\{d_1, d_2\}$ and $\{p_1, p_1\}$). Also, it can capture the interaction information between two proteins via the node set of protein pairs.

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:

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

As can been seen in equation (2), the drug pairwise similarity $R_{2k+1}(d_x, d_y)$ depends on the similarities of protein pairs that are connected to the drug pair (d_x, d_y) 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 $R_0(p_a, p_b)$:

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

where $R_0(p_a, p_b)$ is set as 1 if the pair is self-paired (i.e., a = b) and is set as 0.5 if protein p_a interacts with p_b.

To demonstrate our G^2 graph model, we use the example data in **Figure 2** and assume the two proteins p_1 and p_2 interact with each other. In **Table 1**, we show comparative results $SIM_{target}(d_x, d_y)$ using the proposed G^2 method against the simple method of counting the number of common target proteins $|P(d_x, d_y)|$ and its variant using Pearson's correlation. As can be seen, using either Pearson's correlation or the proposed G^2 method allows one to capture and compare the different strengths of drug similarity. For instance, both

methods find that the drug pair (d_3, d_4) has the highest similarity as they share the exactly same target protein p_2. Moreover, the G^2 method is able to consider the fact that p_1 interacts with p_2 and produces similarity scores accordingly for the two drug pairs (d_1, d_3) and (d_1, d_4) that are assigned with zero similarity otherwise by the two other methods.

Also, one can see from **Table 1** that the similarity scores in our G^2 method are monotonically increasing as k becomes larger. This is because of the propagation of similarities from protein pairs to drug pairs and vice versa. For example, the fact that proteins p_1 and p_2 share one common drug d_2 would contribute to the similarity of protein pair (P_1, P_3) , which in turn would further increase the similarity between drugs that share target proteins p_1 and/or p_2.

	(d_1, d_2)	(d_1, d_3)	(d_1, d_4)	(d_2, d_3)	(d_2, d_4)	(d_3, d_4)
$\lvert P(d_x, d_y)\rvert$	1.00	0.00	0.00	1.00	1.00	1.00
Pearson	0.71	0.00	0.00	0.71	0.71	1.00
$R_1(d_x, d_y)$	0.75	0.50	0.50	0.75	0.75	1.00
$R_3(d_x, d_y)$	0.85	0.71	0.71	0.85	0.85	1.00
$R_5(d_x, d_y)$	0.91	0.83	0.83	0.91	0.91	1.00
$R_7(d_x, d_y)$	0.95	0.90	0.90	0.95	0.95	1.00

Table 1: Comparison of target similarities calculated by different methods.

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

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

Because of the rapid convergence with relative rankings stabilizing as discussed in Jeh and Widom, (2002), we set $S_{\text{target}}(d_x, d_y) = R_5(d_x, d_y)$ when performing this iterative method on large-scale real data.

2.3 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 Eq. (5), which readily integrates drug chemical structure, drug target, and target interaction in one score ranging from 0 to 1.

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

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

2.4 Evaluation of Repositioning Candidates

To assess our method, we first compare the repositioning candidates and their predicted uses with their known uses extracted from the National Drug File-Reference Terminology (NDF-RT). Second, we check evidence of our predictions in published literature and undergoing investigations, respectively. More specifically, given a drug d_x and its predicted uses S_x={s_{x1}, s_{x2}, ... }, we search for the occurrence of drug-disease pair (d_x, s_{xi}) in PubMed and ClinicalTrials.gov. For literature validation, we require the drug-disease pair (d_x, s_{xi}) to be co-mentioned in more than two PubMed abstracts. For trial validation, if the drug-disease pair (d_x, s_{xi}) is co-mentioned in a clinical trial, we would conclude that the drug d_x is being investigated for disease s_{xi}.

3 Materials

The essential information involved in our study includes approved drug uses, drug chemical structures, target proteins, and protein interactions. We collected and integrated all these different types of information from publicly accessible resources.

3.1 Approved Drug List and Target Protein Information

From DrugBank (Wishart et al., 2008), a widely used public database of drug data, we collected 1007 approved small-molecule drugs with their corresponding target protein information. Furthermore, we mapped these drugs to several other key drug resources including RxNorm, PubChem (Wang et al., 2009; Li et al., 2010),

and UMLS in order to extract other drug related information. For instance, we extracted chemical structures of the 1007 drugs from PubChem and used its Score Matrix Service to calculate chemical similarity scores for the 1007*1007 drug pairs. To facilitate collecting target protein information, we mapped target proteins to UniProt Knowledgebase (Magrane & Consortium, 2011), a central knowledge base including most comprehensive and complete information on proteins. In the end, we extracted 3,152 relationships between 1,007 drugs and 775 proteins.

3.2 Drug-Disease Treatment Relationships

We obtained a drug's known use(s) through extracting treatment relationships between drugs and diseases from the National Drug File-Reference Terminology (NDF-RT), which is part of the NLM's Unified Medical Language System (UMLS). One issue of the NDF-RT data set is lack of the management of drug name variants. For instance, disease 'Breast Neoplasm' can be treated by the drugs 'Tamoxifen', 'FULVESTRANT 50MG/ML INJ, SYRINGE, 5ML', and 'CAPECITABINE 150MG TAB'. We overcame this issue by normalizing various drug names to their active ingredients and subsequently mapping ingredient names to unique concept identifiers in UMLS. As a result, the normalized treatment relationships in the above example were 'Tamoxifen'-'Breast Neoplasm', 'Fulvestrant'-'Breast Neoplasm', and 'Capecitabine'-'Breast Neoplasm'. From the normalized NDF-RT data set, we were able to extract therapeutic uses for 799 drugs out of the 1007 drugs, which constructed a gold standard set of 3,250 treatment relationships between 799 drugs and 719 diseases.

3.3 Protein-Protein Interactions

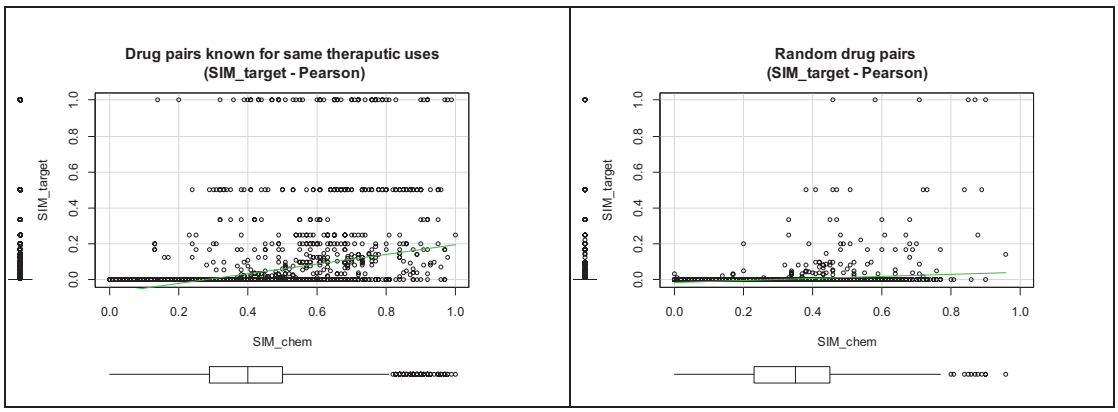
We extracted protein-protein interaction information from the Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009), which contains curated proteomic information pertaining to human proteins. In this study, we used 39,240 binary interactions between 9,673 human proteins in HPRD.

4 Results

4.1 Drug Pairs Known for the Same Therapeutic Uses

In this study, we built our method on the basis that similar drugs are indicated for similar diseases and conditions. To confirm this and to show the strength of our proposed method in boosting the target similarity, we took 177 cardiovascular drugs from our data (e.g., '*Doxazosin'* and *'Terazosin'* are known to treat hypertension) and compared their pairwise chemical/target similarities with those of 4,000 randomly selected drug pairs. In **Figure 3A**, we show the chemical similarity (SIM_{chem}) and target similarity (SIM_{target} computed by Pearson's correlation) for the 4066 drug pairs known for treating cardiovascular diseases. As a comparison, we show in **Figure 3B** the similarities of 4,000 randomly selected drug pairs. It is clear that compared to the random pairs, the drug pairs with similar therapeutic uses have significantly enriched chemical similarity and target similarity (t-test P value $< 2.2 \times 10^{-16}$).

In addition to using Pearson's correlation for computing the target similarity, we show in **Figure 4A** and **4B** two similar scatter plots using the proposed G^2 method. By comparison, we can see that our method significantly boosts (SIM_{target}) of drugs for the same therapeutic uses (t-test P value = 7.67×10^{-8}) (**Figure 4A** vs. **3A**)



(A) **(B)**

Figure 3: Scatter plots of ($SIM_{chem.}$) and ($SIM_{target.}$) computed by Pearson's correlation.

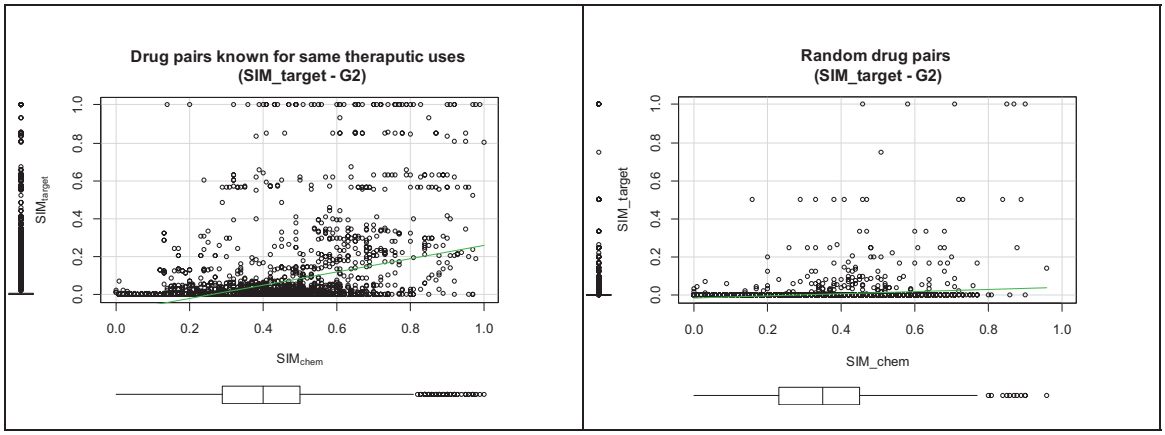


Figure 4: Scatter plots of (SIM_{chem}) and (SIM_{target}) computed by our G^2 method.

while having no significant effects on random pairs (t-test P value = 0.21) (**Figure 4B** vs. **4B**). This suggests that our G^2 method works selectively by only boosting the similarity of related drugs.

4.2 Cross Validation Using Known Drug Uses

To assess our method in predicting novel indications, we used the known therapeutic uses of 799 drugs as the gold standard (see Section 3.2). For each drug, we removed its known uses and attempted to recover them through its top N similar drugs found by our method. To show the performance over the entire dataset of 799 drugs, we plotted ROC curves using both sensitivity and specificity. Five plots are shown in **Figure 5**, each of which represents a different strategy in measuring the drug pairwise similarity depending on: 1) the number of overlapping target proteins ($|P(d_x, d_y)|$), 2) Pearson's correlation of drug targets (Pearson), 3) drug target similarity using the our G^2 method (SIM_{target}), 4) solely chemical structure similarity (SIM_{chem}), and 5) the linear combination of SIM_{chem} and SIM_{target} ($\lambda = 0.8$ is empirically determined). We calculated overall sensitivity and specificity trade-offs by varying N, the number of similar drugs, from 1 to 798. As can be seen, our combination method achieved the best performance with an area under the ROC curve (AUC) of 0.888, better than relying on drug target profile (best AUC = 0.876) or chemical structure similarity (AUC = 0.852) alone. Furthermore, we see that when only using drug target profiles, the performance of our G^2 method was substantially higher (AUC of 0.876) than using Pearson's correlation (AUC of 0.842) or simply counting the overlap (AUC of 0.838). Such results suggest 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.

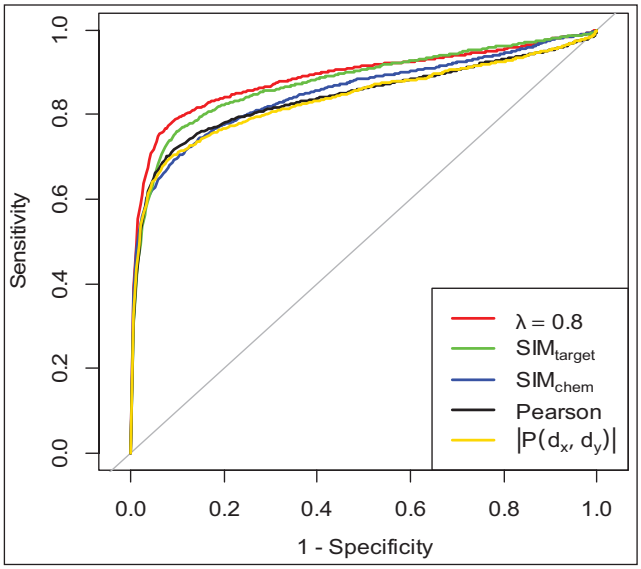


Figure 5: ROC curves of different drug-pairwise similarity strategies.

We compared our method with the guilt-by-association (GBA) (Chiang & Butte, 2009) and PREDICT methods (Gottlieb et al., 2011). The GBA approach assumes that if two diseases share similar therapies, then other drugs that are currently used for only one of the two diseases may also be therapeutic for the other. We applied the GBA approach to the 799 drugs and their known uses in NDF-RT in our data. The GBA approach obtained a sensitivity of 0.74 and specificity of 0.85, which is below the red ROC curve if plotted in **Figure 5**. By comparison, the best cut-off point on the red curve (our combination point) corresponds to a sensitivity of 0.77 and specificity of 0.92 (N = 20), respectively. 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. Gottlieb et al.'s (2011) study evaluated drug use prediction through cross validation on a gold standard set of 1933 associations between 593 drugs and 313 diseases. As reported in Gottlieb et al., (2011), they obtained an AUC of 0.90. For direct comparison, we applied our method to their data and achieved a comparable AUC of 0.89.

4.3 Evidence Validation Using Clinical Trials and Biomedical Literature

After 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. For example, given a drug '*Fluoxetine*', our method would predict 6 indications based on its most similar drug '*Citalopram*'. Two of the predicted uses are known uses (i.e., 'Depressive Disorder' and 'Obsessive-Compulsive Disorder'), thus leaving the other 4 as novel predictions: 'Alcoholism', 'Diabetic Neuropathies', 'Tobacco Use Disorder', and 'Dementia'. 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 treating adolescents with alcohol use disorder and major depression and that the other three uses have been investigated with study results published in the literature (Max et al., 1992; Saules et al., 2004; Mowla et al., 2007). **Table 2** shows 5 examples of novel drug uses predicted by our method and similar drugs supporting these predictions in our method as well as their supporting statements in trials/publications.

Drug	Novel use	Similar drugs supporting this prediction	Evidence in clinical trial/literature
Fluoxetine	Alcoholism	**Citalopram** SIM_{chem} = 0.66 SIM_{target} = 0.53 (Common target P31645 (Sodium-dependent serotonin transporter))	**NCT00027378:** Study fluoxetine (Prozac) versus a placebo in the treatment of adolescents with alcohol use disorder and major depression
Ramipril	Rheumatoid Arthritis	**Enalapril** SIM_{chem} = 0.92 SIM_{target} = 1 (Common target P12821 (Angiotensin-converting enzyme))	**NCT00273533:** Evaluate that Ramipril improves vascular function and reduces markers of low-grade chronic inflammation and oxidative stress in patients with Rheumatoid Arthritis
Sildenafil	Brain Ischemia	**Pentoxifylline** SIM_{chem} = 0.38 SIM_{target} = 0.19 (Common target O76074 (cGMP-specific 3',5'-cyclic phosphodiesterase))	**PMID 20436396:** Study the therapeutic effect of sildenafil citrate on cerebral vasospasm in rat model
Carbidopa	Prostatic Neoplasms	**Flutamide** SIM_{chem} = 0.42 SIM_{target} = 0.33 (Interaction between Carbidopa's target P20711(Aromatic-L-amino-acid decarboxylase) and Flutamide's target P10275 (Androgen receptor))	**PMID 16895983:** Treatment of nude mice containing prostate neuroendocrine cancer with carbidopa plus amiloride and flumazenil leading to significant reductions in tumor growth
Pimecrolimus	Graft-versus-host disease (GVHD)	**Tacrolimus** SIM_{chem} = 0.89 SIM_{target} = 0.78 (Common target P62942 (FK506-binding protein 1A); plus interaction between Pimecrolimus's target P42345 (Serine/threonine-protein kinase mTOR) and the common target P62942)	**PMID 20723118:** Treatment of disfiguring chronic GVHD in a child with topical pimecrolimus: case report

Table 2: Examples of repositioned drugs predicted by our method.

When setting $\lambda = 0.8$ and N = 20 (best performance obtained in cross-validation experiments), our method predicted 30,872 novel indications for the 1007 drugs. 8,564 (~30%) of the predicted novel uses can be found in the literature. In addition, 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 $< 2.2 \times 10^{-16}$). Hence, we conclude that the novel uses predicted by our method are significantly enriched in both scientific literature and clinical trials.

5 Discussion and Conclusion

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, our method was developed based on the hypothesis that a drug can be repositioned to another drug's therapeutic area if two drugs share similar molecular and/or chemical properties. We confirmed this by comparing drug pairs with similar therapeutic uses vs. randomly selected pairs, as shown in **Figures 3** and **4**. From the same set of figures, we can also see that although target similarity somewhat correlates with chemical similarity (correlation coefficients ~ 0.3), many drug pairs with similar therapeutic uses share common targets but do not have similar chemical structures and vice versa. This suggests that either similarity may play its own role in finding similar drugs. Indeed, as shown by our results (**Figure 4**), using either one can already result in good performance. Moreover, **Figure 4** shows that a relatively higher AUC score was obtained with $\lambda = 1$ (i.e., using only target similarity) vs. $\lambda = 0$ (i.e., using only chemical structure similarity), which suggests weighting the former higher than the latter when combining the two. Indeed, we found empirically that the best performance was achieved when λ was set to be 0.8 on our data, confirming our belief that the two similarities can complement each other in identifying similar drugs.

According to **Figure 5**, overall the proposed bipartite graph based method produced significantly better results than the baseline method of considering the overlap of common drug targets (our AUC = 0.876 vs. overlap AUC = 0.838). In particular, when no common target protein exists between two drugs, this method became critical in establishing the target similarity. For instance, as shown in **Table 2**, the predicted drug use (Prostatic Neoplasms) for '*Carbidopa*' would not be found if only common target proteins were considered.

Our method shares some of the same limitations as other drug repositioning methods. First, our method relies on existing knowledge of drug-disease, drug-target, and protein-protein relationships. Unfortunately, such information is currently incomplete from existing resources, thus limiting the prediction power of our method. Second, like any similarity-based approach, 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. Hence, this excludes some drugs that are not a small molecular (e.g., *Rituximab*) or whose protein targets are not known yet (e.g., *Mannitol*).

In conclusion, 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). Furthermore, we found evidence from the literature and clinical trials for many of the novel indications predicted by our method. Note that with significantly fewer features, we were able to obtain similar results to PREDICT. It is possible that adding additional features such as side effects, gene sequences, and disease phenotypes could further improve our performance. We plan to investigate this issue in future work.

6 Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, the Fundamental Research Funds for the Central Universities (Grant No. 13R0101), and the National Key Technology Research and Development Program of China (Grant No. 2013BAI06B01). 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.

7 References

Ashburn, T.T. & Thor, K.B. (2004) Drug repositioning: identifying and developing new uses for existing drug. *Nature Reviews Drug Discovery 3,* pp 673–683.

Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., & Bork, P. (2008) Drug target identification using side-effect similarity. *Science 321*, pp 263–266.

Chiang, A.P. & Butte, A.J. (2009) Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics 86*, pp 507–510.

Dudley, J.T., Deshpande T., & Butte, A.J. (2011) Exploiting drug-disease relationships for computational drug repositioning. *Briefings in Bioinformatics 12*, pp 303–311.

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*, p 496.

Hu, G. & Agarwal P. (2009) Human disease-drug network based on genomic expression profiles. *PLoS One 4*, p e6536.

Hurle, M.R., Yang, L., Xie, Q., Rajpal, D.K., Sanseau, P., & Agarwal P. (2013) Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics 93*, pp 335–341.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences of the United States of America 107*, pp 14621–14626.

Jeh, G. & Widom, J. (2002) SimRank: a measure of structural-context similarity, *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pp 538–543.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S. Mathivanan, S. et al. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Research 37*, pp D767–772.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science 313*, pp 1929–1935.

Li, J. & Lu, Z. (2012) A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity, *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM).*

Li, J. & Lu, Z. (2013) Pathway-based Drug Repositioning Using Causal Inference*. BMC Bioinformatics. 2013. 14(Suppl 16):S3 doi:10.1186/1471-2105-14-S16-S3.*

Li, J., Zhu, X., & Chen J.Y. (2009) Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Computational Biology 5*, p e1000450.

Li, Q., Cheng T., Wang Y., & Bryant, S.H. (2010) PubChem as a public resource for drug discovery. *Drug Discovery Today 15*, pp 1052–1057.

Lu, Z., Agarwal, P., & Butte, A.J. (2013) Computational drug repositioning. *Pacific Symposium on Biocomputing*, pp 1–4.

Magrane, M. & Consortium, U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford) 11,* p bar009.

Max, M.B., Lynch, S.A., Muir, J., Shoaf, S.E., Smoller, B., & Dubner, R. (1992) Effects of desipramine, amitriptyline, and fluoxetine on pain in diabetic neuropathy. *New England Journal of Medicine 326*, pp 1250–1256.

Mowla, A., Mosavinasab, M., Haghshenas, H., & Borhani Haghighi, A. (2007) Does serotonin augmentation have any effect on cognition and activities of daily living in Alzheimer's dementia? A double-blind, placebo -controlled clinical trial. *Journal of Clinical Psychopharmacology 27*, pp 484–487.

Saules, K.K., Schuh, L.M., Arfken, C.L., Reed, K., Kilbey, M.M., & Schuster, C.R. (2004) Double-blind placebo-controlled trial of fluoxetine in smoking cessation treatment including nicotine patch and cognitive-behavioral group therapy. *American Journal on Addictions 13*, pp 438–446.

Scannell, J.W., Blanckley, A., Boldon, H., & Warrington, B. (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery 11*, pp 191–200.

Shaughnessy, A.F. (2011) Old drugs, new tricks. *British Journal of Medicine 342*, p d741.

The U.S. FDA (2013) Orange Book: Approved drug products with therapeutic equivalence evaluations, 33rd Edition. Retrieved from the World Wide Web November 5, 2014: http://www.accessdata.fda.gov/scripts/cder/ob/

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*, pp W623–633.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D. et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research 36*, pp D901–906.

How to cite this article: Li, J and Lu, Z 2015 An Integrative Approach for Discovery of New Uses of Existing Drugs. *Data Science Journal,* 14: 9, pp. 1–11, DOI: http://dx.doi.org/10.5334/dsj-2015-009

Published: 22 May 2015

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS