Towards Drug Repositioning: A Unified Computational Framework for Integrating Multiple Aspects of Drug Similarity and Disease Similarity

Ping Zhang, PhD, Fei Wang, PhD, Jianying Hu, PhD
Healthcare Analytics Research, IBM T.J. Watson Research Center, New York, USA

Abstract

In response to the high cost and high risk associated with traditional de novo drug discovery, investigation of potential additional uses for existing drugs, also known as drug repositioning, has attracted increasing attention from both the pharmaceutical industry and the research community. In this paper, we propose a unified computational framework, called DDR, to predict novel drug-disease associations. DDR formulates the task of hypothesis generation for drug repositioning as a constrained nonlinear optimization problem. It utilizes multiple drug similarity networks, multiple disease similarity networks, and known drug-disease associations to explore potential new associations among drugs and diseases with no known links. A large-scale study was conducted using 799 drugs against 719 diseases. Experimental results demonstrated the effectiveness of the approach. In addition, DDR ranked drug and disease information sources based on their contributions to the prediction, thus paving the way for prioritizing multiple data sources and building more reliable drug repositioning models. Particularly, some of our novel predictions of drug-disease associations were supported by clinical trials databases, showing that DDR could serve as a useful tool in drug discovery to efficiently identify potential novel uses for existing drugs.

Introduction

The inefficiency of pharmaceutical drug development with high expenditure but low productivity has been widely discussed[1, 2]. Drug repositioning, the process of finding additional indications (i.e., diseases) for existing drugs, presents a promising avenue for identifying better and safer treatments without the full cost or time required for *de novo* drug development. Candidates for repositioning are usually either market drugs or drugs that have been discontinued in clinical trials for reasons other than safety concerns. Because the safety profiles of these drugs are known, clinical trials for alternative indications are cheaper, potentially faster and carry less risk than *de novo* drug development. Any newly identified indications can be quickly evaluated from phase II clinical trials. Drug repositioning can reduce drug discovery and development time from 10-17 years to potentially 3-12 years[3]. Therefore, it is not surprising that in recent years, new indications, new formulations, and new combinations of previously marketed products accounted for more than 30% of the new medicines that reach their first markets[4]. Drug repositioning has drawn widespread attention from the pharmaceutical industry, government agencies, and academic institutes. However, current successes in drug repositioning have primarily been the result of serendipitous events based on *ad hoc* clinical observation, unfocused screening, and “happy accidents”. Comprehensive and rational approaches are urgently needed to explore repositioning opportunities.

A reasonable systematic method for drug repositioning is the application of phenotypic screens by testing compounds with biomedical and cellular assays. However, this method also requires additional wet bench work of developing appropriate screening assays for each disease being investigated, and it thus remains challenging in terms of cost and efficiency. Big data analytics for both drugs and diseases provide an unprecedented opportunity to uncover novel statistical associations between drugs and diseases in a scalable manner. Many computational methods have been developed in this direction, including: (1) matching drug indications by their disease-specific response profiles based on the Connectivity Map (CMap) data[5, 6]; (2) predicting novel associations between drugs and diseases by the “Guilt by Association” (GBA) approach[7]; (3) utilizing structural features of compounds/proteins to predict new targets or indications, such as molecular docking[8, 9], and quantitative structure-activity relationship (QSAR) modelling[10]; (4) identifying associations between drugs and diseases in genetic activities, such as genome-wide association study (GWAS)[11], pathway profiles[12], and transcriptional responses[13]; (5) constructing drug network and using network neighbors to infer novel drug uses based on phenotypic profiles, such as side effects[14-16], and gene expression[17, 18]. All of these methods only focus on different aspects of drug/disease activities and therefore result in biases in their predictions. Also, these methods suffer from the noise in the given information source. Recently, several integrative methods which combine chemical, genetic, or phenotypic features were proposed to predict drug indications, for example, PREDICT[19], SLAMS[20], PreDR[21], Li and Lu[22], Huang *et al*[23], and Napolitano *et al*[24].

In this paper, we propose a unified computational framework for drug repositioning hypothesis generation, by integrating multiple **D**rug information sources and multiple **D**isease information sources to facilitate drug **R**epositioning tasks (DDR). DDR utilizes drug similarity network, disease similarity network, and known drug-disease associations to explore the potential associations among other unlinked drugs and diseases. In the experiment, we investigate three types of drug information (i.e., chemical structure, target protein, and side effect) and three types of disease information (i.e., phenotype, ontology, and disease gene). The proposed framework is also extensible, and thus DDR can incorporate additional types of drug/disease information sources.

Compared to prior integrative drug repositioning methods, it is worthwhile to highlight the following novel aspects that DDR can achieve simultaneously: (1) DDR can predict additional drug-disease associations by considering both drug information and disease information. With the exception of PREDICT[19], which integrates drug similarity scores and disease similarity scores using unweighted geometric mean, other integrative methods only consider either some drug information sources or some disease information sources. (2) DDR can determine interpretable importance of different information sources during the prediction. To our knowledge ours is the first study to do so. (3) As by-products, DDR can also discover the drug and disease groups, such that the drugs or diseases within the same group are highly correlated with each other, thus providing additional insights for targeted downstream investigations including clinical trials.

Construction of Drug Similarity and Disease Similarity Measures

In this section we introduce drug/disease similarities to quantify the degree of sharing common characteristics between pairs of drugs/diseases. A drug/disease similarity provided for a pair of drugs/diseases is a score that ranges from 0 to 1, with 0 representing the lowest similarity and 1 standing for the highest similarity. For each drug pair, we calculated three types of similarities based on chemical structures, target proteins, and side effects. For each disease pair, we calculated three types of similarities based on disease phenotypes, disease ontology, and disease genes.

Drug Similarity of Chemical Structures $\mathbf{D^{chem}}$. It is generally believed that drugs with similar chemical structures would carry out common therapeutic function, thus likely treat common diseases. We calculated the first drug pairwise similarity based on a chemical structure fingerprint corresponding to the 881 chemical substructures[25] defined in PubChem database[26]. Each drug *d* was represented by an 881-dimensional binary profile *h(d)* whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Then the pairwise chemical similarity between two drugs *d* and *d'* is computed as the Tanimoto coefficient of their chemical fingerprints:

$$D_{d,d'}^{chem} = \frac{h(d) \bullet h(d')}{|h(d)| + |h(d')| - h(d) \bullet h(d')} \tag{1}$$

where |*h(d)*| and |*h(d')*| are the counts of substructure fragments in drugs *d* and *d'* respectively. The dot product *h(d)* • *h(d')* represents the number of substructure fragments shared by two drugs.

Drug Similarity of Target Proteins $\mathbf{D^{target}}$. A drug target is the protein in the human body whose activity is modified by a drug resulting in a desirable therapeutic effect. Drugs sharing common targets often possess similar therapeutic function. We collected all target proteins for each drug from DrugBank[27]. Then we calculated the pairwise drug target similarity between drugs *d* and *d'* based on the average of sequence similarities of their target protein sets:

$$D_{d,d'}^{\text{target}} = \frac{1}{|P(d)||P(d')|} \sum_{i=1}^{|P(d)|} \sum_{j=1}^{|P(d')|} SW(P_i(d), P_j(d')) \tag{2}$$

where given a drug *d*, we presented its target protein set as *P(d)*; then |*P(d)*| is the size of the target protein set of drug *d*. The sequence similarity function of two proteins *SW* was calculated as a Smith-Waterman sequence alignment score[28].

Drug Similarity of Side Effects $\mathbf{D^{se}}$. Drug side effects, or adverse drug reactions, indicate the malfunction by off-targets. Thus side effects are useful to infer whether two drugs share similar target proteins and treat similar diseases. We obtained side effect keywords from SIDER[29], an online database containing drug side effect information extracted from package inserts using text mining methods. Each drug *d* was represented by 4192-dimensional binary side effect profile *e(d)* whose elements encode for the presence or absence of each of the side effect key words by 1 or 0 respectively. Then the pairwise side effect similarity between two drugs *d* and *d'* is computed as the Tanimoto coefficient of their side effect profiles:

$$D^{se}_{d,d'} = \frac{e(d)\bullet e(d')}{|e(d)|+|e(d')|-e(d)\bullet e(d')} \tag{3}$$

where *|e(d)|* and *|e(d')|* are the counts of side effect keywords for drugs *d* and *d'* respectively. The dot product *e(d)* • *e(d')* represents the number of side effects shared by two drugs.

Disease Similarity of Phenotypes $\mathbf{S^{pheno}}$. Disease phenotypes indicate phenotypic abnormalities encountered in human diseases. We used the phenotypic similarity constructed by van Driel *et al*[30]. The disease phenotypic similarity was constructed by identifying similarity between the MeSH terms[31] appearing in the medical description ("full text" and "clinical synopsis" fields) of diseases from OMIM database[32]. To be specific, each disease *s* in OMIM was represented by *K*-dimensional (*K* is the number of the MeSH terms) MeSH term feature vector *m(s)*: each entry in the feature vector represents an MeSH term, and the counts of the term found for disease *s* are the corresponding feature value. Then the pairwise disease phenotype similarity between two diseases *s* and *s'* is computed as the cosine of the angle between their feature vectors:

$$S^{pheno}_{ss'} = \frac{\sum_{i=1}^{K} m(s)_i\, m(s')_i}{\sqrt{\sum_{i=1}^{K} m^2(s)_i}\sqrt{\sum_{i=1}^{K} m^2(s')_i}} \tag{4}$$

where $m(s)_i$ denotes the *i*-th entry of the feature vector *m(s)*.

Disease Similarity of Disease Ontology $\mathbf{S^{do}}$. The Disease Ontology (DO)[33] is an open source ontological description of human disease, organized from a clinical perspective of disease etiology and location. The terms in DO are disease names or disease-related concepts and are organized in a directed acyclic graph (DAG). Two linked diseases in DO are in an "is-a" relationship, which means one disease is a subtype of the other linked disease. And the lower a disease is in the DO hierarchy, the more specific the disease term is. We calculated the semantic similarity between any pair of the diseases using the tool DOSim[34]. For a disease term *s* in DO, the probability that the term is used in disease annotations is estimated as p_s, which is the number of disease term *s* or its descendants in DO divided by the total number of disease terms in DO. Then the semantic similarity of two diseases *s* and *s'* is defined as the information content of their lowest common ancestor by:

$$S^{do}_{ss'} = -\log \min_{x\in C(s,s')} p_x \tag{5}$$

where *C(s,s')* is the set of all common ancestors of diseases *s* and *s'*.

Disease Similarity of Disease Genes $\mathbf{S^{gene}}$. Disease-causing aberrations in the normal function of a gene define that gene as a disease gene. We collected all disease genes for each disease from "phenotype-gene relationships" field from OMIM database. Then we calculated the pairwise disease similarity between diseases *s* and *s'* based on the average of sequence similarities of their disease gene sets:

$$S^{gene}_{ss'} = \frac{1}{|G(s)||G(s')|}\sum_{i=1}^{|G(s)|}\sum_{j=1}^{|G(s')|} SW(G_i(s),G_j(s')) \tag{6}$$

where given a disease *s*, we presented its disease gene set as *G(s)*; then *|G(s)|* is the size of the disease gene set of disease *s*. The sequence similarity function of two disease genes *SW* was calculated as a Smith-Waterman sequence alignment score.

Methodology

In this section we present the details of the proposed DDR approach. Suppose we have *n* information sources to measure drug similarity and *m* information sources to measure disease similarity. Let $\boldsymbol{D}_k \in \mathbb{R}^{n\times n}$ be the drug similarity matrix measured on the *k*-th information source, and suppose there are in total K_d information sources to measure the drug similarities. Similarly, let $\boldsymbol{S}_l \in \mathbb{R}^{m\times m}$ be the disease similarity matrix measured on the *l*-th information source and suppose there are K_s sources to measure the disease similarities. Let $\boldsymbol{U} \in \mathbb{R}^{n\times C_D}$ be the latent drug grouping matrix with C_D the number of drug groups, and $\boldsymbol{U}_{ij}$ indicates the possibility that the *i*-th drug belonging to the *j*-th drug cluster. $\boldsymbol{V} \in \mathbb{R}^{m\times C_S}$ be the latent disease grouping matrix with C_S the number of disease groups, and $\boldsymbol{V}_{ij}$

indicating the possibility that the *i*-th disease belonging to the *j*-th disease cluster. $\boldsymbol{R}\in\mathbb{R}^{n\times m}$ be the observed (i.e., known) drug-disease association matrix with $\boldsymbol{R}_{ij}$=1 if the association between the *i*-th drug and *j*-th disease is observed, and $\boldsymbol{R}_{ij}$=0 otherwise. Then we aim to analyze the drug-disease network by minimizing the following objective:

$$J = J_0 + \lambda_1 J_1 + \lambda_2 J_2 \tag{7}$$

where the three parts in the objective are:

- The reconstruction loss of observed drug-disease associations:

  $$J_0 = \| \Theta - U\Lambda V^T \|_F^2 \tag{8}$$

  Here $\boldsymbol{\Theta}\in\mathbb{R}^{n\times m}$ is the estimated dense version of $\boldsymbol{R}$, and $\boldsymbol{\Lambda}\in\mathbb{R}^{C_D\times C_S}$ encodes the relationship between drug clusters and disease clusters.

- The reconstruction loss of drug similarities:

  $$J_1 = \sum\nolimits_{k=1}^{K_d} \omega_k \| D_k - UU^T \|_F^2 + \delta_1 \| \omega \|_2^2 \tag{9}$$

  Here the estimated drug similarity matrix is $\boldsymbol{UU}^T$, and $\omega\in\mathbb{R}^{K_d\times 1}$ is the nonnegative weight vector when aggregating the reconstruction loss on different drug information sources. The L_2 norm regularization is added to avoid trivial solution[35] and $\delta_1\geq 0$ is the tradeoff parameter.

- The reconstruction loss of disease similarities:

  $$J_2 = \sum\nolimits_{l=1}^{K_s} \pi_l \| S_l - VV^T \|_F^2 + \delta_2 \| \pi \|_2^2 \tag{10}$$

  Here the estimated disease similarity matrix is $\boldsymbol{VV}^T$, and $\pi\in\mathbb{R}^{K_s\times 1}$ is the nonnegative weight vector when aggregating the reconstruction loss on different disease information sources. The L_2 norm regularization is added for the same reasons in equation (9).

Putting everything together, we obtained the optimization problem to be resolved:

$$\min_{\mathbf{U},\mathbf{V},\boldsymbol{\Lambda},\boldsymbol{\Theta},\omega,\pi} J \tag{11}$$

$$\text{subject to } \boldsymbol{U}\geq 0,\ \boldsymbol{V}\geq 0,\ \boldsymbol{\Lambda}\geq 0,\ \omega\geq 0,\ \omega^T\mathbf{1}=1,\ \pi^T\mathbf{1}=1,\ P_\Omega(\boldsymbol{\Theta})= P_\Omega(\boldsymbol{R})$$

where Ω is the set of indices of the observed associations, and P_Ω is the projection operator on obtaining the entries of a matrix indexed by the indices in Ω. Thus the constraint $P_\Omega(\boldsymbol{\Theta})= P_\Omega(\boldsymbol{R})$ restricts the estimated drug-disease associations should include the ones that are already observed. Note that to enhance the interpretability of the learned model, we require $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\Lambda}$ to be nonnegative, ω and π to be in simplexes. As there are lots of symbols and notations involved in problem (11), we summarize them in Table 1. To further help understanding those symbols as well as their roles in problem (11), we also provide a graphical illustration of the main idea of DDR in Figure 1.

Table 1. Notations and symbols of the methodology

Notation	Size	Meaning
$\boldsymbol{D}_k$	n×n	The *k*-th drug similarity matrix
$\boldsymbol{S}_l$	m×m	The *l*-th disease similarity matrix
$\boldsymbol{U}$	$n\times C_D$	Drug cluster assignment matrix
$\boldsymbol{V}$	$m\times C_S$	Disease cluster assignment matrix
$\boldsymbol{\Lambda}$	$C_D\times C_S$	Drug-disease cluster relationship matrix
$\boldsymbol{R}$	n×m	Observed drug-disease association matrix
$\boldsymbol{\Theta}$	n×m	Densified estimation of $\boldsymbol{R}$
ω	$K_d\times 1$	Drug similarity weight vector
π	$K_s\times 1$	Disease similarity weight vector

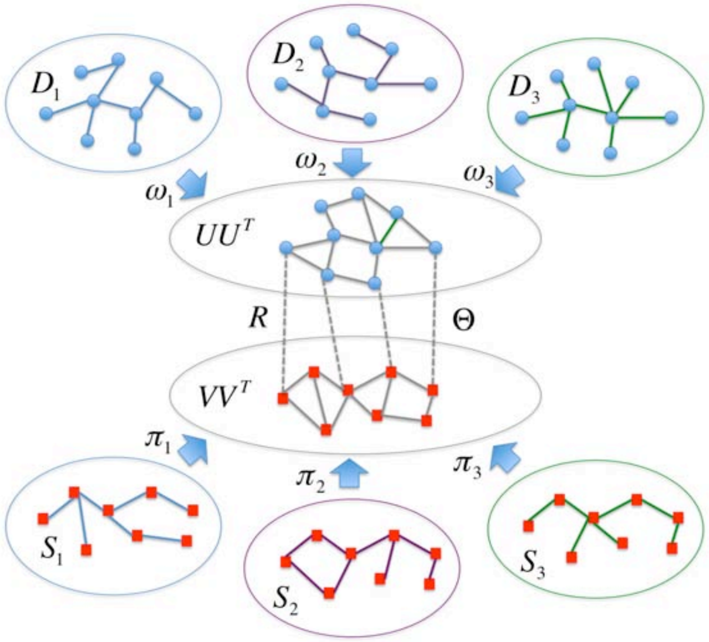

Figure 1. A graphical illustration of the main idea of DDR. There are multiple information sources that we can utilize to construct drug/disease similarities, and the constructed drug/disease similarity matrices are denoted by $\{\boldsymbol{D}_k\}_{k=1}^{K_d}$ or $\{\boldsymbol{S}_l\}_{l=1}^{K_s}$. There is also an observed drug-disease association matrix $\boldsymbol{R}$. Then DDR can learn the drug/disease grouping matrix $\boldsymbol{U}$ or $\boldsymbol{V}$, the estimated drug-disease association matrix $\boldsymbol{\Theta}$ and the importance of different drug/disease information sources ω or π.

Our proposed DDR method integrates multiple drug similarities, multiple disease similarities, and known drug-disease associations to achieve a global estimation on the entire drug-disease network including the intrinsic drug similarity, intrinsic disease similarity, as well as drug-disease associations. DDR formulates such a network estimation problem as a constrained nonlinear optimization problem. Since there are multiple groups of variables involved in the optimization problem (11), we adopt an efficient solution based on the Block Coordinate Descent (BCD) strategy[36]. The BCD approach works by solving the different groups of variables alternatively until convergence. At each iteration, it solves the optimization problem with respect to one group of variables with all other groups of variables fixed. Due to lack of space, details of the BCD solution procedure and its complexity analysis is provided at http://astro.temple.edu/~tua87106/ddr_bcd.pdf.

Results and Discussion

In this section we present experimental evaluation results of the proposed DDR algorithm on a drug repositioning task.

Data Description. The benchmark dataset, which is used to test the performance of DDR using a community standard, was extracted from NDF-RT[37] by Li and Lu[22]. It spans 3,250 treatment associations between 799 drugs and 719 diseases. We considered drug information from three data sources: chemical structure, target protein, and side effect. Thus, three 799×799 matrices were used to represent drug similarities between 799 drugs from different perspectives. Similarly, we considered disease information from three data sources: disease phenotype, disease ontology, and disease gene. Thus, three 719×719 matrices were used to represent disease similarities between 719 human diseases from different perspectives. The presence or absence of known associations between drug and disease was denoted by 1 or 0 respectively. Thus, a 799×719 matrix $\boldsymbol{R}$ used to represent the known drug-disease associations. We plotted the statistic of the known drug-disease associations in Figure 2. In that dataset, most of drugs (75%) treat <5 diseases; 18% of drugs treat 5 to 10 diseases; only 7% of drugs treat >10 diseases (Figure 2(a)). Although the disease hypertension has 78 related drugs, 80% of diseases have only <5 drugs; 10% of diseases have 5-10 drugs; and remaining 10% of diseases have >10 drugs.

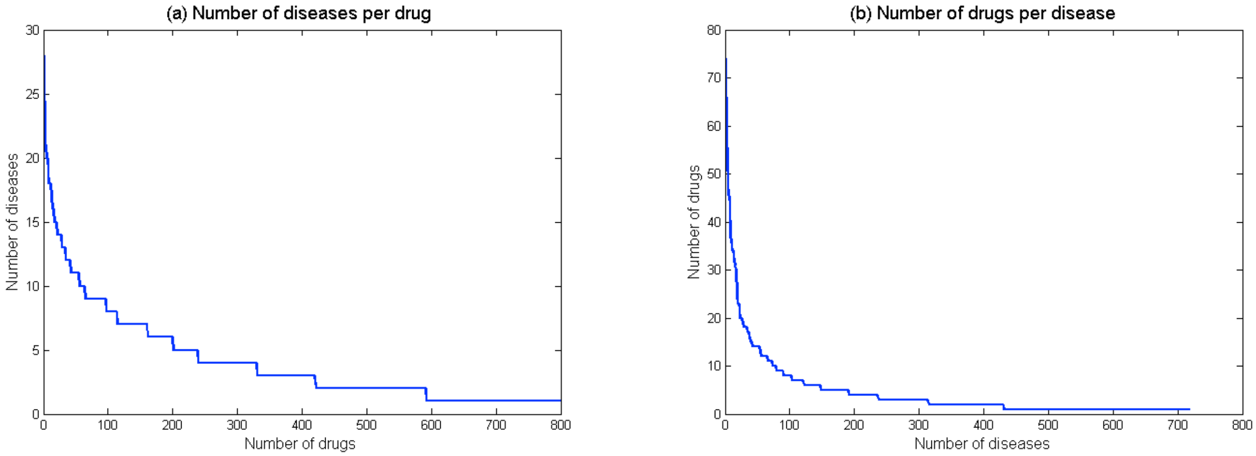

Figure 2. Statistics of the known drug-disease association dataset. (a) The number of indicated diseases per drug. (b) The number of drugs per disease.

Method Comparison. We used a 10-fold cross-validation scheme to evaluate drug repositioning approaches. To ensure the validity of the test cases, we held out all the associations involved with 10% of the drugs in each fold, rather than holding out associations directly. To obtain robust results, we performed 50 independent cross-validation runs, in each of which a different random partition of dataset to 10 parts was used. In our comparisons, we considered five drug repositioning methods: (1) **DDR using Simple Average**. The method only considers reconstruction loss of observed drug-disease associations (i.e., J_0 of objective formula (7) in the methodology section), and assumes each drug/disease source is equally informative. Thus the method uses the average of drug/disease similarity matrices as the integrated drug/disease similarity. (2) **DDR with Weighted Drug Similarity.** The method considers reconstruction losses of observed drug-disease associations and drug similarities (i.e., J_0 and J_1 in objective formula (7)). The method uses the average of disease similarity matrices as integrated disease similarity, and automatically learns drug similarity weight vector (ω) based on the contributions of drug information sources to the prediction. (3) **DDR with Weighted Disease Similarity.** The method considers reconstruction losses of observed drug-disease associations and disease similarities (i.e., J_0 and J_2 in objective formula (7)). The method uses the average of drug similarity matrices as integrated drug similarity, and automatically learns disease similarity weight vector (π) based on the contributions of disease information sources to the prediction. (4) **DDR with Weighted Drug and Disease Similarities.** The method considers all reconstruction losses proposed in the paper (i.e., formula (7) as a whole). The method automatically learns drug similarity weight vector (ω) and disease similarity weight vector (π) together based on the contributions of drug and disease information sources to the prediction. (5) **PREDICT with All Drug and Disease Similarities.** To our knowledge, PREDICT[19] is the only other method could consider both drug and disease information sources. PREDICT uses unweighted geometric mean of pairs of drug-drug and disease-disease similarity measures to construct classification features and subsequently learns a logistic regression classifier that distinguishes between true and false drug-disease associations. PREDICT could not provide weight for each drug/disease information source. Figure 3 shows the averaged ROC curves of 50 runs of the cross-validation for different methods based on the experiment.

Figure 3 shows that our proposed DDR framework is effective for drug repositioning tasks. Without considering reconstruction loss of any similarity measure, DDR using Simple Average obtains an averaged AUC score of 0.7985. When considering weighted drug similarity (i.e., reconstruction loss of drug similarities) or weighted disease similarity (i.e., reconstruction loss of disease similarities), DDRs obtain averaged AUC scores of 0.8508 or 0.8366 respectively. In the experiment, drug-based optimization (i.e., DDR with Weighted Drug Similarity) obtains a higher AUC score than disease-based optimization (i.e., DDR with Weighted Disease Similarity). This could be partially explained with the following reason. The 799 drugs we studied are marketed medications, which usually have rich and precise pharmacological data; thus drug-based optimization might be preferred in this case. For novel drugs or clinical candidates, disease-based optimization might be preferred to overcome missing knowledge in the pharmacology of a drug[38] (e.g., additional targets, unknown side effect). When considering weighted drug similarity and weighted disease similarity together, DDR obtain the highest averaged AUC score (0.8700). The observation indicates that drug-based optimization and disease-based optimization could be complementary, and computational drug repositioning tasks should optimize both drug similarity and disease similarity. Another observation is PREDICT with All Drug and Disease Similarities obtains an averaged AUC score of 0.8301. Although PREDICT considers drug/disease similarity and utilizes a logistic regression to weigh classification features, the result indicates that their strategy of feature construction (assembles all possible combinations of drug/disease similarity measures together as classification features) is less accurate than DDR.

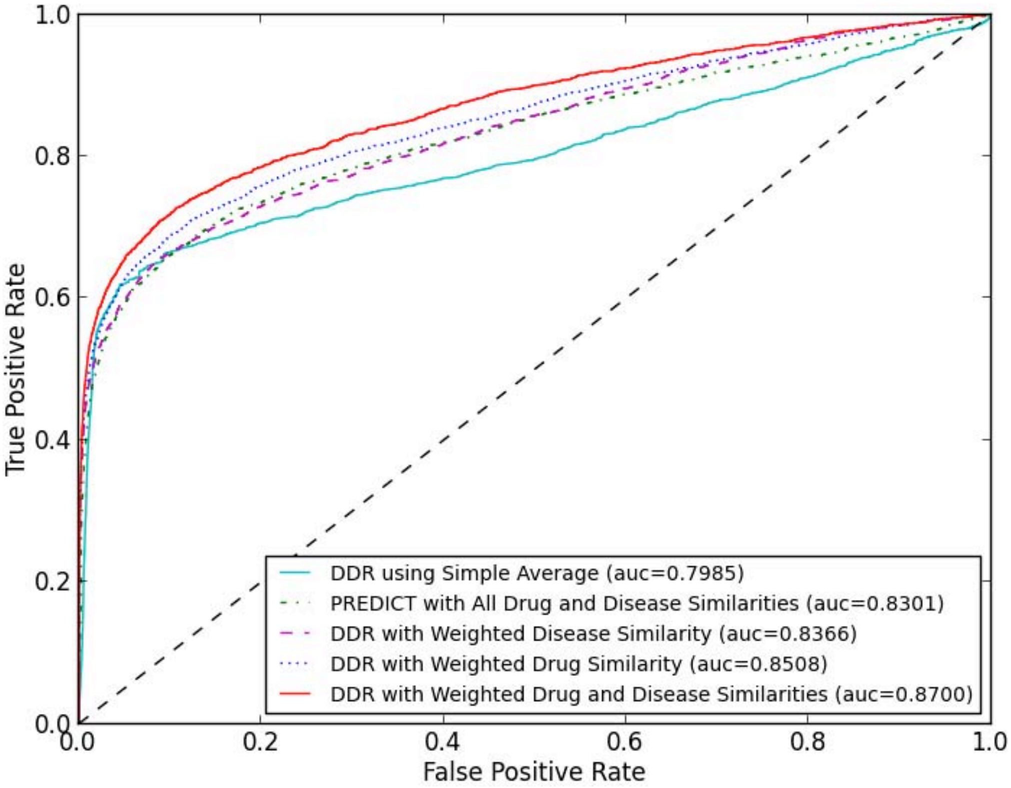

Figure 3. The averaged ROC comparison of five drug repositioning approaches generated from 50 runs of 10-fold cross-validation. Methods are sorted in legend of the figure according to their AUC score.

One “bonus” characteristic of DDR is it provides interpretable importance of different information sources based on their contributions to the prediction. The *i*-th element of drug/disease weight vector ω/π corresponds to the *i*-th drug/disease data sources. Since we constrained ω/π to be in a simplex in problem formula (11), the sum of all elements of ω/π is 1. Obtained from DDR with Weighted Drug and Disease Similarities, the averaged DDR weights of each data source and their standard deviations during the cross-validation experiments are plotted in Figure 4. For drug data sources, chemical structure obtains averaged weight of 0.2744, target protein obtains averaged weight of 0.2295, and side effect obtains a much higher averaged weight of 0.4961 (Figure 4(a)). This could be partially explained with the following reasons. Chemical structure and target protein sources focus on drug’s molecular mechanism of action (MOA) from a genotypic perspective. However, the pre-clinical outcomes based on MOA often do not correlate well with therapeutic efficacy in drug development. It is estimated that of all compounds effective in cell assays, only 30% of them could work in animals. Even worse, only 5% of them could work in humans[39]. Side effects are generated when drugs bind to off-targets, which perturb unexpected metabolic or signaling pathways. For marketed drugs, which have relatively complete side effect profiles, side effect information from clinical patients may be seen as valuable read-outs of drug effects directly on human bodies[40] (i.e., with less translational problems). Thus, side effects could server as a promising perspective for drug repositioning. For disease data sources, phenotype obtains averaged weight of 0.4248, disease ontology obtains averaged weight of 0.3958, and disease gene obtains a lower averaged weight of 0.1794 (Figure 4(b)). The lower weight of disease gene data source may be due to the fact that the gap between phenotype (human disease) and genotype (human gene) is too large[41], and the known associations between diseases and genes (obtained from OMIM) are incomplete.

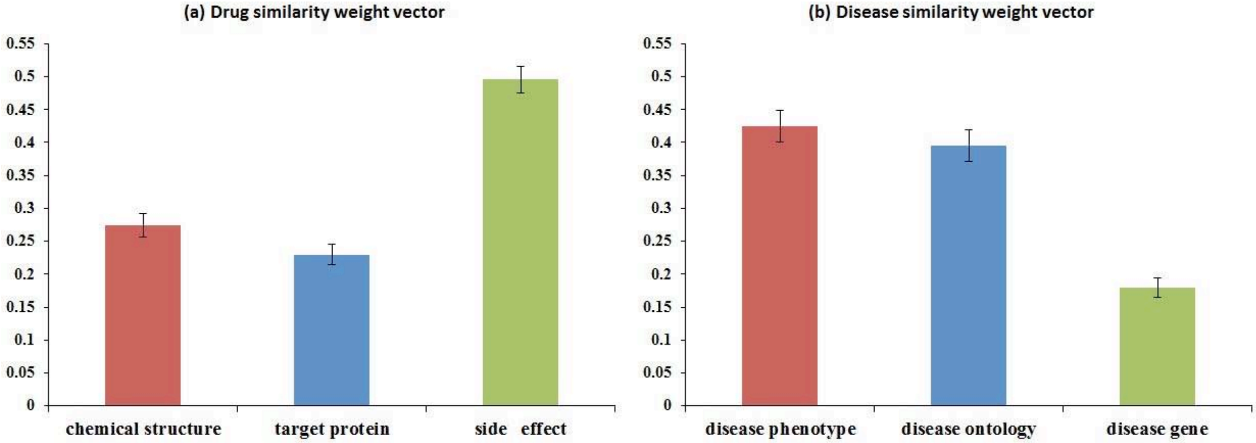

Figure 4. Distribution of averaged weights and standard deviations of the similarity weight vectors obtained by DDR. (a) Drug similarity weight vector ω contains weights of chemical structure, target protein, and side effect data sources. (b) Disease similarity weight vector π contains weights of disease phenotype, disease ontology, and disease gene sources.

Novel Predictions and Case Studies. We performed an additional leave-disease-out experiment to demonstrate the capability of DDR on uncovering drug-disease associations and predicting novel drug candidates for each disease. To ensure the validity of the test cases, we held out all the known drug-disease associations with the tested disease. The validation setting mimics a real-world setting: once rare/unknown diseases without any treatment information arise, a computational drug repositioning method should provide potential drugs based on characteristics (e.g. phenotypes, related genes) of the new diseases and the existing drug/disease similarities. In the experiment, we alternatively leave each disease *i* out and ran DDR (considered weighted drug and disease similarities). More specifically, we set all elements in *i*-th column of matrix ***R*** to 0, and used this ***R*** along with drug/disease similarity matrices as inputs of DDR. Then we used *i*-th column of the densified estimated matrix **Θ** as the drug prediction scores for the disease *i*. In this way, we got prediction scores for all possible associations between the 799 drugs and 719 diseases.

As an example, treatment predictions for Alzheimer's disease (AD) were analyzed. For the six drugs which are known to treat AD, DDR assigned scores of 0.7091 to Selegiline, 0.6745 to Valproic Acid, 0.6348 to Galantamine, 0.5675 to Donepezil, 0.5571 to Tacrine, and 0.5233 to Rivastigmine, which are significantly larger than those of the other 793 drugs (mean and standard deviation are 0.1565±0.1628). Table 2(a) shows the top 10 drugs predicted for AD by our DDR approach. Of the 10 drugs, only three (Selegiline, Valproic Acid, and Galantamine) appear in our known drug-disease association list. The remaining 7 predicted drugs (along with other high-ranked ones in the leave-disease-out experiment) could be considered as drug repositioning candidates for AD. Some predictions are explainable and supported by clinical evidence from ClinicalTrials.gov (i.e., pharmaceutical investigators have been aware of the associations, which are still in the experimental stages). Metformin, a drug commonly used to treat type II diabetes, can help trigger the pathway used to instruct stem cells in the brain to become neural cells[42]. Clinical trial NCT01965756 is under way to evaluate Metformin as a potential therapy for AD. Bexarotene, a skin cancer drug (for cutaneous T-cell lymphoma) that rapidly removed the damaging protein implicated in the progression of the illness from the brains of mice[43], has been tested to treat AD in recent clinical trials (NCT01782742 and NCT02061878). Nilvadipine, a calcium channel blocker (CCB) for treatment of hypertension, also blocks the production of amyloid proteins linked to AD[44]. Nilvadipine has been tested in a clinical trial as a possible treatment for AD in Ireland (NCT02017340).

Another example we analyzed is treatment predictions for Systemic Lupus Erythematosus (SLE). For the three drugs which are known to treat SLE, DDR assigned scores of 0.7269 to Azathioprine, 0.6862 to Triamcinolone, and 0.6374 to Hydroxychloroquine, which are significantly larger than those of the other 796 drugs (mean and standard deviation are 0.1707±0.1617). Table 2(b) shows the top 10 drugs predicted for SLE by our DDR approach. The top 10 predictions include all the three known treatments to SLE, which shows the effectiveness of our method. The remaining 7 predicted drugs (along with other high-ranked ones in the leave-disease-out experiment) could be considered as drug repositioning candidates for SLE. Some predictions are explainable and supported by clinical evidence from ClinicalTrials.gov. Leflunomide, a pyrimidine synthesis inhibitor, is used to treat moderate to severe rheumatoid arthritis and psoriatic arthritis. A genetic link study shows rheumatoid arthritis and SLE sufferers share a variant of the same STAT4 gene, and therapies developed to treat one disease may possibly be able to treat the other[45]. Therefore, it is not surprising to see Leflunomide is tested as a treatment for SLE in a clinical trial (NCT00637819). Nelfinavir, one of the protease inhibitors, has been approved for use in the treatment of human immunodeficiency virus (HIV). Protease inhibitors have been shown to interfere with binding of anti-double stranded DNA antibodies to their targets (some bindings may lead to organ damage) and may decrease inflammation in SLE[46]. Recently, a clinical trial (NCT02066311) has been proposed to evaluate Nelfinavir as a potential therapy for SLE.

Table 2. Top 10 drugs for diseases Alzheimer's Disease (AD) and Systemic Lupus Erythematosus (SLE) based on DDR predictions

(a) Top 10 drugs predicted for AD

Drug	Prediction Score	Clinical Evidence?
Selegiline*	0.7091	—
Carbidopa	0.6924	No
Amantadine	0.6897	No
Procyclidine	0.6826	No
Valproic Acid*	0.6745	—
Metformin	0.6543	Yes
Bexarotene	0.6426	Yes
Neostigmine	0.6385	No
Galantamine*	0.6348	—
Nilvadipine	0.6159	Yes

(b) Top 10 drugs predicted for SLE

Drug	Prediction Score	Clinical Evidence?
Desoximetasone	0.7409	No
Azathioprine*	0.7269	—
Leflunomide	0.7078	Yes
Fluorometholone	0.7054	No
Triamcinolone*	0.6862	—
Beclomethasone	0.6522	No
Etodolac	0.6445	No
Hydroxychloroquine*	0.6374	—
Nelfinavir	0.6371	Yes
Mercaptopurine	0.6150	No

* denotes the drug is known and approved to treat the disease

Conclusion

We have proposed a general computational framework, called DDR, to explore drug-disease association for drug repurposing hypothesis generation. Our method takes into consideration multiple drug similarities, multiple disease similarities, and known drug-disease associations, to uncover the potential additional associations among other unlinked drugs and diseases. Experimental results demonstrate the effectiveness of the proposed method, and suggest that our method could help identify drug repositioning opportunities, which will benefit patients by offering more effective and safer treatments.

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010; 9(3):203-214.
2. Berggren R, Moller M, Moss R, Poda P, Smietana K. Outlook for the next 5 years in drug innovation. Nat Rev Drug Discov 2012; 11(6):435-436.
3. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 2013; 93(4):335-341.
4. Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform 2011; 12(4):346-356.
5. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006; 313(5795):1929-1935.
6. Hu G, Agarwal P. Human Disease-Drug Network Based on Genomic Expression Profiles. PLoS ONE 2009; 4(8):e6536.
7. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 2009; 86(5):507-510.
8. Luo H, Chen J, Shi L, Mikailov M, Zhu H, Wang K, He L, Yang L. DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome. Nucleic Acids Res 2011; 39(Web Server issue):W492-W498.

9. Dakshanamurthy S, Issa NT, Assefnia S, Seshasayee A, Peters OJ, Madhavan S, Uren A, Brown ML, Byers SW. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012; 55(15):6832-6848.
10. Cheng F, Zhou Y, Li J, Li W, Liu G, Tang Y. Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods. Mol Biosyst 2012; 8(9):2373-2384.
11. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V. Use of genome-wide association studies for drug repositioning. Nat Biotechnol 2012; 30(4):317-320.
12. Li J, Lu Z. Pathway-based drug repositioning using causal inference. BMC Bioinformatics 2013; 14(Suppl 16):S3.
13. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci 2010;107(33):14621-14626.
14. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science 2008;321(5886):263–266.
15. Yang L, Agarwal P. Systematic Drug Repositioning Based on Clinical Side-Effects. PLoS ONE 2011;6(12):e28025.
16. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. PLoS One 2014; 9(2):e87864.
17. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, Sage J, Butte AJ. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med 2011; 3(96):96ra77.
18. Wang K, Sun J, Zhou S, Wan C, Qin S, Li C, He L, Yang L. Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS Comput Biol 2013; 9(11):e1003315.
19. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011; 7:496.
20. Zhang P, Agarwal P, Obradovic Z. Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2013; Part III: 579-594.
21. Wang Y, Chen S, Deng N, Wang Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS One 2013; 8(11):e78518.
22. Li J, Lu Z. A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine 2012.
23. Huang YF, Yeh HY, Soo VW. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC Med Genomics 2013; 6(Suppl 3):S4.
24. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, Greco D. Drug repositioning: a machine-learning approach through data integration. J Cheminform 2013; 5(1):30.
25. PubChem substructure description [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf]
26. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009; 37(Web Server Issue):W623-W633.
27. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014; 42(Database Issue): D1091-1097.
28. Smith TF, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. Nucleic Acids Res 1985; 13(2):645-665.
29. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010; 6:343.
30. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet 2006; 14(5):535-542.
31. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000; 88(3):265–266.
32. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005; 33(Database issue):D514-D517.
33. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012; 40(Database issue):D940-D946.
34. Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, Li C, Li X, Rao S, Li X. DOSim: an R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics 2011; 12:266.
35. Wang F, Wang X, Li T. Generalized cluster aggregation. Proceedings of the International Joint Conference on Artificial Intelligence 2009.
36. Bertsekas DP. Block coordinate descent methods. in Nonlinear Programming 2nd Edition 1999.
37. Carter JS, Brown SH, Bauer BA, Elkin PL, Erlbaum MS, Froehling DA, Lincoln MJ, Rosenbloom ST, Wahner-Roedler DL, Tuttle MS. Categorical information in pharmaceutical terminologies. AMIA Annu Symp Proc 2006:116-120.
38. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 2011; 12(4):303-311.
39. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. Nat Rev Drug Discov 2011; 10(6):428-438.
40. Duran-Frigola M, Aloy P. Recycling side-effects into clinical markers for drug repositioning. Genome Med 2012; 4(1):3.
41. Chen Y, Wu X, Jiang R. Integrating human omics data to prioritize candidate genes. BMC Med Genomics 2013; 6:57.
42. Wang J, Gallagher D, DeVito LM, Cancino GI, Tsui D, He L, Keller GM, Frankland PW, Kaplan DR, Miller FD. Metformin activates an atypical PKC-CBP pathway to promote neurogenesis and enhance spatial memory formation. Cell Stem Cell 2012; 11(1):23-35.
43. Cramer PE, Cirrito JR, Wesson DW, Lee CY, Karlo JC, Zinn AE, Casali BT, Restivo JL, Goebel WD, James MJ, Brunden KR, Wilson DA, Landreth GE. ApoE-directed therapeutics rapidly clear β-amyloid and reverse deficits in AD mouse models. Science 2012; 335(6075):1503-1506.
44. Paris D, Quadros A, Humphrey J, Patel N, Crescentini R, Crawford F, Mullan M. Nilvadipine antagonizes both Abeta vasoactivity in isolated arteries, and the reduced cerebral blood flow in APPsw transgenic mice. Brain Res 2004; 999(1):53-61.
45. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, de Bakker PI, Le JM, Lee HS, Batliwalla F, Li W, Masters SL, Booty MG, Carulli JP, Padyukov L, Alfredsson L, Klareskog L, Chen WV, Amos CI, Criswell LA, Seldin MF, Kastner DL, Gregersen PK. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 2007; 357(10):977-986.
46. Bloom O, Cheng KF, He M, Papatheodorou A, Volpe BT, Diamond B, Al-Abed Y. Generation of a unique small molecule peptidomimetic that neutralizes lupus autoantibody activity. Proc Natl Acad Sci 2011; 108(25):10255-10259.