Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources

Ping Zhang[1], Pankaj Agarwal[2], and Zoran Obradovic[3]

[1] Healthcare Analytics Research, IBM T.J. Watson Research Center, USA
pzhang@us.ibm.com
[2] Computational Biology, GlaxoSmithKline R&D, USA
pankaj.agarwal@gsk.com
[3] Center for Data Analytics and Biomedical Informatics, Temple University, USA
zoran.obradovic@temple.edu

Abstract. Drug repositioning helps identify new indications for marketed drugs and clinical candidates. In this study, we proposed an integrative computational framework to predict novel drug indications for both approved drugs and clinical molecules by integrating chemical, biological and phenotypic data sources. We defined different similarity measures for each of these data sources and utilized a weighted *k*-nearest neighbor algorithm to transfer similarities of nearest neighbors to prediction scores for a given compound. A large margin method was used to combine individual metrics from multiple sources into a global metric. A large-scale study was conducted to repurpose 1007 drugs against 719 diseases. Experimental results showed that the proposed algorithm outperformed similar previously developed computational drug repositioning approaches. Moreover, the new algorithm also ranked drug 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.

Keywords: Drug Repositioning, Drug Indication Prediction, Multiple Data Sources, Metric Integration, Large Margin Method.

1 Introduction

In response to the high cost and risk in traditional *de novo* drug discovery, discovering potential uses for existing drugs, also known as drug repositioning, has attracted increasing interests from both the pharmaceutical industry and the research community [1]. Drug repositioning can reduce drug discovery and development time from 10-17 years to potentially 3-12 years [2].

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. Then, any newly identified indications can be quickly evaluated from phase II clinical trials.

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part III, LNAI 8190, pp. 579–594, 2013.

Among the 51 new medicines and vaccines that were brought to market in 2009, new indications, new formulations, and new combinations of previously marketed products accounted for more than 30% [3]. 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 serendipity or clinical observation. Systematic 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 the 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. Data mining and machine learning offer an unprecedented opportunity to develop computational methods to predict all possible drug repositioning using available data sources. Most of these methods have used chemical structure, protein targets, or phenotypic information (e.g., side-effect profiles, gene expression profiles) to build predictive models and some have shown promising results [4-11].

In this study, we proposed a new drug repositioning framework: **S**imilarity-based **LA**rge-margin learning of **M**ultiple **S**ources (SLAMS), which ranks and integrates multiple drug information sources to facilitate the prediction task. In the experiment, we investigated three types of drug information: (1) chemical properties - compound fingerprints; (2) biological properties - protein targets; (3) phenotypic properties - side-effect profiles. The proposed framework is also extensible, and thus the SLAMS algorithm can incorporate additional types of drug information sources.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 describes our SLAMS algorithm. Section 4 presents the conducted experiment and the achieved results. Finally, section 5 concludes the paper.

2 Related Work

Recent research has shown that computational approaches have the potential to offer systematic insights into the complex relationships among drugs, targets, and diseases for successful repositioning. Currently, there are five typical computational methods in drug repositioning: (1) predicting new drug indications on the basis of the chemical structure of the drug [4]; (2) inferring drug indications from protein targets interaction networks [5, 6]; (3) identifying relationships between drugs based on the similarity of their side-effects [7, 8]; (4) analyzing gene expression following drug treatment to infer new indications [9, 10]; (5) building a background chemical-protein interactome (CPI) using molecular docking [11]. All of these methods only focus on different aspects of drug-like activities and therefore result in biases in their predictions. Also, these methods suffer according to the noise in the given drug information source.

Li and Lu [12] developed a method for mining potential new drug indications by exploring both chemical and bipartite-graph-boosted molecular features in similar drugs. Gottlieb *et al.* [13] developed a 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.

This paper differs from the related studies in the following aspects:

1. We consider multiple chemical properties, biological properties, and phenotypic properties at the same time, unlike references [4-11]. Our SLAMS algorithm can also incorporate additional types of drug properties.
2. Li and Lu [12], tried all representative weights for multiple data sources in a brute-force way, but SLAMS assigns weights to all data sources without manual tuning.
3. We use a large margin method (i.e., minimize hinge-rank-loss) to integrate multiple sources, which is usually more optimal than a logistic regression method (i.e., minimize log-loss) [13] from the machine learning theory perspective. Also, the weight vector derived from a large margin method is more interpretable.
4. We use canonical correlation analysis (CCA) [14] to impute missing values of side-effect profiles. Then, we augmented known side-effect profiles with predicted side-effect profiles to build a new side-effect source.
5. We use multiple measures (e.g., precision, recall, F-score) to evaluate the results of drug repositioning experiments. Many previous methods used only area under the ROC curve (AUC) to evaluate their performance, but a high AUC score does not mean much in a highly imbalanced classification task [15] and unfortunately drug repositioning is such a task.

3 Method

In this section, we present the SLAMS algorithm for drug repositioning by integrating multiple data sources. First, we present the algorithmic framework. Second, we present a similarity-based scoring component for each data source. We also introduce the CCA to imputing missing side-effect profiles. Third, we present a large margin method to integrate multiple scoring components.

3.1 Algorithm Overview

The SLAMS algorithm is based on the observation that similar drugs are indicated for similar diseases. In this study, we identify a target drug d_x's potential new indications through 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. There are multiple metrics to measure the similarity between two drugs from different aspects of drug-like activities. The objective of SLAMS is to integrate individual metrics from multiple sources into a global metric.

The SLAMS process framework is illustrated in Fig. 1, where m data sources are involved in the integration process. Each candidate drug d_x queries i-th ($i=1,\dots,m$) data source and gets the prediction score for indicated disease s as $f^{i}(d_x,s)$. Then m

prediction scores from multiple data sources are combined into a single, final score $f^E(d_x,s)$. The details of scoring a single data source via *k*-nearest neighbor classifier and integrating multiple prediction scores via large margin method will be presented next.

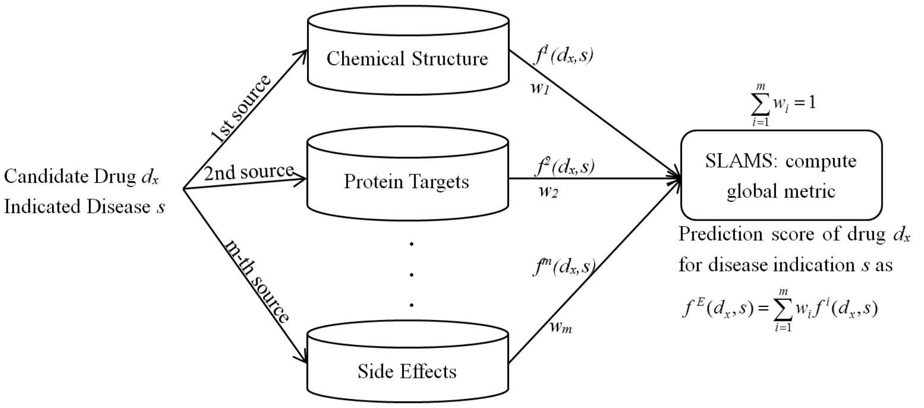

Fig. 1. Illustration of SLAMS Framework

3.2 Similarity Measures

A drug's chemical structure, protein targets, and side-effect profiles are important features in drug design, and evidently associated with its therapeutic use. Also these features are orthogonal to each other and so we consider them in the study.

Computing Similarity of Drug Chemical Structures. Our method for calculating the pairwise similarity $sim_{chem}(d_x,d_y)$ is based on the *2D* chemical fingerprint descriptor of each drug's chemical structure in PubChem [16]. We used chemistry development kit[1] (CDK) [17] to encode each chemical component into an 881-dimensional chemical substructure vector defined in PubChem. That is, each drug *d* is represented by a binary fingerprint *h(d)* 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:

$$sim_{chem}(d_x,d_y) = \frac{h(d_x) \bullet h(d_y)}{|h(d_x)| + |h(d_y)| - h(d_x) \bullet h(d_y)}$$

where $|h(d_x)|$ and $|h(d_y)|$ are the counts of structure fragments in drugs d_x and d_y respectively. The dot product $h(d_x) \bullet h(d_y)$ represents the number of structure fragments shared by two drugs. The sim_{chem} score is in the [0, 1] range.

[1] Available at `http://sourceforge.net/projects/cdk/`.

Computing Similarity of Drug Protein Targets. A drug target is the protein in the human body whose activity is modified by a drug resulting in a desirable therapeutic effect. Our method for calculating the pairwise similarity *$sim_{target}(d_x,d_y)$* is based on the average of sequence similarities of the two target protein sets:

$$sim_{\text{target}}(d_x,d_y)=\frac{1}{|P(d_x)||P(d_y)|}\sum_{i=1}^{|P(d_x)|}\sum_{j=1}^{|P(d_y)|}g(P_i(d_x),P_j(d_y))$$

where given a drug *d*, we present 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 *g* is calculated as a Smith-Waterman sequence alignment score [18]. The sim_{target} score is in the [0, 1] range.

Computing Similarity of Drug Side-Effect Profiles. Clinical side effects provide a human phenotypic profile for the drug, and this profile can suggest additional drug indications. In this subsection, we define the side-effect similarity first. Then, we introduce a method to predict drug side-effect profiles from chemical structure. There are two reasons for this: (1) the current side-effect dataset doesn't cover all drugs. By imputing missing side-effect profiles and using the predicted side-effect profiles with other known data sources, we have more data to train a predictive drug repositioning model; (2) in a real drug discovery pipeline, side-effect information is collected from phase I all the way through phase IV. The candidate drugs for repositioning may not have completed side-effect profiles in the early phases. It is easier to apply the predictive model to the candidate drug with predicted side-effect profiles with other known information.

Definition of Side-effect Similarity. Side-effect keywords were obtained from the SIDER database, which contains information about marketed medicines and their recorded adverse drug reactions [19]. Each drug *d* was represented by 1385-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. The pairwise side-effect similarity between two drugs d_x and d_y is computed as the Tanimoto coefficient of their fingerprints:

$$sim_{se}(d_x,d_y)=\frac{e(d_x)\bullet e(d_y)}{|e(d_x)|+|e(d_y)|-e(d_x)\bullet e(d_y)}$$

where $|e(d_x)|$ and $|e(d_y)|$ are the counts of side-effect keywords for drugs d_x and d_y respectively. The dot product $e(d_x)\bullet e(d_y)$ represents the number of side effects shared by two drugs. The sim_{se} score is in the [0, 1] range.

Predicting drug side-effect profiles. Suppose that we have a set of *n* drugs with *p* substructure features and *q* side-effect features. Each drug is represented by a chemical substructure feature vector $\boldsymbol{x}=(x_1,\ldots,x_p)^T$, and by a side-effect feature vector $\boldsymbol{y}=(y_1,\ldots,y_q)^T$. Consider two linear combinations for chemical substructures and side effects as $\boldsymbol{u}_i=\alpha^T\boldsymbol{x}_i$ and $\boldsymbol{v}_i=\beta^T\boldsymbol{y}_i$ $(i=1,2,\ldots,n)$, where $\alpha=(\alpha_1,\ldots,\alpha_p)^T$ and $\beta=(\beta_1,\ldots,\beta_q)^T$ are weight vectors. The goal of canonical correlation analysis is to find weight vectors α and β which maximize the following canonical correlation coefficient [14]:

$$\rho = corr(u,v) = \frac{\sum_{i=1}^{n} \alpha^T \boldsymbol{x}_i \cdot \beta^T \boldsymbol{y}_i}{\sqrt{\sum_{i=1}^{n} (\alpha^T \boldsymbol{x}_i)^2} \sqrt{\sum_{i=1}^{n} (\beta^T \boldsymbol{y}_i)^2}}$$

Let X denote the $n \times p$ matrix as $X=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]^T$, and Y denote the $n \times q$ matrix as $Y=[\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n]^T$. Then consider the following optimization problem:

$$\max\{\alpha^T X^T Y \beta\} \text{ subject to } \|\alpha\|_2^2 \le 1, \|\beta\|_2^2 \le 1$$

Solving the problem, we obtain m pairs of weight vectors $\alpha_1,\ldots, \alpha_m$ and $\beta_1,\ldots, \beta_m$ (m is the counts of canonical components).

Given the profile of chemical substructure $\boldsymbol{x}_{new}$ for a drug of unknown side effects, we use the following prediction score for its potential side-effect profile $\boldsymbol{y}_{new}$ as:

$$\boldsymbol{y}_{new} = \sum_{k=1}^{m} \beta_k \rho_k \alpha_k^T \boldsymbol{x}_{new} = B \Lambda A^T \boldsymbol{x}_{new}$$

where $A=[\alpha_1,\ldots, \alpha_m]$, $B=[\beta_1,\ldots, \beta_m]$ and Λ is the diagonal matrix whose elements are canonical correlation coefficients. If the j-th element in $\boldsymbol{y}_{new}$ has a high score, the new drug is predicted to have the j-th side-effect ($j=1,2,\ldots,q$).

CCA was showed to be accurate and computationally efficient in prediction of the drug side-effect profiles [20]. Using CCA we augmented the drug-side-effect relationship list with side-effect predictions for drugs that are not included in SIDER, based on their chemical properties. We can use the similarity metric defined in the last subsection to calculate the side-effect similarity.

Computing Prediction Score from a Single Data Source. To calculate the likelihood that drug d_x has the indication s, we use a weighted variant of the k-nearest neighbor (k-NN) algorithm. The optimization of the model parameter k was done in a cross validation setting (k=20 in the study). For the i-th data source, the prediction score f of an indication s for the drug d_x is calculated as:

$$f^i(d_x, s) = \sum_{d_y \in N_k(d_x)} sim^i(d_x, d_y) \cdot C(s \in indications(d_y))$$

where $sim^i(d_x,d_y)$ denotes the similarity score between two drugs d_x and d_y from the i-th source, C is a characteristic function that return 1 if d_y has an indication s and 0 otherwise, and $N_k(d_x)$ are the k nearest neighbors of drug d_x according to the metric sim^i which is determined by the type of i-th data source. The metric sim^i can be one of the similarities defined in the previous subsections (i.e., chemical structure, protein targets, and side effects), or any additional types of drug information sources. Thus, our SLAMS algorithm is extensible. We propose a k-NN scoring component for drug repositioning tasks due to its simplicity of implementation on multiple data sources, straightforward use of multiple scores, and its competitive accuracy with more complex algorithms [21].

3.3 Combining Multiple Measures

We considered multiple data sources and obtained several prediction scores for each pair (*d,s*). Given *m* scores for a drug-disease pair (*d,s*) (i.e., there are *m* different data sources), we propose a large margin method to calculate final score f^E as a weighted average of individual scores:

$$f^E(d_x, s) = \sum_{i=1}^{m} w_i f^i(d_x, s)$$

where w_i is the corresponding weight for the *i*-th (*i=1,...,m*) data source.

We learn the weights from training data using a large margin method as follows. Let us assume that we are given *m* data sources, $\{D_j, j = 1,...,m\}$, and *n* drugs $\{x_i, i = 1,...,n\}$. Each drug is assigned to several indications from the set of *k* indications. Let Y_i denote the set of indications that drug x_i is assigned to, and $\overline{Y_i}$ denote the set of indications that drug x_i is not assigned to. Then, let *f(x,y)* be a vector of length *m*, whose *j*-th element is the score of drug *x* for indication *y* on the data source D_j. A weight vector ***w***, used for integration of *m* prediction, is found by solving the following optimization problem:

$$\begin{aligned}
&\min_{\mathbf{w},\xi} \sum_i \sum_{y\in Y_i,\, \overline{y}\in \overline{Y_i}} \xi_i(y,\overline{y}) \\
&s.t\ \boldsymbol{w}^T(f(x_i,y) - f(x_i,\overline{y})) \ge -\xi_i(y,\overline{y}), \forall i, y\in Y_i, \overline{y}\in \overline{Y_i} \\
&\quad \xi_i(y,\overline{y}) \ge 0, \forall i, y\in Y_i, \overline{y}\in \overline{Y_i} \\
&\quad \boldsymbol{w}^T\mathbf{e} = 1;\ \boldsymbol{w} \ge 0
\end{aligned}$$

where ***e*** is a vector of ones. The resulting convex optimization problem can be solved using standard optimization tools, such as CVX[2]. With the trained weight vector ***w***, the drug-indication scores from different data sources can be integrated by taking their weighted average as $\boldsymbol{w}^T \cdot f^i(x,y)$.

4 Experimental Results

In this section we experimentally evaluate the proposed SLAMS algorithm on a drug repositioning task.

4.1 Data Description

In the experiment, we analyzed the approved drugs from DrugBank [22], which is a widely used public database of drug information. From DrugBank, 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

[2] Available at `http://cvxr.com/`.

PubChem [16] and UMLS [23] in order to extract other drug related information. In the end, we extracted chemical structures of the 1007 drugs from PubChem. Each drug was represented by an 881-dimensional binary profile whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. There are 122,022 associations between 1007 drugs and 881 PubChem substructures.

To facilitate collecting target protein information, we mapped target proteins to UniProt Knowledgebase [24], a central knowledgebase including most comprehensive and complete information on proteins. In the end, we extracted 3152 relationships between 1007 drugs and 775 proteins.

Side-effect keywords were obtained from the SIDER database [19]. SIDER presents an aggregate of dispersed public information on drug side effects. SIDER extracted information on marked medicines and their recorded side effects from public documents and package inserts, which resulted in a collection of 888 drugs and 1385 side-effect keywords. Merging these 888 SIDER drugs to the 1007 DrugBank approved drugs, we obtained 40,974 relationships between 613 drugs and 1385 side effects. A total number of 394 drugs from DrugBank approved list could not be mapped to SIDER drug names. We used the method described in subsection 3.2 to predict their side-effect profiles. Finally we obtained 19,385 predicted relationships between these 394 drugs and 1385 side effects.

We obtained a drug's known use(s) through extracting treatment relationships between drugs and diseases from the National Drug File - Reference Terminology[3] (NDF-RT), which is part of the UMLS [23]. The drug-disease treatment relationship list is also used by Li and Lu [12] as the gold standard set of drug repositioning task. We normalized various drug names in NDF-RT to their active ingredients. 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 3250 treatment relationships between 799 drugs and 719 diseases. We plotted the statistics of the gold drug-disease relationship in Fig. 2. Most of drugs (75%) treat <5 indicated diseases; 18% of drugs treat 5 to 10 diseases; only 7% of drugs treat >10 diseases (Fig. 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 (Fig. 2(b)).

All the data used in our experiments are available at our website[4].

4.2 Evaluation Measures

In the study, we modeled the drug repositioning task as a binary classification problem where each drug either treats or does not treat a particular disease. We measure the final classification performance using four criteria: precision, recall, F-score, and area under the ROC curve. In order to provide the definitions of these four criteria, we first define the classification confusion table for binary classification

[3] NDF-RT found at `http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/`.

[4] Available at `http://astro.temple.edu/~tua87106/drugreposition.html`.

problems where the two classes are indicated as positive and negative, which is constructed by comparing the actual data labels and predicted outcomes (see Table 1).

Then we can define the classification evaluation metrics as: True Positive Rate = *TP / (TP+FN)*, False Positive Rate = *FP / (FP+TN)*, Precision = *TP / (TP+FP),* Recall = *TP / (TP+FN)*, and F-Score = *2 ·Precision ·Recall / (Precision+Recall).*

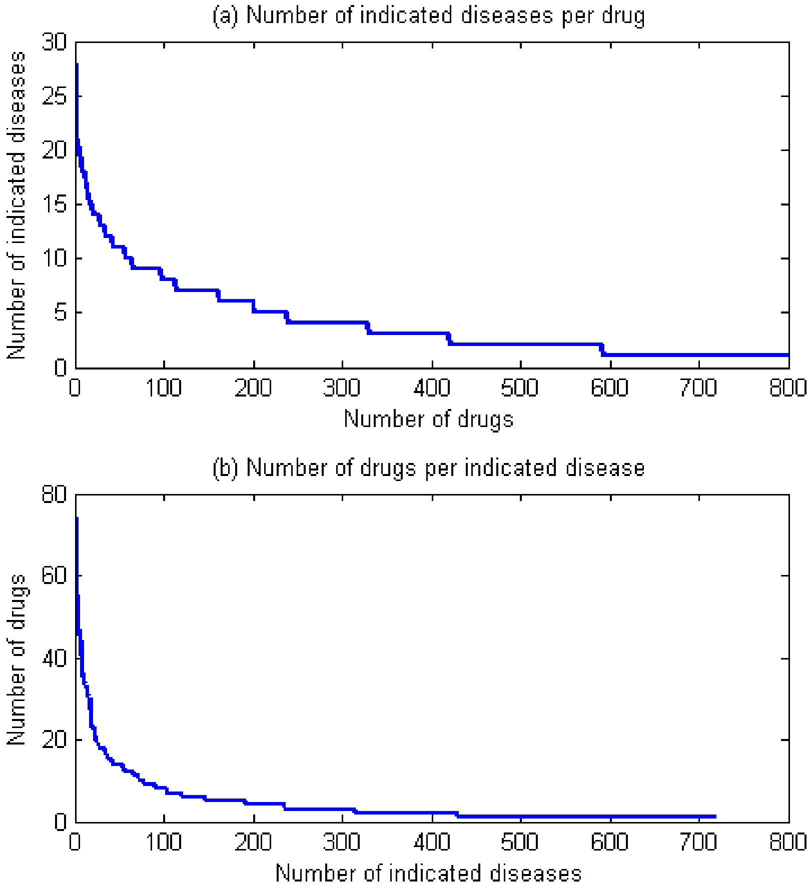

Fig. 2. Statistics of the drug-disease relationship dataset. (a) The number of indicated diseases per drug. (b) The number of drugs per indicated disease.

Table 1. Confusion matrix

	Actual Value	
Predicted Value	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

The confusion matrix can be used to construct a point in the ROC curve, which is a graphical plot of true positive rate against false positive rate. The whole ROC curve can be plotted by varying threshold value for prediction score, above which the output is predicted as positive, and negative otherwise. Then we can use the area under the ROC curve (AUC) as a measure. The other three measures (precision, recall, and F-score) require setting the prediction threshold. In the experiment, a threshold was selected according to the maximum F-score of the predictions. Finally the precision, recall, and F-score are calculated over this specific threshold.

4.3 Method Comparison

To evaluate our SLAMS algorithm, we applied it in a 10-fold cross-validation setting. To avoid easy prediction cases, we hid all the associations involved with 10% of the drugs in each fold, rather than hiding 10% of the associations. In our comparisons, we considered three multiple source integration methods: (1) **PREDICT** [13] that uses similarity measures as features, and learns a logistic regression classifier that weighs the different features to yield a classification score. Replicating the settings of Gottlieb *et al.* [13], the training set used for the PREDICT logistic regression classifier was the true drug-disease associations (positive set), and a randomly generated negative set of drug-disease pairs (not part of the positive set), twice as large as the positive set. (2) **Simple Average** that assumes that each data source is equally informative, thus simply averages all *k*-NN prediction scores from multiple data sources. (3) The **SLAMS** algorithm proposed in this paper that uses a large margin method to automatically weigh and integrate multiple data sources. All evaluation measures are summarized in Table 2.

Table 2. Comparison of SLAMS vs. alternative integration methods according to AUC, precision, recall, and F-score

Method	AUC	Precision	Recall	F-score
Simple Average	0.8662	0.3144	0.6085	0.4146
PREDICT	0.8740	0.3228	0.5987	0.4194
SLAMS	0.8949	0.3452	0.6505	0.4510

As shown in Table 2, our proposed SLAMS algorithm obtained an AUC score of 0.8949. The score was superior to the Simple Average (AUC = 0.8662) and PREDICT (AUC = 0.8740). Also our proposed SLAMS algorithm produced a higher precision of 34.52% and a recall of 65.05% compared with Simple Average (31.44% for precision and 60.85% for recall) and PREDICT (32.28% for precision and 59.87% for recall). The results showed that our proposed SLAMS algorithm, a large-margin method, is better at integrating multiple drug information sources than simple average and logistic regression strategies.

An interesting observation is that for all methods, the AUC score is quite large (around 0.9 in the experiment), but the actual ability to detect and predict positive samples (i.e., the new drug-disease pairs) is low: even for the best method in the experiment - SLAMS, on average 34.52% of its predicted indications will be correct and about 65.05% of the true indications will be revealed for the previously unseen drugs. The reason for this is that the drug repositioning task is a highly imbalanced problem where the dataset has an approximate 1:176 positive to negative ratio. Consequently, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis. Therefore, AUC scores can present an overly optimistic view of an algorithm's performance for the drug repositioning task. Unlike Li and Lu [12] and Gottlieb *et al.* [13], we reported precision, recall, and F-score in addition to AUC.

4.4 Data Source Comparison

In the study, the three data sources we used reveal three different aspects of a drug: (1) chemical properties - compound fingerprints; (2) biological properties - protein targets; (3) phenotypic properties - side-effect profiles. The weight vector $\boldsymbol{w}$ derived from SLAMS is interpretable: the i-th element of $\boldsymbol{w}$ corresponds to the i-th data source, and the sum of all elements of $\boldsymbol{w}$ is 1. The SLAMS weights of each data source and standard deviation during the 10-fold cross-validation are plotted in Fig. 3.

To further characterize the abilities of different data sources and/or their combinations to predict new drug-disease relationships (i.e., drug repositioning), we used SLAMS through a 10-fold validation with different data-source combinations. To conduct a fair and accurate comparison across different data sources, the same experimental conditions were maintained by using the same training drugs and test drugs for each fold. Fig. 4 shows the ROC curves for different data sources based on cross-validation experiments, and Table 3 summarizes the evaluation results.

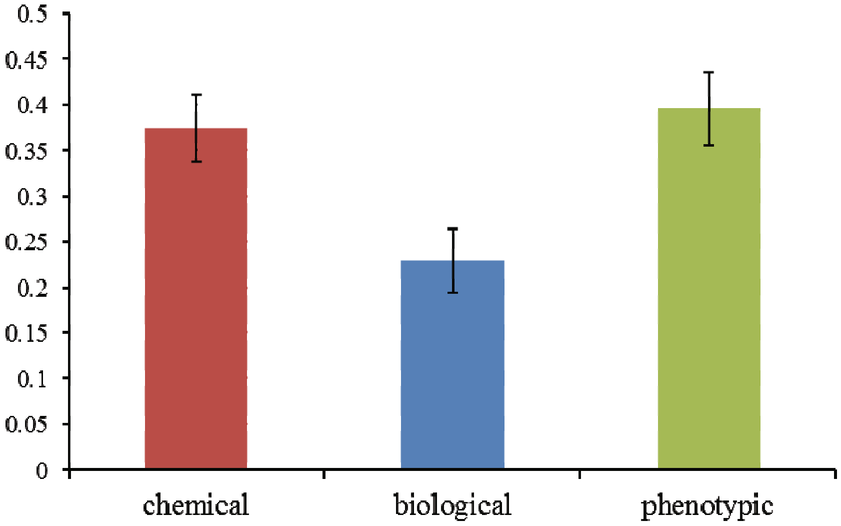

Fig. 3. Distribution of SLAMS weights and standard deviation for chemical, biological and phenotypic data sources in 10-fold cross-validation experiments

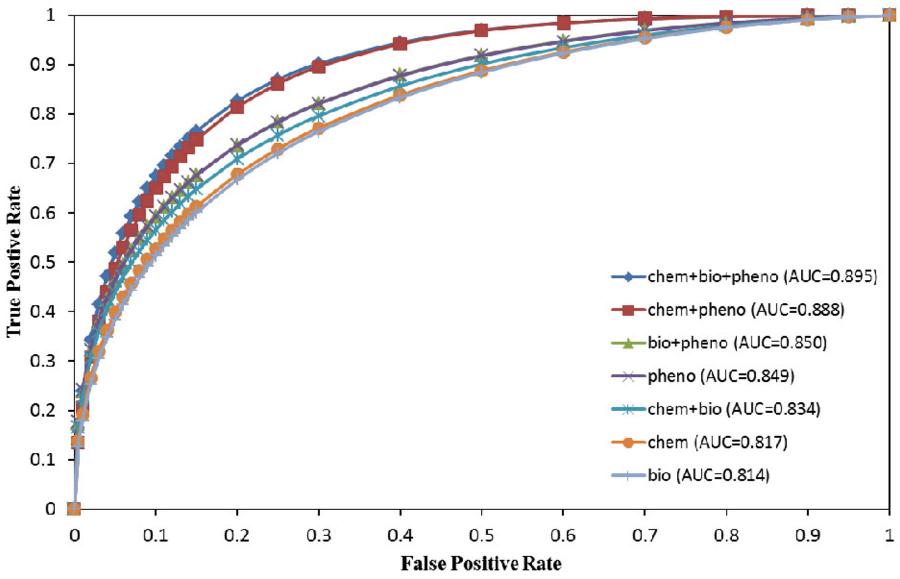

Fig. 4. The ROC comparison in 10-fold cross validation for various data-source combinations using SLAMS. chem: chemical properties; bio: biological properties; pheno: phenotypic properties. Data sources are sorted in the legend of the figure according to their AUC score.

Table 3. Comparison of various data-source combinations according to AUC, precision, recall, and F-score

Data Source	AUC	Precision	Recall	F-score
chem	0.8171	0.2232	0.4633	0.3013
bio	0.8139	0.2166	0.4592	0.2944
pheno	0.8492	0.2685	0.5117	0.3522
chem+bio	0.8339	0.2366	0.5012	0.3215
chem+pheno	0.8876	0.3281	0.6244	0.4302
bio+pheno	0.8503	0.2733	0.5119	0.3563
chem+bio+pheno	0.8949	0.3452	0.6505	0.4510

When the data sources were compared independently, the phenotypic data appeared to be the most informative (highest AUC of 0.8492), and chemical and biological data achieved similar AUC. This could be partially explained with the following reasons. Drug indications (i.e., drug's indicated diseases) and side effects are both measureable behavioral or physiological changes in response to the treatment. Intuitively, if drugs treating a disease share the same side-effects, this may be manifestation of some underlying mechanism-of-action (MOA) linking the indicated disease and the side-effect. Furthermore, both drug indications and side-effects are observations on human in the clinical stage, so there is less of a translational issue. Therefore, phenotypic data is a much more important drug information source with regard to predicting drug indications.

In the experiment while combing any two data sources will improve the AUC, the increase obtained by adding chemical structures on top of phenotypic properties (from 0.8492 to 0.8876) is much more significant than adding biological targets information on it (from 0.8492 to 0.8503). It seems that chemical properties and phenotypic properties are complementary. Combing all three data sources, we obtained the highest AUC score. On the other hand, if we focus on precision and recall, adding chemical properties to phenotypic properties yielded a dramatic increase (~22% in precision and recall). However, in our experiments there was no significant improvement when adding biological properties to phenotypic properties.

4.5 Analysis of Novel Predictions

During the 10-fold cross-validation, our SLAMS method with all chemical, biological, and phenotypic properties produced 3870 false-positive drug-disease associations. In other words, these associations were predicted by our method but they were not present in the gold standard. Some of these associations could be false, but a few associations could be true and can be considered as drug repositioning candidates in the real-world drug discovery. Taking the disease *Rheumatoid Arthritis* as an example, in Table 4 our SLAMS method found 10 drugs to treat it. These 10 drugs don't have associations with *Rheumatoid Arthritis* in the gold standard, and they have their own indications other than *Rheumatoid Arthritis*.

In order to test whether our predictions are in accordance with current experimental knowledge, we checked the extent to which they appear in current clinical trials. In Table 4, the drugs *Ramipril*, *Meloxicam*, and *Imatinib* have been tested for treating the disease *Rheumatoid Arthritis* in some clinical trials. In other words, pharmaceutical investigators have been aware of the associations of the drugs and *Rheumatoid Arthritis*, although they are still in the experimental stage. We downloaded all drug-disease data from registry of federally and privately supported clinical trials conducted around the world[5]. Overall, we acquired 18,392 unique drug-disease associations that are being investigated in clinical trials (phases I-IV). In all, 4798 of these associations involve drugs and diseases that are present in our data set with the exact names, spanning 4066 associations that are not part of our gold standard. Of these 4066 associations, our 3870 false-positive drug-disease associations cover 21% (i.e., 854 associations). It was highly unlikely that our false-positive predictions identified this set of 854 experimental drug-disease associations by chance ($p < 0.0001$, Fisher's exact test [25]). Hence, we conclude that false-positive novel uses predicted by our method attained significant coverage of drug-disease associations tested in clinical trials. All predicted drug-disease associations in our experiments are available at our website[6].

[5] Clinical trials found at `http://clinicaltrials.gov/`.

[6] Available at `http://astro.temple.edu/~tua87106/drugreposition.html`.

Table 4. Repositioned drugs for Rheumatoid Arthritis predicted by our method

Drug Name	Original Uses	Treat Rheumatoid Arthritis in clinical trial
Ramipril	Hypertension Diabetic Nephropathies Heart Failure	NCT00273533 proposed in Jan 2006
Lisinopril	Hypertension Heart Failure	N/A
Mercaptopurine	Lymphoma	N/A
Meloxicam	Osteoarthritis	NCT00042068 proposed in July 2002
Mefenamic Acid	Menorrhagia Fever Dysmenorrhea	N/A
Zileuton	Asthma	N/A
Imatinib	Gastrointestinal Neoplasms Leukemia, Myeloid Blast Crisis	NCT00154336 proposed in Sept 2005
Allopurinol	Gout	N/A
Imiquimod	Condylomata Acuminata	N/A
Masoprocol	Keratosis	N/A

5 Conclusion

In response to the high cost and risk in traditional *de novo* drug discovery, discovering potential uses for existing drugs, also known as drug repositioning, has attracted increasing interests from both the pharmaceutical industry and the research community. From a serendipitous drug repositioning to systematic or rational ways, a variety of computational approaches using single source have been developed. However, the complexity of the problem clearly needs methods to integrate drug information from multiple sources for better solutions.

In this paper, we proposed SLAMS, a new drug repositioning framework by integrating chemical (i.e., compound signatures), biological (i.e., protein targets), and phenotypic (i.e., side effects) properties. Experimental results showed that our method is superior to a few existing computational drug repositioning methods. Furthermore, our predictions statistically overlap drug-disease associations tested in clinical trials, suggesting that the predicted drugs may be regarded as valuable repositioning candidates for further drug discovery research. An important property of our method is that it allows easy integration of additional drug information sources. Moreover, the method ranked multiple drug 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.

Acknowledgment. This project was funded in part under a grant with the GlaxoSmithKline LLC.

References

1. Hurle, M.R., Yang, L., Xie, Q., Rajpal, D.K., Sanseau, P., Agarwal, P.: Computational drug repositioning: from data to therapeutics. Clin. Pharmacol. Ther. 93(4), 335–341 (2013)
2. Ashburn, T.T., Thor, K.B.: Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. Nature Reviews Drug Discovery 3, 645–646 (2004)
3. Sardana, D., Zhu, C., Zhang, M., Gudivada, R.C., Yang, L., Jegga, A.G.: Drug repositioning for orphan diseases. Brief Bioinform 12(4), 346–356 (2011)
4. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., Whaley, R., Glennon, R.A., Hert, J., Thomas, K.L., Edwards, D.D., Shoichet, B.K., Roth, B.L.: Predicting new molecular targets for known drugs. Nature 462, 175–181 (2009)
5. Li, J., Zhu, X., Chen, J.Y.: Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. PLoS Comput. Biol. 5(7), e1000450 (2009)
6. Kotelnikova, E., Yuryev, A., Mazo, I., Daraselia, N.: Computational approaches for drug repositioning and combination therapy design. J. Bioinform Comput. Biol. 8(3), 593–606 (2010)
7. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P.: Drug target identification using side-effect similarity. Science 321, 263–266 (2008)
8. Yang, L., Agarwal, P.: Systematic Drug Repositioning Based on Clinical Side-Effects. PLoS ONE 6(12), e28025 (2011)
9. Hu, G., Agarwal, P.: Human Disease-Drug Network Based on Genomic Expression Profiles. PLoS ONE 4(8), e6536 (2009)
10. Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A., Sage, J., Butte, A.J.: Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci. Transl. Med. 3(96), 96ra77 (2011)
11. 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. 39(Web Server Issue), W492–W498 (2011)
12. Li, J., Lu, Z.: A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity. In: IEEE International Conference on Bioinformatics and Biomedicine (2012)
13. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. 7, 496 (2011)
14. Hotelling, H.: Relations between two sets of variates. Biometrika 28, 321–377 (1936)
15. Davis, J., Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. In: International Conference on Machine Learning (2006)
16. Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H.: PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 37(Web Server Issue), W623–W633 (2009)
17. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E.: The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J. Chem. Inf. Comput. Sci. 43(2), 493–500 (2003)

18. Smith, T.F., Waterman, M.S., Burks, C.: The statistical distribution of nucleic acid similarities. Nucleic Acids Res. 13, 645–656 (1985)
19. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology 6, 343 (2010)
20. Pauwels, E., Stoven, V., Yamanishi, Y.: Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics 12, 169 (2011)
21. Pandey, G., Myers, C.L., Kumar, V.: Incorporating functional inter-relationships into protein function prediction algorithms. BMC Bioinformatics 10, 142 (2009)
22. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 36(Database Issue), D901–D906 (2008)
23. Olivier, B.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32(Database Issue), D267–D270 (2004)
24. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.: UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 32(Database Issue), D115–D119 (2004)
25. Upton, G.: Fisher's exact test. Journal of the Royal Statistical Society, Series A 155(3), 395–402 (1992)