JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 2, 2011

Pp. 133–145
DOI: 10.1089/cmb.2010.0213

Research Articles

Combining Drug and Gene Similarity Measures for Drug-Target Elucidation

*LIAT PERLMAN,[1] *ASSAF GOTTLIEB,[1] NIR ATIAS,[1]
EYTAN RUPPIN,[1,2] and RODED SHARAN[1]

ABSTRACT

Understanding drugs and their modes of action is a fundamental challenge in systems medicine. Key to addressing this challenge is the elucidation of drug targets, an important step in the search for new drugs or novel targets for existing drugs. Incorporating multiple biological information sources is of essence for improving the accuracy of drug target prediction. In this article, we introduce a novel framework—Similarity-based Inference of drug-TARgets (SITAR)—for incorporating multiple drug-drug and gene-gene similarity measures for drug target prediction. The framework consists of a new scoring scheme for drug-gene associations based on a given pair of drug-drug and gene-gene similarity measures, combined with a logistic regression component that integrates the scores of multiple measures to yield the final association score. We apply our framework to predict targets for hundreds of drugs using both commonly used and novel drug-drug and gene-gene similarity measures and compare our results to existing state of the art methods, markedly outperforming them. We then employ our framework to make novel target predictions for hundreds of drugs; we validate these predictions via curated databases that were not used in the learning stage. Our framework provides an extensible platform for incorporating additional emerging similarity measures among drugs and genes. Supplementary Material is available at www.liebertonline.com/cmb.

Key words: computational molecular biology, gene expression, gene networks, genetic variation, machine learning, sequence analysis.

1. INTRODUCTION

DECIPHERING DRUG TARGETS IS A PRIMARY TASK in the development of new drugs, in finding new ways to utilize existing drugs, and in pinpointing their side effects. Experimental identification of drug-target associations remains a laborious and costly task (Haggarty et al., 2003), calling for faster computational prediction methods. Such methods can be used to augment the limited available information on drug targets, which is in sharp contrast to the vast number of compounds existing in chemical databases.

[1]The Blavatnik School of Computer Science and [2]School of Medicine, Tel Aviv University, Tel-Aviv, Israel.
*These authors contributed equally to this work.

Early attempts in computational prediction included docking simulations (Cheng et al., 2007) and text mining (Zhu et al., 2005). The former, however, can be applied only to targets with known three-dimensional (3D) structure. The latter searches for co-occurrences of drugs and genes in texts, and is limited to current knowledge and prone to detection problems due to multiple gene and compound names. Additional attempts were based on reverse engineering of gene regulatory networks, inferring possible targets from cellular responses to administered drugs (di Bernardo et al., 2005; Gardner et al., 2003; Mani et al., 2008). These methods suffer from the complex and noisy nature of molecular networks. Recently, several algorithms have been proposed to predict drug-target associations by combining drug-drug and gene-gene similarity measures (Bleakley and Yamanishi, 2009; Campillos et al., 2008; Keiser et al., 2009; Yamanishi et al., 2008). The key assumption underlying these algorithms is that similar drugs tend to share similar targets (Mitchell, 2001). This has been observed with respect to chemical similarity (Martin et al., 2002; Schuffenhauer et al., 2003), side effect similarity (Campillos et al., 2008), and more.

Several authors had previously predicted drug-target interactions by combining chemical drug-drug similarity and sequence-based gene-gene similarity (Bleakley and Yamanishi, 2009; Yamanishi et al., 2008). Keiser et al. (2009) compared the chemical structure of drugs to a compendium of ligands, known to modulate the function of protein receptors, providing indirect connections between drugs and targets via these ligands. Several approaches, concentrating mainly on indirect drug-gene associations, employed additional similarity measures to gain insights on drugs. Specifically, protein-protein interaction (PPI) network similarity was used in Hansen et al. (2009) to predict drug-gene genetic interactions (termed ''pharmacogenes''), and gene expression data was combined with drug-response data in Kutalik et al. (2008) to infer co-modules of genes and drugs. Last, a recent approach used information on compound-induced fitness defects of yeast deletion strains to predict drug-targets in *S. cerevisiae* (Hillenmeyer et al., 2010).

Different similarity measures utilize different attributes of drugs or genes, hence the use of just one or a few of them may miss information that is relevant to predicting new associations. In particular, the two-dimensional (2D) chemical space does not always comply with three-dimensional (3D) structural similarity (Kuhn et al., 2008). Moreover, sequence similar targets tend to bind the same ligands in the case of G-protein coupled receptor (GPCR) targets but less so for protein kinases (Kuhn et al., 2008). The use of pharmacological information is limited to marketed drugs whose indication/side effect information is available (Campillos et al., 2008).

To overcome these limitations, we have designed a new prediction scheme—Similarity-based Inference of drug-TARgets (SITAR)—that integrates multiple measures to facilitate the prediction task. Our contribution is twofold: (i) We introduce novel drug-drug similarity measures and combine them into the prediction process; and (ii) we propose a way of integrating the drug-drug and gene-gene similarities to create classification features. The result is a new drug-target prediction algorithm, which markedly outperforms previous methods and can cope with new drugs with no known targets.

2. RESULTS AND DISCUSSION

2.1. SITAR: an algorithm for predicting drug targets

We designed a drug-target prediction algorithm with three main components (Fig. 1): (i) drug-drug and gene-gene similarity computations; (ii) combining the drug and gene similarity measures into classification features; and (iii) feature selection and prediction using logistic regression. In the following, we describe these components in detail.

2.2. Similarity measures

In order to overcome the limitations engulfed in using similarity measures of a single type, we set out to incorporate a multitude of similarity measures, including both novel and already published ones. Overall, we considered five drug-drug similarities and three gene-gene similarities from different biological and chemical sources. The drug-drug similarity measures were computed using chemical, registered and predicted drug side effects (Kuhn et al., 2010) of the drug, drug response gene expression profiles, and the Anatomical, Therapeutic and Chemical (ATC) classification system. The gene-gene similarity measures used are based on sequence, closeness in a protein-protein interaction network, and semantic Gene

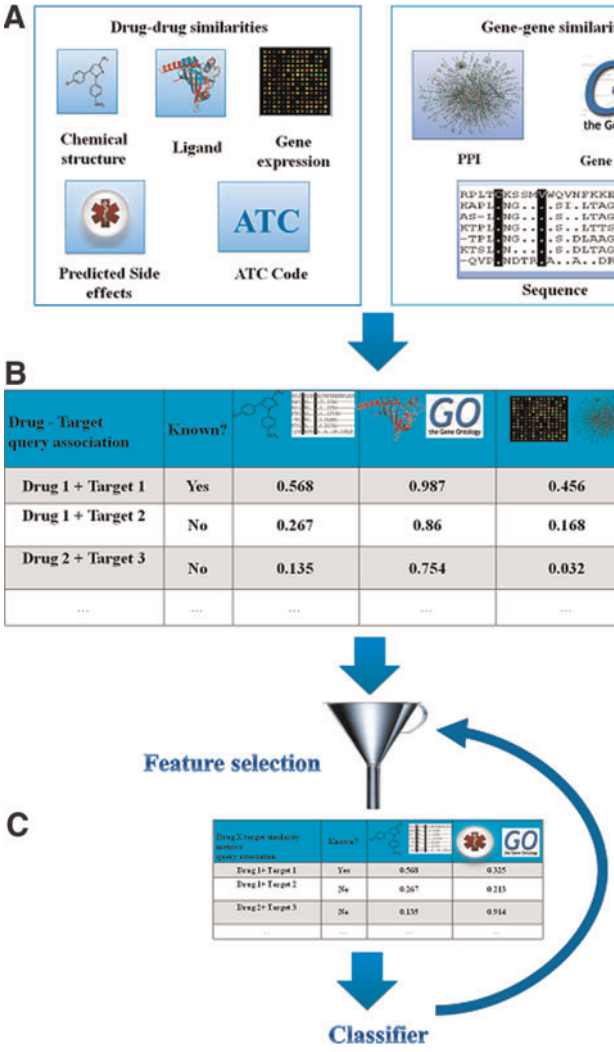

FIG. 1. Algorithm pipeline, comprised of formation of drug-drug and gene-gene similarity matrices **(A),** integration of the similarities to classification features **(B),** and classification with wrapper feature selection **(C).**

Ontology (GO) similarities. Overall, 315 drugs and 250 protein targets were represented in all measures, spanning 782 known associations.

2.3. Feature construction and classification

At the heart of our algorithm lies the process of exhaustive construction of *classification features* that span the entire pairwise space of drug-target measures' combinations. That is, each feature is constructed based on one drug-drug similarity measure and one gene-gene similarity measure. It is calculated by combining the drug-drug similarities between the query drug and other drugs and the gene-gene similarities between the query gene and other target genes across all true drug-target associations. The features are automatically combined using a logistic regression classifier that is coupled with a wrapper feature selection procedure and yield the final classification scores.

2.4. Feature selection and performance evaluation

We performed feature selection using both forward selection and backward elimination, converging to a selected set of ten features, constructed from pairs of drug-drug and gene-gene similarity measures. The area under the precision-recall curve (AUPR) scores before and after the feature selection phase, as well as the AUPR achieved when using each of the ten selected features are listed in Table 1. Similar results were obtained when using an SVM classifier (see Methods below, as well as Table S1 in the Supplementary Material, available at www.liebertonline.com/cmb). Examining the individual contribution of each of the

Table 1. AUPR and AUC Scores Calculated Using Different Feature Sets

Drug similarity	*Gene similarity*	*AUPR*[a]	*AUC*[b]
All features		0.905	0.935
Selected features		0.908	0.935
Ligand	Sequence similarity	0.851	0.867
Ligand	GO semantic similarity	0.845	0.867
Predicted Side Effect	GO semantic similarity	0.832	0.863
ATC similarity	GO semantic similarity	0.81	0.858
Ligand	PPI closeness	0.809	0.844
Chemical	GO semantic similarity	0.805	0.84
ATC similarity	PPI closeness	0.762	0.809
Chemical	Sequence similarity	0.749	0.763
Predicted Side Effect	PPI closeness	0.729	0.759
Co-expression	Sequence similarity	0.724	0.748

[a]All standard deviations are below AUPR of 0.01.
[b]All standard deviations are below AUC of 0.005.

features, we find that the best AUPR scores are achieved using the ligand-based drug-drug similarity of Keiser et al. (2009) together with sequence similarity for genes (AUPR = 0.85). The least contributing combination is the drug co-expression together with PPI network closeness (AUPR = 0.54). Accordingly, we find that the best drug-drug similarity averaged over all gene-gene similarities is the ligand similarity (AUPR = 0.83 ± 0.02), and the worst is the co-expression similarity (AUPR = 0.67 ± 0.11). For genes, the best gene-gene measure across all drug-drug similarities is the GO semantic similarity, and the worst is the PPI closeness measure. We note that the feature selection process did not affect the results significantly, but we expect it to have more impact when additional similarity measures are incorporated as features.

We further assessed the classification quality of different target types (GPCRs, ion channels, enzymes, and nuclear receptors), as in Bleakley and Yamanishi (2009). The results are listed in Table 2, with GPCRs attaining the best scores.

2.5. Comparison to other drug-target prediction methods

We compared our method to two state-of-the-art methods:

(i) The kernel regression-based method (KRM) of Yamanishi et al. (2008) embeds drugs and targets into a unified Euclidean space termed the "pharmacological space," using a regression model. Predicted interacting drug-gene pairs are those that are closer to each other below a certain threshold in the pharmacological space.

(ii) The bipartite local models (BLMs) method of (Bleakley and Yamanishi, 2009) constructs local models to learn drug-target associations based on additional targets of the query drug and additional drugs targeting the query target. We note that the SEA tool of Keiser et al. (2007) provides receptors code names that cannot be mapped to our list of targets, precluding a direct comparison to their method. Figure 2 displays the precision-recall curves of the three methods, and Table 3 summarizes the AUPR and AUC scores between the different methods, overall demonstrating the marked improvement obtained by our new method (AUPR of 0.908, exceeding the KRM and BLM methods by 0.07 and 0.15 AUPR difference, respectively).

Table 2. AUPR Scores Calculated for Different Target Types

Target type	*Number*	*AUPR*	*AUC*
GPCR	49	0.939	0.946
Ion channels	37	0.889	0.927
Enzymes	94	0.877	0.922
Others	56	0.87	0.935
Nuclear receptors	14	0.851	0.863

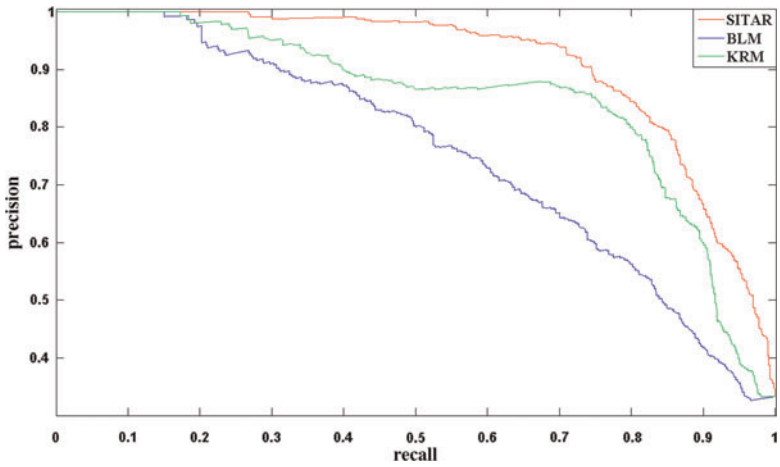

FIG. 2. Precision-recall curves of our logistic regression classifier, KRM and BLM methods.

2.6. *Novel predictions*

After demonstrating the utility of our method in predicting drug target associations, we set out to predict novel targets for drugs in our data set. We focused on the best scoring gene for each drug, obtaining putative novel targets for 307 of the 315 drugs used in this study. To validate these predictions, we compared the predicted associations to those reported in two other sources of drug-target interactions that were not used in the learning process: the Therapeutic Target Database (TTD) (Zhu et al., 2010) and the Matador database (Gunther et al., 2008). TTD included only six interactions that were not part of our original data. Two of them were predicted by us ($p < 2.4e^{-4}$). Matador, on the other hand, reported 219 additional interactions, out of which we predicted thirteen ($p < 8.4e^{-12}$).

For an additional validation, we utilized information on drug-related pathways of KEGG (Kanehisa et al., 2010) and REACTOME (Matthews et al., 2009) databases, as cataloged in the Comparative Toxicogenomics Database (CTD) (Davis et al., 2009). Out of the 315 drugs used in this paper, 217 drugs are associated with known pathways in CTD. For each drug, we constructed a merged list of genes that appear in all the pathways associated with that drug. We then searched for enrichment of the predicted targets of each drug in this merged list. We found 20 drugs whose pathways are enriched for our predicted targets (at a false discovery rate [FDR] of 0.05). For comparison, we randomly permuted the association between drug target predictions and drug pathways and recomputed the number of enriched drugs. Across 100 such permutation tests, 9.2 ± 3.4 drugs were enriched on average, with all tests attaining less than 20 enrichments ($p < 0.01$).

The distribution of target types in our novel predictions is different from the general distribution of known targets. Specifically, nuclear receptors and GPCRs are overrepresented in our novel predictions relative to their general abundance in all the targets (relative abundance in novel predictions is 2.4 and 1.9 times the relative abundance in drug targets for nuclear receptors and GPCRs, respectively). This overrepresentation is attributed to higher scores achieved by drugs associated with nuclear receptors and GPCRs due to higher sequence similarity of nuclear receptors and GPCRs relative to other targets (two and 1.7 times more similar than ion channels, respectively). In addition, nuclear receptors are much closer to each other on the PPI network (average PPI distance between nuclear receptors is smaller by a factor of 0.6 from all other types). We note that dopamine receptor D3 (DRD3), having only two known interactions in our data, had exceptionally high occurrence in our novel prediction set (28 times). Dopamine receptor D3 is primarily predicted to be targeted by drugs indicated for Parkinson's disease and schizophrenia (eight out of

Table 3. Comparison of AUPR and AUC Scores Calculated Using Different Methods

Measure type	*AUPR*	*AUC*
This work	0.908	0.935
KRM (Yamanishi et al., 2008)	0.838	0.884
BLM (Bleakley and Yamanishi, 2009)	0.754	0.814

nine existing in DrugBank). It is noteworthy that half of the predicted drugs targeting DRD3 are also indicated for Parkinson's disease or schizophrenia. These predictions make DRD3 a promising candidate for further investigation.

Next, we applied our method to provide novel predictions for drugs that have no known interacting targets in DrugBank. There are 20 such drugs that are included in all our similarity measures. Applying a cross-validation scheme and choosing a classification threshold maximizing the F1-measure, we could predict targets for 14 drugs. The top 5 scoring predictions are listed in Table S2 in the Supplementary Material. Evaluating our set of top predictions for drugs with unknown targets, we find indirect evidence for the first four of the top five associations. Specifically, Theobromine is predicted to target Adenosine A1 receptor (Adora1). Theobromine, derived from the cacao bean, belongs to the methylxanthine class of chemical compounds. Other methylxanthine derivatives such as Caffeine and Theophylline are reported to target Adenosine A1 receptors (Fredholm and Persson 1982). Cefotetan is predicted to target Paraoxonase 1 (PON1). Cefotetan belongs to the Cephamycin class of antibiotics, which is often classified with second generation Cephalosporins; the latter inhibits PON1 (Sinan et al., 2006). Rolitetracycline is predicted to target Cytochrome C (CYCS) and Caspase 3 (Casp3) with equal score. Rolitetracycline is a member of the tetracycline family of antibiotics. Other members from this family such as Minocycline and Doxycycline are reported to inhibit Cytochrome C (Minocycline [Zhu et al., 2002])) and Caspase 3 (Minocycline and Doxycycline [Chen et al., 2000; Mouratidis et al., 2007]). Last, Sulfamerazine is predicted to target Dihydrofolate reductase (Dhfr). Sulfamerazine belongs to antibacterial sulfonamides whose inhibition of Dihydrofolate reductase has been extensively studied (Rollo, 1970).

Last, we applied our method to detect genes that to date are not recognized as drug targets. In order to reduce the number of tested genes, we constructed a list of the most promising targets in the following way: for each of the 250 known targets, we took the most similar gene in at least one of the three used gene-gene similarity measures (sequence similarity, closeness on the protein-protein interaction network and GO semantic similarity). We thus ended with 135 potential new targets on which we applied a similar procedure as in the case of drugs with no known targets: we performed a 10-fold cross validation to obtain a classification threshold that was used to classify the 135 potential targets. We then considered the top ranked predicted drug for each of the 135 potential targets. The top five drug predictions are listed in Table S3 in the Supplementary Material, all of which are supported by current knowledge as discussed below. Specifically, Fluorometholone is predicted to target calreticulin (CALR). Fluorometholone is a glucocorticoid, which is a class of steroid hormones that binds to the glucocorticoid receptor (GR). Interestingly, it has been shown that CALR binds to the GR, thus inhibiting the effect of glucocorticoids (Burns et al., 1997). Menadione (vitamin K3) is predicted to target coagulation factor III (tissue factor, F3). According to DrugBank, Menadione is involved as a cofactor of vitamin K–dependent coagulation factors II, VII, IX, and X, two of which are known to be part of the tissue factor pathway (Thomson et al., 2002). In addition, tissue factor is known to be targeted by coagulation factor VIIa, which is itself a vitamin K–dependent glycoprotein. Methyldopa, primarily used for hypertension, is predicted to target nitric oxide synthase 1 (neuronal; NOS1). It has been shown that Methyldopa is associated with changes in nitric-oxide synthesis (Podjarny et al., 2001). Clofibrate is predicted to target apolipoprotein E (APOE). According to DrugBank, Clofibrate inhibits the synthesis and increases the clearance of apolipoprotein B, which is part of the low-density lipoproteins (LDL). APOE has a role in the conversion of very-low-density lipoproteins to LDL (Ehnholm et al., 1984). Finally, Bacitracin, an antibiotic, is predicted to target thrombospondin 1 (Thbs1/TSP1). While we could not find a direct connection between the two, it is noteworthy that Bacitracin acts by interfering with bacterial cell wall maintenance of Gram-positive organisms (Stone and Strominger, 1971). Interestingly, it has been shown that TSP1 promotes cellular adherence of Gram-positive pathogens via the recognition of peptidoglycan, the main component of the cell wall (Rennemeier et al., 2007).

3. CONCLUSION

We introduced a novel method, SITAR, for predicting drug-target interactions. Our method incorporates an extensive set of drug-drug and gene-gene similarity measures. Newly incorporated drug-drug similarities are based on predicted side effects, gene expression drug response profiles, and the ATC classification system. The classification features are constructed based on a new score integrating the drug-drug and gene-gene similarity spaces. These features are integrated via a logistic regression classifier, combined with

a feature selection process. Our method is flexible and allows the incorporation of new emerging measures without altering already computed scores on other measures. Using our method, we show marked improvement of classification performance over previous drug-target prediction approaches.

We provide novel predictions of drug-target interactions and validate them against public databases. Last, we predict targets for drugs which to-date have no known targets.

Having shown that our method is robust with respect to different score choices, selected features, and different classification methods, it seems that the primary reason for the increased performance compared to previous methods stems from the use of multiple similarity measures. Each of the resulting features alone does not have enough predictive power, but the combination of multiple features allows the classification procedure to perform well. Accordingly, we noticed that using a low number of features (less than five) deteriorates the results. Nevertheless, our method can be enhanced in several ways. First, one could improve and expand the measures used. Of special interest is improving the gene co-expression similarity based on the Connectivity Map data, which currently exhibits the worst performance. Another extension would be to increase the number of represented drugs and genes shared between the different measures. This could be achieved either by predicting missing similarities from existing ones (Atias and Sharan, 2010) or by incorporating imputation methods to overcome missing information in some of the measures.

4. METHODS

4.1. Similarity measures

We defined and computed five drug-drug similarity measures and three gene-gene similarity measures. All similarity measures were normalized to be in the range [0, 1].

We used the following drug-drug similarity measures:

(1) *Chemical-based:* Canonical simplified molecular input line entry specification (SMILES) of the drug molecules were downloaded from DrugBank (Wishart et al., 2008). Hashed fingerprints were computed using the Chemical Development Kit (CDK) with default parameters (Steinbeck et al., 2006). The similarity score between two drugs is computed on their fingerprints according to the two-dimensional Tanimoto score (Tanimoto, 1957), which is equivalent to the Jaccard score (Jaccard, 1908) of their fingerprints, i.e., the size of the intersection over the union when viewing each fingerprint as specifying a set of elements.

(2) *Ligand-based:* The Similarity Ensemble Approach (SEA) (Keiser et al., 2007) relates protein receptors based on the chemical 2D similarity of the ligand-sets modulating their function. Given a drug's canonical SMILES, the SEA search tool compares it against a compendium of ligand-sets and computes *E*-values for those ligand sets. To compute a drug-drug similarity, we queried drugs using their canonical SMILES on the SEA tool. To obtain robust results, we queried the drug against the two ligand databases provided in the tool (MDL Drug data report and WOMBAT [Olah et al., 2005]) and used two different methods to compute the drug fingerprint (Scitegic ECFP4 and Daylight), resulting in four lists of similar ligand sets. Unifying the four lists and filtering drug-ligand set pairs with E-values $>10^{-5}$, we obtained a list of relevant protein-receptor families for each drug. Finally, the similarity between a pair of drugs was computed as the Jaccard score between the corresponding sets of receptor families. We note that, due to a partial mapping of the receptor families to proteins, we could not use the drug-receptor associations directly as classification features.

(3) *Expression-based:* Gene expression responses to drugs were retrieved from the Connectivity Map project (Lamb et al., 2006). We experimented with three different methods to calculate drug similarity from Connectivity Map ranked gene expression profiles: (i) Spearman rank correlation coefficient; (ii) calculating a Jaccard score between the 500 most differentially expressed genes (250 most up-regulated and 250 most down-regulated genes); and (iii) using the method proposed by (Iorio et al., 2009), employing Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) as a similarity measure. We dealt with multiple experiments per drug as follows: In the Spearman calculation, we averaged over the *d1xd2* different correlation coefficients obtained between the *d1* experiments of one drug against the *d2* repeated experiments of the second drug. In the Jaccard case, we used differentially expressed genes that appeared in at least 50% of the gene expression responses to the same drug. The method proposed by Iorio et al. (2009) handles repeated experiments of the same drug through iterative merging.

(4) *Side-effect based:* Drug side effects were obtained from SIDER (Kuhn et al., 2010), an online database containing drug side effects associations extracted from package inserts using text mining methods. Recently, we developed an algorithmic framework to predict side effects for drugs by combining side effect information on known drugs with their chemical properties (Atias and Sharan, 2010). Following this work, we defined the similarity between drugs according to the Jaccard score between their top ten predicted side effects.
(5) *Annotation-based:* We used the World Health Organization (WHO) ATC classification system (Skrbo et al., 2004). This hierarchical classification system categorizes drugs according to the organ or system on which they act, their therapeutic and their chemical characteristics. ATC codes were obtained from DrugBank. To define a similarity between ATC terms, we used the semantic similarity algorithm of Resnik (1999). This algorithm associates probabilities $p(x)$ with all the nodes (i.e., terms) x in the hierarchy and calculates the similarity of two drugs as the maximum over all their common ancestors c of $-\log(p(c))$.

The gene-gene similarity measures we used include:

(1) *Sequence similarity:* based on a Smith-Waterman sequence alignment score (Smith et al., 1985). Following the normalization suggested in Bleakley and Yamanishi (2009), we divide the Smith-Waterman score between two protein sequences by the geometric mean of the scores obtained from aligning each sequence against itself.
(2) *Closeness in a protein-protein interaction (PPI) network*: Human protein-protein interactions were compiled from the literature (Breitkreutz et al., 2008; Ewing et al., 2007; Rual et al., 2005; Stelzl et al., 2005; Xenarios et al., 2002). The distances between each pair of genes were calculated on their corresponding proteins using an all-pairs shortest paths algorithm. Distances were transformed to similarity values using:

$$S(p,p') = Ae^{-bD(p,p')} \tag{1}$$

where $S(p,p')$ is the computed similarity value between two proteins, $D(p,p')$ is the shortest path between these proteins in the PPI network, and A, b are free parameters.

(3) *Gene Ontology (GO) semantic similarity:* GO annotations (Ashburner et al., 2000) were downloaded from UniProt (Jain et al., 2009). Semantic similarity scores between targets were calculated according to Resnik (1999), using the csbl.go R package (Ovaska et al., 2008). All three ontologies were used in the computation as similar drugs are expected to interact with proteins that act in similar biological processes, or have similar molecular functions or reside in similar compartments.

4.2. Combining measures

We view the drug target prediction problem as a classification problem, where the goal is to learn a classifier that can distinguish true and false drug-target associations. True drug-target interactions were retrieved from KEGG DRUG (Kanehisa et al., 2010), DrugBank (Wishart et al., 2008), and DCDB (Liu et al., 2010). An independent set of drug-target interactions was downloaded from Matador (Gunther et al., 2008) for validation purposes only. The classification features that we use are constructed from scores calculated on pairs combined of one of M drug-drug measures (five in our case) and one of N gene-gene similarity measures (three in our case), resulting in an MxN set of drug-gene measure features for each drug-target association (15 in our case). For a given pair of measures (i.e., feature), the score of a given association (d,t) is calculated by considering the similarity, according to the given pair of measures, of all true drug-target interactions to this association. The computation is done as follows: First, for each true interaction (d',t') we compute the drug-drug similarity $S(d,d')$ and the target-target similarity $S(t,t')$. Next, we combine the two similarities to a single score. Finally, we integrate the scores of all true interactions to form the classification feature.

We experimented with several ways of combining the two similarities and integrating the scores over the various true associations, aiming to compute a score that would differ as much as possible between true and false drug-target interactions. This difference was tested using the Wilcoxon rank sum test for equal medians. To combine a drug-drug and a gene-gene similarity, we considered two types of averaging functions: arithmetic mean and geometric mean. The two options scored similarly ($p < 1e^{-10}$) with a slight advantage to

geometric mean (AUPR difference <0.01), which we used in the sequel. To integrate the scores of the known associations, we studied the effect of averaging the top-k scores for various selections of k and discovered that taking only the top scoring drug-target interaction yields the lowest p-value, as can be seen in Figure 3. This effect was observed also under the arithmetic mean score.

To conclude, we used the maximum value over the following weighted version of geometric mean to score drug target associations:

$$Score(d,t) = \max_{d',t' \neq d,t} S(d,d')^r \cdot S(t,t')^{(1-r)}, 0 \leq r \leq 1 \quad (2)$$

The optimization of the scoring parameter r was done in a cross validation setting.

4.3. Measure construction considerations

The transformation of a data source into a similarity measure can be done in various ways. In order to choose a transformation that will perform well in drug target prediction, we applied the Wilcoxon rank sum test to choose the transformation with greatest separation power. Precisely, a suggested drug-drug similarity measure was combined with all the gene-gene similarity measures to form classification features. A Wilcoxon rank-sum test was then applied to each of these features to test whether the medians of the distributions of feature values on true and false drug-targets interactions differ. The logarithms of the resulting p-values were averaged to yield the separation score for that suggested measure. An analogous procedure was applied to score gene-gene similarity measures.

We used these separation scores to specify two measures: the co-expression based drug similarity and the PPI-based gene similarity. For the co-expression based drug similarity, Jaccard-based similarities scored better (average rank sum $p < 7e^{-89}$) than GSEA-based similarities (average rank sum $p < 2e^{-35}$) and Spearman-based similarities ($p < 3e^{-11}$). The decreased performance of the Spearman-based similarity measure is understandable in light of the fact that gene expression profiles tend to be informative mainly with regard to the top (upregulated) and bottom (downregulated) genes. For the PPI-based gene similarity, we used the separation score in order to choose the optimal parameters A and b (equation 1). Hence, we picked $b = 1$ and chose A such that direct network neighbors would get a value close to 1 ($A = 0.9$). However, the results were robust over a wide range of parameters; similar AUPRs (AUPR difference <0.003) were obtained with b ranging from 0.5 to 2 and with A ranging from 0.5 to 1. Self similarities were set to 1.

4.4. Prediction assessment and parameter optimization

We used a 10-fold cross validation scheme to evaluate the accuracy of our prediction algorithm. The training set used for the cross validation included 782 true drug-target interactions and a randomly generated set of drug-gene pairs (not part of the positive set), twice larger than the positive set. We note that taking negative set of equal size or three times larger did not alter the results significantly (data not shown) (Li and Lai, 2007).

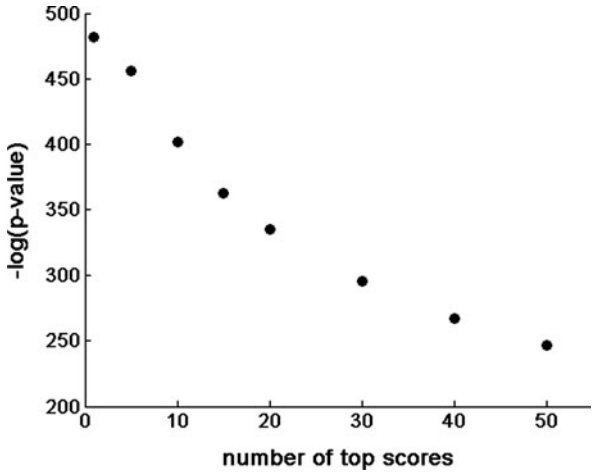

FIG. 3. *-log(p-value)* of the Wilcoxon rank sum test for the distribution of scores of true versus false drug-target associations. The score was computed by averaging the top k scores, k ranging from 1 to 50.

In order to improve the classification accuracy, we optimized the weight factor r in equation 2 and performed feature selection on the 15 features associated with the various pairs of measures. Specifically, we trained a logistic regression classifier where the data was divided into 10 parts and we iteratively selected eight parts to be the training set, a ninth part was considered a test set on which we optimized the parameter r and the 10th part was considered the validation set, on which the performance of the classification was assessed. We found that the AUPR was robust to the selection of the weight r, with slightly better classification performance using $r = 0.4$ (AUPR difference <0.02).

To select features, we used a wrapper feature selection procedure, which is based on a classification accuracy measure (Guyon and Elisseeff, 2003). A common method to avoid exhaustively searching for the optimal set of features in terms of classification accuracy is through an iterative greedy algorithm. One variant, termed *forward selection*, starts with an empty set and adds in each step the most contributing feature, while another, termed *backward elimination*, starts with the full set and removes the least contributing feature in each step. These two algorithms yield nested subsets of features from which a set attaining the highest accuracy is finally chosen.

We considered two different scores as indicators for the classification accuracy: the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve (Fawcett, 2005) and the AUPR curve (Davis and Goadrich, 2006). Standard deviations for both measures were assessed by using 100 random partitions of the training data into cross validation sets. Importantly, when dealing with uneven number of true and false examples, precision-recall (PR) curves give a more informative picture of an algorithm's performance (Davis and Goadrich, 2006). Furthermore, Davis and Goadrich (2006) show that a curve dominates in ROC space if and only if it dominates in PR space. Hence, we focus on the AUPR criterion as an indication for classification performance. We evaluated the classification quality also in regard to different target types, including enzymes, G-protein coupled receptors (GPCRs), ion-channels, and nuclear receptors. The target type was inferred from DrugBank and Uniprot annotations.

The classification was done using MATLAB implemented logistic regression. A cutoff was selected according to the best F1-measure, defined as the harmonic mean between the precision and recall. We also tested a Support Vector Machine (SVM) classifier with different kinds of kernel functions (linear, polynomial, radial basis function and sigmoid) using the LIBSVM package for MATLAB (Chang and Lin, 2001) with default parameters.

4.5. Computing novel predictions

To predict additional targets for drugs in our data set, we used a training set that included all the true associations and a twofold larger randomly generated set of drug-gene pairs that are not known to interact. We applied the trained classifier to all remaining drug-gene pairs to form our prediction set. To assign prediction scores also to the random negative set that we used, we repeated the analysis with another randomly picked negative set, distinct from the first one. Overall, we obtained classification scores for all 77,968 unknown drug-gene pairs. We used a classification threshold that yielded the best F1-measure. Out of the resulting prediction set, we picked for each drug its maximal scoring target as the most promising target. Overall, we could predict targets for 307 drugs.

ACKNOWLEDGMENTS

We thank Martin Kupiec and Gidi Stein for critical comments on the manuscript. L.P, A.G., and N.A. are partially funded by the Edmond J. Safra bioinformatics program. R.S. and E.R. were partially supported by a Converging Technologies grant from the Israel Science Foundation. R.S. was additionally supported by the Israel Science Foundation (research grant no. 385/06).

AUTHORS' CONTRIBUTIONS

L.P., A.G., and R.S. conceived and designed the experiments. L.P., A.G., and N.A. performed the experiments and analyzed the data. A.G., N.A., E.R., and R.S. wrote the article.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Accelrys, Inc. 2009. Available at: http://accelrys.com/products/scitegic/. Accessed November 1, 2010.

Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Atias, N., and Sharan, R. 2010. An algorithmic framework for predicting side-effects of drugs. *RECOMB 2010* (to appear).

Bleakley, K., and Yamanishi, Y. 2009. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403.

Breitkreutz, B.J., Stark, C., Reguly, T., et al. 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* 36, D637-D640.

Burns, K., Opas, M., and Michalak, M. 1997. Calreticulin inhibits glucocorticoid- but not cAMP-sensitive expression of tyrosine aminotransferase gene in cultured McA-RH7777 hepatocytes. *Mol. Cell Biochem.* 171, 37–43.

Campillos, M., Kuhn, M., Gavin, A.C., et al. 2008. Drug target identification using side-effect similarity. *Science* 321, 263–266.

Chang, C.-C., and Lin, C.-J. 2001. LIBSVM: a library for support vector machines. Available at: http://www.csie.ntu.edu.tw/∼cjlin/libsvm. Accessed November 1, 2010.

Chen, M., Ona, V.O., Li, M., et al. 2000. Minocycline inhibits caspase-1 and caspase-3 expression and delays mortality in a transgenic mouse model of Huntington disease. *Nat. Med.* 6, 797–801.

Cheng, A.C., Coleman, R.G., Smyth, K.T., et al. 2007. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75.

Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., et al. 2009. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.* 37, D786-D792.

Davis, J., and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. *Proc. ICML '06* 233–240.

di Bernardo, D., Thompson, M.J., Gardner, T.S., et al. 2005. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* 23, 377–383.

Ehnholm, C., Mahley, R.W., Chappell, D.A., et al. 1984. Role of apolipoprotein E in the lipolytic conversion of beta-very low density lipoproteins to low density lipoproteins in type III hyperlipoproteinemia. *Proc. Natl. Acad. Sci. USA* 81, 5566–5570.

Ewing, R.M., Chu, P., Elisma, F., et al. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3, 89.

Fawcett, T. 2005. An introduction to ROC analysis. *Patt. Recogn. Lett.* 27, 861–874.

Fredholm, B.B., and Persson, C.G. 1982. Xanthine derivatives as adenosine receptor antagonists. *Eur. J. Pharmacol.* 81, 673–676.

Gardner, T.S., di Bernardo, D., Lorenz, D., et al. 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301, 102–105.

Gunther, S., Kuhn, M., Dunkel, M., et al. 2008. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* 36, D919-D922.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.

Haggarty, S.J., Koeller, K.M., Wong, J.C., et al. 2003. Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chem. Biol.* 10, 383–396.

Hansen, N.T., Brunak, S., and Altman, R.B. 2009. Generating genome-scale candidate gene lists for pharmacogenomics. *Clin. Pharmacol. Ther.* 86, 183–189.

Hillenmeyer, M.E., Ericson, E., Davis, R.W., et al. 2010. Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biol.* 11, R30.

Iorio, F., Tagliaferri, R., and di Bernardo, D. 2009. Identifying network of drug mode of action by gene expression profiling. *J. Comput. Biol.* 16, 241–251.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. *Bul. Soc. Vaudoise Sci. Nat.* 44, 223–270.

Jain, E., Bairoch, A., Duvaud, S., et al. 2009. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 10, 136.

Kanehisa, M., Goto, S., Furumichi, M., et al. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355-D360.

Keiser, M.J., Roth, B.L., Armbruster, B.N., et al. 2007. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206.

Keiser, M.J., Setola, V., Irwin, J.J., et al. 2009. Predicting new molecular targets for known drugs. *Nature* 462, 175–181.

Kuhn, M., Campillos, M., Gonzalez, P., et al. 2008. Large-scale prediction of drug-target relationships. *FEBS Lett.* 582, 1283–1290.

Kuhn, M., Campillos, M., Letunic, I., et al. 2010. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 6, 343.

Kutalik, Z., Beckmann, J.S., and Bergmann, S. 2008. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.* 26, 531–539.

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

Li, Q., and Lai, L. 2007. Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics* 8, 353.

Liu, Y., Hu, B., Fu, C., et al. 2010. DCDB: drug combination database. *Bioinformatics* 26, 587–588.

Mani, K.M., Lefebvre, C., Wang, K., et al. 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol. Syst. Biol.* 4, 169.

Martin, Y.C., Kofron, J.L., and Traphagen, L.M. 2002. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358.

Matthews, L., Gopinath, G., Gillespie, M., et al. 2009. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* 37, D619–622.

MDL Drug Data Report. 2006. MDL Information Systems Inc., San Leandro, CA.

Mitchell, J.B. 2001. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J. Chem. Infect. Comput. Sci.* 41, 1617–1622.

Mouratidis, P.X., Colston, K.W., and Dalgleish, A.G. 2007. Doxycycline induces caspase-dependent apoptosis in human pancreatic cancer cells. *Int. J. Cancer* 120, 743–752.

Olah, M., Mracec, M., Ostopovici, L., et al. 2005. WOMBAT: World of Molecular Bioactivity, 221–239. *In* Tudor, I.O., ed., *Chemoinformatics in Drug Discovery*. New York.

Ovaska, K., Laakso, M., and Hautaniemi, S. 2008. Fast Gene Ontology based clustering for microarray experiments. *BioData Min.* 1, 11.

Podjarny, E., Benchetrit, S., Katz, B., et al. 2001. Effect of methyldopa on renal function in rats with L-NAME-induced hypertension in pregnancy. *Nephron* 88, 354–359.

Rennemeier, C., Hammerschmidt, S., Niemann, S., et al. 2007. Thrombospondin-1 promotes cellular adherence of gram-positive pathogens via recognition of peptidoglycan. *FASEB J.* 21, 3118–3132.

Resnik, P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 11, 95–130.

Rollo, I.M. 1970. Dihydrofolate reductase inhibitors as antimicrobial agents and their potentiation by sulfonamides. *CRC Crit. Rev. Clin. Lab. Sci.* 1, 565–583.

Rual, J.F., Venkatesan, K., Hao, T., et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.

Schuffenhauer, A., Floersheim, P., Acklin, P., et al. 2003. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* 43, 391–405.

Sinan, S., Kockar, F., and Arslan, O. 2006. Novel purification strategy for human PON1 and inhibition of the activity by cephalosporin and aminoglikozide derived antibiotics. *Biochimie* 88, 565–574.

Skrbo, A., Begovic, B., and Skrbo, S. 2004. [Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes]. *Med. Arh.* 58, 138–141.

Smith, T.F., Waterman, M.S., and Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.* 13, 645–656.

Steinbeck, C., Hoppe, C., Kuhn, S., et al. 2006. Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* 12, 2111–2120.

Stelzl, U., Worm, U., Lalowski, M., et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.

Stone, K.J., and Strominger, J.L. 1971. Mechanism of action of bacitracin: complexation with metal ion and C55-isoprenyl pyrophosphate. *Proc. Natl. Acad. Sci. USA* 68, 3223–3227.

Subramanian, A., Tamayo, P., Mootha, V.K., et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102, 15545–15550.

Tanimoto, T.T. 1957. IBM internal report. 17 November.

Thomson, A.E., Squires, E.J., and Gentry, P.A. 2002. Assessment of factor V, VII and X activities, the key coagulant proteins of the tissue factor pathway in poultry plasma. *Br. Poult. Sci.* 43, 313–321.

Thor and Merlin, version 4.62. Daylight Chemical Information Systems Inc., Irvine, CA. Available at: www.daylight .com. Accessed November 1, 2010.

Wishart, D.S., Knox, C., Guo, A.C., et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901-D906.

Xenarios, I., Salwinski, L., Duan, X.J., et al. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305.

Yamanishi, Y., Araki, M., Gutteridge, A., et al. 2008. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240.

Zhu, F., Han, B., Kumar, P., et al. 2010. Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.* 38, D787–D791.

Zhu, S., Okuno, Y., Tsujimoto, G., et al. 2005. A probabilistic model for mining implicit "chemical compound-gene" relations from literature. *Bioinformatics* 21, Suppl 2, ii245–ii251.

Zhu, S., Stavrovskaya, I.G., Drozda, M., et al. 2002. Minocycline inhibits cytochrome c release and delays progression of amyotrophic lateral sclerosis in mice. *Nature* 417, 74–78.

Address correspondence to:
Dr. Roded Sharan
The Blavatnik School of Computer Science
Tel-Aviv University
Tel-Aviv, Israel

E-mail: roded@post.tau.ac.il