Bioinformatics, 32(15), 2016, 2338–2345
doi: 10.1093/bioinformatics/btw168
Advance Access Publication Date: 1 April 2016
Original Paper

Systems biology

Drug-induced adverse events prediction with the LINCS L1000 data

Zichen Wang, Neil R. Clark and Avi Ma'ayan*

Department of Pharmacology and Systems Therapeutics, One Gustave L. Levy Place Box 1215, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 20, 2015; revised on March 5, 2016; accepted on March 23, 2016

Abstract

Motivation: Adverse drug reactions (ADRs) are a central consideration during drug development. Here we present a machine learning classifier to prioritize ADRs for approved drugs and pre-clinical small-molecule compounds by combining chemical structure (CS) and gene expression (GE) features. The GE data is from the Library of Integrated Network-based Cellular Signatures (LINCS) L1000 dataset that measured changes in GE before and after treatment of human cells with over 20 000 small-molecule compounds including most of the FDA-approved drugs. Using various benchmarking methods, we show that the integration of GE data with the CS of the drugs can significantly improve the predictability of ADRs. Moreover, transforming GE features to enrichment vectors of biological terms further improves the predictive capability of the classifiers. The most predictive biological-term features can assist in understanding the drug mechanisms of action. Finally, we applied the classifier to all >20 000 small-molecules profiled, and developed a web portal for browsing and searching predictive small-molecule/ADR connections.

Availability and Implementation: The interface for the adverse event predictions for the >20 000 LINCS compounds is available at http://maayanlab.net/SEP-L1000/.

Contact: avi.maayan@mssm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Adverse Drug Reactions (ADRs) are harmful or unpleasant unintended side effects resulting from drug intervention (Edwards and Aronson, 2000). ADRs are a major concern for both public health and the drug development process. Failure to identify severe ADRs in clinical trials can lead to significant morbidity, and drugs withdrawn from the market can carry a substantial negative economic impact (Giacomini *et al.*, 2007). Despite such harmful consequences, ADRs can provide useful information about drug/human–phenotype relationships (Kuhn *et al.*, 2010). ADRs and human diseases can often overlap. For example, common ADR phenotypes, such as neuropathy or long-QT syndrome, can manifest as genetically inheritable diseases (Kuhn *et al.*, 2010). Therefore, better understanding drug/human–phenotype connections from a network perspective can lead to improved understanding of human diseases. Various efforts have been made to predict ADRs using the properties/attributes of drugs. Generally, the attributes of drugs used so far to predict ADRs can be categorized into two types: (i) those that mostly consider the chemical aspect of the drug, and (ii) those that consider the biological aspects of the drug. Once such attributes are organized into attribute tables, a binary classification problem can be established for each ADR. Initially, chemical features of drugs alone were used to predict the association between drugs and their known ADRs under the assumption that ADRs may correlate with the chemical fragments of the drugs that induce them (Pauwels *et al.*, 2011; Scheiber *et al.*, 2009). For example, (Scheiber *et al.*, 2009) were able to map the chemical substructure of drugs, described by the extended connectivity fingerprints (ECFPs), to different system organ classes ADRs defined by the Medical Dictionary for Regulatory Activities (MedDRA) (Brown *et al.*, 1999).

Extending this method a step further (Pauwels *et al.*, 2011) used the 881-bit PubChem/CACTVS (Chemical Algorithms Construction, Threading and Verification System) (Chen *et al.*, 2009) chemical fingerprints as drug attributes, and sparse canonical correlation analysis (SCCA) to improve the predictive classification of ADRs. Using the biological feature space approach (Campillos *et al.*, 2008) showed that drugs with similar ADR profile tend to share similar protein targets. Similarly, gene-ADR connections were elucidated through a text-mining effort by Yang *et al.* (2009). Fukuzaki *et al.* (2009) placed targets of drugs in the context of human-curated biological pathways to predict ADRs; whereas Lee *et al.* (2011) used drug-induced gene expression (GE) changes in human cell lines from the Connectivity Map (Lamb *et al.*, 2006) to predict ADRs.

Datasets generated from large-scale NIH-funded multi-institute projects such as Library of Integrated Network-based Cellular Signatures (LINCS) and CTD2 produce rich information about the response of different types of human cells to drug and small-molecule perturbations. For instance, for the LINCS project, members of the Broad Institute utilized the L1000 technology to profile GE changes before and after treating many different types of human cells with a large panel of FDA-approved drugs and other small-molecule compounds. Integration of such data to predict ADRs can improve the quality of predictions as well as lead to new insights about drug mechanisms of action. The primary aim of this study is to develop an unbiased, high-performance predictive model for ADRs that is scalable to thousands of preclinical compounds. To fulfill this aim, we developed a machine learning classifier that primarily utilizes the GE data from the LINCS L1000 project to predict ADRs. The advantage of using this newly available GE dataset to predict ADRs lies in its comprehensiveness, and the void of literature research focus biases that exist for target, pathway, and protein–protein interaction (PPI) knowledge. In this way, we can circumvent the limitation of the target-based approach to ADR prediction that dominates the field. Moreover, our classification model is scalable to predict ADRs for all small-molecule compounds that have GE data profiled by applying the small molecules to treat human cells in a dish.

2 Methods

2.1 Collection of ADRs, biological and chemical attributes of drugs

Briefly, three sources of drug-ADR connections were used in this study. (i) On-label ADRs of FDA-approved drugs were downloaded from the Side Effect Resource (SIDER) (Kuhn *et al.*, 2010), which covers 996 marketed drugs, 4192 ADRs and 99 423 drug-ADR associations. (ii) Off-label ADRs were retrieved from the PharmGKB Offsides table (Tatonetti *et al.*, 2012) that contains processed data from the FDA Adverse Event Report System (FAERS), covering 1322 drugs, 10 097 ADRs and 438 801 drug-ADR associations. (iii) The reference gold standard for four clinically significant ADRs: upper gastrointestinal ulcer, acute liver failure, acute myocardial infarction and acute kidney failure, were downloaded from the Observational Medical Outcomes Partnership (OMOP) website (Ryan *et al.*, 2012, 2013), with biotech drugs removed, resulting in 62, 78, 60 and 40 positive and negative drugs for the four ADRs, respectively. The ADR terms used in SIDER and FAERS are mapped to Preferred Terms (PTs) coded in MedDRA v16.0. The GE profiles of 20 413 small-molecule compound perturbations were downloaded from the lincscloud.org website that provides access to the LINCS L1000 dataset. The quantile-normalized GE profiles of all replicates were used for further analysis. For each small-molecule compound, the LINCS L1000 project profiled GE in cells treated with different dosages while expression was measured at different time points. To reduce the number of features used for classification, only the strongest signatures were used for each drug regardless of the cell type, dosage, or time point. The strongest signatures are predefined by the meta-data provided from the lincscloud.org API with the variable name 'distil_ss'. This variable quantifies the magnitude of differential expression of the landmark genes when comparing the average drug treatment with the DMSO treatment. A signature for a compound is defined as a vector of continuous values, each representing the direction and magnitude of differential expression between control samples and compound-treated samples. The Characteristic Direction (CD) method (Clark *et al.*, 2014) was used to compute GE signatures for drug perturbations using only the 978 directly measured landmark genes, as well as all genes both measured and imputed. The CD signatures in the space of 978 landmark genes are used as GE signatures for drugs to predict ADRs. A geometric extension of CD called Principal Angle Enrichment Analysis (PAEA) (Clark *et al.*, 2015) was used to compute enrichment p-values for each CD signature in the space of all genes against gene sets created from the Gene Ontology (GO) including Biological Processes, Cellular Components and Molecular Functions and other gene set libraries available from the Enrichr tool (Chen *et al.*, 2013). The cell morphological (MC) profiles were downloaded from the MLPCN project website (Wawer *et al.*, 2014). The MLPCN dataset contains processed images with sample aggregation. Each small-molecule compound has 812 MC feature descriptors with numerical values extracted from the cell images, representing the MC changes of the cell upon treatment with a drug.

The 2D chemical structures (CSs) of small-molecule compounds are represented in simplified molecular-input line-entry system (SMILES) format. The SMILES strings for marketed drugs were collected from PubChem using PubChem Compound IDs provided in SIDER. The SMILES strings for all other small-molecule compounds were provided in the metadata of LINCS L1000 and MLPCN projects. SMILES strings were converted into a binary feature set by the 166-bit Molecular ACCess System chemical fingerprints using the Open Babel cheminformatic toolbox (O'Boyle *et al.*, 2011). To map marketed drugs to small-molecule compounds profiled in the LINCS L1000 and MLPCN projects, salts were first stripped from SMILES strings of drugs. These SMILES strings were then converted to canonical SMILES strings with Open Babel, which were further used to map to the 'pert_id' of small-molecule compounds profiled in the LINCS L1000 and MLPCN projects. Drugs that were unable to be mapped to their pert_id were then mapped using names and synonyms retrieved from DrugBank (Law *et al.*, 2014). Collectively, 826 small-molecule compounds were mapped to marketed drugs in SIDER. All of these drug attributes from the various sources were combined to predict ADRs.

2.2 Prediction of ADRs using multi-label classification

The task of predicting multiple ADRs for each drug was implemented as a multi-label classifier. Formally, the goal of a multi-label classifier is to find a model that maps inputs x to a binary vector y; where x represents the attribute vector of a drug and the label vector y is a vector of Boolean values describing whether the drug causes the corresponding ADRs. The Binary Relevance (BR) method (Tsoumakas and Katakis, 2006) was employed to transform the multi-label classification problem to a set of independent

single label classification tasks coupled with various binary classifiers.

Drug-attribute-sets, including GE, CS, cell MC profiles and GO or other enrichment vectors were combined for feature selection. Because the BR approach was chosen among others (see Supplementary Methods) for multi-label classification, feature selection was also performed independently for each ADR. Briefly, stability selection (Meinshausen and Bühlmann, 2010) was used for each ADR to prioritize the top 50 most predictive features. 75% of the drugs were randomly sampled, and an L1 regularized logistic regression (Logit) model was learned from the subset of the data with a regularization parameter equal to 1; class weights are set to the inverse of class frequency in the loss function to handle class imbalance. The coefficients of the decision function for each drug feature from the Logit model were recorded. This procedure was repeated 200 times. The coefficients were then averaged and used as the measurement of feature importance.

Extra Trees (ETs) (Geurts *et al.*, 2006) classifiers were trained for each ADR and evaluated using 3-fold cross-validation. The number of trees was set to 100 and class weights were also set to the inverse of class frequency in the loss function to handle class imbalance. All other parameters of ET were set to default as recommended by the Scikit-Learn Python package (Pedregosa *et al.*, 2011). This classification method was selected after evaluating various alternatives including Random Forest (RF), Support Vector Machines (SVMs) and a few others (see Supplementary Methods). To evaluate the multi-label classification model, both instance-based and label-based metrics have been applied. Micro-averaged Receiver Operating Characteristic (ROC) curves and Hamming loss were employed for instance-based evaluations to evaluate the performance of the multi-label classifiers in assigning the correct ADRs to individual drugs. To plot the micro-averaged ROC curves, the multi-label classification problem was considered as a single binary classification problem asking whether a certain drug causes a certain ADR. Then, False Positive Rate and True Positive Rate were calculated under different discrimination thresholds. Hamming loss was defined as the fraction of incorrectly predicted ADRs for drugs compared with the total number of ADRs. To assess the performance of the multi-label classifiers in predicting individual ADRs, area under the ROC curves (AUROC) and area under the Matthews Correlation Coefficient (MCC)-recall curves for each ADR were calculated. MCC-recall curves were generated by plotting the MCC, and then the area under the MCC-recall curve was calculated as another measure of classification performance.

2.3 Construction of the ADR-GO term network

To associate well-predictable ADRs with GO terms, we selected the top 25 percentile of the most predictable ADRs based on their AUROC, and made connections with their most predictive GO terms with feature importance score of >0.35. The resulting network contains 112 ADRs, 323 GO terms and 420 associations.

2.4 Development of the web portal for predicted ADRs

The ET classifier was applied to all of the 20 413 small-molecule compounds that are profiled within the LINCS L1000 project. We constructed an ADR–ADR similarity network at the PT level using the Sets2Network algorithm(Clark *et al.*, 2012) and visualized this network using the packed bubble plot implemented in the JavaScript library D3 (Bostock *et al.*, 2011). The packed bubble plot visualization, together with the predicted ADRs for all the compounds in LINCS L1000, was deployed as a web portal available at http://maayanlab.net/SEP-L1000/. The website was designed with the Bootstrap HTML5 framework as the frontend and PHP 5.4 and MySQL database at the server-side backend.

3 Results

3.1 The strongest L1000 GE signatures of drugs are most predictive of ADRs

The LINCS L1000 dataset contains drug/small-molecule compounds that induce changes in GE once applied to human cell lines at different dosages while expression levels are measured at different time points. Since not all possible combinations of drug/cell/dose/time-point possibilities are covered, for practical considerations, we only selected one representative GE signature for each drug/small-molecule. We show that, in general, the strongest signatures are most predictive of ADRs (empirically computed P-value ≤ 0.0001) (Supplementary Figure S1).

3.2 Benchmarks of machine learning pipelines to predict ADRs

We formulated the problem of using drug attributes to predict multiple ADRs as a multi-label classification problem. We first compared the predictive performance of various multi-label classification models. We found that BR outperformed classifier chain-based methods (Read *et al.*, 2011) as well as hierarchical multi-label classifiers (Tsoumakas *et al.*, 2008) (Supplementary Figure S2); the former attempt to handle the correlation between ADRs and the latter incorporate the prior hierarchy in ADRs.

The imbalanced class size of many ADRs posed a challenge because many ADRs are disproportionally rare with many 0s and only a few 1s in the response vector. To balance the class sizes, we adjusted the weights for the two classes (0 and 1) to the inverse of class frequencies in the loss functions of both feature selection and the classifier to improve the predictability of rare ADRs (Supplementary Figure S3). However, we are aware that rare events are essentially more difficult to predict because of the scarcity of positive examples, as suggested in a previous study (Liu and Altman, 2015).

Feature selection was performed to prioritize predictors among drug attributes for each ADR. We observed that the features selected are stable across random subsets of drugs (Supplementary Figure S4). We also observed that the top $\sim$50 features yield optimal predictive performance (Supplementary Figure S5).

Because the features of drugs include both continuous and categorical variables, we reasoned that non-parametric classifiers should performance better. We observed the best performance for ETs classifiers over several other classification models including: RF, SVM, Logit and few others (Supplementary Figure S6). Further analyses are carried out with ET classifiers. Since Bagging ensemble models such as ET and RF generate out-of-bag estimates during the training process, we also compared the out-of-bag estimates with the cross-validation estimates and found that the two performance measures strongly agree with each other (Supplementary Figure S7). Additionally, the number of folds in cross-validation does not have significant influence on the evaluation (Supplementary Figure S8).

3.3 Integration of GE and CS can improve predictions of ADRs

To build a classifier to predict ADRs, we collected drug attributes representing the following properties of drugs: changes in GE upon drug treatment of human cells from the LINCS L1000 dataset; MC

profiles upon drug treatment of human cells from the MLPCN project; the CS molecular fingerprint of the same drugs; and known side effects of drugs from two independent sources: SIDER and FAERS (Fig. 1A). We then organized the various resources to construct a classifier for ADRs (Fig. 1B). Table 1 provides a summary of data sets used for specific analyses.

In our first attempt to predict ADRs, we created a classifier using the 251 drugs that are shared among all the three attribute table sources. Three-fold cross-validation was used to construct a multi-label ET classifier. To avoid the absence of positive instances during training, we eliminated rare ADRs, which we defined as ADRs with <10 drug associations. ADRs associated with >90% of the drugs in the training set were also removed. By applying several instance-based or label-based benchmarking metrics, we observed that the GE profiling data was the best predictor of drug-ADR associations. Chemical structures and MC profiles display similar instance-based AUROC, while the Hamming loss is greater in CS than MC (Supplementary Figure S9A and B), suggesting that CS and MC have comparable ability in assigning ADRs to drugs. In comparison, the MC profiles were much less predictive of ADRs compared with GE and CS in terms of label-based evaluation metrics (Supplementary Figures S9C and D) regardless of classification algorithms used (Supplementary Figure S10), suggesting that MC is not very helpful in predicting individual ADRs. Importantly, combining GE attributes with CS attributes significantly improves the performance of the classifier (Supplementary Figure S9C and D, Wilcoxon rank sum P-value = 0.0006 for differences in AUROCs, and P-value = 0.025 for area under the MCC-recall curve).

However, the integration of MC profiles with other types of attributes does not significantly contribute to improved predictions, indicating that the MLPCN dataset does not capture much useful knowledge about adverse reactions in human cells. Because the morphology profiling data were found to be not highly predictive of ADRs, we excluded those data from further analyses. This enabled the inclusion of additional drugs into the classifier. Although the additional drugs do not significantly improve the predictive performance on the originally predicted ADRs (Supplementary Figure S11), the resultant classifier without the MC data is more scalable because it only requires the drug CS and the GE data for each drug. The classifiers that contain the CS features combined with GE features achieve higher predictive performance in predicting the ADRs listed on the drug package inserts (SIDER), as well as off-label ADRs mined from the FAERS, compared with using each type of data alone (Fig. 2 and Supplementary Figure S12). It is worth noting that the majority of the L1000 GE data used to construct these classifiers is collected from cancer or immortal human cell lines treated with 10 μM of the drugs, and where GE changes were measured after 6 h (Supplementary Figure S13). Our positive results are encouraging; they suggest that the L1000 GE signatures profiled in human cancer cell lines by a unified protocol can be predictive of the phenotypic effects in humans for most of the drugs.

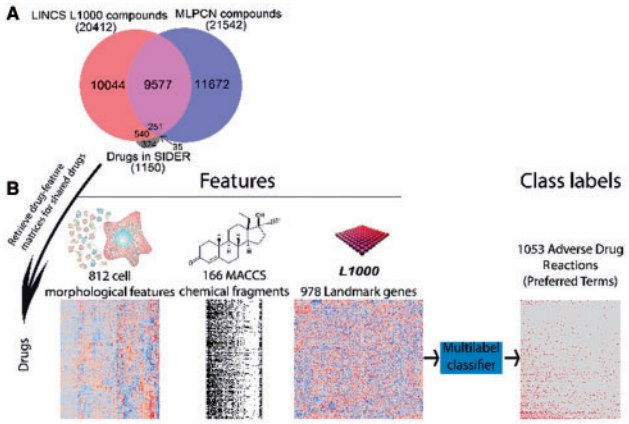

Fig. 1. Overview of the workflow of this study. **(A)** Venn diagram shows the overlap between drugs and small-molecule compounds across the different data sources. **(B)** Representation of the organization of drug-feature matrices of different feature types to predict ADRs using multi-label classification (Color version of this figure is available at *Bioinformatics* online.)

3.4 GO enrichment vectors improve prediction of ADRs

To further improve the predictive performance of the ADRs ET classifier, we tested whether inclusion of prior biological knowledge to transform the GE signatures to enrichment vectors can improve performance. Enrichment vectors are created by computing gene set enrichment for a set of differentially expressed genes, and then using the enrichment scores of the terms as attributes instead of the GE values (Duan *et al.*, 2013). This approach can potentially filter the signal from the GE data and potentially improve the quality of the classifiers. To achieve this, we transformed the GE signatures to enrichment vectors of terms using different types of gene-set libraries including: gene sets created from up-regulated genes in cancer cell lines (Barretina *et al.*, 2012), or up-regulated genes in a variety of human diseases; gene sets created from PPIs for hub proteins; transcription factor targets from ChIP-seq studies (Lachmann *et al.*, 2010); gene sets associated with GO terms; genes the protein products of which belong to curated pathways; and gene sets in which genes in each set share phenotypes from two sources: uberPheno (Köhler *et al.*, 2013) and the Mouse Genome Informatics (MGI) Mammalian Phenotype Ontology (Supplementary FigureS14). We also tested other gene set enrichment methods to transform GE data to GO enrichment vectors including ssGSEA (Barbie *et al.*, 2009) and gene set variation analysis (Hänzelmann *et al.*, 2013); the latter has comparable performance with PAEA (Supplementary Figure S15).

We then attempted to predict ADRs using these enrichment vectors instead of the GE data. With this approach, we found that few of the gene-set library vectors, transformed GE signatures, are more predictive of ADRs compared with GE alone. The most predictive transformations were for the cancer cell line signatures, GO and uberPheno (Supplementary FigureS14). GO-transformed GE signatures were the most predictive among all of these (Fig. 2A and B,

Table 1. Summary of data sets for different analyses in this study

Figures	Intersection of datasets	ADRs	Drugs
Figures 2, 4 and all other Supplementary Figures	LINCS L1000 and SIDER	1053	791
Figure 3	LINCS L1000 and OMOP	4	122
Supplementary Figures S9 and S10	LINCS L1000 and MLPCN morphology and SIDER	539	251
Supplementary Figure S12	LINCS L1000 and FAERS Offsides	1275	693

Supplementary Figures S12 and S14). Intriguingly, the overall predictability for individual ADRs does improve when converting GE signatures to enrichment vectors. However, for individual ADRs, there is either significant improvement or the predictability worsen (Fig. 2C and D). The average AUROC of GO-transformed GE is significantly better than for the GO transformed GE (rank sum P-value $= 2.1e^{-20}$). Additionally, the average area under the MCC-recall curve for the GE alone is also significantly lower than the GO enrichment vectors (P-value $= 1.1e^{-11}$). Scatter plotting the individual ADRs with GE alone versus GO enrichment vectors shows that those ADRs that are more predictive with GE become even more predictable with GO enrichment vectors (Supplementary Figure S16 and Supplementary Table S8). This result implies that the biological knowledge in GO may be able to explain some of the molecular mechanisms for a subset of ADRs.

The fact that other ADRs can be better predicted by the GE data alone (Supplementary Table S1) may suggest that our current knowledge about the genes affected by these drugs in cancer cells is more limited, or less predictive of the human phenotype. Following this hypothesis, we explored those ADRs that are most predictable by the enriched GO terms (Supplementary Table S2). We found that the ADRs that are more predictable by GO terms' enrichment vectors are enriched for neoplasms and for menstrual cycle and uterine bleeding disorders, whereas ADRs that are more predictable by GE alone are enriched for adrenal gland disorders and metabolic disorders. Overall, we see that GO term enrichment vectors may further explain the molecular mechanisms of a distinctive subset of ADRs. Interestingly, we found that the ADRs predictable by CS are enriched for chemical injuries and poisoning (Supplementary Table S3), suggesting that certain chemical substructures in drugs are toxic. To further improve our performance in predicting ADRs, we first attempted to combine GE and GO term enrichment vectors (Supplementary Figure S17), which did not yield significant improvement beyond the GO term enrichment vector. We then combined enrichment vectors with CS. We were able to significantly improve the performance when comparing these predictions using GO enrichment vectors alone (rank sum test, P-value $< 2e^{-7}$). Although we found that endocrine disorders, metabolic disorders, and white blood cell disorders are highly predictable; side effects such as spinal cord disruption, nerve root disorders and musculoskeletal and connective tissue deformities are unpredictable (Supplementary Tables S4 and S5). We also highlight top predictive and least predictive ADRs by the combination of GOt + CS (Supplementary Tables S6 and S7). This suggests that the underlying mechanisms of those ADRs may involve processes that are more systemic and go beyond transcriptional changes at the cellular level. To

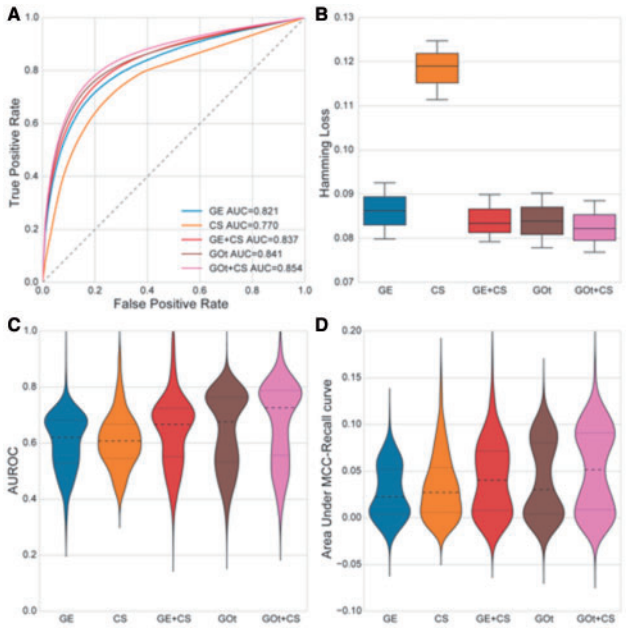

Fig. 2. Conversion of GE features to enrichment vectors to predict ADRs. **(A)** micro-averaged instance-based ROC curves showing the performance of multi-label classifiers using different drug attributes including GE alone, CS alone, GE + CS, and GO enrichment vectors alone; as well as combinations of GO + CS for predicting ADRs. AUROC are indicated in the legend. **(B)** Box plots showing the distribution of the hamming loss across each fold of cross-validation of the classifiers using different drug attributes. **(C,D)** Violin plots showing the distributions of label-based evaluation statistics AUROC and area under the MCC-recall curve for classifiers using different drug attributes (Color version of this figure is available at *Bioinformatics* online.)

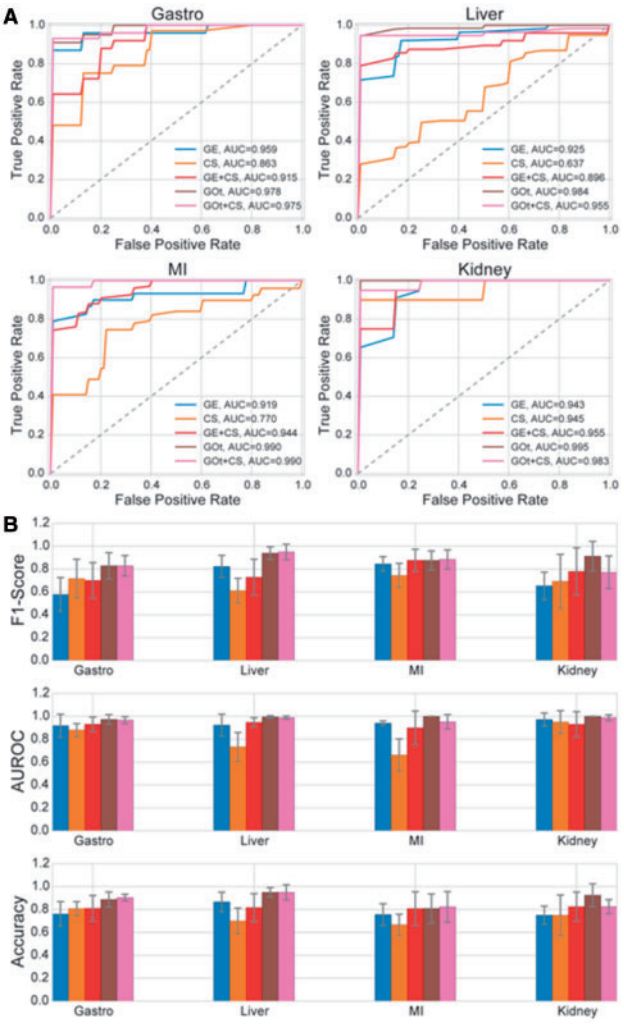

Fig. 3. Performance of ADR predictions measured using the OMOP gold standard. **(A)** ROC curves showing the performance of classification models using different drug attributes in predicting the four individual ADRs (upper gastrointestinal ulcer, acute liver failure, acute myocardial infarction and acute kidney failure). The AUROC values are indicated in the legend. **(B)** Bar plots showing other evaluation metrics including F1-score, AUROC and accuracy of the classification models applied to the four ADRs. Error bar indicates the SD of the scores over cross-validation. Drug attributes: GE, CS, GE + CS, GOt and GOt + CS are plotted in blue, orange, red, brown and pink in both (A) and (B) (Color version of this figure is available at *Bioinformatics* online.)

further benchmark the classifiers, we trained and tested classifiers on the OMOP gold standard dataset to predict the four clinically significant ADRs offered by this testbed (Fig. 3).

Evaluating the predictions using cross-validation, we were able to achieve higher AUROC, F1-score, and accuracy when combining GO enrichment vectors and CS features for all of these four ADRs compared with the performance of other classifiers developed and published in previous studies (Lorberbaum *et al.*, 2015; Ryan *et al.*, 2013).

3.5 Interpretation of the most predictive classifiers to shed lights on the mechanisms of ADRs

In order to interpret the improved predictive quality of the GO term enrichment vectors classifier, we filtered the highly predictive ADRs (AUROCs > 75 percentile) and connected them with their most predictive GO terms. In this way, we were able to construct a network of ADRs connected to their related GO terms. This network consists of 112 ADRs, and 323 GO terms (Fig. 4). In order to validate the associations identified within this network, we counted the co-occurrence of each pair of ADRs and GO terms in PubMed articles and measured the significance of the number of articles that mentioned both terms using Fisher's exact test. We found that ADR-GO connections in the network are significantly more likely to co-occur (rank sum P-value < $1e^{-5}$) (Supplementary Figure S18). We also highlighted some ADR-GO connections that have strong literature support (Fig. 4). With this network, we are able to recover common known causes for ADRs. For example, Dysuria, which is painful urination, is most often caused by infection or inflammation in the urinary tract. We find that this ADR is associated with the GO term 'regulation of inflammatory response to antigenic stimulus' (GO:0002863), pointing to genes that become dis-regulated by the drugs that cause this side effect. We also see that genes involved in the GO term: 'heme transport' are connected to Anaemia (Yuan *et al.*, 2013). ADRs characterized by a decrease in the number of white blood cells such as Leukopenia and Agranulocytosis are associated with cytokine-related GO terms, which are known to be important for regulating the immune system. We also found connections between ventricular fibrillation and cardiac fibrillation (Olshansky *et al.*, 2008; Vaseghi and Shivkumar, 2008) with neuronal projection guidance-related GO terms. The genes that establish this connection can be linked to the autonomic nerve innervation process that is the known mechanistic cause of this side effect. Recovery of such known molecular mechanistic relationship is neither trivial nor necessarily expected because those are enriched terms of differentially expressed genes in cancer cell lines rather than ADR associations found in genes in the relevant cells/tissues.

3.6 Known ADR-phenotype connections can be recovered by the predictive model

To further examine molecular mechanisms of ADRs with an approach that is less biased by a literature research-focus, we transformed the GE signatures of the drugs to enrichment vectors of cross-species phenotypes covering human, mouse and zebrafish. The enrichment vectors of phenotypes are also very predictive of ADRs, similarly to the GO terms' enrichment vectors, but these come from a less biased data source. We asked if the most predictive phenotypes are directly related to the ADRs they predict. We connected the top 235 most predictive ADRs with the top 228 most informative phenotypes. We then mapped the mouse phenotype terms to cross-species phenotypes (Köhler *et al.*, 2013) to expand the number of associated phenotypes with ADRs to 423 phenotypes. Using ElasticNet with cross-validation, we filtered the initial network to only contain 52 ADRs

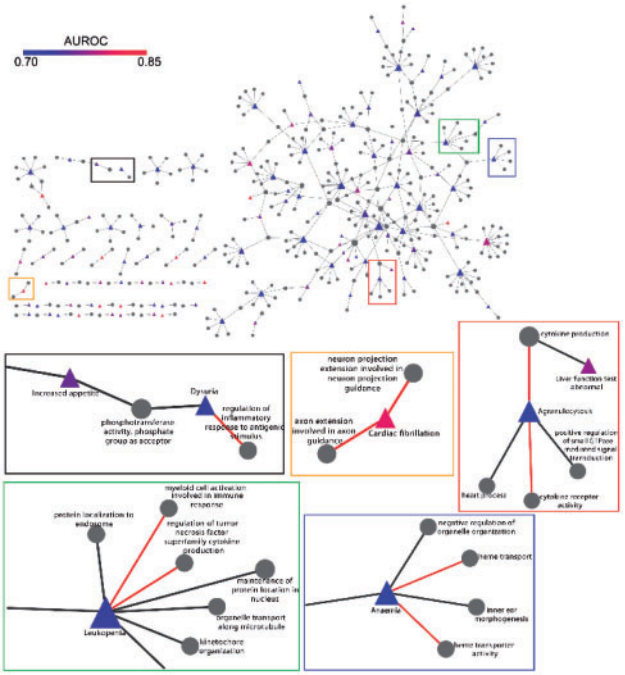

Fig. 4. ADR-GO network connecting the most predictive ADRs with their most predictive GO-related features. The network representation displays ADR and GO terms as triangles or circles, respectively. The size of nodes is proportional to the node's degree. The color of ADRs corresponds to their predictability measured by their AUROC. Recovered ADR-GO term associations supported by literature are highlighted in red (Color version of this figure is available at *Bioinformatics* online.)

and 102 phenotypes (Supplementary Figure S19). As expected, the network recovered known connections between the pulmonary oedema ADR and the human phenotype 'pulmonary odema' (HP:0100598), Arthritis ADR and the mouse phenotype 'increased susceptibility to induced arthritis' (MP:000372), as well as the ADR Asthma and human phenotype 'recurrent pneumonia' (HP:0006532). These connections suggest that the differentially expressed genes in cancer cell lines turn on or off genes associated with these phenotypes and, as such, can be used to predict the human phenotype that they may induce when administrated systemically.

3.7 Development of a web portal to share predicted ADRs for marketed drugs and experimental compounds

The best ET classifier that uses GOt + CS was applied to predict ADRs for all of the 20 412 drugs and small-molecule compounds profiled by the LINCS L1000 project. To share these predictions, we developed a web portal, which is freely available at http://maayanlab.net/SEP-L1000/. A network of predictive ADRs was constructed based on their drug similarity and visualized using a stacked bubble chart. Each drug and ADR has a dedicated page with a list of the relevant predictions and external links to sites such as: PubChem, lincscloud.org, LINCS Information Framework (LIFE) and SIDER. A screenshot of the web portal is shown in Figure 5.

4 Discussion

Drug safety is central to public health and drug development. In this study, we presented a scalable computational approach that improves the prediction of ADRs using the newly available LINCS L1000 dataset. The classifier we developed only requires the molecular structure

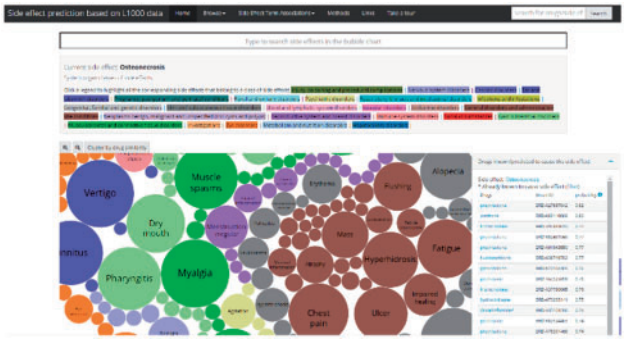

Fig. 5. Screenshot from the SEP-L1000 website. This website serves the results of the predicted ADRs for all of the 20 413 drugs and small-molecule compounds profiled in the LINCS L1000 project (Color version of this figure is available at *Bioinformatics* online.)

and GE profiles for small molecules, so it can be applied to a large panel of investigational drug-like compounds to prioritize their potential ADRs. This alleviates the requirement of knowing the targets of the potential drug which are difficult to identify, and when they are known, such knowledge is often partial. Therefore, we reason that target-based predictive models are heavily dependent on the size, completeness and quality of drug-target datasets. However, target-based information may help to alleviate the incapability of GE-based models to predict certain ADRs. Despite the scalability of the classifiers we developed, we are aware that some ADRs cannot be predicted by our model. This can be due to several reasons: the ADRs can be a result of a systemic malfunction that cannot manifest at the cellular GE level; the receptors for the drugs are not present in the profiled cell lines, the ADRs are highly dose-dependent, and the drug may be processed and metabolized into various forms before inducing its global ADR effect. Although in our approach we do not predict which individuals may be more susceptible to manifest a side effect for a specific drug, our enrichment vector approach takes a step in this direction. The genes that are differentially expressed by the drugs that have known association with the side effect could contain single nucleotide variants that may predispose individuals to manifest the ADRs, or expression signatures from individual patients in their response to drugs can be mapped to predict ADRs in individuals. Such hypotheses remain to be evaluated once we have sufficient data about associations between individuals, their sequenced DNA and RNA, and their propensity to manifest ADRs for specific drugs. It should be noted that although we used SIDER and FAERS as gold standards to predict ADRs, such resources are not set in stone and should not be considered as 100% reliable. The classifiers that we developed can potentially systematically fill in the gaps between established drug-ADR associations and new ones predicted from the data. Overall, we have developed a potentially useful resource, and we present ideas on how to link cellular molecular expression signatures with the human phenotype. Although we have focused on the prediction of ADRs, with some simple modifications, a similar approach can be used to predict novel indications for drug repurposing as well as to identify new drugs and small molecules for the treatment of many diseases.

Funding

This work was partially supported by National Institutes of Health grants U54HL127624, U54CA189201, and R01GM098316 to AM.

Conflict of Interest: none declared.

References

Barbie,D.A. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–307.

Bostock,M. *et al.* (2011) D^3: Data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, **17**, 2301–2309.

Brown,E. *et al.* (1999) The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, **20**, 109–117.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Chen,B. *et al.* (2009) PubChem as a Source of Polypharmacology. *J. Chem. Inform. Model.*, **49**, 2044–2055.

Chen,E. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.

Clark,N. *et al.* (2012) Sets2Networks: network inference from repeated observations of sets. *BMC Syst. Biol.*, **6**, 89.

Clark,N. *et al.* (2014) The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics*, **15**, 79.

Clark,N.R. *et al.* (2015) Principal Angle Enrichment Analysis (PAEA): Dimensionally Reduced Multivariate Gene Set Enrichment Analysis Tool. *Proc. IEEE Int. Conf Bioinformatics Biomed.*, **2015**, 256–262.

Duan,Q. *et al.* (2013) Metasignatures identify two major subtypes of breast cancer. *CPT Pharmacometrics Syst. Pharmacol.*, **2**, 1–10.

Edwards,I.R. and Aronson,J.K. (2000) Adverse drug reactions: definitions, diagnosis, and management. *Lancet*, **356**, 1255–1259.

Fukuzaki,M. *et al.* (2009) Side effect prediction using cooperative pathways. In: *Bioinformatics and Biomedicine, 2009. BIBM'09. IEEE.* pp. 142–147.

Geurts,P. *et al.* (2006) Extremely randomized trees. *Mach. Learn.*, **63**, 3–42.

Giacomini,K.M. *et al.* (2007) When good drugs go bad. *Nature*, **446**, 975–977.

Hänzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, **14**, 1–15.

Köhler,S. *et al.* (2013) Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000 Res.*, **2**, 30.

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

Lachmann,A. *et al.* (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Law,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

Lee,S. *et al.* (2011) Building the process-drug-side effect network to discover the relationship between biological Processes and side effects. *BMC Bioinformatics*, **12**(Suppl 2), S2.

Liu,T. and Altman,R.B. (2015) Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *J. Chem. Inform. Model.*, **55**, 1483–1494.

Lorberbaum,T. *et al.* (2015) Systems pharmacology augments drug safety surveillance. *Clin. Pharmacol. Ther.*, **97**, 151–158.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 417–473.

O'Boyle,N. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminformatics*, **3**, 33.

Olshansky,B. *et al.* (2008) Parasympathetic nervous system and heart failure: pathophysiology and potential implications for therapy. *Circulation*, **118**, 863–871.

Pauwels,E. *et al.* (2011) Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, **12**, 169.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Read,J. *et al*. (2011) Classifier chains for multi-label classification. *Mach. Learn*., **85**, 333–359.

Ryan,P.B. *et al*. (2012) Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat. Med*., **31**, 4401–4415.

Ryan,P.B. *et al*. (2013) Medication-wide association studies. *CPT Pharmacometrics Syst. Pharmacol*., **2**, 1–12.

Scheiber,J. *et al*. (2009) Mapping adverse drug reactions in chemical space. *J. Med. Chem*., **52**, 3103–3107.

Tatonetti,N.P. *et al*. (2012) Data-driven prediction of drug effects and interactions. *Sci. Trans. Med*., **4**, 125ra131.

Tsoumakas,G. and Katakis,I. (2006) Multi-label classification: an overview. *Department of Informatics*, Aristotle University of Thessaloniki, Greece.

Tsoumakas,G. *et al*. (2008) Effective and Efficient Multilabel Classification in Domains with Large Number of Labels. In: *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*. Antwerp, Belgium.

Vaseghi,M., and Shivkumar,K. (2008) The role of the autonomic nervous system in sudden cardiac death. *Prog. Cardiovasc. Dis*., **50**, 404–419.

Wawer,M.J. *et al*. (2014) Automated structure–activity relationship mining: connecting chemical structure to biological profiles. *J. Biomol. Screen*., **19**, 738–748.

Yang,L. *et al*. (2009) A CitationRank algorithm inheriting Google technology designed to highlight genes responsible for serious adverse drug reaction. *Bioinformatics*, **25**, 2244–2250.

Yuan,X. *et al*. (2013) Heme transport and erythropoiesis. *Curr. Opin. Chem. Biol*., **17**, 204–211.