Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2013 ; : 440–445. doi:10.1109/BIBM.

2013-6732532

An Ensemble Approach for Drug Side Effect Prediction

Md Jamiul Jahid and

Department of Computer Science, University of Texas at San Antonio, San Antonio, Texas
78249, mjahid@cs.utsa.edu

Jianhua Ruan

Department of Computer Science, University of Texas at San Antonio, San Antonio, Texas
78249, jianhua.ruan@utsa.edu

Abstract

In silico prediction of drug side-effects in early stage of drug development is becoming more popular now days, which not only reduces the time for drug design but also reduces the drug development costs. In this article we propose an ensemble approach to predict drug side-effects of drug molecules based on their chemical structure. Our idea originates from the observation that similar drugs have similar side-effects. Based on this observation we design an ensemble approach that combine the results from different classification models where each model is generated by a different set of similar drugs. We applied our approach to 1385 side-effects in the SIDER database for 888 drugs. Results show that our approach outperformed previously published approaches and standard classifiers. Furthermore, we applied our method to a number of uncharacterized drug molecules in DrugBank database and predict their side-effect profiles for future usage. Results from various sources confirm that our method is able to predict the side-effects for uncharacterized drugs and more importantly able to predict rare side-effects which are often ignored by other approaches. The method described in this article can be useful to predict side-effects in drug design in an early stage to reduce experimental cost and time.

Keywords

adverse side-effect; drug development; chemical substructure; uncharacterized drug

I. INTRODUCTION

Drug can be considered as molecules that interact with an appropriate target protein to perturb biological systems of different molecular interactions such as metabolic pathway, protein interaction network and signal transduction network to prevent or cure a disease of interest [1], [2]. Conventional drug design focuses on one drug one target paradigm but this may overlook overall effects in human body [3]. Besides primary targets, drug molecules may also interact with other off target proteins which may result in potential side-effects [2]. It is known that about 35% drug molecules have two or more targets [4]. With few exceptions, unexpected drug activities are harmful and may cause severe side-effects, which is the fourth leading cause of death in United States [5]. Note that, more than 2 million people each year are affected by drug adverse effects in United States which cause about 100,000 deaths [1].

Besides major public health concerns, drug development is costly and time consuming. For example, current gold standard (rodent bioassay) for identifying the carcinogenicity of novel chemical compounds takes almost two years [1]. Furthermore, *in vitro* safety profiling is used to predict drug adverse effects by testing compounds with cellular and biochemical assays but target proteins (mainly off-targets) of many drugs have not been characterized yet, which causes the side-effect prediction extremely challenging [6], [2]. For example, around 30% drugs fail clinical trials mainly because of adverse side-effect and drug efficacy [1]. Therefore, *in silico* based prediction tools for drug side-effects in the early drug development phase are of urgent need and may reduce drug development time, cost and fatality before drug reaches clinical trials. These tools should include data such as protein-protein interactions, pathways of drugs, drug chemical profiles, signaling pathways, drug action pathways and metabolism information [1].

Recently, a number of side-effect prediction tools have been proposed, which mainly fall into two broad categories based on the data they use - pathway based and chemical profile based approaches. Pathway based approaches mainly focus on the idea that drugs with similar side-effects have similar profiles of target proteins, which eventually relate side-effects to biological pathways or sub-pathways that are targeted by different drugs [7], [8]. Fukuzuki et al identify side-effects using cooperative pathways by combining gene expression profiles and pathways that function together under same condition [8]. Another recent work for side-effect prediction and understanding the underlying side-effect mechanism by Xie et al, generates off-target binding network by comparing the structure of ligand-binding sites in known proteins to identify possible off targets for torcetrapib [9]. However, the applicability of this approach is limited by the availability of 3D protein structures and known biological pathways [9], [6]. Chemical profile based approaches uses chemical structures to predict drug side-effects. The method developed by Scheiber and colleagues attempts to predict side-effects by extracting highly correlated chemical features related to particular adverse drug reactions [10]. Another recent work by Pauwels et al extracts correlated sets of chemical structures to predict drug side-effects and can be applicable to large scale data [6]. However, this approach is not effective for rare side-effects because applying this approach to a set of 2883 uncharacterized drug molecules results in predicting very few side-effect terms, mostly only high frequency side-effect terms.

In this article we develop an ensemble classification approach to predict drug side-effects of drug molecules based on their chemical structures. Our method is applicable to large scale datasets. A unique feature of our approach is using chemical structure to identify a small subset of similar drugs to predict drug side-effects which makes it less sensitive to the side-effect frequency count and able to predict rare side-effects which is very important in drug development process. For 888 drugs and 1385 side-effect terms, our method outperformed previous approaches. More importantly, for 2883 uncharacterized drug molecules in DrugBank database our approach can predict a large variety of side-effect terms especially rare side-effects; many of them are ignored by previous methods. Besides these, from independent data source we confirm that our predictions for these uncharacterized drug molecules are indeed meaningful and can have practical use.

II. MATERIALS AND METHODS

A. Data

The drugs and their side-effects data used in this article were obtained from SIDER database which have recorded side-effect information of marketed drugs [11]. From the database we obtained 1385 side-effect terms for 888 marketed drugs. Therefore, we have a matrix (S) of size 888×1385 where

$$S_{ij} = \begin{cases} 1 & \text{if drug } i \text{ has side-effect } j \text{ recorded in SIDER,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Figure 1a shows the number of side-effect terms for each drug. From the figure it can be seen that the majority of the drugs have fewer than 100 side-effect terms recorded in the database. The average number of side-effect for each drug is 68.81. Figure 1b shows the number of drugs associated with each side-effect term. From the figure it can be seen that only ~300 side-effects are present in at least 50 drugs, and the other side-effects are relatively rare, which makes the prediction more challenging. The median number of drugs recorded for each side-effect is 8. There are a total of 61,102 connections between drugs and side-effect terms.

The chemical properties of drugs used in this work were obtained from PubChem database, where 881 chemical substructures were defined for each drug [12]. Therefore we have a matrix (C) of size 888×881 where

$$C_{il} = \begin{cases} 1 & \text{if drug } i \text{ has substructure } l \text{ present,} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

There are in total 107,292 substructures present for all drugs and each drug has 120.8 substructures on average.

Additionally, we predicted the side-effect profiles for the drug molecules present in DrugBank database [13] whose side-effect information are not available in SIDER database [11]. In this step we concentrated on 2883 small molecule drugs. Chemical properties of these 2883 drugs were also obtained from PubChem database which is incorporated in matrix T of size 2883×881 .

B. Ensemble classification approach

Our ensemble classification approach has two main steps. In the first step to predict side-effect profiles we construct a number of classifiers for each drug and each side-effect term. From our observation (see section III(A)) that similar drugs have similar side-effects, we built a set of classifier for each drug with its most similar k drugs. To identify similar drugs we calculate Euclidean distance to all other drugs on their chemical substructure (C) property and then to predict the probability of this drug having this side-effect we construct a number of classifiers with its most similar (e.g. $k = 50, 100, 150, 200, 250, 300, 350$ and

400) drugs. The rationale behind our ensemble classification approach is that drugs can vary a lot by their structures and most of the time some substructures are responsible for particular side-effects. Therefore it may not be optimal solution if we construct a single model for all drugs for a particular side-effect because for some rare side-effects, only a few drugs have those side-effects and by incorporating unnecessary drugs into the training set makes the training dataset extremely unbalanced that ended up in low predictive performance model. Thus including fewer similar drugs into the model makes our approach less sensitive to rare side-effect cases which is very important in drug development. Furthermore, to make our classifier more robust we combine results from multiple classification models for each drug where each model is constructed with different training set drugs (different k) because the value of k that gives optimal results depends upon the dataset we are using, partly because of the variety of drugs in the dataset and also the side-effect terms we want to predict. Thus for each drug and each side effect profile we construct N classifiers and have the output probability P_{ijk} which means the probability of drug i having side-effect j by k most similar drugs.

In the next step of our ensemble classification approach to predict whether drug i has side-effect j we consider the output from the previously constructed N classifiers for this drugs for this side-effect terms. To predict the probability that drug i has j side-effect we calculate the mean probability

$$P_{ij} = \frac{\sum_{k \in K} P_{ijk}}{N}, \quad (3)$$

where, $K = \{k_1, k_2, \dots, k_N\} = \{50, 100, 150, \dots, 400\}$ and $|K| = N$.

Thus finally we have a matrix P_{ij} of 888×1385 dimension that has the output of our method. For predicting the side-effect profiles of 2883 uncharacterized drugs we follow the same procedure where we consider each uncharacterized drug at a time and construct models for different number of similar drugs (K) from the available pool of the 888 drugs that have side-effect information. Thus for uncharacterized drugs we have a matrix P_{ij} of 2883×1385 dimension that has the output of our method. We use support vector machine (SVM) as a base classifier in our ensemble classification method but it can be replaced by any other standard classifiers (e.g. decision tree). We used the implementation of support vector machine (SVM) classifier in WEKA [14] (version 3.6.3) and used its linear kernel with all default parameter settings. Classification output was fitted using a logistic regression model, which is included in the WEKA implementation, to predict the probability of side-effects [14]. A pseudocode of our ensemble classification approach is given in Algorithm 1.

C. Performance evaluation

Classification performance was measured using several commonly used evaluation methods including receiver operating characteristic (ROC) curve, area under ROC curve (AUC), and accuracy. We compare our results with another method, sparse canonical correlation analysis (SCCA), which considers chemical structure to predict drug side-effects [6]. To compare, we run their program (available online) to generate the results based on the parameters they used in the article. In this paper we generate results for SCCA based on the

two sets of their parameters (set 1: $c_1 = c_2 = 0.05$ and $m = 20$; set 2: $c_1 = c_2 = 0.2$ and $m = 500$) used in their results [6]. For details see the result and method section in [6]. The predictive performance of SCCA method on parameter set-1 is much worse than parameter set-2. Therefore, we only compare our performance with the performance of SCCA method based on parameter set-2.

III. RESULTS AND DISCUSSIONS

A. Similar drugs have similar side effects

First we validate our hypothesis that similar drugs have similar side-effects. To this end for each drug we identify its closest k neighbors based on Euclidean distance on chemical space. Thus we have k most similar drugs (according to chemical similarity) and based on our hypothesis, they should have similar side-effects. To define side-effects similarity for these k nearest drugs with the interested drug, we calculate a score based on Euclidean distance for the side-effects of each of these neighbor drugs with the interested drug. Mathematically to find the side-effect similarity of drug i with its k most similar drugs we calculate the similarity in the following way,

Algorithm 1

Ensemble classification algorithm

```

1: procedure ENSEMBLECLASSALGO(Drug  $i$ , Side-effect  $j$ ,  $K$ ,  $C$ ,  $S$ )
2:    $t \leftarrow 0$ 
3:   while  $t < |K|$  do
4:      $k \leftarrow K_t$ 
5:      $\mathcal{N}_i^k \leftarrow$  identify  $k$  most similar drugs to drug  $i$  based on  $C$ 
6:      $baseClassifier_t \leftarrow$  create base classifier based on  $C$  and  $S$  of drugs in  $\mathcal{N}_i^k$ 
7:      $P_{ijk} \leftarrow$  predict the probability of drug  $i$  having side-effect  $j$  by  $baseClassifier_t$ 
8:      $t \leftarrow t + 1$ 
9:   end while
10:   $P_{ij} \leftarrow$  calculate the overall probability of drug  $i$  having side-effect  $j$  by Equation 3.
11:  return  $P_{ij}$ 
12: end procedure

```

$$similarity(i, k) = \frac{1}{\sum_{n \in \mathcal{N}_i^k} \frac{1}{k} \times \sqrt{\sum_{m=1}^{|M|} (S_{im} - S_{nm})^2}} \quad (4)$$

where $|M|$ is the number of the side-effect terms and $\mathcal{N}_i^k$ is the k most similar neighbors of drug i .

To identify whether similar drugs have similar side-effects, we calculate the side-effect similarity defined for all the drugs for different number of neighbors. The average similarity for different k values are shown in Figure 2. From the figure it can be seen that drugs with

fewer neighbors have higher similarity compared to drugs with more neighbors and the similarity decreases significantly when increasing the number of neighbors to 850. This verifies our hypothesis that similar drugs have similar side effects because increasing the number of neighbors includes more dissimilar drugs and reduces the similarity. To further validate our hypothesis for each drug we randomly choose k neighbors and find the similarity based on equation 4. From Figure 2 it is evident that with randomly choosing neighbors it is impossible to find as much similarity as we get for real cases. Thus from the above discussion we find that similar drugs (similar chemical structures) have similar side effects which motivates our method because based on the observation a single model for all drugs is not the optimal solution, especially for rare side-effects where the dataset is extremely unbiased. We find that for $k = 50$ the average similarity is highest which drops significantly after $k = 400$. Therefore, to make our method more robust we construct a number of classifiers (from $k = 50, 100, 150$ up to 400) to predict side-effect profiles for each drug and each side-effect term. In this way we make our model less sensitive to rare side-effect terms.

B. Performance evaluation

To find the predictive performance of our method, we compare our results with one previously published approach known as sparse canonical correlation analysis (SCCA) [6]. First, we compare the predictive performance of each method based on AUC scores on individual side-effects. Note that, performance based on each side-effect exhibit the reliability of each model which is important for novel drug side-effect prediction. Figure 3 shows the boxplot of AUC scores for 1385 side-effect terms. From the figure it can be seen that the predictive performance of our method outperformed SCCA significantly (mean AUC 0.57 vs 0.62). For each individual AUC score for 1385 side-effects our method has higher AUC scores for 62% cases compared to SCCA. Directly comparing the AUC scores of our method with NN (nearest neighbour) and single SVM from the result in [6] shows that our method has superior performance to these conventional approaches.

Next, Figure 4 shows the ROC curve for our method and SCCA. To draw the ROC curve we merged the prediction scores for all side-effects and a single ROC curve for each method is drawn. The area under the curve score (AUC) for our method and SCCA is 0.84 and 0.74 respectively. The AUC score shows that our method significantly outperformed SCCA. Based on the curve we see that at up to 55% sensitivity both methods have similar predictive performance while our method outperformed SCCA by increasing the sensitivity beyond this point.

To further compare our method with SCCA, we find for some of the side-effect terms the AUC score is below 0.5 which is lower than random guessing (Figure 3). The possible reason for such low AUC scores may be due to the fact that only a few drugs in the dataset have these side-effects which makes the prediction difficult. For SCCA method 284 side-effect terms have AUC score below 0.5 while for our method 199 side-effects terms have AUC score below 0.5. This implies that for the same dataset our method can predict side-effects more reliably than SCCA. Next, we categorize side-effect terms in different classes based on their significance in SCCA and our method. Note that, the models for side-effects

with AUC = 0.5 can be considered as significant while AUC < 0.5 considered as insignificant. We find that for 949 cases AUC score for both methods are significant while 47 cases have AUC score below 0.5 in both methods. For 237 cases AUC score is significant only in our method and insignificant in SCCA while 152 cases it is opposite (only significant in SCCA). Furthermore, Figure 5 shows the frequency of different side-effects that fall into these different classes. From the figure it can be seen that there is a strong correlation between the side-effect frequency and the AUC score. All those side-effect terms with low AUC have low recorded side-effect frequency which makes the prediction problem more challenging. Importantly, the average frequency count for the side-effect terms that are only significant in our method is 7.1 while for SCCA it is 9.0. It means that our method is less sensitive to the side-effect frequency count and will perform better for rare side-effects.

C. Prediction on uncharacterized drugs

To further evaluate the real life application of our method we predict the side-effects for 2883 drugs in Drug-Bank database with no side effect information in SIDER database [13], [11]. After applying our method for these 2883 uncharacterized drugs, very interestingly we notice almost similar number of side-effect terms per drug predicted by our method that is recorded in SIDER database for 888 training drugs (69.1 vs 68.81 side-effect terms per drug respectively). For further validation in this section we only consider the top scoring 40,454 drug-side-effect predictions by our method.

Analyzing the prediction by our method we observe that very common side-effects terms in SIDER appear top in our prediction. Among top 10 frequent side-effect terms in SIDER database our method predicts 90% side-effects correctly. It is also important to notice that these side-effect terms are common reaction for many drugs with less specificity with particular drug categories. Thus a more important criteria to evaluate any prediction algorithm may be how it performs or predicts the less frequent side-effect terms. Therefore, we identify among the 1385 side-effect terms how many appear in our prediction. We find that 974 side-effect terms present in our prediction while only 232 side-effect terms present in prediction from SCCA method (Figure 6). Further analysis revealed that many side-effect terms present in our prediction have low frequency in SIDER database (data now shown) compared to SCCA method. Therefore, our method predicts 53% more side-effect terms, especially those relatively rare side-effects that are missed by SCCA.

To further investigate the predictive power of our method we find that various drugs with known side-effects can be predicted by our method. (Note that, in this subsection we only limit our discussion to rare side-effects disregarding common side-effects even if it is predicted by our method. For example, Promazine (DB00420) is used for short term treatment of disturbed behaviors [13]. Our method correctly identified one side-effect of this drug termed as 'gynecomastia'. Carphenazine (DB01038) used for hospitalized patients to manage chronic schizophrenic psychoses has a side effect called dyskinesia, which is predicted by our method [13]. For more example, drug 3-Methylthiofentanyl (DB01439) has side-effects that cause serious respiratory problems, while our method identifies side-effect termed as 'respiratory failure' [13]. For drug Dihydrocodeine (DB01551), a common side-

effect is difficulty in passing urine while our approach identifies 'urinary hesitancy'. Drug Rimonabant (DB06155) in Europe was withdrawn from the market for various side-effects related to complex psychic mechanisms [6], [13]. Our method identifies nervousness which is likely for this drug. Therefore, our method can identify side-effects correctly for different uncharacterized drugs.

IV. CONCLUSION

In this work we propose an ensemble classification approach to predict drug side-effects that is based on the decision from a number of base classifiers, where each classifier is developed from drugs with similar chemical structures. We applied our method to predict 1385 side-effect profiles for 888 drugs in DrugBank database. Our method significantly outperformed previously published approaches; especially for rare side-effect cases that are very important for predicting side-effect profiles for novel drugs. Furthermore, after applying on 2883 small molecule uncharacterized drugs our approach is able to predict a large variety of side-effects, notably which are missed by other previous methods. Based on drug side-effect information from different sources, we verify that the prediction by our approach is indeed correct. Therefore, the ensemble classification approach proposed in this article can be useful to predict side-effects in drug design to reduce experimental cost, time and especially fatality that is occurring each year by adverse drug reactions.

Acknowledgments

The authors would like to thank Pauwels et al [6] for making the dataset and code available online. This research was supported in part by the National Science Foundation (IIS-1218201), National Institutes of Health (SC3GM086305, U54CA113001, G12MD007591 (Computational Systems Biology Core)), and a UTSA Tenure-track Research Award.

REFERENCES

1. Tatonetti N, Liu T, Altman R. Predicting drug side-effects by chemical systems biology. *Genome Biology*. 2009; 10(no. 9):238. [PubMed: 19723347]
2. Takarabe M, Kotera M, Nishimura Y, Goto S, Yamanishi Y. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*. 2012; 28(no. 18):i611–i618. [PubMed: 22962489]
3. Hopkins A. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*. 2008; 4:682–690.
4. Keiser M, Roth B, Armbruster B, Ernsberger P, Irwin J, Shoichet B. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007; 25:197–206. [PubMed: 17287757]
5. Giacomini K, Krauss R, Roden D, Eichelbaum M, Hayden M, Nakamura Y. When good drugs go bad. *Nature*. 2007; 446(no. 7139):975–977. [PubMed: 17460642]
6. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*. 2011; 12(no. 1):169. [PubMed: 21586169]
7. Campillos M, Kuhn M, Gavin A, Jensen L, Bork P. Drug target identification using side-effect similarity. *Science*. 2008; 32(1):263–266. [PubMed: 18621671]
8. Fukuzaki M, Seki M, Kashima H, Sese J. Side effect prediction using cooperative pathways. *IEEE International Conference on Bioinformatics and Biomedicine 2009 (IEEE BIBM 2009)*. 2009:142–147.
9. Xie L, Li J, Xie L, Bourne P. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of cetyl inhibitors. *PLoS Comput Biol*. 2009; 5:e1000387. [PubMed: 19436720]

10. Scheiber J, Jenkins J, Sukuru S, Bender A, Mikhailov D, Milik M, Azzaoui K, Whitebread S, Hamon J, Urban L, Glick M, Davies J. Mapping adverse drug reactions in chemical space. *J Med Chem.* 2009; 52(no. 9):3103–3107. [PubMed: 19378990]
11. Kuhn M, Campillos M, Letunic I, Jensen L, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol.* 2010; 6:343. [PubMed: 20087340]
12. Chen B, Wild D, Guha R. Pubchem as a source of polypharmacology. *Journal of chemical information and modeling.* 2009; 49(no. 9):2044–2055. [PubMed: 19708682]
13. Wishart D, Knox C, Guo A, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research.* 2006; 34:D668–D672. [PubMed: 16381955]
14. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. *SIGKDD Explor. Newsl.* 2009; 11(1):10–18.

(a) Number of side-effects for each drug

(b) Number of drugs associated with each side-effect

Fig. 1.
Characteristics of the drug and side-effect data

Fig. 2. Average similarity of drug side-effects based on different number of neighbors. Blue (solid) line for actual neighbors and red (dotted) line for random neighbors.

Fig. 3.
Boxplot for the AUC scores of side-effects obtained from different methods.

Fig. 4.
ROC curve for SCCA and our method.

Fig. 5.
Frequency of side-effect terms for different categories and methods.

Fig. 6.
Pie chart showing percentage of side-effect terms covered by different methods.