Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Exploring the associations between drug side-effects and therapeutic indications

Fei Wang *[,1], Ping Zhang [1], Nan Cao, Jianying Hu, Robert Sorrentino

IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

ARTICLE INFO

Article history:
Received 7 November 2013
Accepted 27 March 2014
Available online xxxx

Keywords:
Theraputic indications
Side-effects
Associations

ABSTRACT

Drug therapeutic indications and side-effects are both measurable patient phenotype changes in response to the treatment. Inferring potential drug therapeutic indications and identifying clinically interesting drug side-effects are both important and challenging tasks. Previous studies have utilized either chemical structures or protein targets to predict indications and side-effects. In this study, we compared drug therapeutic indication prediction using various information including chemical structures, protein targets and side-effects. We also compared drug side-effect prediction with various information sources including chemical structures, protein targets and therapeutic indication. Prediction performance based on 10-fold cross-validation demonstrates that drug side-effects and therapeutic indications are the most predictive information source for each other. In addition, we extracted 6706 statistically significant indication-side-effect associations from all known drug-disease and drug-side-effect relationships. We further developed a novel user interface that allows the user to interactively explore these associations in the form of a dynamic bipartitie graph. Many relationship pairs provide explicit repositioning hypotheses (e.g., drugs causing postural hypotension are potential candidates for hypertension) and clear adverse-reaction watch lists (e.g., drugs for heart failure possibly cause impotence). All data sets and highly correlated disease-side-effect relationships are available at http://astro.temple.edu/~tua87106/druganalysis.html.

1. Introduction

Drug discovery is a slow and expensive process. By conservative estimation, it takes at least 10–15 years and USD 500 million to USD 2 billion to bring a single drug to market [1]. Although the research on drug development has increased significantly in recent years, the number of new therapeutic chemical and biological entities approved by the United States Food and Drug Administration (US FDA) has been declining since the late 1990s. There are two most important reasons for drugs fail clinical trials: (1) lack of efficacy; (2) adverse side-effect. And each of these two reasons accounts for around 30% of clinical trials failures [2]. Therefore it is highly desirable to develop tools that can predict drug therapeutic indications and side-effects accurately.

Therapeutic indication is a valid reason to use a certain medication. Inferring potential novel therapeutic indications for new or approved drugs is one important problem in drug development. Accurate indication prediction can drastically reduce the risk of attrition in clinical phases. In recent years, a number of computational methods have been developed to predict drug indications including.

- Inferring novel drug usage based on shared treatment profile using a network-based, guilt-by-association method [3].
- Predicting drug indications using their chemical structures [4].
- Inferring drug indications from protein targets interaction networks [5,6].
- Identifying relationships between drugs based on the similarity of their phenotypic profiles (e.g., side-effects [7,8] and connective map gene expression [9,10]).
- Integrating multiple information (e.g., chemical, biological, or phenotypic information) of drugs and diseases to predict drug indications [11–13].

With the exception of Yang et al. [8] which used side-effects, these strategies focus primarily on using preclinical information. However, clinical therapeutic effects are not always consistent with preclinical outcomes.

Drug side effect is a secondary, typically undesirable effect of a drug or medical treatment. Predicting drug side-effects, or adverse

* Corresponding author.
E-mail address: feiwang03@gmail.com (F. Wang).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.jbi.2014.03.014
1532-0464/

drug reactions, is another important aspect of drug development. According to the statistics, serious drug side-effects has been the fourth leading cause of death in US, resulting in 100,000 deaths per year [14]. One approach for identifying potential adverse drug side-effects is preclinical in vitro safety profiling, which tests compounds with biomedical and cellular assays. However this experimental methodology is very expensive and labor intensive. Therefore, developing effective computational methods for accurate drug side-effect prediction is of vital importance. There have been some prior studies on this topic, which can be categorized into three classes:

- Linking drug side-effects to their chemical structures [15–17], following the spirit of QSAR (quantitative structure–activity relationship).
- Relating drug side-effects to its protein targets [18,19] because drugs with similar in vitro protein-binding profiles tend to exhibit similar side-effects.
- Predicting drug side-effects by integrating multiple data sources (e.g., chemical, biological, or phenotypic properties) [20–22].

From these existing studies we can see that, although therapeutic indications and side-effects are both measurable behavioral or physiological changes in response to the treatment, they have mostly been researched independently in the past. Intuitively, if drugs treating a disease share some common side-effects, this could suggest some underlying mechanism-of-action (MOA) linkage between the indicated disease and the side-effects. Moreover, many side effects are extensions of a drug's intended phenotypic effect (e.g., hyper- and hypo-tension), so it is logical that there is a correlation between indication and side effect. However, there is a lack of systematic study on exploring the associations between drug therapeutic indications and side-effects, which could be of broad interests in drug development and repositioning.

In this paper, we conducted a comprehensive investigation on building effective computational models for predicting drug therapeutic-indications and drug side-effects. We compared the predictive power of different sources of information (drug chemical structure, protein target, as well as disease indication and side-effects themselves), which shows that, indeed, drug side-effects and therapeutic indications are strong predictors of each other. This confirms the hypothesis that there exist strong associations between drug indications and side-effects. To quantize the strength of those associations, we performed Fisher's exact test with the prediction results [23]. Note that some preliminary evaluations on known associations between drug indications and side-effects are presented in our conference paper [24]. In this paper, we did a much more thorough investigation on all possible (both known and unknown) drug indication and side-effects associations. We also built a visualization tool to facilitate the user's exploration of those detected associations, which can be used to provide repositioning hypotheses (e.g., drugs causing postural hypotension are potential candidates for hypertension), as well as adverse-effect watch lists (e.g., drugs for heart failure possibly cause impotence).

The key differences between this paper and prior studies are:

- We evaluate effectiveness of both drug therapeutic indications and side-effects when predicting each other. Most prior work does not explicitly leverage the relationship between indications and side-effects, in combination with other drug properties. The prior work that is most closely associated with ours is Yang et al. [8]. However they used side-effects alone to predict drug indications. Moreover, their approach was only evaluated on a small data set (145 diseases and 584 side-effects). The data set we used in this paper is much larger, which includes 719 diseases and 1385 side-effects.
- We build disease-side-effect profiles to elucidate interesting relationships between drug side-effects and therapeutic indications with clinical implications, which provides a systematic way to generate drug indication hypotheses and adverse-effect watch lists. To the best of our knowledge, there is no prior work on this topic.
- We propose a novel visualization approach to support the interactive exploration of indication and side-effect associations in the form of a dynamic bi-partite graph.

The rest of this paper is organized as follows. In Section 2 we will introduce the details of the data set we used for this study. The methodology is presented in Section 3, followed by the experimental results in Section 4. Finally we will conclude in Section 5.

2. Data set

We performed our study on approved drugs from DrugBank [25], which is a widely used public drug information database. From DrugBank, we collected 1447 FDA-approved small-molecule drugs, and mapped them to PubChem [26] to get their chemical structure information. After matching by the DrugBank provided PubChem Compound ID for the drugs, we extracted chemical structures of the 1103 drugs. To encode the drug chemical structure, we used a fingerprint corresponding to the 881 chemical substructures defined in the PubChem. Each drug was represented by an 881-dimensional binary profile, within which the entry is 1 if the corresponding PubChem substructure is present, otherwise it is 0. Take the drug calcium as an example, its chemical formula is just Ca, which only meets the requirement of the bit 52 (>=1 Ca). Thus drug calcium only has 1 association with chemical substructures >=1 Ca. Similarly, aspirin has 115 associations with chemical substructures, ibuprofen has 84 associations with chemical substructures. A description of the 881 chemical substructures can be found at the website of PubChem (http://pubchem.ncbi.nlm.nih.gov/). Adding up together, we identified 132,092 associations between drugs and chemical substructures in the dataset, i.e., each drug has 119.8 substructures on average.

From DrugBank, we can also obtain the protein target information for each drug. To facilitate collecting such information, we mapped those target proteins to UniProt Knowledgebase [27], a central knowledgebase including the most comprehensive and complete information on proteins. After matching with the DrugBank provided UniProt ID for the drugs, we extracted 3152 relationships between 1007 drugs and 775 protein targets, so each drug has 3.1 protein targets on average. Similar to the chemical structure representation, each drug was represented by a 775-dimensional binary profile whose elements encode the presence or absence of their corresponding target proteins.

The third type of information we are interested in is drug side-effects. We extract side-effect keywords from the SIDER database [28], which contains information about medicines that are in market and their recorded adverse drug reactions. SIDER uses STITCH compound ids as its drug id, but can be easily matched to PubChem Compound ID via this rule (ftp://sideeffects.embl.de/SIDER/2012-10-17/README). This dataset contains 888 small-molecule drugs and 1385 side-effect keywords. Similar to the representations we mentioned above, each drug can be represented by a 1385-dimensional binary profile whose elements encode the presence or absence of each side-effect keyword. We plotted the cumulative counts of side-effect data in Fig. 1, from which we can observe that 1.69% of drugs have between 10 and 100 different side effects; 22% of drugs have more than 100 side-effects; only 9% of drugs have

Please cite this article in press as: Wang F et al. Exploring the associations between drug side-effects and therapeutic indications. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.03.014