Dydruma: A Dynamic Drug Map for Exploring the Associations Among Drug Therapeutic Indications and Side-Effects

Fei Wang, PhD, Ping Zhang, PhD, Nan Cao, PhD,
Jianying Hu, PhD, Robert Sorrentino, MD
IBM T.J. Watson Research Center, New York, USA

Abstract

Inferring potential therapeutic indications and identifying clinically important side-effects are both important and challenging tasks in modern drug development. Previous studies have utilized drug chemical structures and protein targets to construct predictive models for both tasks. According to our study, the drug therapeutic information itself is highly predictive for side-effects, and drug side-effect information highly predictive of therapeutic indication. This confirms that there exist underlying associations among drug therapeutic indications and side-effects. Exploring these associations can lead to better understanding of the drugs as well as more informed hypotheses for drug repositioning and adverse effect monitoring. In practice however, it is impractical to check all possible associations using an exhaustive list. In this study, we present **Dydruma***, a dynamic drug map which encodes drug therapeutic indications and side-effects as well as their associations. The map can be dynamically adjusted based on a significance value attached to each association, where the significance value is derived from a statistical test such as Fisher's exact test. We describe an optimization-based approach for dynamic bipartite graph layout to ensure visual continuity among successive layouts at different significance thresholds, to help the user maintain a consistent mental map throughout the exploration. We demonstrate the effectiveness of dytruma for exploring the associations among drug indications and side effects and for hypothesis generation using real world data sets.*

Introduction

Predictive modeling of therapeutic indications and side-effects of drugs holds is an important computational approach for reducing the drug attrition rate and improving the drug discovery process. In the past many works have been done along this line. For example, prediction with drug chemical structure [1][2], protein target information [3][4], and fusion of multiple information sources [5][6]. In our recent study [7], we developed models to incorporate drug therapeutic indication and side-effect information into the prediction procedure of each other. We showed that drug side-effects are important information for therapeutic indication prediction; and drug therapeutic indications are also predictive for their side-effects. This confirms that there are some underlying associations between them, and understanding those associations can be very helpful to drug development. For example, those discovered associations can 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). We also compiled a list of associations among all known drug-disease and drug-side-effect to build disease-side-effect profiles and identified statistically significant associations between drug side-effects and therapeutic indications.

However, because the size of the potential association list is very large, it is exhausting to check the list one by one to identify the interesting ones. Therefore in this study we present *dydruma*, a bi-partite graph representation of the drug therapeutic indication, side-effects and their associations. We will present the details on the visualization design in next section, followed by case study and conclusions.

Visualization Design

On dydruma there are two types of nodes, drug therapeutic indications and side-effects, and the edges correspond to the discovered associations between them. Attached to each association there is a strength (which is represented by the p-value of Fisher's exact test, the smaller the value the stronger the association [7]). We do not want to encode those values into the edge colors because that will make the graph messy. Therefore we come up with a dynamic design. We provide a sliding bar which the user can use to adjust the association strength threshold: only the associations whose p-value below the threshold will be shown. With the user dragging that bar, the bipartite graph will change and we

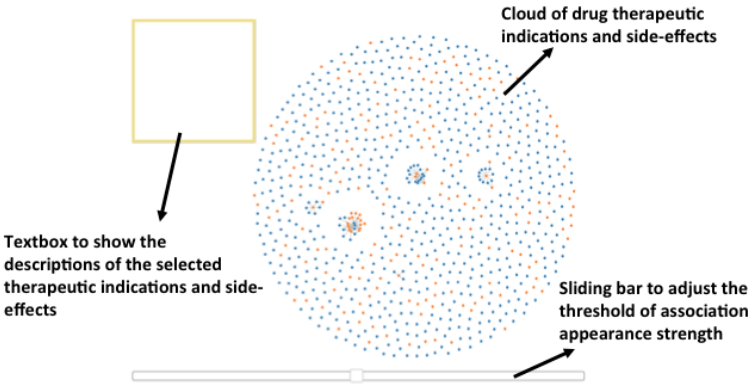

Figure 1: An overview of the *dydruma* system, which consists of three parts: (I) the generated bipartite graph of the drug therapeutic indications (orange) and side-effects (blue); (II) a sliding bar for adjusting the strength threshold (p-value) for showing the discovered associations; (III) a text box displaying the descriptions of the selected nodes (the user can select a bunch of nodes with mouse and their descriptions will be depicted in this box).

need to make that change as smooth as possible for visual pleasure. In the following section we will introduce the details of our visualization design. Before going into that, we give an overview of *dydruma* in Fig. 1.

The dynamic bi-partite graph layout is based on an optimization approach. Specifically, in order to help maintain a user's mental map, successive layouts of similar graphs should have minimal changes (*stability*). Furthermore, each of such layouts should still effectively convey the properties of the underlying graph (*readability*). Thus, our goal is to produce a sequence of graph layouts that optimize both the stability and readability of the resulted visualization. To achieve this goal, we develop a spectral layout algorithm.

Given a dynamic graph $\mathcal{G}_t = \langle \mathcal{V}_t, \mathcal{E}_t \rangle$ at time t, consisting of a set of nodes $\mathcal{V}_t$ and links $\mathcal{E}_t$, we define an energy function to model the desired graph layout as follows:

$$\min \left[\sum\nolimits_{i<j} \omega_{ij} \alpha (||X_i - X_j|| - d_{ij})^2 + \sum\nolimits_{i \in C_k} (1-\alpha)(X_i - X_i')^2 \right] \tag{1}$$

Where X_i' and X_i represent the previous and new position of node $v_i \in \mathcal{V}_t$, respectively. The first term of the objective in Eq.(1) is from the Kamada and Kawai method [8], which maximizes the readability of a graph visualization by preserving the pairwise distances, where d_{ij} is the shortest distance between two nodes v_i and v_j. The second item, which we have added, attempts to minimize the changes in successive layouts.

Instead of stabilizing all the unchanged nodes (which consists of a set $\mathcal{U}$), we extract a representative set of unchanged nodes $\mathcal{C}_k \in \mathcal{U}$ to improve the performance of the algorithm. The final layout model is constructed by optimizing Eq.(1) with a spectral method. Here $\alpha \in (0, 1)$ is the weight that is dynamically computed to achieve the desired balance between readability and stability. An online demo of the *dydruma* system can be found on `http://nancao.org/demos/druggraph/`.

Case Study

Data. In the experiment, we analyzed the approved drugs from DrugBank [9], where we collected 1,447 FDA-approved small-molecule drugs. We mapped these drugs to several other key drug resources including PubChem [10] and UMLS [11] in order to extract other drug related information. In the end, we extracted chemical structures of the 1103 drugs from PubChem. 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 whose elements encode for the presence or absence of each PubChem substructure. There are 132,092 associations between drugs and chemical substructures in the dataset, and each drug has 119.8 substructures on average.

From DrugBank, we also got target information of each drug. To facilitate collecting target protein information, we mapped target proteins to UniProt Knowledgebase [12], a central knowledgebase including most comprehensive and complete information on proteins. In the end, we extracted 3,152 relationships between 1007 drugs and 775 proteins, and each drug has 3.1 protein targets on average. Each drug was represented by a 775-dimensional binary profile whose elements encode for the presence or absence of each target protein by 1 or 0, respectively.

Side-effect keywords were obtained from the SIDER database [13] which contains information about marketed medicines and their recorded adverse drug reactions. This led to a dataset containing 888 small-molecule drugs and 1385 side-effect keywords. Each drug was represented by a 1385-dimensional binary profile whose elements encode for the presence or absence of each of the side-effect keywords by 1 or 0, respectively. Altogether, there are 61,102 associations between drugs and side-effect terms in the dataset, and each drug has 68.8 side-effects on average.

Drugs' known usages were obtained by extracting treatment relationships between drugs and indications from the National Drug File - Reference Terminology (NDF-RT), which is part of the UMLS [11]. This list is also used by Li et al [14] as the standard set for drug repositioning. After normalizing various drug names in NDF-RT to their active ingredients, we were able to extract therapeutic indications for 799 drugs out of the 1103 drugs, which constructed 3250 treatment relationships between 799 drugs and 719 indications. Thus each drug was represented by a 719-dimensional binary profile whose elements encode for the presence or absence of each of the therapeutic indications.

Prediction Results.

As stated in the introduction, the whole motivation of dydruma is the high prediction ability between drug therapeutic indications and side-effects. Based on the approach in [7], we got 167,392 predicted associations between 567 drugs and 1262 side-effect terms, and 22,639 predicted associations between 567 drugs and 612 indications. Then we combine both predicted and ground truth indication-side-effects associations to construct the confusion table for Fisher's exact test [15], from which we can get the strength (p-value) of those predicted associations [7].

Example Clique. We demonstrate those discovered associations in our *dydruma* system and show an example clique in Fig.2, where there are four diseases: *Diabetic Nephropathies, Heart Failure, Hypertension, and Ventricular Dysfunction*. The last three are cardiovascular diseases, and Diabetic Nephropathies is a common comorbidity and one of the causes of cardiovascular diseases. This clique also contains 30 highly correlated side-effects. Some of the side-effects are physiologically linked to the cardiovascular diseases and the mechanism of action (MOA) can be explained. For example, some hypertension drugs may result in a sudden drop in blood pressure when a person stands up, thus the side-effect postural hypotension happens. Some cardiac drugs (e.g., β-blockers) hits α-adrenergic receptors protein target in penile tissue, which will cause side-effect impotence. The decreased blood pressure caused by some cardiac drugs (e.g., β-blockers) also cause side-effects cold extremities, dizziness, vertigo, and weakness. Side-effect pemphigus is related to ACE inhibitors, which is also one kind of cardiac drug. Some popular cardiac medications (e.g., Diuretics) cause human body to lose salt and water, potentially causing side-effect gout.

Conclusions

We present *dydruma*, a dynamic bi-partite graph laying out system for visualizing the predicted associations between drug therapeutic indications and side-effects, which could be potentially very interesting because of their high predictability of each other according to our previous study. We finally present a real world case study to demonstrate its validity.

References

1. Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, et al. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

2. Kathleen M Giacomini, Ronald M Krauss, Dan M Roden, Michel Eichelbaum, Michael R Hayden, and Yusuke Nakamura. When good drugs go bad. *Nature*, 446(7139):975–977, 2007.

3. Jiao Li, Xiaoyan Zhu, and Jake Yue Chen. Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. *PLoS computational biology*, 5(7):e1000450, 2009.

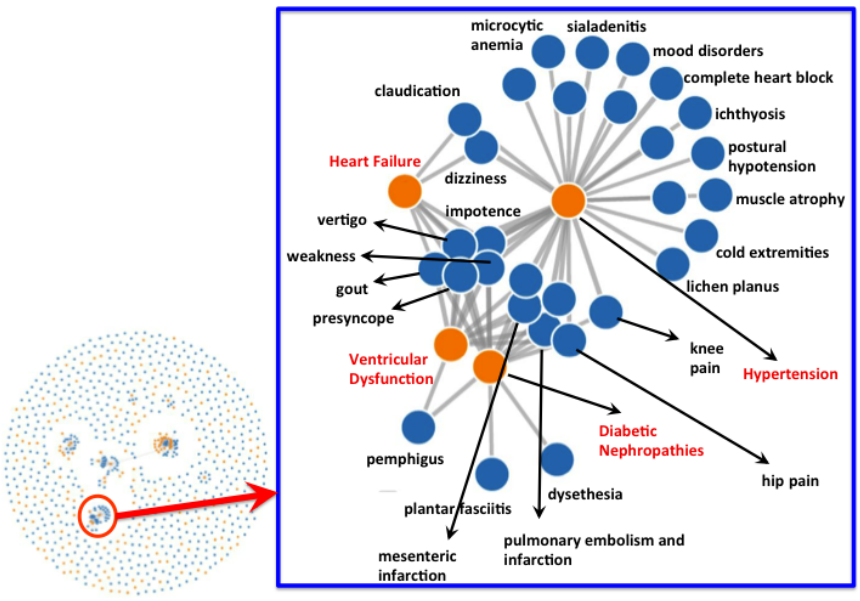

Figure 2: An example clique of drug therapeutic indications and side-effects.

4. Li Xie, Jerry Li, Lei Xie, and Philip E Bourne. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. *PLoS computational biology*, 5(5):e1000387, 2009.

5. Ping Zhang, Pankaj Agarwal, and Zoran Obradovic. Computational drug repositioning by ranking and integrating multiple data sources. In *Machine Learning and Knowledge Discovery in Databases*, pages 579–594. Springer, 2013.

6. Mei Liu, Yonghui Wu, Yukun Chen, Jingchun Sun, Zhongming Zhao, Xue-wen Chen, Michael Edwin Matheny, and Hua Xu. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association*, 19(e1):e28–e35, 2012.

7. Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Exploring the relationship between drug side-effects and therapeutic indications. In *Proceedings of American Medical Informatics Association Annual Symposium*, 2013.

8. Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989.

9. David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008.

10. Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, 37(suppl 2):W623–W633, 2009.

11. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

12. Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119, 2004.

13. Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1), 2010.

14. Jiao Li and Zhiyong Lu. A new method for computational drug repositioning using drug pairwise similarity. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–4. IEEE, 2012.

15. Graham JG Upton. Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in society)*, pages 395–402, 1992.