DrugPathSeeker: Interactive UI for Exploring Drug-ADR Relation via Pathways

Janu Verma* Heng Luo† Jianying Hu ‡ Ping Zhang §

IBM TJ Watson Research Center, Yorktown Heights, NY - 10598

ABSTRACT

Biological interpretation and understanding of machine learning based predictive models are highly desirable in healthcare analytics. Predicting Adverse Drug Reactions (ADRs) is extremely important for safe and precision medicine. There are various machine learning based approaches to predict adverse reactions for drugs. These models, though effective, lack biological interpretation and are treated as black-boxes. We propose DrugPathSeeker, a novel interactive user interface that integrates the machine learning model, database query API, statistical analysis, and visualization for exploring and understanding of the association between drugs and ADRs. The proposed UI can take a query drug, and provide a visual interface designed to support exploration of the predictions from the machine learning model for further understanding and interpretation. DrugPathSeeker uses a machine learning model, Small Molecular Risk Profiler, to make ADR predictions for a given drug. The visualization uses Sankey type flow diagrams for highlighting the relation between the drugs and ADRs. The main goal of DrugPathSeeker is to mine the gene-pathways from public databases and analyze them in a visual manner to generate a biological hypothesis. DrugPathSeeker's effectiveness is demonstrated with two use cases: mechanisms of action for *carbamazepine-induced dystonia*, and *fluorometholone-induced diabetes mellitus*

Index Terms: H.5.2 [Information Interface and Principles]: User Interfaces—User-centered design; D.2.2 [Software Engineering]: Design Tools and Techniques—User interfaces;

1 INTRODUCTION

Adverse drug reactions (ADRs) are the clinical conditions resulted from taking medications at normal doses. They cause 700,000 emergency department visits and 120,000 hospitalizations per year and are one of the major causes of death among hospitalized patients [7]. The *antihistamine* drug *astemizole* (Hismanal) and the gastrointestinal-disorders drug *cisapride* (Propulsid) were withdrawn from the market in 1999 and 2000, respectively, after the discovery that both could cause fatal *arrhythmias* when given in combination with certain other drugs. It is, therefore, very important to be able to predict the ADRs for drugs even in the early development stage. To identify and predict ADRs, researchers have developed various methods e.g. machine learning models on structural descriptors [4], similarity analysis of molecular docking profiles [18], network-based approaches [9] and data mining of electronic health records [1]. While some of these methods can easily generate ADR predictions, they do not contain any information about the underlying biological mechanisms. Such models often produce a score quantifying the association of the drug with the ADR. What is not clear, is the biological interpretation of the models and why a particular ADR has higher association score than the other for the drug. These models are black-boxes from the biological perspective, and don't exude a strong biological evidence. Deeper understanding and interpretation of drug-ADR associations are highly desirable. In this work, we make a step forward towards proposing biological evidences in support of machine learning predictions.

We introduce a model based on the analysis of biological (gene) pathways of the drugs and the ADRs. The pathways contain genes of potential interest and the underlying biological machanisms of action. With the development of systems biology and omics, more and more biological pathways related to drugs and ADRs are being discovered and collected in publicly available databases. We mine the Drug-Path database [24] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [16] for the pathways for the drug and the ADR under consideration. In order to support flexible exploration, we present *DrugPathSeeker*, an interactive user-interface designed to generate reasonable biological hypotheses for the mechanism of action by integrating visualization with KEGG database API, statistical analysis and machine learning models.

2 DRUGPATHSEEKER

DrugPathSeeker is designed to study the association of drugs and adverse drug reactions. It provides a unified framework to harvest and visualize pathways that are associated with drugs and clinical conditions, i.e., ADRs in this study. The drug-ADR associations are obtained by analyzing pathways, and are ranked by statistics. The ranking is shown explicitly using visual encodings. There are various interactions for further exploration of the association. The flow chart of the framework is shown in Figure 1.

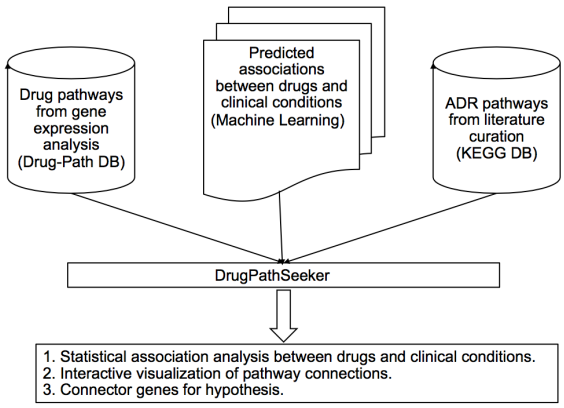

Figure 1: Flowchart of the system.

2.1 User Interface

The visual-analytical system starts with an input query drug and provides an interactive framework for exploring and understanding

*e-mail: jverma@us.ibm.com
†e-mail:heng.luo@ibm.com
‡email: jyhu@us.ibm.com
§e-mail:pzhang@us.ibm.com

of the ADRs associated with the drug. *DrugPathSeeker* employs a machine learning model, Small Molecule Risk Profiler (SMRP), to predict the ADRs for the query drug. The model is described in the Section 2.2. After the user submits a query drug and the number of top ADRs to be retrieved, the results of SMRP model are shown in an interactive visual-analytical system. To provide a biological interpretation of the SMRP results, the proposed system queries the Drug-Path database and the KEGG database API for gene pathways of the drug and the predicted ADRs. The pathways are then analyzed to obtain association scores between the drug and the ADR. The pathways and the associations between them are displayed in an interactive visualization of Sankey type flow diagrams. The visualization is described in Section 2.4.

2.2 Small Molecule Risk Profiler

We used a machine learning model called *Small Molecule Risk Profiler* [19] to predict ADRs from drug structures. We harvested the SIDER database [17], which contains drug-ADR relationships, as our labels of ADR prediction. We filtered the ADR endpoints to make sure each ADR was caused by at least 5 known drugs to ensure a sufficient number of positive samples. Given any drug, a binary vector of 881 PubChem pre-defined structural descriptors were generated as features using R package *rcdk* package [14]. Therefore, for each ADR, we have 1,034 drugs as rows, 881 binary features as independent variables (X) and whether this drug causes the ADR as dependent variable (Y). A *logistic regression* model with L2 regularization was developed using *scikit-learn* package in *python 2.7* for each of the 1,358 ADR. To make sure the prediction scores across the 1,358 ADRs are comparable, we utilized an *empirical Bayes method* [8] to normalize them. The details of the implementation were described in the paper [19].

2.3 Database Mining and Pathway Analysis

To analyze the pathway connections of a predicted drug-ADR pair, we harvested drug pathways from the Drug-Path database [24], a database for drug-induced pathways by analyzing gene expression, and queried ADR pathways from the KEGG database [16], which contains a comprehensive collection of pathways curated from the literature. A pathway is a sequence of genes which can be affected by the drug related to the ADR. For example, in this paper [23], the small molecule activated the cancer-related pathway.

We built a network of pathways with two sets of nodes - drugs pathways and ADR pathways, and there are edges from drug pathways to the ADR pathways. The edge from a drug pathway to an ADR pathway is weighted by the the *Jaccard Index* of the sets of the genes in the corresponding pathways. More explictly, if $S_{drugPath}$ is the set of genes in a pathway $drugPath$ of the drug, and $S_{adrPath}$ is the set of genes in a pathway $adrPath$ of the ADR, then the *Jaccard Index* of the pathways is defined as

$$Jaccard\ Index = \frac{size(S_{drugPath} \cap S_{adrPath})}{size(S_{drugPath} \cup S_{adrPath})} \quad (1)$$

The rationale for using Jaccard Index is that the genes shared by both the drug and the ADR pathways are potential connectors and can be used for hypothesis generation. Larger the *Jaccard Index*, stronger is the biological connection between the pathways, which offers an explanation for the association between the drug and the ADR.

2.4 Interactive Visualization

DrugPathSeeker attempts to not only extract and analyze the pathways but also to present the information in an interactive user interface so that the results of the machine learning and pathway mining can be easily explored and utilized. Thus, a key component of the *DrugPathSeeker* system is a hierarchical Sankey type flow visualization. We use the implementation of Sankey layout in D3.js library [6].

The visualization supports three levels of hierarchy. At the first level, the results of SMRP are presented (see Figure 2). In this view, the drug and the predicted ADRs are represented as nodes, and there are edges from the drug to each of the ADRs. The nodes and the edges are positioned using a modified Sankey diagram layout [20].The width of the edges quantifies the confidence score of the machine learning model.

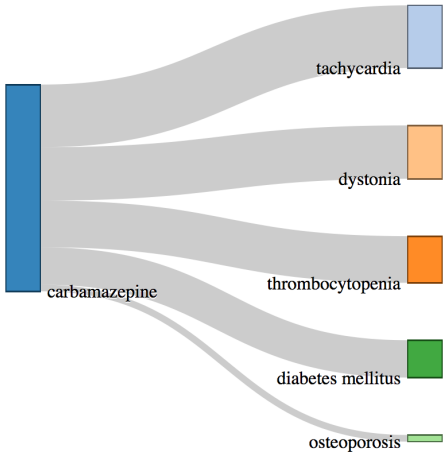

Figure 2: Results of SMRP

Going a level deeper, we have the pathway view, which contain drug pathways and ADR pathways nodes in addition to the drug and the ADR nodes. The two column view of Figure 2 expands into a four column view in this view. There are edges between the drug (or the ADR) and all its pathways. There are also edges between the drug pathways and the ADR pathways if they share genes. The thickness of the edges between pathways encapsulates the *Jaccard Index*. The flow of information is as follows:

$$Drugs \rightarrow Drug\ Pathways \rightarrow ADR\ Pathways \rightarrow ADRs \quad (2)$$

For example, we have a sequence

$$Carbamazepine \rightarrow MAPK\ signaling\ pathway \rightarrow Osteoclast\ differentiation \rightarrow Immune\ thrombocytopenia \quad (3)$$

(see Figure 3) which asserts that the drug *carbamazepine* has a pathway *MAPK signaling pathway*, which has non-null overlap with the pathway *Osteoclast differentiation* of the ADR *Immune thrombocytopenia*. Also, one drug-pathway can have non-zero *Jaccard similarity* with more than one ADR-pathways and vice-versa, as seen in the Figure 3, *MAPK signaling pathway* has an edge to *Phagosome* and *Fc gamma R-mediated phagocytosis*. All the edges from each pathway of a drug to each pathway of the ADR contribute to the confidence score for the association between the drug and the ADR.

The edges are comprised of the gene overlap between the pathways, *DrugPathSeeker* supports an even finer level comprising of genes. The genes responsible for an edge can be viewed by clicking on the edge. The common genes will be displayed on the right side of the window (Figure 4). These genes shared by the drug and the ADR pathways are potential trigger genes can can be used for further hypothesis generation and testing.

The interaction provided in the system include tooltip on mouseover to the edges which shows the number of overlapping

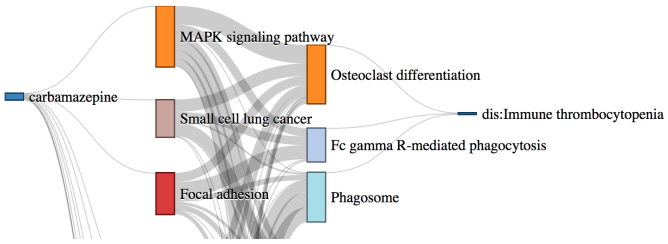

Figure 3: Pathway Connections between drug and ADRs

MAPK signaling pathway --> Osteoclast differentiation
Jaccard index: 0.142

AKT3; AKT serine/threonine kinase 3
CHUK; conserved helix-loop-helix ubiquitous kinase
TAB1; TGF-beta activated kinase 1 (MAP3K7) binding protein 1
MAP2K1; mitogen-activated protein kinase kinase 1
MAPK9; mitogen-activated protein kinase 9
MAPK11; mitogen-activated protein kinase 11
MAPK13; mitogen-activated protein kinase 13
MAPK10; mitogen-activated protein kinase 10

Figure 4: Genes shared by the pathways

genes and the *Jaccard similarity* of the edge (Figure 5). One of the limitations of the current system is that the overlap of labels can impede readability. We provide dragging interaction in the visualization so that the edges and the nodes can be dragged for better readability.

The system also supports tuning of the threshold of *Jaccard similarity* to filter out the non-significant edges, which helps in quick selection and filtration. This makes the system more scalable and tamed, even for a large number of drugs and ADRs.

This visualization makes the results easy to observe and compare by explicitly displaying the ranking of various comparisons using the thickness of the edges. Though we focussed on one query drug, *DrugPathSeeker* can also be used to study many drugs simultaneously to compare the outcomes, thereby making the results more transparent and explainable. Such function enables the applications of the tool in the fields of drug-drug combination and drug-drug interactions.

3 Use Case

3.1 Mechanism of action for carbamazepine-induced thrombocytopenia

We would like to explore the ADRs and related pathways for *carbamazepine*. As the first step, we submitted the drug structure to the Small Molecule Risk Profiler, the machine learning model of *DrugPathSeeker*, to predict ADRs and found that it may cause *tachycardia*, *thrombocytopenia* and *dystonia* with 74%, 71% and 64% confidence, respectively (Table 1). This prediction is also supported by other evidences reported in literature [21, 13, 3]. For

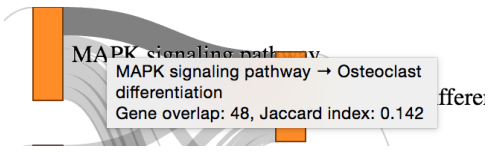

Figure 5: Tooltip

further insights, we analyzed the pathways via *DrugPathSeeker* visualization, which showed connections between *carbamazepine* and the three ADRs (Figure 6). Upon investigation, we found that there are a number of genes such as *mitogen-activated protein kinases (MAPKs)* that are shared by the *MAPK signaling pathway* (drug pathway) and the *osteoclast differentiation* (ADR pathway) between *carbamazepine* and *immune thrombocytopenia* (Figure 7). It was reported that *MAPK activation* is related with *thrombocytopenia* [5]. Also known is that *carbamazepine* has association with p38 MAPK activation [15]. This piece of information may be helpful for understanding the mechanism of *carbamazepine-induced thrombocytopenia*. We believe that the connections highlighted by *DrugPathSeeker* can be used as potential evidence to further investigate the mechanism of action in biological experiments.

Table 1: ADR prediction for carbamazepine

UMLS code	ADR name	Confidence score
C0039231	Tachycardia	74%
C0040034	Thrombocytopenia	71%
C0393593	Dystonia	64%

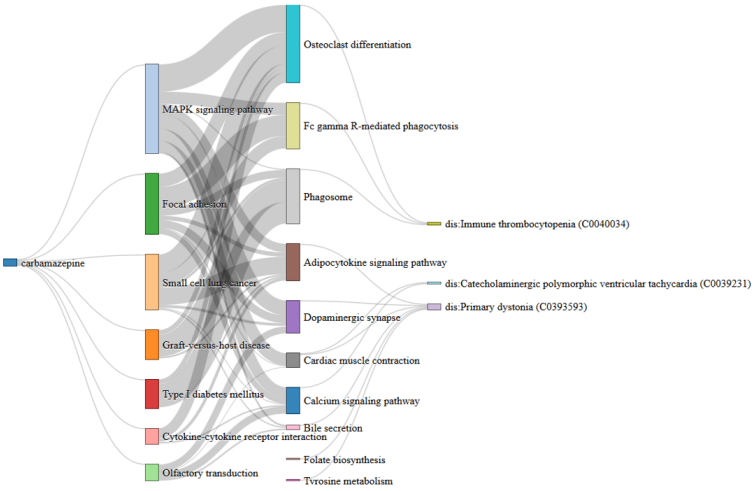

Figure 6: Pathway connection visualization for carbamazepine

3.2 Fluorometholone-induced diabetes mellitus

Fluorometholone is a corticosteroid drug for eye inflammation. The Small Molecule Risk Profiler found that it may lead to *diabetes mellitus*, *osteoporosis* and *osteonecrosis* with 71%, 59% and 57% confidence, respectively (Table 2). These findings are partially supported by prior research [2, 11]. The *DrugPathSeeker* Sankey visualization of the pathway connections between *fluorometholone* and the three ADRs is shown in Figure 8. By using the interactions in the system, we found that the *MAPKs* are among the common genes shared by the pathways between *fluorometholone* and *diabetes mellitus*. It has been reported that corticoids may affect the *MAPK signal pathway* [22]. The prior research also showed that the *MAPKs* have an association to diabetes [12]. We believe the MAPK genes are a good start to study *fluorometholone-induced diabetes mellitus*.

In both the case studies, *DrugPathSeeker's* analysis via the pathway-connection visualization was able to provide biological clues for the mechanisms of action and to generate hypotheses for drug-ADR associations. We were able to validate past research findings with high confidence. At the same time, the framework can

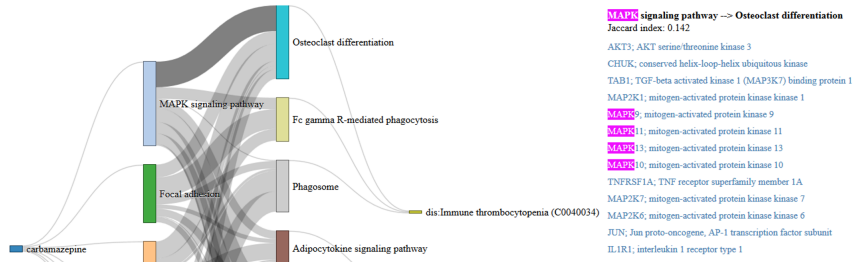

Figure 7: Genes shared between MAPK signaling pathway and Osteoclast differentiation

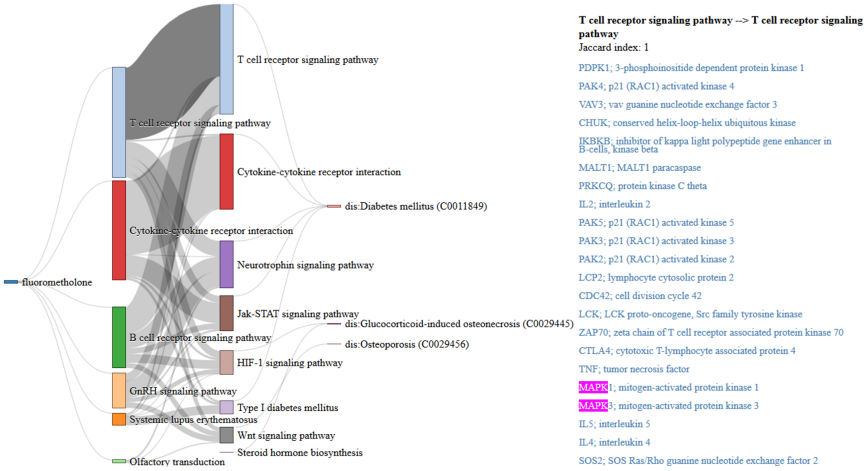

Figure 8: Pathway connection visualization for fluorometholone

Table 2: ADR prediction for fluorometholone

UMLS code	ADR name	Confidence score
C0039231	Diabetes mellitus	71%
C0393593	Osteoporosis	59%
C0029445	Osteonecrosis	57%

further be used to guide wet-lab experiments such as gene knock-out or knock-down experiments to understand the mechanisms of how the drugs induce the ADRs. Instead of directly using cell lines or animal models to test drug reactions, it provides molecular interpretations and insights which makes the hypothesis generation and testing easier and more targetted. Note that we used ADRs as an example of clinical conditions in this paper to demonstrate the tool. We believe the tool can be further applied to related areas such as understanding of drug-drug combinations and drug-drug interactions.

4 Limitations and Future Work

There are some obvious limitations of the current system. Most notably, the overlap of labels with the visualization and potentially with other labels impedes readability. Using drag interaction, this problem can be somewhat circumvented, but this is far from an ideal solution. We used *Jaccard Index* as the measure to quantify the association of a drug and an ADR. This choice is rudimentary, we believe a more sophisticated metric (e.g. based on gene set enrichment analysis [10]) would yield better results. We plan to address this in future work. This system can be adapted for facilitating analysis of connections within similar type of hierarchical data. Finding an appropriate example case or presenting the system in its full generality is another future extension of this work. Although the system provides for filtering based on a threshold on *Jaccard Index*, more filtering operations, highlighting, and zooming can greatly help in readability.

5 Conclusion

This paper presents *DrugPathSeeker*, a novel visual-analytical interface designed to explore and understand the association between drugs and adverse drug reactions (ADRs). *DrugPathSeeker* includes a machine learning model to predict the ADRs for drugs, and the goal is to offer a biological interpretation and understanding of the prediction. The system is designed to have interactive visualization based on Sankey type flow diagrams, which supports views at three differect hierarchies. The interactive analytics supported by *DrugPathSeeker* can help to explore and generate hypotheses for the mechanism of action of the associations between drugs and ADRs, which aids a better understanding and safer evaluation of drug candidates.

References

[1] D. Backenroth, H. Chase, C. Friedman, and Y. Wei. Using rich data on comorbidities in case-control study design with electronic health record data improves control of confounding in the detection of adverse drug reactions. *PloS one*, 11(10):e0164304, 2016.

[2] I. Bahar, I. Rosenblat, M. Erenberg, I. Eldar, D. Gaton, R. Avisar, and D. Weinberger. Effect of dexamethasone eyedrops on blood glucose profile. *Current eye research*, 32(9):739–742, 2007.

[3] S. Bansal, M. Gill, C. Bhasin, et al. Carbamazepine-induced dystonia in an adolescent. *Indian Journal of Pharmacology*, 48(3):329, 2016.

[4] A. Bender, J. Scheiber, M. Glick, J. W. Davies, K. Azzaoui, J. Hamon, L. Urban, S. Whitebread, and J. L. Jenkins. Analysis of pharmacology

data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, 2(6):861–873, 2007.

[5] D. Bluteau, A. Balduini, N. Balayn, M. Currao, P. Nurden, C. Deswarte, G. Leverger, P. Noris, S. Perrotta, E. Solary, et al. Thrombocytopenia-associated mutations in the ankrd26 regulatory region induce mapk hyperactivation. *The Journal of clinical investigation*, 124(2):580–591, 2014.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011.

[7] D. S. Budnitz, D. A. Pollock, K. N. Weidenbach, A. B. Mendelsohn, T. J. Schroeder, and J. L. Annest. National surveillance of emergency department visits for outpatient adverse drug events. *Jama*, 296(15):1858–1866, 2006.

[8] S. Chen, J. Kang, and G. Wang. An empirical bayes normalization method for connectivity metrics in resting state fmri. *Frontiers in neuroscience*, 9, 2015.

[9] X. Chen, H. Shi, F. Yang, L. Yang, Y. Lv, S. Wang, E. Dai, D. Sun, and W. Jiang. Large-scale identification of adverse drug reaction-related proteins through a random walk model. *Scientific Reports*, 6, 2016.

[10] C. A. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma. The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17(6):353–364, Apr. 2016.

[11] K. L. Gebhard and H. I. Maibach. Relationship between systemic corticosteroids and osteonecrosis. *American journal of clinical dermatology*, 2(6):377–388, 2001.

[12] S. Gogg, U. Smith, and P.-A. Jansson. Increased mapk activation and impaired insulin signaling in subcutaneous microvascular endothelial cells in type 2 diabetes: the role of endothelin-1. *Diabetes*, 58(10):2238–2245, 2009.

[13] J. Goraya and V. Virdi. Carbamazepine-induced immune thrombocytopenia. *Neurology India*, 51(1):132, 2003.

[14] R. Guha et al. Chemical informatics functionality in r. *Journal of Statistical Software*, 18(5):1–16, 2007.

[15] H. J. Jeon, S. R. Han, M. K. Park, K. Y. Yang, Y. C. Bae, and D. K. Ahn. A novel trigeminal neuropathic pain model: compression of the trigeminal nerve root produces prolonged nociception in rats. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 38(2):149–158, 2012.

[16] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, page gkv1070, 2015.

[17] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork. The sider database of drugs and side effects. *Nucleic acids research*, page gkv1075, 2015.

[18] H. Luo, J. Chen, L. Shi, M. Mikailov, H. Zhu, K. Wang, L. He, and L. Yang. Drar-cpi: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. *Nucleic acids research*, page gkr299, 2011.

[19] H. Luo, P. Zhang, X. H. Cao, D. Du, H. Ye, H. Huang, C. Li, S. Qin, C. Wan, L. Shi, et al. Dpdr-cpi, a server that predicts drug positioning and drug repositioning via chemical-protein interactome. *Scientific Reports*, 6, 2016.

[20] H. M. Riehmann, P. and B. Froehlich. Interactive sankey diagrams. In *IEEE InfoVis*, page 233240, 2005.

[21] C. R. Rodríguez, J. Mejias, and J. A. Hidalgo. Unmasking false epilepsy: Catecholaminergic polymorphic ventricular tachycardia. *Cardiol J*, 17(1):98–99, 2010.

[22] K. Stuhlmeier and C. Pollaschek. Glucocorticoids inhibit induced and non-induced mrna accumulation of genes encoding hyaluronan synthases (has): hydrocortisone inhibits has1 activation by blocking the p38 mitogen-activated protein kinase signalling pathway. *Rheumatology*, 43(2):164–169, 2004.

[23] L. T. Vassilev, B. T. Vu, B. Graves, D. Carvajal, F. Podlaski, Z. Filipovic, N. Kong, U. Kammlott, C. Lukacs, C. Klein, et al. In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science*, 303(5659):844–848, 2004.

[24] H. Zeng, C. Qiu, and Q. Cui. Drug-path: a database for drug-induced pathways. *Database*, 2015:bav061, 2015.