Translational Knowledge Discovery between Drug Interactions and Pharmacogenetics

Aditi Shendre, PhD\textsuperscript{1}, Heng-Yi Wu, PhD\textsuperscript{2}, Shijun Zhang, MS\textsuperscript{1}, Pengyue Zhang, PhD\textsuperscript{1}, Lei Wang, PhD\textsuperscript{1}, Desta Zeruesenay, PhD\textsuperscript{3}, Luis M. Rocha, PhD\textsuperscript{4}, Hagit Shatkay, PhD\textsuperscript{5}, Sara K. Quinney, PhD\textsuperscript{3}, Xia Ning, PhD\textsuperscript{1,6}, Lang Li, PhD\textsuperscript{1}

\textsuperscript{1}Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH
\textsuperscript{2}Genentech, Inc., San Francisco, CA
\textsuperscript{3}Division of Clinical Pharmacology, School of Medicine, Indiana University, Indianapolis, IN
\textsuperscript{4}School of Informatics and Computing, Indiana University, Bloomington, IN
\textsuperscript{5}Department of Computer and Information Sciences, University of Delaware, Newark, DE
\textsuperscript{6}Department of Computer Science and Engineering, The Ohio State University, Columbus, OH

Introduction

Drug-drug interactions (DDIs) are one of the major causes of adverse drug events (ADEs) and have been demonstrated as a public health burden.\textsuperscript{1,2} However, barriers among diverse scientific domains makes it difficult to provide a comprehensive understanding of the pharmacokinetics (PK), pharmacodynamics (PD) as well as the pharmacogenetics (PG) and molecular mechanisms of drug interactions.\textsuperscript{3} Despite the existence of several databases built to tackle these issues, none of them sufficiently address this knowledge gap. Informatics research provides an opportunity to bridge this gap between the different domains, enabling translational research as well as the study of a large number of drugs simultaneously. In this context, text and data mining approaches have been applied to mine drug interaction signals from the literature and efficient knowledge discovery tools.\textsuperscript{4,5} As a result, we developed a novel translational research method by using text mining to screen in\textit{-vitro}, in\textit{-vivo} and clinical DDI as well as PG studies.\textsuperscript{6} Additionally, we determined overlaps and subsequently discovered knowledge gaps between the DDI and PG studies.

Methods

Text mining was accomplished in two phases: information retrieval (IR) and information extraction (IE). For this, a lexica of drug names, enzymes, action terms, and ADEs was first prepared. Separate corpora for the IR and IE phases were then constructed for each of the four study types. The IR phase included the training and testing of document-level classifiers to build IR models that maximized the recall rate of identifying relevant abstracts. The recall rates were 0.98, 0.99, 0.86, and 0.97 for the optimal in\textit{-vitro} PK DDI, clinical PK DDI, clinical PD DDI, and PG models. These IR models were then used to screen 25 million PubMed abstracts to identify relevant DDI and DGI abstracts. The IE phase included training and testing of relation-level classifiers that maximized F-measure of extracting relation pairs. The F-measure for optimal in\textit{-vitro} PK DDI was 0.83 using feature group G5 and the Naïve Bayes classifier, for clinical PK DDI was 0.85 using feature group G1 and the Iterative Classifier Optimizer, for clinical PD DDI was 0.73 using feature group G1 and the Naïve Bayes classifier, and for DGI was 0.82 using 50 features and logistic linear regression classifier. These IE models were then used to extract DDI and DGI pairs from the retrieved abstracts. An overlapping analysis was conducted to identify common DDI pairs. Validation was performed using manual review and comparison with DrugBank of the top 20 DDI pairs from all the different areas of DDI evidences. Finally, knowledge gaps between the DDI and DGI evidences were determined using results from the overlapping analysis.

Results

Figure 1 shows the number of abstracts retrieved in the IR phase and the number of DDI and DGI pairs extracted in the IE phase for each of the four study types. Among these, 986 unique drug pairs across all three DDI study types were identified. In case of the DGI pairs, a total of 217,562 drug pairs were further generated given shared enzyme relationships between the two drugs. Overlapping analysis performed to compare “extracted DDIs” from DDI IE to “predicted DDIs” from DGI IE revealed that 94.8% of the 986 unique drug pairs can be predicted by DGI results. Comparison with DrugBank DDI showed 88% overlap of DDI pairs shared between all three DDI studies. Almost all of the top 20 frequently reported DDI pairs as well as those identified in the overlapping analysis were validated as true DDIs through manual review, but only some of them were reported in the DrugBank database.
**Figure 1.** Text mining results from the IR and IE phases and the overlapping analysis

**Hypotheses generation**

Two different types of hypotheses with respect to ADEs were generated based on the knowledge gaps revealed through our text mining approach: 1) genetic hypotheses were generated by translating DDI signals. For example, 150 and 31 genetic hypotheses were generated from 68 CYP3A and 25 CYP2D6 substrates, respectively, of which 119 and 18 were new genetic hypotheses, and 2), molecular mechanistic hypotheses were generated by translating PG signals. For example, 68, 38, and 12 DDI pairs were found to share common molecular pathways, genes, and genetic variants, respectively. Seven of the 38 DDI pairs with shared genes were validated to have DDI effects, of which three had *in-vitro* DDI evidence. Thus, our text mining approach provides integrated drug interaction knowledge that can be used to facilitate discoveries by generating novel hypotheses.

**Conclusion**

We developed a translational research method using text mining approaches to help discover knowledge gaps that exist in drug interaction studies. One of the unique aspects of our method is that it helps categorize the evidence based on the type of DDI studies available, something most databases have not yet done. Secondly, the 986 unique DDI pairs shared across all three study types have potential for future clinical applications. Therefore, this text mining approach will enable us to investigate drug-ADE and genetic/molecular interactions that may not have been considered through conventional drug interaction research methods.

**References**