Emily Doughty	Brian Kirk
University of Maryland, Baltimore County	Rice University, Houston, TX

Mentor: Olivier Bodenreider

ABSTRACT

Integrating drug-gene associations mined from the literature with drug information sources -- A comparison with PharmGKB.

E. Doughty[1], B. Kirk[2], M. G. Kann[1], O. Bodenreider[3]
[1]University of Maryland, Baltimore County, Baltimore, MD 21250, USA
[2]Rice University, Houston, TX 77005
[3]National Library of Medicine, Bethesda, MD 20894, USA
ED and BK shared first authorship

BACKGROUND: While pharmacogenomics is interested in finding the connections between drugs and human genomic variation from the perspective of personalized medicine, mutational information is still primarily locked in the literature. Fortunately, high-throughput text mining approaches are being developed to facilitate the identification of pharmacogenomic knowledge. The principal pharamacogenomics database is PharmGKB, in which mutation-drug associations are manually curated. Text mining and curated resources have different strengths and can be used in combination, where text mining is used to identify areas for curation and curated data serve as a reference for the evaluation of text mining methods.
METHODS: We developed a system for the purpose of comparing drug-gene associations between one text mining tool, the Extractor of MUtations (EMU), and PharmGKB. Our system integrates drug information (from the National Drug File – Reference Terminology, NDF-RT), drug-associated mutations automatically extracted from PubMed abstracts by EMU, and various protein databases, including UniProt. It also integrates drug-gene relations from PharmGKB for comparison purposes. All integrated data are stored in RDF triples that are queried using the SPARQL query language. Mutationally-relevant drug-gene annotations extracted from the literature were compared against drug-gene pairs related to point mutations in the variant annotations from PharmGKB. Annotations for select drugs were reviewed manually.
RESULTS: We found a total of 556 unique drug-gene pairs from EMU and PharmGKB. Thirty-four pairs (6 percent) were found in both EMU and PharmGKB. 334 were only found using EMU (60 percent). Finally, 118 were only found in PharmGKB (34 percent). In addition, there were 484 drugs linked to mutations extracted by EMU that were not listed in PharmGKB's variant annotations. These drugs were linked to 1,279 genes. From a qualitative perspective, of the seven paclitaxel-related citations, all but one were deemed relevant and revealed three genes related with direct effects, two of which (PIK3CA and TP53) were not genetically annotated in PharmGKB. Analogously, PharmGKB has no genetic information on mitoxantrone, but ten citations were identified by EMU as being mutationally relevant, of which two were related with effect.
DISCUSSION: Select EMU-only genes were inspected to confirm they included some relevant mutational drug-gene relationships. From this inspection, genes related to clozapine and paclitaxel were easily identified as needing further curation. With a manageable amount of citations per drug, evaluation of the majority of the citations should be a relatively quick process. From a previous evaluation, EMU's precision in extracting correct mutational information and genes was about .70. In terms of recall, EMU misses some genes found in PharmGKB due to its inability to find non-standard gene names in the text.

Using drug class information from NDF-RT, we were able to determine that most drugs lacking genetic annotation are antimicrobials.
CONCLUSION: The majority of citations in which EMU extracted mutations are relevant to curation. We have shown that the use of EMU can provide new citations for current PharmGKB curations and provides citations for drugs not genetically annotated in PharmGKB.

OVERVIEW

We built a data-integrated system to be query-able using the SPARQL query language. This system was built to be able to link drugs to mutations to protein structure and function. We integrated several databases: NDF-RT, UniProt, Reactome, BioCyc, and Gene Ontology. We also included proteins linked to domains provided by Dr. Maricel Kann's lab at University of Maryland, Baltimore County. Furthermore, we compiled a list of PubMed citations that were linked to drugs via UMLS concepts and to mutations via MeSH="mutation" and ran this list through a high-throughput text-mining program called Extractor of MUtations (EMU, developed at Dr. Kann's lab) that extracts mutational information and genes from text. With this system, we were able to query mass amounts of data for post-processing. We also examined mutational gene-drug pairs extracted by EMU and compared to the mutational gene-drug pairs already annotated in PharmGKB. Future work will look to compare drugs linked to genes via literature (through extracted mutations) to drugs linked to genes via domains.

CONTRIBUTION

This project involved several steps: acquisition of datasets and EMU output, changing datasets into RDF triples for storing in Virtuoso, integrating data resources into a single graph, creation of efficient queries, and abstract/manuscript preparation and authorship. I first gathered and prepared all needed datasets (EMU output on all PubMed citations linked to drugs, protein linked to domains table). I then converted these datasets to RDF triples, so all data had the form of <*subject* > <*predicate*> <*object*>. Once all data was stored as triples (including database triples or RDF provided by Brian Kirk were also stored in Virtuoso), Brian and I worked to integrate all of our data pieces together. We went through several graph revisions and different integration strategies before settling to use UniProt data as a connecting node between multiple protein/gene resources. Once the working system was in place, I worked to create efficient queries in order to traverse and obtain results within reasonable amount of time (under ten minutes). Creating queries proved more difficult than initially thought due to the complexity of the data and links in our system. The solution was to go through the system one-way and use grounded queries when possible (start query from a given drug, domain, pathway, etc.) My final work on this project has been on writing abstract for submission to the Pacific Symposium on Biocomputing workshop on Mining the Pharmacogenomics Literature. With this abstract, I will also be working on the writing of actual manuscript.

LIMITATIONS

Our system has several limitations, some of which being easily resolvable in future work. These limitations include technical and system limitations. The technical limitations revolve around the

utilization of UniProt, interoperability between data resources, and efficient query strategies. These technical limitations can easily be dealt with in future work. The system limitations, however, are more complex. These limitations involve not having concrete mutation-drug relationships and not having the knowledge depth needed to make valid hypotheses.

The technical limitations of our system involved the integration of UniProt and interoperability between two pathway resources. However for both limitations, we were able to find relatively simple solutions. UniProt was integrated into our system due to its sheer amount of protein knowledge and links to other databases. This greatly helped with integrating the other databases into our system. However, the UniProt RDF distribution is massive and contains the entirety of UniProt. UniProt was stored in a 100 gigabyte file and contained about 1.5 billion triples. The other pieces of our system contained in total several million triples, so the majority of our actual stored data was from UniProt. However, the majority of the UniProt data was of no use to our system as we were only concerned with human proteins and genes. With all this additional data, it became difficult to query effectively, and we had to adapt and create queries that traversed our system one-way and started with anchors (drug class, pathway, domain, etc). The challenge of interoperability involved the two pathway databases Reactome and BioCyc , and these two resources were unable to link to one another. However, we were able to use UniProt to go between them. In the future, we plan to modify our data to only include the human portion of UniProt.

The system limitations involve the quality of our mutation-drug links and the depth of the entire system. Due to time constraints, the mutations extracted by EMU were unable to be curated. Due to this lack of curation, we were only able to say a drug was associated with a mutation, and we could not say if a drug caused or interacted with a mutation. This lack of depth hindered our system in the kinds of questions we were able to ask. Along with this lack of mutational depth, the system lacked depth when linking clinical drug properties to protein structure and function. Due to the complexity of biological systems and interactions, it was difficult for us to be able to say “drug A is associated with mutation B, which has a gene involved in pathway C. Therefore, is there a relationship between drug A and another drug associated with a mutation in pathway C?” We don’t know how drug A is involved with mutation B, and we don’t know how gene affects pathway C. We also do not know which part of the pathway the two genes are involved in, which can be in two completely different subpathways. This knowledge-depth limitation is partially due to integrated databases and lack of time for the entire project (only the summer). We’d like to address these issues in the future.