Presented at the 2011 PSB Workshop on Mining the Pharmacogenomics Literature (electronic proceedings).

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.