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

Exploring automatic approaches to extracting pharmacogenomic information from the biomedical literature.

Bastien Rance, Dina Demner-Fushman, Thomas C. Rindflesch, Olivier Bodenreider
National Library of Medicine, Bethesda, MD 20894, USA

BACKGROUND: One aspect of personalized medicine is better adaptation of therapeutic drugs to the specific situation of a given patient, part of which is determined by his or her unique genetic make-up. Pharmacogenomics attempts to assess the influence of genetic variation on drug response. The biomedical literature is the primary vehicle for reporting the association between gene variants and drugs. This information is generally extracted from text and curated manually in order to create reference knowledge bases, such as PharmGKB, which offers several levels of curation for the articles it references (in depth, curated and non-curated). Information extraction can be also automated using natural language processing (NLP) tools. Here, we explore two NLP approaches (one recall- and the other precision-oriented) to extracting pharmacogenomic information from PubMed/MEDLINE citations, which we compare to PharmGKB.

METHODS: On the one hand, we extract drug-gene associations in a given article using MetaMap to identify drugs and links provided by the Entrez system between PubMed (articles) and Entrez Gene (genes). The second approach leverages SemRep for extracting named relations between genes and drugs from the title and abstract of articles. The two approaches were applied to a corpus of 47,315 articles indexed with "mutation" and exhibiting at least one drug name in the period 2001-2008. Drugs are restricted to ingredients in RxNorm and genes to human genes. Articles selected by our approaches are compared to articles listed as evidence by PharmGKB's curators. We reviewed some of the articles selected by our approaches for the drug warfarin.

RESULTS: Number of article-drug-gene associations identified in our corpus: 23,264 (MetaMap), 6504 (SemRep) and 1340 (PharmGKB). Proportion of reference articles in PharmGKB (N=470) also identified by our approaches: 6% (MetaMap), 2% (SemRep) overall; 0.9% (MetaMap), 0.3% (SemRep) for the in-depth curated variants (196 articles).

The two genes whose variants are associated with warfarin (in-depth curation in PharmGKB) are VKORC1 and CYP2C9. These associations are also identified by MetaMap and SemRep. The gene CYP2C19 identified by our approaches is discussed in the context of warfarin, albeit not positively linked to this drug.

DISCUSSION: The automatic approaches only identified a fraction of the reference articles in PharmGKB. A failure analysis is necessary to elucidate why this is the case. In contrast, the automatic methods identified drug-gene associations for many drugs not currently curated in PharmGKB. SemRep contributes to identifying a smaller number of reference articles in PharmGKB than the approach based on MetaMap. However, unlike MetaMap, SemRep qualifies gene-drug associations with named relationships. The precision of the automatic methods remains to be evaluated. Anecdotal evidence suggests that automatic methods could help identify many relevant documents.

CONCLUSION: The two automatic methods presented can help support the manual curation of pharmacogenomic knowledge. They allow high-throughput processing of the biomedical literature, helping to expand the scope of PharmGKB to additional drugs and to prioritize the manual curation.