MINING THE PHARMACOGENOMICS LITERATURE

KEVIN BRETONNEL COHEN
University of Colorado Denver

YAEL GARTEN
Stanford University

UDO HAHN
Friedrich Schiller University of Jena

NIGAM H. SHAH
Stanford University

Abstract:

The promise of pharmacogenomics for individualized medicine is on the crest of realization as a result of advances that allow us to predict beneficial, non-beneficial, and deleterious drugs for specific individuals based on aspects of both the individual and the drug. In spite of these advances, information management in this field relies on fairly traditional means, which do not scale to the available volume of full text publications. The aim of this workshop is to bring together researchers working on the automatic or semi-automatic extraction of relationships between biomedical entities from the pharmacogenomic research literature. The workshop will focus particularly on methods for the extraction of genotype-phenotype, genotype-drug, and phenotype-drug relationships and the use of the relationships for advancing pharmacogenomic research. Efforts aimed at creating benchmark corpora as well as comparative evaluation of existing relationship extraction methods are of special interest.

Goals of the workshop:

The promise that pharmacogenomic holds for individualized medicine may be on the crest of realization due to technical advances such as large genotyping arrays and analytical advances that allow us to predict beneficial, non-beneficial, and deleterious drugs for specific individuals based on salient features of both the individual and the drug.

However, information management in this field relies on fairly traditional means, e.g. curated databases, which do not scale to (1) the rapid expansion of the pharmacogenomics literature in recent years and (2) the increasingly available volume of full text publications, which contain more specific and (potentially) informative facts than Medline abstracts. Hence, although there is a large demand and significant utility of text analytics to the study of pharmacogenomics, its potential is not fully realized; in part because the work to date has failed to bridge the two distinct worlds—that of (bench) molecular biology and that of (clinically oriented) pharmacology—and because the developers of text analytics are not fully aware of this challenging field.

Last year's workshop (Genotype-Phenotype-Drug relationship extraction from text) examined the current state-of–the-art and reported ongoing research of labs already involved in this area of research. The steady stream of work on extracting interactions from text, the increasing attention in the Semantic Web towards capturing facts as "nano-publications" (individual assertions that are attributable to authors and traceable in their publications), and representing scientific discourse in a structured manner, all indicate that the time seems to be ripe for research that goes even beyond the mere extraction of explicitly stated knowledge in documents, to linking text-mined and database data through formal reasoning to uncover implicit and in some sense "new" knowledge. In order to advance this agenda, it is essential that existing relationship extraction methods be compared to one another and that a community-wide sharable benchmark corpus emerges against which such efforts can be compared. The goal of the workshop is to utilize a corpus put forth by PharmGKB to compare different relationship extraction methods and the corresponding "new" knowledge discovery those methods might drive.

This workshop aims to address the gap in coverage of text mining for pharmacogenomics. The technical area of the workshop will particularly focus on extraction of genotype-phenotype-drug relationships. Work on named entity recognition (e.g. gene taggers) would not be considered for inclusion. Approaches that combine text-mining and knowledge-based systems are of special interest. We invite researchers working on text mining and reasoning to submit applications of their research efforts to the area of pharmacogenomics, and particularly genotype-phenotype-drug relationships. Topics solicited include:

- Relation extraction between genotypes, phenotypes, and drugs, and other semantic classes relevant to pharmacogenomics
- Corpus development for pharmacogenomics text mining
- Associating gene variants (mutations, alleles, rs/ss numbers) to the associated gene name
- Work on the corpus of documents linked to by PharmGKB
- Reasoning systems applied over the PharmGKB knowledge base

We anticipate that this year's workshop will build on the success of last year's workshop and seed a community around the shared goal of computationally collecting and distributing pharmacogenomics knowledge.