PSB 2010 Special Workshop

GPD-Rxn Workshop: Genotype-Phenotype-Drug Relationship Extraction from Text

Thursday, January 7, 2010, 2:00pm-5:30pm

Plaza III, the Fairmont Orchid, HI, USA

Co-chairs: Adrien Coulet [1], Nigam Shah [1], Larry Hunter [2], Chitta Baral [3], Russ Altman [1]

[1]Stanford University
[2]University of Colorado School of Medicine
[3]Arizona State University

The GPD-Rxn Worksop is sponsored by the
National Center for Biomedical Ontology

Schedule

Thursday, January 7, 2010, 2:00pm-5:30pm
Plaza III

Co-chairs: Adrien Coulet, Nigam Shah, Larry Hunter, Chitta Baral, Russ Altman

2:00-2:10 Workshop Introduction

2:10-3:10 First Session: Corpus Selection and Entity Recognition
4 talks of 15 minutes (including questions)

Tool performance and semantic type distribution for genotype-phenotype-drug relationship
Kevin Bretonnel Cohen, HL Johnson, K Verspoor, C Roeder, and L Hunter

A Dictionary Approach to the Identification of Small Molecules and Drugs in Free Text
Kristina M Hettne, RH Stierum, MJ Schuemie, PJM Hendriksen, BJA Schijvenaars, EM van Mulligen, J Kleinjans, and J Kors

Pharmacokinetics Ontology Development and Pharmacokinetics Parameter Text Mining
Zhiping Wang, SK Quinney, and L Li / Yuming Zhao, W Zhiping, and L Li

Taking a SNPShot of PubMed - a repository of genetic variants and their drug response phenotypes
Jörg Hakenberg, D Voronov, VH Nguyen, S Liang, B Lumpkin, S Anwar, R Leaman, LN Tari, and C Baral

3:10-3:30 Break

3:30-4:00 Keynote Lecture
Christopher Manning – Natural Language Processing Group, Stanford University

4:00-5:00 Second Session: Relationship Extraction and Comparison
4 talks of 15 minutes (including questions)

Linking a Gene/Protein-Focused Relation Extractor to the Pharmacogenomic Domain
Ekaterina Buyko, and U Hahn

Extraction and Integration of Pharmacogenomic Relationships from Medline Abstracts
Adrien Coulet, N Shah, and RB Altman

Identifying Novel Drug Indications using Ontology-based Mashups of Disease, Phenotype and Drug Information
Cartic Ramakrishnan, K Kundu, A Qu, A Jegga, EK Neumann, and BJ Aronow

Towards mining drug-phenotype data: context based calculation of probabilistic semantic similarity
Lixia Yao, J. Evans, and A. Rzhetsky

5:00-5:30 Open Discussion with Panelists

TOOL PERFORMANCE AND SEMANTIC TYPE DISTRIBUTION FOR GENOTYPE-PHENOTYPE-DRUG RELATIONSHIP MINING FROM TEXT

K. BRETONNEL COHEN, HELEN L. JOHNSON, KARIN VERSPOOR, CHRISTOPHE ROEDER, AND LAWRENCE HUNTER

Biomedical Text Mining Group, Center for Computational Pharmacology, University of Colorado School of Medicine

Mining data on genotype-phenotype-drug (GPD) relationships requires the ability to locate mentions of entities of interest within free text. The entities of interest for this task include gene names, mutations, diseases, and drugs. We evaluated the performance of three gene name taggers, a mutation tagger, two drug taggers, and a disease tagger on relevant texts. Finding GPD information in particular may require the ability to process the full text of journal articles. However, to date almost all research in biomedical text mining has focused on the abstracts of journal articles. For this reason, we examined the distributions of the entities of interest in these two types of text and compared the performance of the tools on abstracts and on article bodies within a benchmark corpus that we are preparing. The corpus contains both linguistic annotations, such as syntactic parses, and biological annotations, such as genes and Gene Ontology terms. Gene name taggers were found to perform substantially better in abstracts than in article bodies, with F-measure as much as 14 points higher for abstracts depending on the combination of tagger and statistical model. Mutation taggers were found to perform differently, but at high levels (above .95 F-measure), in both text types. Both of the dictionary-based drug taggers that we evaluated required substantial editing of the input dictionary to avoid rampant false positives. We found marked distributional differences in the entity types. Gene mentions were present at a higher density in abstracts than in article bodies. In contrast, there is independent reason to suspect that more gene types are present in the article bodies. Mutation mentions were present far more frequently in article bodies than in abstracts, with most articles that mentioned mutations not mentioning them in the abstract at all. Disease mentions also were present more frequently in article bodies than in abstracts, with almost all bodies mentioning a disease but only two thirds of abstracts mentioning one. The average number of mentions of drugs was quite different in abstracts and article bodies, though the density of mentions was similar in both. Thus, the distributional findings on mutations, diseases, and drugs suggests that the ability to process full-text articles will be crucial to GPD relation mining, but the differential performance of gene mention taggers demonstrates that this will present some technical challenges. The incidence of all of the entity types of interest in our corpus suggests that it might be a suitable part of a benchmark corpus for investigating genotype-phenotype-drug relations.

Acknowledgments

We thank Robert Leaman and Graciela Gonzalez for giving us access to a pre-release version of BANNER Disease.

A DICTIONARY APPROACH TO THE IDENTIFICATION OF SMALL MOLECULES AND DRUGS IN FREE TEXT

KRISTINA M HETTNE[1,2,3], ROB H STIERUM[3,4], MARTIJN J SCHUEMIE[2], PETER JM HENDRIKSEN[5], BOB JA SCHIJVENAARS[6],
ERIK M VAN MULLIGEN[2], JOS KLEINJANS[1,3] AND JAN A KORS[2]

[1] *Department of Health Risk Analysis and Toxicology, Maastricht University, Maastricht, The Netherlands*
[2] *Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands*
[3] *Toxicoinformatics, Netherlands Toxicogenomics Centre, Maastricht, The Netherlands*
[4] *Business unit Biosciences, Physiological Genomics, TNO Quality of Life, Zeist, The Netherlands*
[5] *Safety and Health, RIKILT Institute of Food Safety, Wageningen, The Netherlands*
[6] *Collexis Holdings Inc, Columbia SC, USA*

The identification of biomedical terms in natural language is essential for finding associations and directed relations between biomedical concepts such as genes, proteins, drugs and diseases. From the scientific community, much effort has been spent on the correct identification of gene and protein names in text, while less effort has been spent on the correct identification of chemical names [1]. Dictionary-based term identification has the power to recognize the diverse representation of chemical information in the literature and map the chemicals to their database identifiers. We developed a dictionary for the identification of small molecules and drugs in text, combining information from the Unified Medical Language System metathesaurus, Medical Subject Headings, the ChEBI ontology, DrugBank, KEGG compound and KEGG drug, the Human Metabolome database, and ChemIDplus. The PubChem database was initially included but later removed from the combined dictionary due to poor performance. Rule-based term filtering, manual review of highly frequent terms, and disambiguation rules were applied. We evaluated the combined dictionary and the individual dictionaries derived from each resource on an annotated corpus [2], and conclude the following: (1) each of the different processing steps increases precision with a minor loss of recall; (2) the overall performance of the combined dictionary is average (precision 0.67, recall 0.40 (0.80 for trivial names); (3) the combined dictionary performs better than the dictionary in the chemical recognizer OSCAR3; (4) the performance of a dictionary based on ChemIDplus alone is comparable to the performance of the combined dictionary. The combined dictionary is freely available as an XML file in Simple Knowledge Organization System format from http://www.biosemantics.org/chemlist and has been incorporated into the text mining and knowledge discovery tool Anni [3], available at http://www.biosemantics.org/anni.

[1] Erhardt, R.A.A., Schneider, R. and Blaschke, C. (2006) Status of text-mining techniques applied to biomedical text, Drug Discov Today, 11, 315-325.
[2] Kolarik, C., Klinger, R., Friedrich, C.M., Hofmann-Apitius, M. and Fluck, J. (2008) Chemical names: terminological resources and corpora annotation., Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference).
[3] Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biology* 2008 Jun 12, 9(6):R96

PHARMACOKINETICS ONTOLOGY DEVELOPMENT

ZHIPING WANG, SARA K. QUINNEY AND LANG LI

Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN, 46202.

Introduction

Model based drug development has becoming a main theme in academia, industry, and regulation agencies. It facilitates the translational research all the way from the *in vitro* drug screen, to animal model, and to clinical studies. The pharmacokinetics (PK) model is the central component of this modeling system, and it predicts drug exposure. The PK model construction relies on the data collected from its published studies. One significant hurdle is the non-standard expressions on drug names, PK terminology, PK study description and etc. We develop a PK ontology. It provides a standardized PK knowledge structure and annotations.

Methods

The PK ontology is primarily composed of three components: pharmacokinetics parameters and study designs, metabolic and transportation enzymes, and environmental effects. Pharmacokinetics parameters and study designs are divided into two categories: *in vitro* and *in vivo*. They were manually assembled from text books, publications, and our local experts. Metabolic and transportation enzymes annotations were collected from multiple public domain databases. The environmental effects are composed of disease populations, physiological conditions, and demographic variables. They were all collected from multiple public domain databases. This PK ontology was organized and presented with Protégé.

Applications

To demonstration the application of this PK ontology, three PK studies were tested: a single dosing PK clinical study, a steady state PK and pharmacogenetic clinical study, and *in vitro* PK study. We showed that PK ontology sufficiently annotated these studies.

Discussion

In this research, we presented the development and application of a PK ontology for drug study annotations. To our knowledge, this is the first PK ontology attempt. This ontology can be extended or linked to other potential ontology for pharmacodynamics studies, which have much diverse phenotypic data and genomic data. The ontology will be public available through BioPortal (National Center for Biomedical Ontology, NCBO).

PHARMACOKINETICS PARAMETER TEXT MINING

YUMING ZHAO, WANG ZHIPING, AND LANG LI

Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, IN, 46202.

Background: Model based drug development has become a force in both pharmaceutical industry and FDA. It integrates both pharmacokinetics (PK) and pharmacodynamics (PD) mechanisms from *in vitro* to *in vivo*, and provides quantitative guidance to clinical trial designs and decision making. To fulfill its modeling potential in large scale, there is an enormous need for PK/PD parameter text mining.

Method: A text mining system was developed for PK data extraction. It is composed of 7 steps: information retrieval (IR), sentence boundary detecting (SBD), tokenization, part of speech tagging (POS), shallow parsing (SP), information extraction (IE), and mixture model for outlier detection. Both SBD and POS steps were implemented with Cognitive Computation Group (CCG) tools, “Sentence Segmentation” and the “SNoW-based Part of Speech Tagger”. A set of knowledge based finite state automata (FSA) was developed to identify “DRUG-PK” and “PK data” phrase in the SP module. In the IE step, a set of classification rules were established to extract PK parameters and their numerical values. In the final step, a mixture model was established to identify outliers (potential false positives) among mined data. Our text mining system was evaluated by extracting clearance data from PubMed abstracts for 8 drugs: efavirenz, ticlopidine, rifampin, dextromethorphan, quinidine, dexamethasone, midazolam, and ketoconazole, which covered three major druf metabolism pathways (CYP3A, CYP2B6 and CYP2D6). These three pathways are responsible for more than 60% drugs on the market. We compared our current method, which is named as Version II, to Version I (Wang *et al.* J Biomed Inform, 2009, 42, 726-735), and to the manually accumulated PK parameter database DiDB (http://www.druginteractioninfo.org/). Version I didn’t have NLP and FSA technologies. Among all 8 drugs, positive information content (PIC, the number of mined clearance data), and precision were compared. The recall rate was bench marked by the midazolam alone.

Results: PIC of Version II was almost the same as Version I, and they were higher than DiDB. Version II had noticeable improvement on precision over Version I among 4 out 8 drugs. Their performances stayed the same for the others. The average precision improvement of the 8 drugs was 301.38%. The recall rate for midazolam was the same, 92%, in both version I and II.

Conclusion: Text mining approaches are more powerful than manually data accumulation in getting relevant PK data, and they have high recall rate. NLP improves the precision in PK data text mining.

TAKING A SNPSHOT OF PUBMED - A REPOSITORY OF GENETIC VARIANTS AND THEIR DRUG RESPONSE PHENOTYPES

JÖRG HAKENBERG, DMITRY VORONOV, VÕ HÁ NGUYÊN, SHANSHAN LIANG, BARRY LUMPKIN, SAADAT ANWAR, ROBERT LEAMAN, LUIS NG TARI, AND CHITTA BARAL

BioAI Lab, Department of Computer Science and Engineering
Arizona State University, Tempe, AZ 85281-8809, USA

Motivation:
Genetic factors determine differences in pharmacokinetics, drug efficacy, and adverse drug responses between individuals and sub-populations. Wrong dosages of drugs can lead to severe adverse drug reactions in individuals whose drug metabolism drastically differs from the 'assumed average'. Poor or ultra-rapid metabolizers of the drug have one or even two inefficient/non-functional copies of the gene coding for the drug-metabolizing enzyme, or a variant that leads to highly efficient enzymes, respectively. Databases such as PharmGKB and DrugBank are excellent sources of pharmacogenetic information on enzymes, gene variants, and drug response affected by changes in enzymatic activity. They build on manual annotation of publications (we found more than 100,000 publications to be relevant), and might thus be 'incomplete' for a long time to come.

Approach:
In this paper, we present an approach to automatically populate a repository of information on genetic variants, their drug response phenotypes, occurrence in sub-populations, and associations with disease. We mine textual data from PubMed abstracts to discover such genotype-phenotype associations, focusing on SNPs that can be associated with variations in drug response. The overall repository covers all relations found between genes, variants, alleles/haplotypes, drugs, diseases, populations, and frequencies. We cross-reference these data to EntrezGene, PharmGKB, and DrugBank, for single entities and relations.

Results:
We show that the performance regarding entity recognition and relation extraction yields high F1-scores of around 85-92% for the major entity types (gene, drug, disease), and around 79-83% for relations involving these types. For other, "simpler" categories, F1-scores are between 90 and 100%. We compare the data in our repository to PharmGKB and DrugBank, to estimate the coverage of the 55,000 abstracts currently indexed by SNPshot. This reveals a coverage of 90% of the gene-drug associations in DrugBank and 64% in PharmGKB. We can also recover about 68% of the gene-variant mappings by searching these abstracts only.

Availability:
An interface to search SNPShot by genes and drugs can be found at http://bioai4core.fulton.asu.edu/SNPshot. We are also integrating the SNPShot data with a platform for collaborative curation, in which users can search PubMed or browse SNPShot by extracted information, and then add manual annotations. This will help correcting erroneously extracted information as well as building better models for extraction because each annotation will provide a training example to be used when training models for future use.

Keywords: genetic variation; drug response; drug metabolism; pharmacogenetics; text mining

LINKING A GENE/PROTEIN-FOCUSED RELATION EXTRACTOR TO THE PHARMACOGENOMIC DOMAIN

EKATERINA BUYKO AND UDO HAHN

Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany

http://www.julielab.de

We have developed **JREX**, a relation extraction system focused on genes/proteins and events they take part such as `Localization`, `Binding`, `Gene expression`, `Transcription`, `Protein catabolism`, `Phosphorylation`, positive or negative or neutral `Regulation` (Buyko et al., 2009). This extractor, originally built for the *BioNLP'09 Shared Task on Event Extraction*,[1] was recently linked to the pharmacogenomic domain. We extended the definitions for selected types of events, i.e. positive, negative and neutral `Regulation`, by expanding the set of allowed argument and adapted the feature generation system. Rather than allowing arguments to refer to genes and proteins (and related events) only, we enhanced this criterion such that at least one argument of the identified event was allowed to be a drug or a pharmaceutical ingredient (as regulator).

We started our experiment on a set of approximately 15,000 Medline abstracts MeSH-term-selected for the *osteoporosis* disease as one of the most frequent age-related diseases.[2] We cut down this collection to 2,000 abstracts containing co-occurrences of relevant protein and drug/pharmaceutical ingredient names based on entries in the *'Orange Book'*[3] extended with corresponding term variants from the MeSH.

Using **JREX** we then identified about 5,000 events from which 120 contained drugs and pharmaceutical ingredients as arguments with a causal role in regulation events. One expert biologist classified the extracted events according to the qualitative criteria of correctness and novelty. Our results show that half of these 120 relations are biologically invalid, while almost all of the other half are correctly identified causal relations between pharmaceutical ingredients and proteins that can even be classified as new ones (i.e., the article deals with this relation as a novel finding). Furthermore, we assessed term co-occurrences of one frequently used drug for *osteoporosis*, viz. *alendronate*, and molecular events (rather than the much tighter relational encoding). The results show that we achieve a significant increase in recall for causal relations by considering simple co-occurrences. On a larger scale, we plan to enhance **JREX**'s gene/protein-focused information extraction functionality with an additional coverage of diseases and drugs.

References

Ekaterina Buyko, Erik Faessler, Joachim Wermter and Udo Hahn. Event extraction from trimmed dependency graphs. In *Proceedings of the NAACL HLT 2009 Workshop on BioNLP 2009 – Companion Volume: Shared Task on Event Extraction.*, Boulder Colorado, USA, June 2009, pp.19-27

[1] `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/`

[2] This reflects our recent involvement in **JENAGE**, a large research initiative at Friedrich-Schiller-Universität Jena, which is devoted to mild stressors and their influence on aging (`cf. http://www.jenage.org/`).

[3] `http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm`

EXTRACTION AND INTEGRATION OF PHARMACOGENOMIC RELATIONS FROM MEDLINE ABSTRACTS

ADRIEN COULET[1,2], NIGAM SHAH[2], RUSS B. ALTMAN[1,3]

[1]Department of Genetics, [2] Department of Medicine, [3]Department of Bioengineering,
Stanford University, Stanford, CA 94305, USA

Background

Knowledge pieces that contribute to the understanding of pharmacogenomics can be represented by direct relations between *genes*, *drugs*, *phenotypes* but also by relations between entities relative to genes, drugs and phenotypes such as *gene polymorphism*, *drug dose* or *disease risk*. The approach presented in this work focuses on the extraction of such relations (or events) from scientific literature with the goal of facilitating the understanding of pharmacogenomics.

Approach

According to this goal, we describe an approach that improves existing methods by two key aspects: *(1)* by extracting relations between recognized entities (*genes*, *drugs*, *phenotypes* names) and also interaction between entities relative to recognized entities (*gene polymorphism* and *drug dose* for instances); *(2)* extracted relations are integrated in conformity with a precisely defined semantic. To extract relations, our approach makes use of a probabilistic natural language parser (the *Stanford Parser*) and of manually defined patterns. The grammatical structure of sentences, computed by the parser, is compared to patterns to identify relations to extract. Valid relations are represented as assertions of concepts and roles in Description Logics, what enables their comparison and their integration in a Knowledge Base.

Results

We conducted a preliminary experiment that consists in the extraction, from all Medline abstracts, of relations that involve *important* pharmacogenomic genes (*important* according to PharmGKB http://www.pharmgkb.org/search/annotatedGene/index.jsp). The first step of our approach led to the extraction of 32,637 raw relations which are secondly used to instantiate a Knowledge Base. A manual evaluation shows that a random subset of our relations presents 74% of true and complete relations, 88% of true completely or incompletely described relations), and 12% of false positives.

Conclusion

We adapted existing methods of relation extraction to pharmacogenomic relations and propose a scaling and necessary new approach that extracts and integrates relations between complex entities of this domain.

IDENTIFYING NOVEL DRUG INDICATIONS USING ONTOLOGY-BASED MASHUPS OF DISEASE, PHENOTYPE AND DRUG INFORMATION

CARTIC RAMAKRISHNAN[2,3] , KUNAL KUNDU[2,3], ANGELA QU[1,2,3], ANIL JEGGA[1,2,3] ERIC K NEUMANN[4], BRUCE J ARONOW[1,2,3]

[1]Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH, USA, [2]Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA, [3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA, [4]Clinical Semantics Group, Lexington, MA, USA

The re-positioning, re-purposing, and novel combination of established drugs for new applications is an important strategy for both pharmaceutical companies and next-generation personalized medicine. Hypothesizing and validating novel disease uses for drugs already approved for another disease can reduce cost, mitigate drug development risks, and provide a huge benefit for patients and physicians. Current approaches to accomplishing this include combinations of text mining, pathway reasoning, as well as clinical / preclinical serendipity. The potential for improving all of these approaches via integrated knowledge modeling and intelligent applications is great.

To approach this we have developed a knowledge and data model framework in order to connect therapeutic agents to diseases via mechanistic representations of entities and relationships that include therapeutic agents, diseases, phenotypes to genes, mRNAs, miRNAs, proteins, pathways, and entity-associated features (i.e. domains, binding sites, interactions, regulatory elements and mechanisms, and genetic variations and mutations). To do this, we have constructed a Systems Biology Disease Ontology (SBDO) in OWL to model and interconnect existing drug knowledge with disease mechanisms, genomics, phenotypes, and other biological contexts. We have used a semi-automated approach to the development of SBDO class structure and relationships and have extensively used multiple ontologies and vocabularies such as MeSH, FMA, GO Ontology, BIOPAX, OBO, Reactome, ICD9, MGI, NCI Metathesaurus, Sequence Ontology and UMLS using RDF-represented XREF methodologies. In addition, we have class and relationship feature-linked to data schemas of the diverse data sets aggregated from various pharmacological and biological domains. At our present level of normalization and aggregation we have established more than 31,000 classes, 111,000 xrefs, and 549 UMLS-defined relationships.

Based on an earlier version of this ontology, we analyzed and ranked several diseases based on their genetic or phenotypic associations with gene, pathway, and ontology-based relations to the phenotypes associated with other diseases for which therapeutics exist. This provided the ability to rapidly aggregate and inter-relate all mechanism-associated disease features as well as the ability to rank all therapeutic action associated drugs using graph network analysis approaches. Since our knowledge of underlying mechanisms remains relatively incomplete, we are exploiting available data relationships to extend mechanistic knowledge by utilizing evidence from phenotypic and gene associations. This aggregate set of knowledge was used to uncover potentially significant and novel relations.

Using the complex disease Systemic Lupus Erythematosus (SLE) as an example, a high-dimensional pharmacome-diseasome graph network was generated as RDF XML, and subjected to graph-theoretic proximity and connectivity analytic approaches to rank drugs versus the compendium of SLE-associated genes, pathways, and clinical features. Tamoxifen, a current candidate therapeutic for SLE, was the top-ranked drug. Other use cases that we are focusing on include specific cancers and chronic disease syndromes with and without known gene defects.

This early-stage demonstration highlights critical directions to follow that will enable translational pharmaco-therapeutic research. We would like to partner with other organizations and advance this approach with resources available to pharmaceutical companies. Uniform and scalable applications of Semantic Web technologies and methodologies to problems in drug R&D are becoming more wide-spread. These approaches are viewed to combine and improve data integration, knowledge representation, and analyses so that knowledge mining of drug action and disease mechanism relationships can be scaled significantly and applied in diverse challenging situations including the design and analysis of novel clinical trials and ultimately clinical practice. Further improvements in semantic representation of mechanistic relationships will provide a fertile basis for accelerated development of therapeutics.

TOWARDS MINING DRUG-PHENOTYPE DATA: CONTEXT BASED CALCULATION OF PROBABILISTIC SEMANTIC SIMILARITY

LIXIA YAO[1], JAMES EVANS[2,3], ANDREY RZHETSKY[3,4]

[1]*Department of Biomedical Informatics, Center for Computational Biology and Bioinformatics,*
Columbia University, New York, NY 10032, USA
[2]*Department of Sociology, University of Chicago, Chicago, IL 60637, USA*
[3]*Computation Institute, University of Chicago, Chicago, IL 60637, USA*
[4]*Department of Medicine, Department of Human Genetics, Institute for Genomics and Systems Biology,*
University of Chicago, Chicago, IL 60637, USA

Available clinical and biological databases contain valuable information for drug discovery. For example, electronic medical records can help identify unanticipated drug side effects and novel indications, and pharmacogenomic databases such as PharmGKB and DrugBank can be used to predict novel drug targets and combinatory drug therapies. To realize these possibilities, we must develop methods that automatically process large quantities of textual data from those databases to predict new knowledge.

We propose a statistical method to measure semantic similarity, which is expressed as the probability that a word can be substituted by a particular synonym in a domain-specific corpus. We implement this model using several English thesauri. Using the approach and resource, we will be able to define and calculate semantic similarity for biomedical concepts, especially diseases, symptoms and other human phenotypes. We also suggest a metric for evaluating the “fitness" of those thesauri relative to a biomedical corpus, which account for both topical coverage and precision.

Posters Related to the Topic of the Workshop
(…to see before the workshop!)

Towards mining drug-phenotype data: context-based calculation of probabilistic semantic similarity
Lixia Yao, James Evans, Andrey Rzhetsky

Automatic method for detecting pharmacogenomically "hot" drugs
Yael Garten, Nicholas P. Tatonetti, Russ B. Altman

Biological Document Clustering using a Gene-Gene Network
Hong-Woo Chun, Shinobu Okamoto, Yasunori Yamamoto, Atsuko Yamaguchi

A Dictionary Approach to the Identification of Small Molecules and Drugs in Free Text
Kristina M Hettne, Rob H Stierum, Martijn J Schuemie, Peter JM Hendriksen, Bob JA Schijvenaars, Erik M van Mulligen, Jos Kleinjans, Jan A Kors

Protein Interaction Databases: How Accurate Are They
Ziegler, Kirby; Frolkis, Alexandra; Wishart, David S.

Ontology Web Services for Semantic Applications
Patricia L. Whetzel, Nigam H. Shah, Natalya F. Noy, Clement Jonquet, Adrien Coulet, Nicholas Griffith, Cherie Youn, Michael Dorf and Mark A. Musen

Uncovering drug-drug interactions through synthesis and reasoning on pharmacokinetic pathways
Shanshan Liang, Luis Tari, Jörg Hakenberg, Chitta Baral

…and probably others.

Poster sessions are in the Salon 1 & Ballroom Courtyard, on

Wednesday, January 6
1:00-2:30 Poster Session (A-L, last name of the first author presents)

Thursday, January 7
12:15-2:00 Poster Session (M-Z, last name of the first author presents)