TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY

GRACIELA GONZALEZ

Department of Biomedical Informatics,
Arizona State University
Scottsdale, AZ 85259, USA
Email: ggonzalez@asu.edu

KEVIN BRETONNEL COHEN

Computational Bioscience Program
U. Colorado School of Medicine
Aurora, CO
Email: kevin.cohen@gmail.com

MARICEL G. KANN

Department of Biological Science
University of Maryland, Baltimore County
Baltimore, MD, 21250, USA.
Email: wspc@wspc.com

CASEY S. GREENE

Department of Genetics
Geisel School of Medicine at Dartmouth
Hanover, NH 03755, USA
Email: Casey.S.Greene@dartmouth.edu

ROBERT LEAMAN

National Center for Biotechnology Information
Bethesda, MD 20894, USA
Email: robert.leaman@nih.gov

UDO HAHN

Friedricdh Schiller University Jena
Language and Information Engineering Lab
Jena, Germany
Email: Udo.Hahn@uni-jena.de

NIGAM SHAH

Stanford Center for
Biomedical Informatics Research
Stanford, CA 94305, USA
Email: nigam@stanford.edu

JIEPING YE
Computer Science and Engineering,
Arizona State University, Tempe, AZ 85287
Email: jieping.ye@asu.edu

The biggest challenge for text and data mining is to truly impact the biomedical discovery process, enabling scientists to generate novel hypothesis to address the most crucial questions. Among a number of worthy submissions, we have selected six papers that exemplify advances in text and data mining methods that have a demonstrated impact on a wide range of applications. Work presented in this session includes data mining techniques applied to the discovery of 3-way genetic interactions and to the analysis of genetic data in the context of electronic medical records (EMRs), as well as an integrative approach that combines data from genetic (SNP) and transcriptomic (microarray) sources for clinical prediction. Text mining advances include a classification method to determine whether a published article contains pharmacological experiments relevant to drug-drug interactions,

a fine-grained text mining approach for detecting the catalytic sites in proteins in the biomedical literature, and a method for automatically extending a taxonomy of health-related terms to integrate consumer-friendly synonyms for medical terminologies.

1. Introduction

The explosion of genomic data available to researchers from countless gene expression and sequencing experiments, coupled with the abundance of knowledge in the published literature and curated databases, fuels the need for novel and transformative methods for knowledge extraction, visualization, and analysis that take advantage of all of these sources to elicit new and meaningful hypotheses. The biggest challenge for text and data mining is to truly impact the biomedical discovery process, enabling scientists to generate novel hypothesis to address the most crucial questions. Formulation of a flexible and general approach for integrating heterogeneous data and knowledge sources for discovery is elusive and highly dependent upon the specific underlying scientific question. The true impact of text and data mining is only realized if it goes beyond a focus on the methods for extraction and storage, and into the true impact they can have on enabling understanding of the molecular underpinnings of biological processes.

This session seeks to bring together researchers with a strong text or data mining background who are collaborating with bench scientists for the deployment of integrative approaches in translational bioinformatics. It serves as a unique forum to discuss novel approaches to text and data mining methods that respond to specific scientific questions, enabling predictions that integrate a variety of data sources and can potentially impact scientific discovery.

In order to find the optimal way to integrate relevant information that will help translational and clinical researchers pinpoint novel findings, a thorough understanding of the decision process through which active researchers vet the discoveries proposed by automated systems is required. However, very little is presently known about how scientists actually interpret this information. Cohen and Hersh argue that the "major challenge of biomedical text mining over the next 5-10 years is to make these systems useful to biomedical researchers" [1]. Langley notes the tendency for such tools to be developed for the use of professional data-miners rather than active researchers, and argues for the development of discovery systems with a greater degree of user interactivity [2].

There have been increased efforts to develop such systems and approaches, but there is no single place to present them. This session attracted cross-discipline collaborators with focused applications of discovery and prediction methods. Given the ever increasing deluge of data and knowledge that overwhelms bench scientists around the world; interest in such systems will only increase over time. Some examples of topics of interest to this session include novel approaches that integrate empirical data with knowledge extracted from the literature, curated databases or ontologies to perform discovery-related tasks such as:

- Gene prioritization
- Binding site prediction
- Gene/protein function prediction,
- Prediction of associations (protein-protein, gene-drug, gene-disease, drug-drug)
- Pathway generation or validation

for translational applications such as pharmacogenomics, genome-phenome validation, or detection, diagnostic and prognosis of disease.

2. Challenges

Improving text and data mining methods for any task requires careful consideration and evaluation. The biomedical domain presents specific challenges given the diversity, complexity and volume of the information being mined. This section presents a brief overview of the fundamental challenges faced by researchers in these areas.

2.1. *Text Mining*

Although in general there are challenges such as summarization and question answering, for the type of applications focus of this session, two text mining tasks seem to be specifically relevant: named entity recognition and association extraction.

Named entity recognition (NER) is the problem of finding references to entities (mentions) such as genes, proteins, diseases, drugs, adverse reactions, or organisms in natural language text, and tagging them with their location and type. NER is also referred to as "entity tagging". This is a basic building block for all other extraction tasks. While there has been significant progress into named entity recognition in the biomedical domain, research has been primarily focused on genes and proteins. Attempts to recognize other entities of interest have concentrated on dictionary matching or statistical approaches. Machine-learning based systems overcome this limitation to a certain extent, given it is possible to retrain such systems to recognize different entity classes. Retraining requires, however, considerable effort in annotation to create a suitable corpus for training the engine, as well as some feature analysis.

Tagging specific entities is of interest as a fundamental step towards the true goal: extracting true associations between terms, such as genes and diseases. Information extraction (IE) from the biomedical literature is usually developed around the extraction of such relationships of interest from text. A typical architecture is composed of special-purpose programs that perform a pipeline of processing modules, including sentence splitters, tokenizers, named entity recognizers, shallow or deep syntactic parsers, and finally extraction based on a collection of patterns. Such systems are usually file-based, so large amounts of processed data can be passed from one module to the next. Relational databases would play a limited role at the end of the extraction pipeline to store the extracted relationships.

This session includes specialized examples of entity recognition and association extraction, showing the trend towards finer granularity in the type of information needed for meaningful applications of text mining for biomedical discovery, requiring a tighter collaboration between the text mining community and domain experts.

2.2. *Data Mining*

In 2006, a paper in the International Journal of Information Technology & Decision Making explored “10 Challenging Problems in Data Mining Research” [3], based on the replies of 14 experts (organizers of the most prestigious Data Mining conferences). It is interesting to note that not only “Data mining for biological and environmental problems” is listed specifically as one of these challenging problems, but that 8 out of the 9 other challenges apply specifically to biomedical data, namely:

- Scaling up for high dimensional data and high speed data streams
- Mining sequence data and time series data
- Mining complex knowledge from complex data
- Data mining in a network setting
- Distributed data mining and mining multi-agent data
- Data Mining process-related problems
- Security, privacy and data integrity
- Dealing with non-static, unbalanced and cost-sensitive data

Of particular interest to this session is the recognized need, when mining complex data, ”for integrating data mining and knowledge inference” and to “to incorporate background knowledge into data mining”. Included in this session are works that address precisely these aspects.

3. Overview of Contributions

Holzinger et al present an integrative approach, ATHENA, used to combine data from genetic (SNP) and transcriptomic (microarray) sources to predict a clinically important feature (HDL-C level). The combined data are capable of predicting HDL-C levels better than either of the individual data sources. Methods capable of connecting measurements of the genome to measurements of transcript and protein abundance for prediction of a clinically relevant phenotype are expected to play a key role in precision medicine.

Hu et al explores the application of the statistical epistasis networks (SEN) approach as filters in the discovery of 3-way genetic interactions. Genetic epistasis is considered an important factor that is related to the etiology of complex diseases. Exhaustive search for high-order interaction is unrealistic due to the large data volume. The authors show that SEN can significantly reduce the number of candidates that need to be considered in a high-order interaction model with improved accuracy.

Kolchinski et al describe a document triage task (binary text classification) for biomedical (Pubmed) articles to determine whether the article contains pharmacological experiments relevant to drug-drug interactions. This joint work between a BioNLP lab (Rocha’s) and a lab doing research in pharmacokinetics (Li’s) exemplifies the type of collaborations likely to result in fruitful advances in biomedical discoveries. The approaches used are variations of approaches known to perform well on similar tasks. The sort of dimensionality reduction and feature transforms performed in the paper are not used as often in BioNLP as they probably should be.

Verspoor et al discuss a fine-grained text mining approach for detecting the catalytic sites in proteins in the biomedical literature. The authors create a silver standard corpus, apply a machine learning technique, and achieve reasonable results. The work has application in computational prediction of the functional significance of protein sites as well as in curation workflows for databases that capture this information.

Bush et al describe workflows for the analysis of genetic data in the context of electronic medical records (EMRs). Using EMR data in conjunction with genetic data is an important step in the study of both genetic and environmental factors related to complex human diseases, but analyses combining these data pose substantial privacy concerns. This contribution discusses such concerns, as well as a system that has been developed to allow such analyses via a web server while maintaining appropriate privacy for individuals participating in the study.

Seedorff et al seek to extend a taxonomy of health-related terms, the Mayo Consumer Health Vocabulary (MCV), that helps customers understand the terminology used by healthcare professionals. The authors argue for the importance of integrating synonyms for medical terminologies as well as both genetic risk factors and non-genetic risk factors for diseases into MCV, and present a method for automatically extending it using text mining. The successful extension of MCV can then form a basis to build consumer- oriented products and sophisticated search and information retrieval standards for patient-facing applications.

References

1. Cohen AM, Hersh WR, A survey of current work in biomedical text mining, Briefings in Bioinformatics, 2005, 6: 57-71

2. Langley, P., Lessons for the computational discovery of scientific knowledge. In Proceedings of First International Workshop on Data Mining Lessons Learned, 2002, p. 9-12.

3. Yang, Q., Xindong Wu, 10 Challenging Problems in Data Mining Research, International Journal of Information Technology & Decision Making, Vol. 5, No. 4 (2006) 597–604.