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

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

CASEY S. GREENE

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

MARICEL G. KANN

Department of Biological Science
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
Email: mkann@umbc.edu

ROBERT LEAMAN

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

NIGAM SHAH

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

JIEPING YE

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

Text and data mining methods constantly advance and are applied in different fields. In order for them to impact the biomedical discovery process, it is necessary to thoroughly engage scientists at both ends, and conduct thorough empirical evaluations as to their ability to suggest novel hypotheses and address the most crucial questions. The PSB 2014 Session on Text and Data Mining for Biomedical Discovery presents eight papers that advance the field in this mutually reinforcing fashion. Work presented in this session includes data mining and analysis techniques that are applicable to a broad spectrum of problems, including the analysis and visualization of mass spectrometry based proteomics data and longitudinal data, as well as gene function, protein function and protein fold prediction. Text mining approaches selected for presentation include a method for predicting genes involved in disease or in drug response, a method for extracting events relevant to biological pathways, and an approach that mixes text and data mining techniques to predict important milestones in the female reproductive lifespan.

1. Introduction

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.

Successes in the application of computational approaches that solve biological problems have led to the broad application of these methods to an ever-growing set of specific problem areas. Consequently it is no longer possible to enumerate the biological questions targeted by computational approaches. These questions include, but are not limited to, the problems addressed by papers in this session. Broadly though, we can discuss trends in the field.

While data mining approaches have previously been applied to biological questions in ways that assume the functions of genes are constant, advances in underlying computational platforms and methodology are now allowing computational biologists to begin to address problems in a context specific manner. This means that instead of asking about the overall function of a gene, we are now identifying the role of a gene in a given environment, cell lineage, or individual. We anticipate that approaches that embrace rather than ignore such underlying biological complexities will provide the next generation of advances in personalized medicine.

2. Challenges

The biomedical domain presents specific challenges to text and data mining given the diversity, complexity and volume of the information being mined. The submissions to this session allowed us a unique glimpse at these challenges, which can perhaps be summarized as the constant call to fully incorporate the richness of the available resources and tackle the analysis of data of ever-growing complexity.

Thus, an overarching challenge for biomedical text mining is to incorporate the many knowledge resources that are available to us into the natural language processing pipeline. In the biomedical domain, unlike the general text mining domain, we have access to large numbers of extensive, well-curated ontologies and knowledge bases. However, we have, in general, failed to take advantage of them for tasks like coreference resolution, semantic typing of possible subjects and objects of predicates in information extraction, and the like.

Biomedical ontologies provide an explicit characterization of a given domain of interest and can enhance biomedical discovery significantly when used in a pragmatic manner. Using existing ontologies (from the UMLS and BioPortal) as sources of terms in building lexicons, for figuring out what concept subsumes what other concept, and as a way of normalizing alternative names to one identifier, would likely increase the quality of data-mining efforts. For example, using ontologies as described enabled the use of unstructured clinical notes for generating practice-based evidence on the safety of a highly effective, generic drug for peripheral vascular disease (PubMed 23717437).

Among the papers in this session, there are several examples where important advances to biomedical discovery are based on precisely this expansion on the use of knowledge and literature resources. For example, Funk et al predict pharmacogenomic genes on a genome-wide scale using Gene Ontology annotations and simple features mined from the biomedical literature. Ravikumar et al, present a rule-based literature mining system to extract pathway information from text to assist human curators.

Today, the data being generated is massive, complex, and increasingly diverse due to recent technological innovations. However, the impacts of this data revolution on our lives are being hampered by the limited amount of data that has been analyzed. This necessitates data mining tools and methods that can match the scale of the data and support timely decision-making through integration of multiple heterogeneous data sources. We see in this session numerous contributions to methods and approaches, better outlined in the next section.

Finally, another area in which the field has fallen short and that the papers in this session can only begin to address, is that of making text mining applications that are easily adaptable by end users. Many researchers have developed systems that can be adapted by other text mining specialists, but applications that can be tuned by bench scientists are mostly lacking.

3. Overview of Contributions

Funk et al. describe a method for predicting genes involved in disease or in drug response based on combining heterogenous data, including curated Gene Ontology annotations, text-mined Gene Ontology annotations, and surface linguistic features. These feature types are combined and passed as input to a classifier.

Ravikumar et. al. develop a system to extract events relevant to biological pathways from the literature by combining named entity recognition and normalization with pattern templates to detect event mentions and the role of each entity. Notably, the system resolves both entity and event anaphora with discourse analysis. The authors evaluate their system against PharmGKB pathway annotations, and manually examine a subset of the results.

Malinowski et al report on development and performance of data-mining techniques to identify the age at menarche (AM) and age at menopause (AAM), which are important milestones in the reproductive lifespan; and are often recorded in free-text notes. The authors demonstrate the ability to discriminate age at naturally-occurring menopause (ANM) from medically-induced menopause. Their ultimate goal is to apply the methods to data from the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study, in an attempt to support clinical studies that incorporate these female reproductive milestones.

Han describes an application of a dimensionality reduction technique, called derivative component analysis (DCA) for the analysis and visualization of mass spectrometry based proteomics data. As an implicit feature selection algorithm, DCA enables to extract true signals by capturing subtle data characteristics and removing built-in data noises for input proteomics profiles.

Zupan and Zitnik develop a general matrix factorization-based data integration approach for gene function prediction that fuses heterogeneous data sources, such as gene expression data, known protein annotation, interaction and literature data. The fusion is achieved by simultaneous matrix tri-factorization that shares matrix factors between data sources. The proposed approach is applicable for any number of data sources, which can be expressed in a matrix form.

Liu et al. describe a method for analyzing longitudinal data. Functional regression is a popular approach for longitudinal data analysis, as it is capable of identifying the relationship between features and outcomes along with time information by assuming features and/or outcomes as random functions over time rather than independent random variables. The proposed approach empowers basic functional regression models to simultaneously identify features with significant predictive power across time points, enforce smoothness of functional coefficients, and achieve interpretable estimations of functional coefficients using a novel sparsity-inducing penalty.

Clark and Radivojac develop a novel machine learning algorithm for protein function and fold prediction. In particular, their method introduces a kernel function on time series data that can be obtained from protein sequences and structures. The proposed kernel showed high performance in the task of classifying proteins in SCOP classes. Accurate functional classification of proteins is critical for understanding the molecular mechanisms involved in all biological process across species, which translates into advances in biomedical research. Furthermore, this methodology is applicable to problems beyond computational biology.

Vembu and Morris demonstrate *LMGraph*, two step approach to construct binary predictors from gene networks and features. The first step extracts informative features from the network. The second step combines these network-extracted features with node features to construct predictors. The authors demonstrate that this two-step approach outperforms related algorithms, suggesting that such combined approaches could offer benefits to other methods.