CANCER PATHWAYS: AUTOMATIC EXTRACTION, REPRESENTATION, AND REASONING IN THE 'BIG DATA' ERA

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

CHITTA BARAL
School of Computing, Informatics, and Decision Systems Engineering,
Arizona State University, Tempe, AZ 85287, USA
Email: chitta@asu.edu

JEFF KIEFER
Knowledge Mining Laboratory,
Translational Genomics Research Institute (TGen), Scottsdale, AZ, 85259, USA
Email: jkiefer@tgen.org

SEUNGCHAN KIM
Integrated Cancer Genomics Division, Biocomputing Unit
Translational Genomics Research Institute (TGen), Phoenix, AZ 85004, USA
Email: skim@tgen.org

JIEPING YE
School of Computing, Informatics, and Decision Systems Engineering,
Arizona State University, Tempe, AZ 85287, USA
Email: jieping.ye@asu.edu

There has been great interest and research initiatives in the biomedical community around harnessing "big data", including data from the literature, high-throughput gene expression experiments, array CGH and high-throughput siRNA and many other types of data to generate novel hypothesis to address the most crucial biomedical questions and aid in the discovery of more effective and improved therapeutic options for the treatment of complex and pervasive diseases such as cancer. Cancer research has progressed rapidly in the last decade with the implementation of high-dimensional genomic technologies. The large amount of data generated over the years has enabled a systems-based approach to uncovering and elucidating the complex signaling networks associated with cancer. However, even though new technologies have advanced our understanding of cancer biology beyond what could be imagined even a decade ago, there still exist unique challenges associated precisely with the amount of data that is now routinely generated from even a single patient. The data must be stored and processed, with novel analysis strategies called for to uncover new insights into cancer biology that are literally hidden in 'big data'. Interest in taming 'big data' through methods and systems to extract, represent, and transform it into knowledge that can effectively be used for reasoning and question answering will only increase over time,

enabling scientists to finally use the data for personalized treatment, discovery and validation.

Work presented in this session includes novel approaches to explore cancer gene expression data, applying algebraic topology (Lockwood and Krishnamoorthy) and Denoising autoencoders (Tan et al) to identify significant properties of genomic data that cannot be found by traditional algorithms. There is also a novel methodology for leveraging somatic mutation data for predicting survival in cancer samples (Kim et al), a computational system for automated gene expression pattern annotation on mouse brain images that could prove to be key to understanding the pathogenesis of brain tumors and their early detection (Yang et al). With respect to knowledge extraction, this session includes work on a weakly supervised machine learning approach for automatic pathway extraction from PubMed abstracts (Poon et al), and on the use protein interaction data from multiple sources to investigate mutations in 125 genes that were earlier identified as driving tumorigenesis when mutated (Engin et al).

1. Introduction

This session brings together researchers in text and data mining, knowledge representation and reasoning with bench scientists, geneticists and translational scientists. It serves as a unique forum to discuss novel approaches to the complex text extraction and knowledge representation and reasoning problems that come with dealing with big data. Given that computational approaches often draw upon disease-specific resources and expertise, we focus the session on disseminating approaches to extract, represent, or reason with knowledge derived from the literature, databases, or experimental data that respond to biological questions about the signaling pathways in cancer pathophysiology.

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. *Extraction, Representation, and Reasoning*

While there exists significant research in information extraction from biomedical text, little has been done on the extraction of more general knowledge that requires the integrated use of information from many different publications, databases, and experimental results into a coherent story, a (proven or hypothetical) biological pathway that can be used to drive new research directions. For this task, accurate named entity recognition and named entity identification (also referred to as normalization) are important, as well as the extraction of (binary) relationships or events involving any two of these entities. However, when the goal is to go beyond atomic events and into establishing an ordered sequence of these events to reconstruct the “biological stories” they are telling us (pathways), the problem has strong similarity with the classical planning and scheduling problem in Artificial Intelligence (AI) and shares similar challenges. Furthermore, to answer questions about such pathways, one needs to start with a query representation that is equally comprehensible to human as to machines.

This session includes specialized examples that make some advances into the mostly uncharted territory of pathway extraction, showing the trend towards finer granularity in the type of information needed for meaningful advances, requiring a tighter collaboration between the text mining community and domain experts.

2.2. *Data Mining*

Recent advance in technologies has generated a large amount of high-dimensional genomic data. These data are a key to illuminate fundamental principles of cancer biology. However, analysis of these data poses unique challenges for data mining. One of the key issues is the curse of dimensionality, i.e., an enormous number of samples is required to perform accurate prediction on problems with large numbers of features. Feature selection, which selects a small number of features by removing the irrelevant, redundant, and noisy information, is an effective way to overcome the curse of dimensionality. However, existing methods for feature selection are less effective when there are strong empirical correlations among the features. Furthermore, the presence of missing values which are ubiquitous in biomedical data further complicates the problem. Several recent works apply unsupervised feature learning algorithms to address these challenges. The goal of these methods is to identify highly correlated features and construct representative features from the data.

Included in this session are works that address some of these challenges.

3. Overview of Contributions

Lockwood and Krishnamoorthy apply tools from algebraic topology to explore cancer gene expression data. Specifically, the proposed method selects a small relevant subset from tens of thousands of genes while simultaneously identifying higher order topological features. By employing tools from algebraic topology, the proposed method is capable of identifying geometric properties of the data that cannot be found by traditional algorithms such as clustering.

Poon et al applied distant supervision, a state-of-the-art weakly supervised machine learning approach, for automatic pathway extraction from PubMed abstracts. From 22 million PubMed abstracts, they extracted 1.5 million pathway interactions at a precision of 25%. More than 10% of interactions are mentioned in the context of one or more cancer types. The paper reports interesting results, exploring the effectiveness of utilizing the information available in the pathway database (PID) for automatic annotation of the training data. The applied machine learning approach does not rely on the manually annotated sentences and can potentially be applied to other problems.

Tan et al present a method based on Denoising autoencoders (DAs) for feature extraction from biological data. DAs aim to learn compact and efficient representations from input data. DAs serve as building blocks for deep networks, which have been applied successfully for image and audio processing. The authors present an interesting application of DAs to, for the first time, identify and extract complex patterns from genomic data.

Kim et al touches on a number of themes appropriate for this session. They present a novel methodology for leveraging somatic mutation data for predicting survival in cancer samples. As a test case, they apply their methodology to Renal Cell Carcinoma data from the TCGA. The authors take advantage of their previous tool and methods, BioBin, ATHENA, and Grammatical Evolution Neural Networks (GENN) in the current study based on prior knowledge resources.

Yang et al present a computational system for automated gene expression pattern annotation on mouse brain images. The annotation system aims to provide an accurate description of the locations where the genes are active. Since genetic factor is one of the significant risk factors for brain cancer, such a description is crucial for understanding the pathogenesis of brain tumors and for early detection.

Engin et al. use protein interaction data from multiple sources to investigate mutations in 125 genes that were earlier identified as driving tumorgenesis when mutated. They do structural enrichment of driver protein-protein interactions (PPIs) and map chromosomal coordinates to PDB coordinates. They use TCGA mutation data and RNAseq data as source of patient mutation and expression profiles. Thus by integrating protein interaction networks with protein structure, and using patient related mutation and gene expression data, they observe “patient specific network wiring” and analyze the HRAS subnetwork. Overall, this paper is a good example of cancer network analysis and the integration and mining of various different kinds of data.