Report on ACM Text Mining in Bioinformatics (TMBIO 006)

Min Song
New Jersey Institute of Technology
min.song@njit.edu

Zoran Obradovic
Temple University
zoran@core.ist.temple.edu

Abstract

Effective use of knowledge gained by mining such heterogeneous and unstructured data plays a crucial role in all stages of integrative biological research studies. Text Mining in Bioinformatics has drawn attention from both biomedical researchers and computational biologists. In this article we report on a workshop conducted in conjunction with the ACM CIKM Conference in Washington, D.C. in November 2006. Participants from both academic and industry in Biomedical Text Mining joined the workshop to discuss the issues and trends in the field.

1 Introduction

Biological researchers increasingly rely on a very large amount of biomedical electronic data to correctly interpret and understand their complex biological systems. The biomedical data repositories deliver different types of data: gene expression data from Microarray experiments, protein identification and quantification data from proteomics experiments, genomic sequence data gathered by the Human Genome Project, SNP data from high-throughput SNP arrays, and facts extracted from biomedical scientific publications. Increasingly such facts from the scientific literature are automatically extracted and integrated into complex analytical tools that bring together different types of information, e.g. data from Microarray experiments and sequence annotations.

The focus of this workshop is to bring together researchers that work in the field of text mining and computational biology and that want to integrate such heterogeneous unstructured data, which is a challenging task, and want to better embed literature information into bioinformatics solutions. The workshop seeks in particular:

- Bioinformatics text mining solutions that identify relevant background knowledge in textual documents, such as scientific publications, or in database annotations. Such approaches currently being studied range from term recognition to extraction of complex relationships of interaction between proteins.
- Ambitious knowledge discovery solutions that process heterogeneous biomedical data collected from electronic bulletin boards, scientific publications, and any type of experiments. It is important to identify and solve issues as to how to consolidate information extracted from textual documents into other types of data in structured form.

It is difficult and requires profound understanding of both text mining and computational biology to identify problems and optimization criteria which, when maximized by text mining algorithms, actually contribute to a better understanding of biological systems. Identification of appropriate text mining problems in Bioinformatics and development of evaluation methods for text mining results are ongoing efforts.

2 Workshop

The first international workshop on Text Mining in Bioinformatics (TMBIO) was successful. The number of papers received was 21 and we accepted 8 papers. The acceptance rate was 38%. We invited one keynote speaker from South Korea and also had one invited talk. The keynote speaker is Dr. Doheon Lee, associate professor of Department of BioSystems at KAIST in South Korea. He is the head of Bio-Information System National Research Lab at KAIST and is also the director of General Affairs, Korean Society for Bioinformatics. The invited talk was given by professor Slobodan Vucetic from Temple University. As the first workshop, we are proud that TMBIO 2006 was successful in many ways. It drew attention from researchers in both Text Mining and Bioinformatics research area. The selected papers were invited to two prestigious journals: BMC Bioinformatics and International Journal of Data Mining and Bioinformatics.

TMBIO 2006 was held in a day. The day was divided into two technical paper sessions, a keynote speech, and an invited talk. We summarize each below

2.1.1 Keynote Speech

Dr. Doheon Lee, the director of Korean Society for Bioinformatics, delivered his keynote speech on "BioCAD: An Information Fusion Platform for Bio-Network Inference and Analysis." After giving us a brief summary of the field of Text Mining, he described his current research on bio-network inference and analysis. He proposed the BioCAD system, the information fusion software, to provide a framework for optimization for bio-network inference. He concluded his presentation with issues and future works in bio-network inference and how Text Mining can play a pivotal role in bio-network inference.

2.1.2 Morning Session

The morning session was divided into two themes: 1) Biomedical Knowledge Acquisition and 2) Applications to Biomedical Data. In the Biomedical Knowledge Acquisition session, Tasha Inniss led the talk with her presentation on "Towards Applying Text Mining and Natural Language Procession for Biomedical Ontology Acquisition." She described the research effort of her research group on building automatic ontology in Age-Related Macular Degeneration. She reported the experimental results of comparing three different ontology construction methods: 1) Expert-Retinal Specialist, 2) NLP-NSP, and 3) SAS-Text Miner.

In the Applications to Biomedical Data session, Melissa Rogers presented her research on "Emerging Genome Data to Identify Conserved Bone Morphogenetic Protein(Bmp)2 Gene Expression Mechanisms". As a bio-chemist, she emphasized that understanding the evolution o the complex transcriptional and post-transcriptional mechanisms required novel and interdisciplinary approaches. As an effort in this direction, she described her effort towards defining Bmp2 gene expression determinants by applying text mining techniques for computational analyses of emerging genome data. Another presenter, Brian Westwood, talked about his research on "Application of Correlate Summation to Data Clustering in the Estrogen-and Salt-Sensitive Female mRen2.Lewis Rat." He attempted to do so by using a new congenic rodent model of hypertension. Specifically he compared two statistical methods designed to identify changes in experimental parameters that are significantly linked and reported the positive impact of clustering on the Estrogen-and Salt-Sensitive Female mRen2.Lewis Rat.

2.1.3 Afternoon Session

The afternoon session was broken into two sub-sessions: 1) Ontology for Text Mining and Information Extraction. In the Ontology for Text Mining session, Darren Natale gave a presentation on “Conceptual Framework for a Protein Ontology”. He described a framework for the protein ontology (PRO) and illustrated its use on human proteins from the TGF-beta signaling pathway with the goal to facilitate use of computer-based inference about proteins. Illhoi Yoo presented his work on “Integrating Biomedical Literature Clustering and Summarization Approaches using Biomedical Ontology”. He introduced a technique for combined document clustering and text summarization. Based on his previous work on ontology-enriched graphical representation for documents for document clustering, he used the graphical model for document clustering and text summarization. As a last presenter of the session, Padmini Srinivasan gave a talk on her research “GO for Gene Documents”. Finding information in huge and growing biological datasets is an enormous challenge. She stressed that Gene Ontology organizes experimental findings in a classification scheme that is allows automated data mining. To build an automatic Gene Ontology, she adopted support vector machine classifiers to test various forms of data analyses to define criteria for accurate classification of terms.

In the Information Extraction session, Manabu Torii presented his and his colleagues work on “a Comparison Study of Biomedical Short Form Definition Detection Algorithms.” He described a comparative study of abbreviation extraction algorithms together with ways for improving the algorithms. By providing a solid description of how various abbreviation extraction works, he suggested the better way of extracting abbreviation. The last presenter, Hyunsoo Kim gave a talk on “Extracting indirect gene relationships from biomedical literatures via matrix factorizations.” He explored utility of the singular value decomposition (SVD) and non-negative matrix factorization (NMF) in discovering gene relationships from MEDLINE citations. The paper attempted to validate use of NMF as an alternative to SVD and showed that SVD and NMF are effective retrieval of related genes from literature.

3 Acknowledgement

We give our sincere thanks to SIGIR, to the authors and presenters, to the keynote speaker, and to the members of the program committee for their contributions to organize an outstanding workshop.