BIOKDD 2008: A Workshop Report on Data Mining in Bioinformatics

Jake Y. Chen
School of Informatics
Indiana University
Indianapolis, IN 46202
jakechen@iupui.edu

Mohammed Zaki
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590
zaki@cs.rpi.edu

Stefano Lonardi
Dept. of Computer Science and Eng.
University of California
Riverside, CA 92521
stelo@cs.ucr.edu

INTRODUCTION

Bioinformatics is an interdisciplinary study that aims to collect, manage, interpret, and disseminate biological information, primarily those at molecular and cellular levels. The ongoing influx of data from high-throughput Genomics, Proteomics, Metabolomics, and Structural Genomics projects have made their interpretation both rewarding and challenging for data mining researchers. On one hand, data mining approaches and bioinformatics problems are ideally suited for each other because of the availability of large amounts of "Omics" data and presence of innate biological signals from experimental measurements. On the other hand, bioinformatics problems tend to suffer from high data noises due to inherent inaccuracies in observing complex biology with high-throughput instruments, large sample variability, high data dimensionality, and limited sample size from many experiments. To highlight recent advances in the use of data mining techniques to solve biological problems, we organized the 2008 International Workshop on Data Mining in Bioinformatics (BIOKDD '08), held as a half-day event in conjunction with the 2008 ACM SIGKDD Conference, Las Vegas, NV, August 24-27, 2008. This was the 8th year for the BIOKDD workshop series. To promote this year's program, we established an Internet web site at http://bio.informatics.iupui.edu/biokdd08.

The theme of this year's workshop was focused on research that aims to integrate complex biological systems and knowledge discovery. Different from analyzing single molecules, complex biological systems consist of components that are in themselves complex and interacting with each other. Understanding how the various components work in concert, using modern high-throughput biology and data mining methods, is crucial to the ultimate understanding of complex biological phenomena.

PROGRAM

We accepted 8 papers out of 24 submissions into the workshop program and proceedings due to exceptionally high-quality submissions. Each paper was peer reviewed by three members of the program committee and papers with declared conflict of interest were reviewed blindly to ensure impartiality. All papers, whether accepted or rejected, were given detailed review forms as a feedback. The 8 accepted papers were organized in two workshop sessions and given 20 minutes each to present at the workshop.

In the paper "*Function Prediction Using Neighborhood Patterns*", Petko Bogdanov and Ambuj K. Singh from the University of California, Santa Barbara, USA reported how to predict functions of uncharacterized proteins by comparing their functional neighborhoods to proteins of known function. The method accomplished a good balance between prediction accuracy and coverage of proteins with predicted functions.

In the paper "*Statistical Modeling of Medical Indexing Processes for Biomedical Knowledge*

Information Discovery from Text", Markus Bundschus, Mathaeus Dejori, Shipeng Yu, Volker Tresp, and Hans-Peter Kriegel from the University of Munich, Germany, presented a Topic-Concept Model as a general framework to index the topic structure of documents by learning the statistical dependencies between words, topics and MeSH (Medical Subject Headings) concepts. The enriched topic representation provides important additional information absent from conventional MeSH term abstract classifications.

In the paper "*Information Theoretic Methods for Detecting Multiple Loci Associated with ComplexDiseases"*, Pritam Chanda, Aidong Zhang, Lara Sucheston and Murali Ramanathan from the State University of New York, Buffalo, USA, developed two information theoretic metrics in identifying gene-gene interactions, based on thousands of single-nucleotide polymorphism data for complex diseases, using rheumatoid arthritis successfully as a case study.

In the paper *"A Fast, Large-scale Learning Method for Protein Sequence Classification",* Pavel Kuksa, Pai-Hsi Huang, and Vladimir Pavlovic from Rutgers University, USA, applied a class of efficient string-based kernels, sparse spatial sample kernels, to classify protein sequence. Compared with the conventional profile-based searches, the method is promising in becoming more computationally efficient, scalable, and accurate.

In the paper "*Catching Old Influenza Virus with A New Markov Mode*l", HamChing Lam and Daniel Boley from the University of Minnesota, USA, devised a Markov model to model the genetic distance between avian influenza viruses based on the Hemag-glutinin (HA) gene, a major surface antigen. Their model predicts that recent surging of similar viruses to those decades ago is extremely unlikely and may arise from a reservoir of dormant aviant viruses.

In the paper "*GPD: A Graph Pattern Diffusion Kernel for Accurate Graph Classification with Applications in Cheminformatics*", Aaron Smalter, Luke Huan, Gerald Lushington and Yi Jia from the University of Kansas, USA, described a novel graph mining technique based on frequent pattern discovery methods and kernel methods to classify chemical compounds.

In the paper "*Reinforcing Mutual Information-based Strategy for Feature Selection for Microarray*", Jian Tang, Shuigeng Zhou, Feng Li and Jiang Kai, Fudan University, China, presented an integrative method, in which substantial relevance boosting was used to combat microarray data noise and increasing likelihood of feature interactions was used to compensate for gene-gene interactions.

In the paper "*Graph-based Temporal Mining of Metabolic Pathways with Microarray Data*", Chang hun You, Lawrence B. Holder, Diane J. Cook, Washington State University, USA, described a new dynamic graph-based relational learning approach, which overcomes the weakness of conventional static analysis of biological networks in revealing new patterns in biological networks.

The collection of workshop papers presented at the workshop reflected the recent progress in our community, in which the conventional theme of "sequence to structure to function" is replaced by the new theme of a diverse set of research questions, for example, protein function prediction, protein classification, protein-protein interaction networks, genetic analysis of genome-wide genetic association data, literature mining, and microarray data analysis. We were excited by the emergence of new research accomplishments in the data mining community in response to changing biology needs in the post-genome era.

The invited keynote talk titled "Link Mining: exploring the power of links" was given by Dr. Philip Yu, Professor and Wexler Chair in Information Technology, Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA.

WORKSHOP CO-CHAIRS

- Stefano Lonardi, University of California, Riverside
- Jake Y. Chen, Indiana University – Purdue University, Indianapolis
- Mohammed J. Zaki, Rensselaer Polytechnic Institute (General Chair)

PROGRAM COMMITTEE

Alberto Apostolico (Georgia Tech & University of Padova), Ann Loraine (University of North Carolina, Charlotte), Chad Myers (University of Minnesota), Chandan K. Reddy (Wayne State University), Dong Xu (University of Missouri), Giuseppe Lancia (University of Udine, Italy), Isidore Rigoutsos (IBM T. J. Watson Research Center), Jason Wang (New Jersey Institute of Technology), Jie Zheng (NCBI), Jing Li (Case Western Reserve University), Knut Reinert (Freie Universitt Berlin, Germany), Li Liao (University of Delaware), Luke Huan (University of Kansas), Mehmet Koyuturk (University of Georgia), Muhammad Abulaish (Case Western Reserve University), Natasa Przulj (University of California, Irvine), Michael Brudno (University of Toronto), Muhammad Abulaish (Jamia Millia Islamia, India), Natasa Przulj (University of California, Irvine), Phoebe Chen (Deakin University, Australia), Rui Kuang (University of Minnesota), Seungchan Kim (Arizona State University), Si Luo (Purdue University), Simon Lin (Northwestern University), Walid G. Aref (Purdue University), Wei Wang (University of North Carolina, Chapel Hill), Xiaohua Hu, (Drexel University), Yaoqi Zhou (Indiana University), Yves Lussier (University of Chicago).

ACKNOWLEDGEMENT

We would like to thank all the program committee members, contributing authors, invited speaker, and attendees for contributing to the success of the workshop. Special thanks are also extended to the SIGKDD ’08 conference organizing committee, particularly Eamonn Keogh (UC Riverside) and Ying Li (Microsoft), for coordinating with us to put together the excellent workshop program on schedule.