Data Mining in Bioinformatics: Report on BIOKDD'03

Mohammed J. Zaki
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
zaki@cs.rpi.edu

Jason T. L. Wang
Computer Science Department
New Jersey Institute of Technology
Newark, NJ 07102, USA
wangj@njit.edu

Hannu T. T. Toivonen
Computer Science Department
University of Helsinki
Helsinki, FIN-00014, Finland
htoivone@cs.helsinki.fi

1. FOREWORD

Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Genome sequencing projects have contributed to an exponential growth in complete and partial sequence databases. The structural genomics initiative aims to catalog the structure-function information for proteins. Advances in technology such as microarrays have launched the subfield of genomics and proteomics to study the genes, proteins, and the regulatory gene expression circuitry inside the cell. What characterizes the state of the field is the flood of data that exists today or that is anticipated in the future; data that needs to be mined to help unlock the secrets of the cell.

While tremendous progress has been made over the years, many of the fundamental problems in bioinformatics, such as protein structure prediction or gene finding, are still open. Data mining will play a fundamental role in understanding gene expression, drug design and other emerging problems in genomics and proteomics. Furthermore, text mining will be fundamental in extracting knowledge from the growing literature in bioinformatics.

BIOKDD'03 was held in conjunction with the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in Washington, DC, in August 2003. The goal of this workshop was to encourage KDD researchers to take on the numerous challenges that Bioinformatics offers. The workshop featured keynote talks from noted experts in the field, and the latest data mining research in bioinformatics; it encouraged papers that proposed novel data mining techniques for tasks such as:

- Gene expression analysis
- Protein/RNA structure prediction
- Phylogenetics
- Sequence and structural motifs
- Genomics and Proteomics
- Gene finding
- Drug design
- Text mining in bioinformatics

The workshop proceedings contain 9 papers (out of 24 submissions) that were accepted for presentation at the workshop. Each paper was reviewed by three members of the program committee. Along with 2 keynote talks by Steven Salzberg (from TIGR) and Sorin Istrail (from Celera), we were able to assemble a very exciting program. A brief overview of the invited talks and contributed papers appears below.

This workshop followed the previous two highly successful workshops: BIOKDD02 [2], held in Edmonton, Canada, and BIOKDD01 [1] held in San Francisco, CA. Online proceedings of the BIOKDD workshops are accessible at the following links:

- http://www.cs.rpi.edu/∼zaki/BIOKDD01/
- http://www.cs.rpi.edu/∼zaki/BIOKDD02/
- http://www.cs.rpi.edu/∼zaki/BIOKDD03/

2. KEYNOTE TALKS

The opening keynote talk for the morning session, entitled *genome paleontology: discoveries from complete genomes*, was delivered by Steven Salzberg, Senior Director of Bioinformatics at The Institute for Genomic Research (TIGR). Steven defined genome paleontology as the task of comparing genomes to uncover: i) history of species, ii) genome transformations, iii) recent mutations such as Single Nucleotide Polymorphisms (SNPs), and iv) evolution. He described the work his group has done on genome scale sequence alignments, the hunt for genome scale duplications, symmetric chromosomal inversions, Human genome analysis, horizontal gene transfers, and other interesting problems.

The opening keynote talk for the afternoon session, entitled *the minimum informative subset problem*, was given by Sorin Istrail, Senior Director for Informatics Research at Celera Genomics/Applied Biosystems. Sorin introduced a recurring problem within bioinformatics, which he termed the minimum informative subset problem, defined as follows: given a set of objects, find a minimum set of attributes that are informative. He highlighted several applications of this problem, for example, SNP selection and haplotype tagging, assay design, literature survey, and so on.

3. CONTRIBUTED PAPERS

The first session in the morning consisted for four papers. Zhenzhen Kou, William Cohen, and Robert F. Murphy (Carnegie Mellon University) presented their work on *extracting information from text and images for location proteomics*. They introduced new text mining and OCR techniques for caption understanding, and figure-text matching, with applications to protein sub-cellular localization, from the text and images found in online journals. In the next paper Mark A. Krogel and Tobias Scheffer (Humboldt University, Germany) studied the *effectiveness of information extraction, multi-relational, and multi-view learning for predicting gene deletion experiments*. To build a biologically relevant model, they utilized all available data such as unlabeled data, relational data and abstracts from research papers. Jinyan Li, Huiqing Liu and Limsoon Wong (Institute for Infocomm Research, Singapore) claimed that *mean-entropy discretized features are effective for classifying high-dimensional biomedical data*. The paper studied empirical feature selection heuristics, utilizing entropy, for classification of biomedical datasets. They show results on reduction in features obtained in practice. In the final paper of the session, Marios Skounakis and Mark Craven (University of Wisconsin, Madison) discussed their work on *evidence combination in biomedical natural-language processing*. In many cases the individual entities or relations of interest are found in multiple contexts in a corpus. They present a statistical approach to combine evidence across such multiple contexts. They apply the work to protein names and protein-protein interactions from MEDLINE.

The second session in the afternoon, had two papers. Aleksander Icev, Carolina Ruiz, and Elizabeth F. Ryder (Worcester Polytechnic Institute), proposed *distance-enhanced association rules for gene expression*. They mine association rules among motifs extracted from promoter regions of genes. The rules are extended with distances between motifs, and show improvements in classification versus use of traditional association rules. Kevin Y. Yip, David W. Cheung and Michael K. Ng (University of Hong Kong) presented *a highly-usable projected clustering algorithm for gene expression profiles*. Most existing projected clustering algorithms depend on several user specified parameters, which are often ill understood. Their new algorithm overcomes such limitations, and they study applications in mining gene expressions.

In the final session, there were three papers. Li Zhang, Aidong Zhang and Murali Ramanathan (University of Buffalo), opened with *enhanced visualization of time series through higher Fourier harmonics*. They studied methods for visualizing gene expression datasets using higher Fourier harmonics projections. Hsiao-Mei Lu, Sumeet Gupta, and Yang Dai (University of Illinois, Chicago) gave methods for *reducing large diagonals in kernel matrices through semidefinite programming (SDP)*. Application of Support Vector Machines in Bioinformatics, may result in kernel matrices with large diagonal elements. The paper used SDP to reduce large kernels. Finally, Xintao Wu, Yong Ye, Kalpathi, and R. Subramanian (University of North Carolina, Charlotte), presented their work on *interactive analysis of gene interactions using graphical gaussian model*. The datasets came from DNA microarray studies; they propose an interactive system to explore gene relations.

4. ACKNOWLEDGMENT

We would like to thank all the authors, invited speakers, and attendees for contributing to the success of the workshop. Special thanks are due to the program committee for help in reviewing the submissions.

5. WORKSHOP CO-CHAIRS

Mohammed J. Zaki, Rensselaer Polytechnic Institute, USA
Jason T.L. Wang, New Jersey Institute of Technology, USA
Hannu T.T. Toivonen, University of Helsinki, Finland

6. PROGRAM COMMITTEE

Srinivas Aluru, Iowa State University
Pierre Baldi, University of California, Irvine
Yi-Ping Phoebe Chen, Queensland University of Technology, Australia
Mark Craven, University of Wisconsin
Hasan Jamil, Mississippi State University
George Karypis, University of Minnesota
Ross D. King, University of Wales, UK
Stefan Kramer, Technical University of Munich, Germany
Simon M. Lin, Duke University
Zoran Obradovic, Temple University
Srini Parthasarathy, Ohio State University
Luc De Raedt, Albert-Ludwigs University, Germany
Tobias Scheffer,Otto-von-Guericke University, Germany
Mona Singh, Princeton University
Shin-Mu Vincent Tseng, National Cheng Kung University, Taiwan
Alfonso Valencia, National Center for Biotechnology, Spain
Limsoon Wong, Institute for Infocomm Research, Singapore
Jiong Yang, University of Illinois, Urbana-Champaign

7. REFERENCES

[1] M. J. Zaki, H. Toivonen, J. T. L. Wang. BIOKDD01: Workshop on Data Mining in Bioinformatics. In *SIGKDD Explorations*, Volume 3, Issue 2, pp 71-73, January 2002.

[2] M. J. Zaki, H. Toivonen, J. T. L. Wang. BIOKDD02: Recent Advances in Data Mining in Bioinformatics. In *SIGKDD Explorations*, Volume 4, Issue 2, pp 112-114, December 2002.