

# BIOKDD03: Workshop on Data Mining in Bioinformatics

## August 27th, 2003

### Washington, DC, USA

in conjunction with

9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

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

### Opening Remarks

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.

The goal of this workshop is to encourage KDD researchers to take on the numerous challenges that Bioinformatics offers. The workshop features keynote talks from noted experts in the field, and the latest data mining research in bioinformatics. We encouraged papers that propose 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

These 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, we were able to assemble a very exciting program.

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.

This workshop follows the previous two highly successful workshops: BIOKDD02, held in Edmonton, Canada, and BIOKDD01 held in San Francisco, CA. We expect BIOKDD03 to be equally successful.

### Workshop Co-Chairs

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

### 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

## Workshop Program & Table of Contents

*8:50-9:00am: Opening Remarks*

*9:00-10:00am: Keynote Talk I*

- Genome Paleontology: Discoveries from Complete Genomes, Steven L. Salzberg, Senior Director of Bioinformatics, The Institute for Genomic Research (TIGR) page 1

*10:00-10:30am: Break*

*10:30-12:00pm: Session I*

- Extracting information from text and images for location proteomics, Z. Kou, W. Cohen, R. Murphy, Carnegie Mellon University. pages 2-9
- Effectiveness of information extraction, multi-relational, and multi-view learning for prediction gene deletion experiments, M. Krogel (University of Magdeburg), T. Scheffer (Humboldt University), Germany. pages 10-16
- Mean-entropy discretized features are effective for classifying high-dimensional biomedical data, J. Li, H. Liu, L. Wong, Inst for Infocomm Research, Singapore. pages 17-24
- Evidence combination in biomedical natural-language processing, M. Skounakis, M. Craven, University of Wisconsin. pages 25-32

*12:00-1:30pm: Lunch*

*1:30-2:30pm: Keynote Talk II*

- The Minimum Informative Subset Problem, SORIN ISTRAIL, Senior Director, Informatics Research, Celera Genomics/Applied Biosystems. page 33

*2:30-3:15pm: Session II*

- Distance-enhanced association rules for gene expression, A. Icev, C. Ruiz, E. Ryder, Worcester Polytechnic Institute. pages 34-40
- A highly-usable projected clustering algorithm for gene expression profiles, K. Yip, W. Cheung, M. Ng, University of Hong Kong. pages 41-48

*3:15-3:45pm: Break*

*3:45-5:00pm: Session III*

- Enhanced visualization of time series through higher fourier harmonics, L. Zhang, A. Zhang, M. Ramanathan, University of Buffalo. pages 49-56
- Reducing large diagonals in kernel matrices through semidefinite programming, H-M Lu, S. Gupta, Y. Dai, University of Illinois, Chicago. pages 57-62
- Interactive Analysis of Gene Interactions Using Graphical gaussian model, X. Wu, Y. Ye, K. Subramanian, L. Zhang, University of North Carolina, Charlotte. pages 63-69

*5:00-5:30pm: Discussion Session.*