Report on BIOKDD04: Workshop on Data Mining in Bioinformatics

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

Shinichi Morishita
Department of Computational Biology
University of Tokyo
Kashiwa City, Chiba Pref., 277-8562, Japan
moris@k.u-tokyo.ac.jp

Isidore Rigoutsos
Computational Biology Center
IBM Thomas J Watson Research Center
Yorktown Heights, NY 10598, USA
rigoutso@us.ibm.com

1. FOREWORD

BIOKDD'04 was held in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in Seattle, WA, in August 2004. There are numerous sources of biological data that provides challenging opportunities for data mining. For example, 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. Other sources of data include the rapidly growing literature in bioinformatics (e.g., PubMed), the data on biochemical pathways, the evolutionary relationships among organisms in the Tree of Life, high throughput drug design combinatorial libraries, and so on. In addition to the data from biology and genomics, there are rich sources of data from other biosciences, including biomedical, and neuroscience data.

The goal of this workshop was to encourage KDD researchers to take on the numerous challenges that Bioinformatics offers. The workshop featured a keynote talk, 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
- RNAi and microRNA Analysis
- Text mining in bioinformatics
- Modeling of biochemical pathways

The workshop proceedings contain 10 papers (6 long and 4 short), out of 26 submissions. Each paper was reviewed by three members of the program committee. Along with a keynote talk, we were able to assemble a very exciting program. A brief overview of the invited talk and contributed papers appears below.

This workshop followed the previous three highly successful workshops: BIOKDD03 [3], held in Washington, DC, 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/
- http://www.cs.rpi.edu/∼zaki/BIOKDD04/

2. KEYNOTE TALK & CONTRIBUTED PAPERS

The keynote talk was delivered by Mark Boguski, M.D, Ph.D., Senior Director, Development and Research, Allen Institute for Brain Science; and affiliate faculty, Fred Hutchinson Cancer Research Center and Department of Medicine/Genetics, University of Washington. Mark talked about the data mining challenges in the Allan Brain Atlas project. Mark explained that our current understanding of how the brain is organized and how it works is still in the very early stages. Basic mechanisms and procesees of cognition and memory are yet a mystery. It is estimated that the human brain contains a trillion nerve cells or neurons, each capable of making up to a thousand different connections. However, we don't know how many subtypes of neurons exist, how they are linked up in circuits, or how they work. The Brain Atlas is intended to be the foundation for neuroscience research by integrating genomic-based methodologies with computer science and traditional neuroanatomy. It seeks to develop a new brain atlas that emphasizing the relationships between molecular signatures of gene expression and neural structure and function. This brain atlas and signature data will offer numerous opportunities for data mining research, when it is made public.

The first session in the morning consisted of two long papers. Hongyuan Li, Keith Marsolo, Srinivasan Parthasarathy and

Dmitrii Polshakov (Ohio State University) presented their work on *A New Approach to Protein Structure Mining and Alignment*. They presented a new method that combines both substructure mining and sequence alignment information to determine the similarity between protein molecules. The second paper was by Zhengdeng Lai and Yang Dai (University of Illinois, Chicago), entitled *A Novel Approach for Prediction of Protein Subcellular Localization from Sequence Using Fourier Analysis and Support Vector Machines*. They apply the fast Fourier transform on a numerical encoding of the protein sequence, which is then fed into the SVM classifier to predict the subcellular localization.

The second session had two short papers. Jake Chen, Andrey Sivachenko, and Lang Li (Indiana University/Prolexys Pharmaceuticals), talked about *High-throughput Protein Interactome Data: Minable or Not?* They looked at the problem of mining the collections of thousands of protein interaction pairs (called the protein interactome), derived from high-throughput experiments. Their comprehensive study of over 70,000 protein interactions derived from systematic Yeast 2-Hybrid (SY2H) method reveals high quality data that is suitable for mining, as opposed to traditional Y2H data which is typically noisy. In the second paper Ruggero Pensa, Claire Leschi, Jeremy Besson, and Jean-Francois Boulicaut (INSA/INRA, Lyon, France) presented an *Assessment of Discretization Techniques for Relevant Pattern Discovery from Gene Expression Data*. The compared the clustering results obtained via different discretizations, and they were able to determine the best choice of a discretization technique and its parameters for each specific dataset.

The third session consisted of two long and two short papers. The first long paper was by Benny Fung and Vincent Ng (Hong Kong Polytechnic University), entitled *Metaclassification of Multi-type Cancer Gene Expression Data*. It uses expression level histograms for significant genes to represent profiles, and uses the differences between profiles to obtain dissimilarity measures and indicators of the predictive class. The second long paper by Mario Medvedovic and Junhai Guo (University of Cincinnati), presented work on *Bayesian Model-Averaging in Unsupervised Learning From Microarray Data*. They show that clustering via infinite mixture model offers more robust performance than the finite mixture model approach. The first short paper was by Miles Trochesset and Anthony Bonner (University of Toronto) on *Clustering Labeled Data and Cross-Validation for Classification with Few Positives in Yeast*. They use gene expression and phenotype data from *S. cerevisiae* for predicting the biological functions of essential genes. They used a hierarchical clustering approach with labeled data to identify positives and make predictions for unlabeled genes. The final short paper by Yi-Feng Lin, Tzong-Han Tsai, Wen-Chi Chou, Kuen-Pin Wu, Ting-Yi Sung, and Wen-Lian Hsu (IIS, Academia Sinica) was on *A Maximum Entropy Approach to Biomedical Named Entity Recognition*. They adopt a hybrid method using maximum entropy along with dictionary-based and rule-based methods for post-processing, for finding the named entities like Proteins, RNA, DNA, etc.

The fourth and final session had two long presentations. Keith Marsolo, Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta (Ohio State University), presented work on *Discovering Spatial Relationships Between Approximately Equivalent Patterns in Contact Maps*. They develop a criteria for determining whether contact patterns are approximately equivalent, to then discover relationships between connected patterns. The last paper was by Christopher Besemann, Anne Denton, and Ajay Yekkirala (North Dakota State University) on *Differential Association Rule Mining for the Study of Protein-Protein Interaction Networks*. They further identify differences between networks, and use the rules to compare with protein annotations.

3. 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.

4. WORKSHOP CO-CHAIRS

- Mohammed J. Zaki, Rensselaer Polytechnic Institute
- Shinichi Morishita, University of Tokyo
- Isidore Rigoutsos, IBM T.J. Watson Research Center

5. PROGRAM COMMITTEE

• **Srinivas Aluru**, Iowa State U., USA • **Alberto Apostolico**, Prudue U., USA • **Tatsuya Akutsu**, Kyoto U., Japan • **Charles Elkan**, UC San Diego, USA • **Jayant Haritsa**, Indian Inst. of Science, India • **Hasan Jamil**, Wayne State U., USA • **Andreas Karwath**, U. Freiburg, Germany • **George Karypis**, U. Minnesota, USA • **Ross D. King**, U. of Wales, UK • **Jinyan Li**, Inst. for Infocomm Research, Singapore • **Lance A. Liotta**, NIH/NCI, USA • **Ambuj Singh**, UC Santa Barbara, USA • **David Page**, U. Wisconsin, USA • **Srinivasan Parthasarathy**, Ohio State U., USA • **Jignesh M. Patel**, U. Michigan, USA • **Daniel E. Platt**, IBM TJ Watson, USA • **Luc De Raedt**, U. Freiburg, Germany • **Tobias Scheffer**, Humboldt U., Germany • **Karlton Sequeira**, RPI, USA • **Hannu Toivonen**, U. Helsinki, Finland • **Jason Wang**, NJIT, USA • **Wei Wang**, UNC Chapel-hill, USA • **Jiong Yang**, Case Western Reserve U., USA • **Aidong Zhang**, U. Buffalo, USA

6. 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.

[3] M. J. Zaki, H. Toivonen, J. T. L. Wang. Data Mining in Bioinformatics: Report on BIOKDD'03. In *SIGKDD Explorations*, Volume 5, Issue 2, pp 198-199, December 2003.