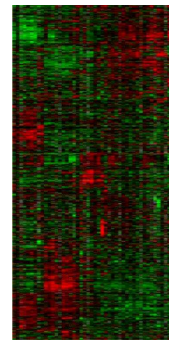


Data Mining for Biomedical Informatics

Saturday, April 28, 2007

A full-day workshop, to be held in conjunction with the 7th SIAM International Conference on Data Mining.



Organizing Committee

Petros Drineas

Rensselaer Polytechnic Institute

Vipin Kumar

University of Minnesota

Michael W. Mahoney

Yahoo! Research

Scope of the workshop

Biology is rich in data, and is getting richer all the time. Recent advances in DNA sequencing, microarray data generation, high-throughput, gene-function studies, medical imaging, and electronic medical records (EMR) have resulted in the automatic generation of new, vast, and exciting databases. Deriving “big pictures from this sea of biomedical data,” as described in the July 2005 issue of *Science*, is a major scientific challenge that will require the close collaboration of computer scientists, biologists, and mathematicians. This workshop will provide a venue to facilitate the exchange of ideas between the three disciplines by bringing together researchers to discuss and present sources of data, research topics that may be addressed by such data, and data mining algorithms that may be used to analyze them.

The long term goal of this workshop is to foster more interaction between the SIAM Data Mining community and the numerous organizations that generate biomedical data, in order to promote joint research on topics that are relevant to both communities.

Program Committee

<i>Dimitris Achlioptas</i>	UC Santa Cruz
<i>Christos Faloutsos</i>	Carnegie Mellon University
<i>Ananth Grama</i>	Purdue University
<i>Boulos Harb</i>	University of Pennsylvania
<i>Kenneth Kidd</i>	Yale University
<i>Francois Meyer</i>	Princeton University
<i>Yi Pan</i>	Georgia State University
<i>Peristera Paschou</i>	Democritus University of Thrace
<i>Mohammed Zaki</i>	Rensselaer Polytechnic Institute

Program

Section 1 (8:30 – 10:00)

- 8:30 - 8:40 *Welcome comments*
- 8:40 - 9:00 ***F. El Khettabi and P. Kyriakidis***
The L_2 Discrepancy Framework to Mine High-Throughput Screening Data for Targeted Drug Discovery: Application to AIDS Antiviral Activity Data of the National Cancer Institute
- 9:00 - 9:20 ***N. Wale, G. Karypis, and I. Watson***
Methods for Effective Scaffold Hopping in Chemical Compounds
- 9:40 - 10:00 ***J. Pandey, M. Koyuturk, W. Szpankowski, and A. Grama***
A Statistical Model for Functional Characterization of Regulatory Pathways

Coffee Break (10:00 – 10:30)

Keynote Talk (10:30 – 11:15)

Inferencing Across Clinical and Genomic Data: Mining Madness, Statistical Folly, and the Joy of Systems Biology

Christopher G. Chute

Mayo Clinic, College of Medicine

Abstract. The advent of the human genome and its integration into clinical medicine poses one of the great challenges for biomedicine into the 21st century. Colloquially referenced as translational medicine, the challenges from data management, data representation, and inferencing perspectives are formidable. Recent technology is able to generate over one million SNPs per specimen per analytic work frame. Correlating this volume of data with hundreds or thousands of phenotypic expression characteristics causes traditional stochastic models and statistical methods to collapse. The historical work around to the “statistical fishing

expedition” problem has been hypothesis driven research. However, in a data discovery model, hypothesis generation has almost equal importance in this brave new world. One mechanism is to leverage the knowledge resources implicit (explicitly not yet explicit) in systems biology and dynamic pathway networks of metabolic systems, sub-cellular physiology, and the new integrated biology. This talk will outline some of the data representation issues, frameworks for biological workflow analysis, and introduce the promise of systems biology.

Section 2 (11:15 – 12:00)

11:15 - 11:40

Orly Alter

Discovery of Principles of Nature From Matrix and Tensor Modeling of DNA Microarray Data

11:40 - 12:00

R. Gupta, T. Garg, G. Pandey, M. Steinbach, and V. Kumar

Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements Using Association Analysis

Lunch Break (12:00 – 13:15)

Keynote Talk (13:15 – 14:00)

DNA sequence variation around the genome and around the world

Kenneth K. Kidd, Ph.D.

Prof. of Genetics, Psychiatry, and Ecology & Evolutionary Biology

Yale University

Abstract. Even before the explosion of genetic data on humans in the past two and a half decades, two facts were clear: considerable normal genetic variation exists in all human populations and the frequencies of the different variants (alleles) vary from population to population. The molecular data on DNA sequence variation has confirmed those in spades and allowed us to begin to quantify many aspects of variation on a genome- and species-wide basis. Numerous interesting questions exist and many are analytically and/or computationally challenging.

However, the data often exist in diverse locations and are often fragmentary, e.g., different populations studied for different genetic variants precluding any direct comparison between populations. Some resources are available and others will be available soon. The HapMap has data on the largest number of SNPs but is limited to only four populations that cannot fully represent global genetic variation. Data sets on larger numbers of populations have much more limited data. However, enough data exist now that the global pattern of human population similarity is becoming clear and will be presented along with an overview of online resources currently available.

Section 3 (14:00 – 15:00)

- | | |
|---------------|--|
| 14:00 - 14:20 | <i>A. Javed and P. Paschou</i>
Extracting tagging SNPs from Genome-wide Datasets |
| 14:20 - 14:40 | <i>M. Maggioni and R.R. Coifman</i>
Multiscale Analysis of Data Sets with Diffusion Wavelets |
| 14:40 - 15:00 | <i>F. Meyer and X. Shen</i>
Exploration of high dimensional biomedical datasets with low-distortion embeddings |

Coffee Break (15:00 – 15:30)

Section 4 (15:30 – 16:30)

- | | |
|---------------|--|
| 15:30 - 15:50 | <i>C. Besemann and A. Denton</i>
Mining Edge-disjoint Patterns in Graph-relational Data |
| 15:50 - 16:10 | <i>W. Li and Y. Liu</i>
Generalized Replicator Dynamics for Efficient Phylogenetic Inference |
| 16:10 - 16:30 | <i>A. Denton and A. Kar</i>
Finding Differentially Expressed Genes Through Noise Elimination |
| 16:30 - 16:50 | <i>C. Ding, C. Wang, and S. Holbrook</i>
Computing Overlapping Protein Interaction Modules |