BIOKDD06: Data Mining in Bioinformatics

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

George Karypis
Department of Computer Science
University of Minnesota
Minneapolis, MN 55455, USA
karypis@cs.umn.edu

Jiong Yang
Electrical Engineering and Computer Science Department
Case Western Reserve University
Cleveland, OH 44106 USA
jiong@eecs.cwru.edu

Data Mining is the process of automatic discovery of novel and understandable models and patterns from large amounts of data. Bioinformatics is the science of storing, analyzing, and utilizing information from biological data such as sequences, molecules, gene expressions, and pathways. Development of novel data mining methods will play a fundamental role in understanding these rapidly expanding sources of biological data. The extensive databases of biological information create both challenges and opportunities for developing novel data mining methods.

The 6th BIOKDD Workshop was held on August 20th, 2006, Philadelphia, PA, USA, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The goal of the workshop was to encourage KDD researchers to take on the numerous challenges that Bioinformatics offers. The BIOKDD workshops have been held annually in conjunction with the ACM SIGKDD Conferences, since 2001. Additional information about the workshop can be obtained online at: http://www.cs.rpi.edu/~zaki/BIOKDD06.

Out of the 18 submissions, 6 were selected for presentation at the workshop. An overview of each paper is given below.

Contributed Talks

In their paper, *Signal Transduction Model Based Functional Module Detection Algorithm for Protein-Protein Interaction Networks*, Woochang Hwang, Young-Rae Cho, Aidong Zhang and Murali Ramanathan, describe the unexpected properties of the protein-protein interaction (PPI) networks and their use in a clustering method to detect biologically relevant functional modules. They propose a new method called STM (signal transduction model) to detect the PPI modules, and compare it with previous approaches to demonstrate its effectiveness in discovering large and arbitrary shaped clusters.

In *Protein Folding Trajectories: Summarization, Event Detection and Consensus Partial Folding Pathway Identification*, Hui Yang, Srinivasan Parthasarathy and Duygu Ucar, describe a method to mine protein folding molecular dynamics simulations datasets. They describe a spatio-temporal association discovery approach to mine protein folding trajectories, to identify critical events and common pathways.

In the paper titled *Automatic Layout and Visualization of Biclusters*, Gregory A. Grothaus, Adeel Mufti and T. M. Murali, present a novel method to display biclusters mined from gene expression data. The approach allows querying and visual exploration of the clusters/sub-matrices. The software is also available as open-source.

In *ExMotif: Efficient Structured Motif Extraction*, Yongqiang Zhang and Mohammed J. Zaki, describe a new algorithm called ExMotif to extract frequent motifs from DNA sequences. The method can mine structured motifs and profiles which have variable gaps between different elements. The demonstrate the efficiency of the method compared to state-of-the-art methods, and also demonstrate an application in mining composite transcription factor binding sites.

In the paper *Motif Refinement using Hybrid Expectation Maximization based Neighborhood Profile Search*, Chandan K. Reddy, Yao-Chung Weng and Hsiao-Dong Chiang, show how one can refine the profile motifs discovered via Expectation Maximization/Gibbs Sampling based methods. They search the neighborhood regions of the initial alignments to obtain locally optimal solutions, which improve the information content of the discovered profiles.

In *Protein Classification using Summaries of Profile-Based Frequency Matrices*, Keith Marsolo and Srinivasan Parthasarathy, describe a new method for classifying protein sequences by creating a wavelet-based summary of the frequency scores obtained via PSI-BLAST. They demonstrate that the new method is competitive with state-of-the-art methods for remote homology detection and fold recognition. The approach can also be used for indexing protein sequences.

Invited Talks

There were two invited talks at the workshop. David Roos gave a talk on *Designing and Mining Pathogen Genome Databases: Targets for Drugs, Vaccines and Diagnostics.* David highlighted the development of new algorithms for comparative genomic analysis, and databases enabling the integration and mining of diverse large-scale post-genomics datasets, especially in the context of a variety of pathogen genome databases his team has developed. Sridhar Hannenhalli gave a talk on *Deciphering Gene Regulatory Networks by in silico Approaches*. Sridhar outlined various computational problems pertaining to transcriptional regulation, namely, (1) representation and identification of transcription factor binding sites, (2) PolII promoter prediction, (3) Predicting interaction among transcription factors, (4) Transcriptional modeling, i.e. identifying arrangements of TFs that co-regulate a set of transcripts.