BIOKDD 2005 Workshop Report

Srinivasan Parthasarathy
The Ohio State University
srini@cse.ohio-state.edu

Wei Wang
University of North Carolina
weiwang@cs.unc.edu

Mohammed Zaki
Rennselaer Polytechnic University
zaki@cs.rpi.edu

1. REPORT

Bioinformatics is the science of managing, mining, and interpreting information from biological entities. 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. Knowledge extracted from such analysis can be used effectively to better design new drugs, offer better medical care via diagnostic tests that combine information from multiple sources, and improve scientific and clinical practice.

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 was to encourage KDD researchers to take on the numerous challenges that Bioinformatics offers. The workshop featured an invited talk from a noted expert in the field, and the latest data mining research in bioinformatics from world class researchers. 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; RNAi and microRNA Analysis; Text mining in bioinformatics; Modeling of biochemical pathways; and Biomedical and clinical informatics.

10 papers (5 long and 5 short), out of 20 submissions, were accepted for presentation at the workshop. Each paper was reviewed by at least three members of the program committee. In some cases where there was a wide variance in reviews a fourth review was sought. Each long paper selected had at least two strong supporters and no strong detractor. Each short paper selected has at least one strong support and typically no strong detractor. As a result along with a distinguished invited talk, we were able to assemble a very exciting program.

Progress in biomedical research has reached a level that it is now critical to merge the activities of computer scientists, mathematicians, statisticians, and biomedical scientists for the purpose of addressing grand challenges in medical research. The invited talk, related to this theme, was delivered by Aidong Zhang (SUNY Buffalo) on **Computational Approaches for Bridging Genomics and Health. Dr. Zhang** discussed research issues in biomedical computing that will contribute to mid- and long-term research, development, and experimental deployment of applications supporting the integration and analysis of genomic, proteomic, and clinical data for diseases and treatment effects. Current research trends in genomics and proteomics was also discussed and the importance of identifying correlations across heterogeneous datasets drawing from clinical as well as genomic data was underscored as a key aspect to improve our knowledge of both disease detection and as well as treatment of disease.

Finding motifs in proteins are important for enormous applications in bioinformatics. In the first paper "Motif discovery for proteins using subsequence clustering" by Hardik A. Sheth and Sun Kim, the authors propose an algorithm to extract sequence motifs using pair-wise sequence alignment and sequence clustering. Applied to PROSITE sequence families, the proposed method identified comparable motifs to other methods and achieved better results in recognizing conserved motif cross non-homogenous sequences.

In the second paper, "Graphical models of residue coupling in protein families" by John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg, the authors devised a method to identify amino acid residue coupling in a protein family. They model residue coupling by an undirected graphical model where nodes represent amino acid residues and edges connect two residues if they are strongly coupled. Learning a graphical model is an NP-hard problem and in this paper, various heuristics have been proposed to speed up the learning process. The method was applied to two protein families to demonstrate its viability in detecting residue coupling.

Predicting cancer susceptibility is important for early diagnosis/treatment of cancer. In the third paper "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma", Michael Waddell, David Page, Fenghuang Zhan, Bart Barlogie and John Shaughnessy, Jr showed that single-nucleotide polymorphism (SNP) profile and support vector machine, when combined together, can build a reasonably reliable prediction model for cancer susceptibility. The experimental results are particularly encouraging since only 3000 SNPS were utilized in profiling in this paper but there are millions of known SNPs that may be profiled to build more accurate model.

DNA sequencing techniques are at the core of the genome projects where the sequences of billions of DNA base pairs need to be determined. Traditional technology utilize labeling and resource consuming steps to segment a genome into pieces, sequence each piece, and assemble the whole genome sequence. Recently sequencing-by-hybridization has received much attention because of high throughput and cost-effectiveness. The kernel of this technique is an automated sequence assembling module that derives the full genome sequence from a massive number of small sequence segments. In the fourth paper,

"Accelerating DNA sequencing-by-hybridization with noise", by Chen Chen, Dong Xin, and Jiawei Han, the authors propose a new algorithm to significantly speed up the current sequence reconstruction method. The method has been applied to human genome database and achieved significant performance improvements with almost the same degree of output accuracy.

Microarray array technology considered revolutionary in the biological domain as it allows one to study the behavior of all the genes within a cell in only one experiment. One main objective of biologists is to develop a deeper understanding of how cells regulate gene expression and other cellular tasks. However, most existing algorithms proposed to mine expression relationships alone rely on the support measure to prune the search space. This is a major shortcoming as it results in the pruning of many potentially interesting rules which have low support but high confidence. In the fifth paper "On Discovery of Maximal Confident Rules without Support Pruning in Microarray Data", Tara Mclntosh and Sanjay Chawla proposed the MaxConf algorithm which exploits the weak downward closure of confidence to directly mine for high confidence rules. They also provide a means to evaluate the biological significance of the gene relationships identified. A head to head comparison with alternative algorithms shows that MaxConf can efficiently discover more high confidence and potentially interesting rules.

In the sixth paper "A Datamining Approach to Cell Population Deconvolution from Gene Expressions using Particle Filters", Sushmita Roy, Terran Lane, and Margaret Werner-Washburne focused their study on estimation subpopulation from gene expressions. Estimation of these subpopulation proportions is important for measuring the extent of synchrony in the entire population. Based upon the gene expression specific to individual subpopulations, genes can be clustered and assigned functions. The relative abundance of the cellular subpopulations also reveals phenotypic information of mutant populations that is valuable for studies of genetic diseases such as cancer. A novel approach was developed to model a biological process that provides (i) a maximum a posteriori estimate of the subpopulations given the gene expression, (ii) stage-specific gene expression values and (iii) a gene clustering method based on their stage-specific expression. The utility of this approach was demonstrated in modeling the yeast cell-cycle.

Sequence analysis has been the core of many bioinformatics applications. In the seventh paper "siRNA Off-target Search: A Hybrid q-gram Based Filtering Approach", Wenzhong Tao and Terran Lane investigated the problem of designing highly effective and gene-specific short interfering RNA (siRNA) sequences. A critical requirement for applying RNAi process in therapeutic applications is the ability to predict and to avoid side effect interactions with unintended transcripts (mRNA). In their paper, a new siRNA Off-target Search (SOS) program is devised. This algorithm uses a hybrid, q-gram based approach, combining two filtering techniques based on overlapping and non-overlapping q-grams. Experimental study demonstrated that SOS achieves better performance than BLAST in detecting siRNA off-targets.

Genome wide protein networks have become reality in recent years due to high throughput methods for detecting protein interactions. Recent studies show that a networked representation of proteins provides a more accurate model of biological systems and processes compared to conventional pair-wise analyses. Complementary to the availability of protein networks, various graph analysis techniques have been proposed to mine these networks for pathway discovery, function assignment, and prediction of complex membership. In the eighth paper "Analysis of Protein-Protein Interaction Networks Using Random Walks", Tolga Can, Orhan Camoglu, and Ambuj K. Singh proposed using random walks on graphs for the complex/pathway membership problem and demonstrated that the random walk technique achieves similar or better accuracy with more than 1,000 times speed-up compared to the best competing technique.

In the ninth paper "Finding Cliques in Protein Interaction Networks via Transitive Closure of a Weighted Graph", Chris Ding, Xiaofeng He, and Hanchuan Peng tackled the challenge of detecting protein functional modules in protein interaction networks. This amounts to finding densely connected subgraphs. Standard methods such as cliques and k-cores produce very small subgraphs due to highly sparse connections in most protein networks. Furthermore, standard methods are not applicable on weighted protein networks. To overcome the sparsity problem, the concept of transitive closure on weighted graphs was introduced, which is based on enforcing a transitive affinity inequality on the connection weights. The resulting algorithm demonstrated promising results in yeast protein network.

The tenth paper "Boosting Performance of Bio-Entity Recognition by Combining Results from Multiple Systems" falls into the area of biomedical literature mining. While numerous algorithms have been proposed for this task, biomedical named-entity recognition remains a challenging task and an active area of research, as there is still a large accuracy gap between the best algorithms for biomedical named-entity recognition and those for general newswire named-entity recognition. The reason for such discrepancy in accuracy results is generally attributed to inadequate feature representations of individual entity recognition systems and external domain knowledge.

In order to take advantage of the rich feature representations and external domain knowledge used by different systems, Lou Si, Tapas Kanungo, and Xiangji Huang proposed a collection of biomedical named-entity recognition algorithms that combine recognition results of various recognition systems. Empirical results on the GENIA biomedical corpus showed significant improvement measured by the F score.

2. ACKNOWLEDGMENTS

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

PC Co-Chairs

Srinivasan Parthasarathy, The Ohio State University

Wei Wang, University of North Carolina

General Chair

Mohammed Zaki, Rensselaer Polytechnic Institute

International Program Committee

Raj Acharya (USA)
Srinivas Aluru (USA)
Jean-Francois Boulicaut (France)
Michael Berthold (Germany)
Chris Ding (USA)
Hakan Ferhatosmanoglu (USA)
Vasant Honavar (USA)
Melanie Hilario (Switzerland)
George Karypis (USA)
Stefan Kramer (Germany)
Jayant Haritsa (India)
Kum Hoe Tung (Singapore)
Jaewoo Kang (USA)
Mitsunori Ogihara (USA)
Yi Pan (USA)
Isidore Rigoutsos (USA)
Kotagiri Rammohanrao (Australia)
Ambuj Singh (USA)
Hannu Toivonen (Finland)
M. Vidyasagar (India)
Ke Wang (Canada)
Jason T. Wang (USA)
Jiong Yang (USA)

External Reviewers

Amol Ghoting (USA)
Keith Marsolo (USA)
Mathew Otey (USA)
Duygu Ucar (USA)