BIOKDD01: Workshop on Data Mining in Bioinformatics

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

Jason T. L. Wang
College of Computing Sciences
New Jersey Institute of Technology, Newark, NJ
jason@cis.njit.edu

Hannu T.T. Toivonen
Department of Computer Science University of Helsinki and Nokia Research Center
Helsinki, Finland
Hannu.Toivonen@cs.helsinki.fi

ABSTRACT

In this report we provide a summary of the BIOKDD01 Workshop on Data Mining in Bioinformatics, held in conjunction with the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26, 2001 at San Francisco, California, USA.

1. INTRODUCTION

Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting, and utilizing information from biological sequences and molecules. It has been mainly fueled by advances in DNA sequencing and mapping techniques. The Human Genome Project has resulted in an exponentially growing database of genetic sequences. Knowledge Discovery and Data mining (KDD) techniques will play an increasingly important role in the analysis and discovery of sequence, structure and functional patterns or models from large sequence databases. High performance techniques are also becoming central to this task.

Some of the grand challenges in bioinformatics include protein structure prediction, homology search, multiple alignment and phylogeny construction, genomic sequence analysis, gene finding and gene mapping, as well as applications in gene expression data analysis, drug discovery in pharmaceutical industry, etc. In protein structure prediction, one is interested in determining the secondary, tertiary and quaternary structure of proteins, given their amino acid sequence. Homology search aims at detecting increasingly distant homologues, i.e., proteins related by evolution from a common ancestor. Multiple alignment and phylogenetic tree construction are inter-related problems. Multiple alignment aims at aligning a whole set of sequences to determine which subsequences are conserved. This works best when a phylogenetic tree of related proteins is available. Gene finding aims at locating the genes in a DNA sequence. Finally, in gene mapping the task is to identify potential gene loci for a particular disease, typically based on genetic marker data from patients and controls.

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich, but lacks a comprehensive theory of life's organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for developing novel KDD methods. To highlight these avenues we organized the Workshop on Data Mining in Bioinformatics (BIOKDD01), held in conjunction with the ACM SIGKDD 2001 Conference.

The workshop attracted approximately 100 participants, from academia, industry and government labs., underscoring the surge of interest in this exciting and rapidly expanding field. The program of the workshop included 11 contributed papers out of 19 submissions, and two invited talks. The online proceedings for the contributed papers and invited talks is available at the workshop URL:
http://www.cs.rpi.edu/~zaki/BIOKDD01

2. INVITED TALKS

In the opening talk Bruce Shapiro (Center for Cancer Research, National Cancer Institute) talked about *Determination of RNA folding pathway functional intermediates using a massively parallel genetic algorithm*. Simulating folding pathways and determining intermediate states have proven to be important in the determination of RNA structure and function. Since simulations can be very expensive, Bruce presented a generalizable methodology that uses a massively parallel genetic algorithm that operates on thousands of RNA structures. His experiments have illuminated some novel phylogenetically conserved interactions.

In his invited talk *Shared Challenges in Data Mining and Computational Biology*, Charles Elkan (University of California, San Diego) discussed a number of fundamental research issues for a continued progress both in computational biology and in data mining: (1) learning when the important class is very rare, (2) estimating accurate probabilities, (3) measuring the statistical significance of rare predictive features, (4) feature selection where only combinations of features are predictive, (5) learning where the number of features is far higher than the number of examples, (6) deciding which examples to find labels for experimentally, when labeling examples is expensive (active learning), and (7) learning when test examples and training examples are drawn from different populations. Charles illustrated many of these issues by using the KDD Cup 2001 competition as an example.

3. CONTRIBUTED PAPERS

The contributed papers spanned four main topics including microarray expression mining, sequence assembly, protein classification and structure prediction, and sequence modeling and clustering.

3.1 Microarray Expression Mining

Simon M. Lin, Sumeer Dhar, and Rose-Mary Boustany (Duke University Medical Center) used Batten Disease as a case study for mining gene expression data (*Extracting Knowledge from Gene Expression Data: A Case Study of Batten Disease*). Batten disease refers to a group of neurodegenerative disorders and expression studies can help provide clues to the molecular mechanisms involved. They proposed a prototype KDD system to analyze massive microarray data, and were able to identify the genes implicated in a new form of Batten disease (CLN9), which does not have similarity with other variants, but shares common disease mechanisms.

Goutham Kurra, Wen Niu and, Raj Bhatnagar (University of Cincinnati) presented work that extracts gene-cores from expression data (*Mining Microarray Expression Data for Classifier Gene-Cores*). The goal of their work is to classify two acute leukemia classes using gene expression data. Utilizing class scatter metrics and heuristic search they first extract *gene-sets*, the minimal combinations of genes with discriminatory capabilities. After refinement using perceptron learning, they mine the gene-sets to discover *core* patterns. These gene-cores may reveal valuable information about inter-gene dependencies and gene functions.

Paul Pavlidis, Christopher Tang, and William S. Noble (Columbia University) presented a probabilistic approach for microarray data classification (*Classification of Genes Using Probabilistic Models of Microarray Expression Profiles*). They summarize the profile of a class of co-expressed genes using a probabilistic model similar to position-specific scoring matrices. Their model gives insight into a class' expression characteristics as well as accurate recognition.

Silvio Bicciato and Carlo Di Bello (University of Padova), and Mario Pandin and Giuseppe Didoné (Cittadella Hospital, Italy) presented work on pattern identification in microarray data using autoassociative memory neural networks (*Analysis of an associative memory neural network for pattern identification in gene expression data*). Such patterns are useful in inferring gene-networking relationships; they were able to extract relationships among different genes involved in major metabolic pathways and to relate specific genes to different classes of leukemia.

3.2 Sequence Assembly

Mark K. Goldberg, Darren T. Lim, and Malik Magdon-Ismail (Rensselaer Polytechnic Institute) presented a learning approach to DNA shotgun sequencing, or the String Assembly Problem (*A Learning Algorithm for String Assembly*). The novelty of their approach is in using a parameterized form of a sequencing algorithm, and a set of already sequenced DNA strands to learn the optimal sequencing parameters. With this supervised learning strategy, the sequencing algorithm can be tuned for a particular domain, such as DNA from a certain species. Experimental results by the authors show that considerable speed-up can be obtained without significant loss in accuracy.

Sun Kim (Indiana University), Li Liao, and Jean-Francois Tomb (DuPont) presented a new method for validating sequence assemblies, in particular for identifying probable misassembled regions in shotgun sequence assemblies (*A Probabilistic Approach to Sequence Assembly Validation*). Their probabilistic approach is based on three steps: (1) analysis of the distributions of randomly selected patterns in fragments, (2) using these distributions to estimate the probability that a fragment contributes to misassembly, and finally (3) computing the entropy at each base position, derived from the misassembly probabilities. Experiments with real shotgun data of *Mycoplasma genitalium* show very good promise for the approach.

3.3 Protein Structure and Classification

Mikael Huss, Henrik Boström, Lars Asker, and Joakim Cöster (Virtual Genetics Laboratory) showed how low-level properties of biological sequences, available from several on-line sources, can be combined to construct a classifier that recognizes a high-level property (*Learning to Recognize Brain Specific Proteins Based on Low-level Features from On-line Prediction Servers*). The authors addressed the largely unsolved task of predicting brain specificity of a protein, for which features were gathered from eight servers. For learning, decision tree and rule induction methods were used, with extensions covering inductive logic programming, bagging, boosting, and randomization. The experimental success of the approach hints that it might be possible to construct a whole range of classifiers for high level properties of proteins based on their low level properties.

Nitesh Chawla, Thomas E. Moore Jr., Kevin W. Bowyer, and Lawrence O. Hall (University of South Florida) together with Clayton Springer and Philip Kegelmeyer (Sandia National Laboratories) analyzed the effect of bagging in protein secondary structure prediction (*Bagging-Like Effects for Decision Trees and Neural Nets in Protein Secondary Structure Prediction*). The authors examined various ways of partitioning a large data set into smaller subsets to form a voting committee of multiple classifiers. The conclusion was that a simple disjoint partitioning, useful especially when the data does not fit in main memory, performs at least as well as standard bagging and is expected to outperform a single classifier built from all of the data.

3.4 Sequence Modeling & Clustering

Eugen C. Buehler and Lyle H. Ungar (University of Pennsylvania) described ways for modeling amino acid sequences by utilizing maximum entropy methods (*Maximum entropy methods for biological sequence modeling*). These methods allow information based on the entire history of a sequence to be considered. This is in contrast to the commonly used Markov models, whose predictions are based on some fixed number of previous emissions. Experimental results show that there is significant "long-distance" information in amino acid sequences, suggesting that maximum entropy techniques may be beneficial for biological sequence analysis and modeling.

Raymond T. Ng and Monica C. Sleumer (University of British Columbia) together with Jörg Sander (University of Alberta) studied data cleaning and data reduction problems for the purpose of clustering SAGE (Serial Analysis of Gene Expression) data (*Hierarchical cluster analysis of SAGE data for cancer profiling*). The data is extremely high dimensional, and has many errors and missing values. The authors presented several preprocessing techniques to reduce errors, restore missing data, normalize the data, and identify a relevant subset of genes (attributes) using the Wilcoxon test. Experimental studies indicate that, with the preprocessing techniques, clustering results are much better than those without the preprocessing steps. In addition to gene expression data analysis, their techniques can also be applied to high dimensional data clustering whenever some high-level categories of interest are known for the data.

Finally, Valerie Guralnik and George Karypis (University of Minnesota) presented a new approach to clustering protein sequences (*A scalable algorithm for clustering protein sequences*). Their approach is based on projecting the sequences onto a space of frequent motifs and then using a K-means based clustering algorithm to find protein clusters in that space. The authors applied their approach to three different data sets containing sequences from the SWISS-PROT database. Experimental results show that this approach is promising and leads to reasonably good clusters.

Acknowledgments

We would like to thank all the invited speakers, authors and participants for contributing to the success of the workshop. Special thanks are due to the program committee for their support and help in reviewing the submissions.

About the Authors

Mohammed J. Zaki is an assistant professor of Computer Science at Rensselaer Polytechnic Institute. He received his M.S. and Ph.D. degrees in computer science, both from the University of Rochester in May 1995 and July 1998, respectively. His research interests include the design of efficient, scalable, interactive and parallel algorithms for various data mining tasks. He is specially interested in developing novel data mining techniques for applications in bioinformatics and web mining. He has co-edited 3 books and published over 50 papers on data mining and its applications. He received an NSF CAREER Award (2001) for his work on "Application-Oriented Large-Scale Parallel Data Mining."

Hannu T.T. Toivonen is a professor of computer science at the University of Helsinki and a principal scientist at Nokia Research Center, Helsinki, Finland. He received his M.Sc. and Ph.D. degrees in computer science from the University of Helsinki in 1991 and 1996, respectively. His research interests include data mining and computational methods for data analysis, with applications in genetics, ecology, and mobile communications. He has published over 50 papers on data mining and analysis, and coauthored the Best Applied Research Award paper in KDD-98.

Jason T.L. Wang received the B.S. degree in mathematics from National Taiwan University, and the Ph.D. degree in computer science from the Courant Institute of Mathematical Sciences, New York University, in 1991. He is a full professor of computer science in the College of Computing Sciences at New Jersey Institute of Technology and director of the university's Data and Knowledge Engineering Laboratory. Dr. Wang's research interests include data mining and databases, pattern recognition, bioinformatics, and Web information retrieval. He is coauthor of over 100 papers and 2 books, entitled *Pattern Discovery in Biomolecular Data* (Oxford University Press, 1999) and *Mining the World Wide Web* (Kluwer Academic, 2001), respectively.

Program Committee

- Chuck Baldwin, Lawrence Livermore National Labs
- Chris Bystroff, Rensselaer Polytechnic Institute
- Shi-Kuo Chang, University of Pittsburgh
- Wesley W. Chu, University of California, Los Angeles
- Diane J. Cook, University of Texas at Arlington
- Charles Elkan, University of California, San Diego
- Janice Glasgow, Queen's University, Canada
- Richard Hughey, University of California, Santa Cruz
- Hasan Jamil, Mississippi State University
- Minoru Kanehisa, Kyoto University
- Simon M. Lin, Duke University Medical Center
- Jacob V. Maizel, Jr., National Institutes of Health
- Sharad Mehrotra, University of California at Irvine
- Shinichi Morishita, University of Tokyo
- Jane Richardson, Duke University
- Isidore Rigoutsos, IBM T.J. Watson Research Center
- Bruce Shapiro, National Institutes of Health
- Vassilis J. Tsotras, University of California, Riverside
- Alex Tuzhilin, New York University
- Jeff Vitter, Duke University
- Cathy H. Wu, Georgetown University Medical Center
- Michael Zucker, Rensselaer Polytechnic Institute