BIOKDD 2002: Recent Advances in Data Mining for Bioinformatics

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

Jason T. L. Wang
Computer Science Department
New Jersey Institute of Technology
Newark, NJ 07102, USA
wangj@njit.edu

Hannu T. T. Toivonen
Computer Science Department
FIN-00014, University of Helsinki
Finland
Hannu.Toivonen@cs.helsinki.fi

1. FOREWORD

Bioinformatics provides opportunities for developing novel data mining methods. 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. Following the greatly successful 1st BIOKDD Workshop [1], the 2nd BIOKDD Workshop on Data Mining in Bioinformatics was held in July 2002, at Edmonton, Canada, in conjunction with 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

The workshop aimed to present latest results in this important area at the intersection of biology and KDD. The workshop proceedings contained 10 papers (out of 16 submissions) that were accepted for presentation at the workshop. Each paper was reviewed by two or three members of the program committee. Along with one invited speaker, we were able to assemble a very exciting program. A brief overview of the invited talk and contributed papers appears below.

Online proceedings of both the BIOKDD workshops are accessible at:

- http://www.cs.rpi.edu/∼zaki/BIOKDD01/
- http://www.cs.rpi.edu/∼zaki/BIOKDD02/

A book entitled, "Data Mining in Bioinformatics," co-edited by Jason Wang, Mohammed Zaki, Hannu Toivonen and Dennis Shasha, will be published in late 2003 by Springer-Verlag London Ltd. The book will be based on selected papers from the previous BIOKDD workshops and several invited chapters from leading authorities. It will be an invaluable resource for researchers and practitioners in this exciting field.

2. INVITED TALK

In the opening talk *How Can Data Mining Help Bio-Data Analysis*, Jiawei Han (University of Illinois at Urbana-Champaign) listed a number of data mining themes that are relevant to bio-data analysis:

- data cleaning, data preprocessing, and semantic integration of heterogeneous, distributed bio-medical databases,
- exploration of existing data mining tools for bio-data analysis,
- similarity search and comparison in bio-data,
- association analysis: identification of co-occurring bio-sequences or other correlated patterns,
- frequent pattern-based cluster analysis,
- path analysis: linking genes or proteins to different stages of disease development,
- data visualization and visual data mining, and
- privacy preserving mining of bio-medical data.

Jiawei showed many examples of applying existing data mining techniques to analyzing a wide variety of bio-data.

3. STRUCTURE MINING

In the first paper, Jingjing Hu, Xiaolan Shen, Yu Shao, Chris Bystroff, and Mohammed Zaki (Rensselaer Polytechnic Institute), presented some mining tasks for contact maps

(*Mining Protein Contact Maps*). Contact maps are binary matrices that note the contact/non-contact between all pairs of amino acids in a protein. They focused on two problems: to discover the frequent 2D patterns of contacts/non-contacts that occur in real proteins, and to cluster the mined patterns into distinct groups. They discussed how the mined clusters can help in improving the prediction of contact maps for proteins whose structure is not known.

In the next paper, Mukund Deshpande, Michihiro Kuramochi, and George Karypis (University of Minnesota), describe feature mining techniques for classifying structures (*Automated Approaches for Classifying Structures*). They propose an approach that first mines discriminating substructures for different classes, and then uses them as features for classifiers. They applied the approach to a chemical compound dataset.

In the third paper, Li Liao, Jean-Francois Tomb (DuPont Experimental Station), Sen Zhang, and Jason Wang (New Jersey Institute of Technology), present some preliminary results on classifying and clustering tasks in metabolic pathway analysis (*Clustering and Classifying Enzymes in Metabolic Pathways: Some Preliminary Results*). They observed that pathways co-occurring in many organisms tend to have common enzymes; they propose new tree matching techniques for clustering the pathways and classifying the enzymes.

In the last paper of the session, Steven Eschrich, Nitesh Chawla, and Larry Hall (University of South Florida), develop generalization techniques for mining protein structure data and high throughput drug discovery data (*Generalization Methods in Bioinformatics*). They show that an ensemble classifier, combining classifiers on small subsamples, can yield better results than a single classifier on the entire dataset. They show the effectiveness of the proposed approaches on the KDDCup 2001 drug discovery dataset, and on protein secondary structure prediction data.

4. DATA MINING AND DRUG DESIGN

In the first paper of the session, George Forman (Hewlett-Packard Labs) proposed active, incremental classifier learning for problems where the goal is to identify positive cases with minimal cost (*Incremental Machine Learning to Reduce Biochemistry Lab Costs in the Search for Drug Discovery*). At each step, a classifier is trained with the known examples. It is then used to select the most likely positive example from the unknown ones, to be analyzed chemically and added to the set of known examples. Experimental results with 2001 KDD Cup data demonstrate the power of the approach.

In the next paper, Huma Lodhi and Yike Guo (Imperial College, University of London) addressed structure activity relationship analysis (*Gram-Schmidt Kernels Applied to Structure Activity Analysis for Drug Design*). They apply Gram-Schmidt kernel and support vector machines to predict the inhibition of dihydrofolate reductase by pyrimidines or by triazines. Their experimental results show great promise.

In the last paper of the session, John L. Pfaltz and Christopher M. Taylor (University of Virginia) used closed sets to discover deterministic relationships, with applications in biological data (*Closed Set Mining of Biological Data*). The main contribution is an algorithm that incrementally creates the closure lattice, so the results are applicable widely outside biology, too.

5. GENE EXPRESSION

In the first paper of the session, Abdelghani Bellaachia, David Portnoy (George Washington University) and Yidong Chen, Abdel G. Elkahloun (National Institutes of Health) presented techniques for improving the cluster affinity search technique (CAST) used in clustering gene expression data (*E-CAST: A Data Mining Algorithm for Gene Expression Data*). The authors showed experimentally that their new approach, called E-CAST, outperforms previously published methods on gene expression data from melanoma, pheochromocytoma and brain cell tissue samples generated using micro-array technology.

In the next paper, Li Zhang, Aidong Zhang, and Murali Ramanathan (State University of New York at Buffalo) expanded their previous VizCluster technique to classify multiple types of samples (*Visualized Classification of Multiple Sample Types*). VizCluster combines the merits of both high dimensional projection scatter plot and parallel coordinate plot, and takes advantage of graphical visualization methods for pattern discovery. Their experimental results demonstrated the feasibility and usefulness of the proposed approach for gene expression datasets.

In the final paper, Jessica M. Phan, Raymond Ng, Man Saint Yuen (University of British Columbia), and Steve Jones (British Columbia Genome Sequence Center) presented the gene expression analyzer (GEA) for performing cluster analysis on gene expression data (*GEA: A Toolkit for Gene Expression Analysis*). In contrast to other methods, GEA provides (i) algebraic operators for manipulating the data, and (ii) facilities to help identify candidate genes for further clinical analysis. Furthermore, GEA is optimized to handle the high dimensionality of the data.

6. ACKNOWLEDGMENT

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 and external referees for help in reviewing the submissions.

7. WORKSHOP CO-CHAIRS

Mohammed J. Zaki, Rensselaer Polytechnic Institute, USA
Jason T.L. Wang, New Jersey Institute of Technology, USA
Hannu T.T. Toivonen, University of Helsinki, Finland

8. PROGRAM COMMITTEE

Henrik Bostrom, Stockholm University, Sweden
Julie Dickerson, Iowa State University, USA
Mark Embrechts, Rensselaer Polytechnic Institute, USA
Hasan Jamil, Mississippi State University, USA
George Karypis, University of Minnesota, USA
Sun Kim, Indiana University, USA
Simon Lin, Duke University, USA
Hiroshi Mamitsuka, Kyoto University, Japan
Shinichi Morishita, University of Tokyo, Japan
William S. Noble, Columbia University, USA
Zoran Obradovic, Temple University, USA
David Page, University of Wisconsin, USA
Laxmi Parida, IBM T.J. Watson Research Center, USA
Srinivasan Parthasarathy, Ohio State University, USA
William H. Piel, University at Buffalo, USA

Joerg Sander, University of Alberta, Canada, USA
Bruce Shapiro, National Cancer Institute, USA
Limsoon Wong, Labs for IT, Singapore
Cathy Wu, Georgetown University Medical Center, USA
Aidong Zhang, University at Buffalo, USA

9. ABOUT THE AUTHORS

Mohammed J. Zaki is an assistant professor of Computer Science at RPI. He received his Ph.D. degree in computer science from the University of Rochester in 1998. His research interests include the design of efficient, scalable, and parallel algorithms for various data mining techniques and he is specially interested in developing novel data mining techniques for bioinformatics. Dr. Zaki has published over 70 papers on data mining, he has co-edited 5 books, and served as guest-editor for *Information Systems* (special issue on bioinformatics and biological data mining), *SIGKDD Explorations* (special issue on online, interactive, anytime data mining), and *Distributed and Parallel Databases: An International Journal* (special issue on parallel and distributed data mining). He is the founding co-chair for the ACM SIGKDD Workshop on Data Mining in Bioinformatics (2001, 2002), and has co-chaired several workshops on High Performance Data Mining. He is on the program committee of more than 40 conferences and workshops. He received the prestigious National Science Foundation Career Award in 2001 and the Department of Energy Early Career Principal Investigator Award in 2002.

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 refereed papers and 2 books, entitled *Pattern Discovery in Biomolecular Data* (Oxford University Press, 1999) and *Mining the World Wide Web* (Kluwer Academic, 2001), respectively, and on the editorial boards of 5 journals including *Information Systems*, *Information Sciences*, *Knowledge and Information Systems*, *Intelligent Data Analysis*, and *Pattern Recognition*.

10. 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.