Guest Editors' Introduction: Special Issue on Mining Biological Data

Wei Wang and Jiong Yang

MINING biological data is an emerging area of intersection between data mining and bioinformatics. Bioinformaticians have been working on the research and development of computational methodologies and tools for expanding the use of biological, medical, behavioral, or health-related data. Data mining researchers have been making substantial contribution to the development of models and algorithms to meet challenges posed by the bioinformatics research. Some successful examples are frequent pattern discovery on biological molecules, text mining in biomedical literature, information integration, probabilistic modeling of genome sequences, etc. This special issue of the *IEEE Transactions on Knowledge and Data Engineering* features a collection of 11 papers, selected from 54 submissions, representing recent advances at the frontier of mining biological data.

Mining frequent trees is very useful in bioinformatics applications. The first paper, "Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications" by Mohammed J. Zaki, formulates the problem of mining (embedded) subtrees in a forest of rooted, labeled, and ordered trees. The author presents TreeMiner, a novel algorithm to discover all frequent subtrees in a forest, using a new data structure called scope-list. TreeMiner has been proven to be superior to previous methods such as PatternMatcher and has shown promising results in analyzing RNA structure and phylogenetics data sets.

In the second paper, "Frequent Substructure-Based Approaches for Classifying Chemical Compounds" by Mukund Deshpande, Michihiro Kuramochi, Nikil Wale, and George Karypis, the authors devise a substructure-based classification algorithm that decouples the substructure discovery process from the classification model construction and uses frequent subgraph discovery algorithms to find all topological and geometric substructures present in the data set. The advantage of this approach is that, during classification model construction, all relevant substructures are available and thus allow the classifier to intelligently select the most discriminating ones. This approach is employed to build models to correctly assign chemical compounds to various classes of interests, which have many applications in pharmaceutical research and are used extensively at various phases during the drug development process.

- *W. Wang is with the Computer Science Department, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599. E-mail: weiwang@cs.unc.edu.*
- *J. Yang is with the Electrical Engineering and Computer Science Department, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106. E-mail: jiong@eecs.cwru.edu.*

For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org.

Glycans, or carbohydrate sugar chains, are regarded as the third class of biological molecules, subsequent to DNA and proteins, and the recent advent of glycome informatics has generated an increasing number of glycan structures and annotation data. Glycans play important roles in the development and functioning of multicellular organisms and their structures can be represented by labeled ordered trees. The third paper, "A Probabilistic Model for Mining Labeled Ordered Trees: Capturing Patterns in Carbohydrate Sugar Chains" by Nobuhisa Ueda, Kiyoko F. Aoki-Kinoshita, Atsuko Yamaguchi, Tatsuya Akustu, and Hiroshi Mamitsuka, proposes a probabilistic model for mining labeled ordered trees and an EM algorithm for efficient learning.

Proteins are the machinery of life. A number of techniques have been developed to classify proteins according to important features in their sequences, secondary structures, or three-dimensional structures. The fourth paper, "Finding Patterns on Protein Surfaces: Algorithms and Applications to Protein Classification" by Xiong Wang, introduces a novel approach to protein classification based on significant geometric patterns on the surface of a protein.

The binding in protein-protein interactions exhibits a kind of biochemical stability in cells, which can be described by the mathematical notion of the fixed points. In the fifth paper, "Using Fixed Point Theorems to Model the Binding in Protein-Protein Interactions" by Jinyan Li and Haiquan Li, the authors define a point as a protein motif pair consisting of two traditional protein motifs. They propose a method to discover stable motif pairs of a given function from a large protein interaction sequence data set.

With the rapid growth of articles on genomics research, it has become a challenge for biomedical researchers to access this ever-increasing quantity of information to understand the newest discovery of functions of proteins they are studying. To facilitate functional annotation of proteins by utilizing the huge amounts of biomedical literature and transforming the knowledge into easily accessible database formats, the text mining technique thus becomes essential. The sixth paper, "Literature Extraction of Protein Functions Using Sentence Pattern Mining" by Jung-Hsien Chiang and Hsu-Chun Yu, proposes the method of sentence pattern mining to extract protein functions from biomedical literature.

Identifying concepts that have already been patented is essential for undertaking new biomedical research. Traditional keyword-based search on patent databases may not be sufficient enough to retrieve all the relevant information, especially for the biomedical domain. The seventh paper, "Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web" by Sougata Mukherjea, Bhuvan Bamba, and Pankaj Kankar, presents BioPatentMiner, a system that facilitates information retrieval and knowledge discovery from biomedical patents. BioPatentMiner first identifies biological terms

and relations from the patents and then integrates the information from the patents with knowledge from biomedical ontologies to create a Semantic Web. Besides keyword search and queries linking the properties specified by one or more RDF triples, the system can determine the similarity between patents based on the associated biological terms and can discover Semantic Associations between the Web resources.

The Internet has emerged as an ever-increasing environment of multiple heterogeneous and autonomous data sources that contain relevant but overlapping information on microorganisms. Microbiologists might therefore benefit from the design of intelligent software agents that assist in navigation through this information-rich environment, together with the development of data mining tools that can aid in the discovery of new information. These applications heavily depend upon well-conditioned data samples that are correlated with multiple information sources; hence, accurate database merging operations are desirable. Information systems designed for joining the related knowledge provided by different microbial data sources are hampered by the labeling mechanism for referencing microbial strains and cultures, which suffers from syntactical variation in the practical usage of the labels, whereas, additionally, synonymy and homonymy are also known to exist among the labels. This situation is even complicated by the observation that the label equivalence knowledge is itself fragmentarily recorded over several data sources which can be suspected of providing information that might be both incomplete and incorrect. The eighth paper, "Knowledge Accumulation and Resolution of Data Inconsistencies during the Integration of Microbial Information Sources" by Peter Dawyndt, Marc Vancanneyt, Hans De Meyer, and Jean Swings, presents how extraction and integration of label equivalence information from several distributed data sources has led to the construction of a so-called integrated strain database, which helps to resolve most of the above problems.

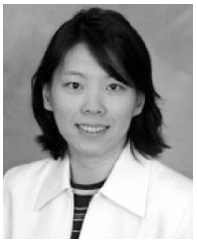
Case-based reasoning (CBR) is a suitable paradigm for class discovery in molecular biology, where the rules that define the domain knowledge are difficult to obtain and the number and the complexity of the rules affecting the problem are too large for formal knowledge representation. To further extend its capabilities, the ninth paper, "Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains" by Niloofar Arshadi and Igor Jurisica, proposes a mixture of experts for case-based reasoning (MOE4CBR), a method that combines an ensemble of CBR classifiers with spectral clustering and logistic regression. This approach not only achieves higher prediction accuracy, but also leads to the selection of a subset of features that have meaningful relationships with their class labels.

Classification based on decision trees is well studied in data mining and has applications in many fields. In recent years, database systems have become highly distributed and distributed system paradigms such as federated and peer-to-peer databases are being adopted. The tenth paper, "Hierarchical Decision Tree Induction in Distributed Genomic Databases" by Amir Bar-Or, Daniel Keren, Assaf Schuster, and Ran Wolff, considers the problem of inducing decision trees in a large distributed network of genomic databases. This work is motivated by the existence of distributed databases in healthcare and in bioinformatics and by the emergence of systems which automatically analyze these databases and by the expectancy that these databases will soon contain large amounts of highly dimensional genomic data. The authors present an algorithm that sharply reduces the communication overhead by sending just a fraction of the statistical data and demonstrate the communication efficiency and scalability of the algorithm.

Translation initiation sites (TISs) are important signals in cDNA sequences. Many research efforts have tried to predict TISs in cDNA sequences. In the eleventh paper, "Translation Initiation Sites Prediction with Mixture Gaussian Models in Human cDNA Sequences," Guoliang Li, Tze-Yun Leong, and Louxin Zhang propose using mixture Gaussian models for TIS prediction. Using both local features and some features generated from global measures, the proposed method predicts TISs with sensitivity 98 percent and specificity 93.6 percent. This method outperforms many other existing methods in sensitivity while keeping specificity high.

In closing, we would like to thank the authors for their high-quality contributions to this special issue and the referees for their generous help and valuable comments and suggestions. We also appreciate Philip Yu and Xindong Wu for offering the opportunity to publish this special issue.

Wei Wang
Jiong Yang
Guest Editors



Wei Wang received the MS degree from the State University of New York at Binghamton in 1995 and the PhD degree in computer science from the University of California at Los Angeles in 1999. She is an assistant professor in the Department of Computer Science and a member of the Carolina Center for Genomic Sciences at the University of North Carolina (UNC) at Chapel Hill. She was a research staff member at the IBM T.J. Watson Research Center between 1999 and 2002. Dr. Wang's research interests include data mining, bioinformatics, and databases. She has filed seven patents and has published one monograph and more than 70 research papers in international journals and major peer-reviewed conference proceedings. Dr. Wang received the IBM Invention Achievement Awards in 2000 and 2001. She was the recipient of a UNC Junior Faculty Development Award in 2003 and a US National Science Foundation Faculty Early Career Development (CAREER) Award in 2005. She was named a Microsoft Research New Faculty Fellow in 2005. Dr. Wang is an associate editor of the *IEEE Transactions on Knowledge and Data Engineering* and an editorial board member of the *Journal of Data Management*. She serves on the program committees of prestigious international conferences such as ACM SIGMOD, ACM SIGKDD, VLDB, ICDE, ACM CIKM, IEEE ICDM, and SSDBM.

Jiong Yang received the BS degree from the Electrical Engineering and Computer Science Department at the University of California at Berkeley in 1994, and the MS and PhD from the Computer Science Department at the University of California at Los Angeles in 1996 and 1999, respectively. He is a Schroeder Assistant Professor in the Department of Electrical Engineering and Computer Science, an assistant professor in the Epidemiology and Biostatistics Department, and a member of the Comprehensive Cancer Center at the Case Western Reserve University. Dr. Yang's research interests include data mining, bioinformatics, networking, and database systems. He has published one monograph and more than 50 research papers in various top peer-reviewed international journals and conference proceedings. Dr. Yang also has given tutorials on data mining and bioinformatics at top international conferences. He has served on the programming committees of various prestigious international conferences and workshops.