Data Mining in Bioinformatics

Jinyan Li, *Institute for Infocomm Research*

Limsoon Wong, *National University of Singapore*

Qiang Yang, *Hong Kong University of Science and Technology*

Bioinformatics and data mining provide exciting and challenging research and application areas for computational science. Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Advances such as genome-sequencing initiatives, microarrays, proteomics, and functional and structural genomics have pushed the frontiers of human knowledge. In addition, data mining and machine learning have been advancing in strides in recent years, with high-impact applications from marketing to science. Although researchers have spent much effort on data mining for bioinformatics (see the sidebar), the two areas have largely been developing separately.

1541-1672/05/$20.00 © 2005 IEEE
Published by the IEEE Computer Society

Bridging the gap

This special issue aims to bridge the gap between bioinformatics and data mining by presenting research integrating the two. We believe that data mining will provide the necessary tools for better understanding of gene expression, drug design, and other emerging problems in genomics and proteomics. For this special issue, we encouraged papers that propose novel data mining techniques for tasks such as

- gene expression analysis,
- searching and understanding of protein mass spectroscopy data,
- 3D structural and functional analysis and mining of DNA and protein sequences for structural and functional motifs, drug design, and understanding of the origins of life, and
- text mining for biological knowledge discovery.

The large number of submitted papers (24) supports our belief that data mining for bioinformatics is an area of intense interest. We're quite pleased with the quality of the five accepted papers, and we're especially happy to see this special issue come out in a timely fashion.

Five different views

Each accepted article deals with a different type of biomedical data.

In "Choosing the Optimal Hidden Markov Model for Secondary-Structure Prediction," Juliette Martin, Jean-François Gibrat, and François Rodolphe handle the secondary structural data of proteins. Their approach assumes that a model is good if it can achieve the best compromise between the number of parameters and prediction accuracy.

In "Finding Protein Domain Boundaries: An Automated, Non-Homology-Based Method," Brian Gurbaxani and Parag Mallick mine protein sequence data to reveal subtle variations of the sequences' amino acid composition. They have developed a Bayesian algorithm that identifies structural domains in proteins by cataloging the occurrence of groups of amino acids.

In "Building Innovative Representations of DNA Sequences to Facilitate Gene Finding," Jianbo Gao, Yinhe Cao, Yan Qi, and Jing Hu propose a way to determine the best discrimination of noncoding and coding regions of genomic DNA sequences. To do this, their approach devises two codon indices based on a new representation of DNA sequences.

In "MicroCluster: Efficient Deterministic Biclustering of Microarray Data," Lizhuang Zhao and Mohammed Zaki present their algorithm for mining gene expression data. MicroCluster first constructs a range multigraph from the microarray data and then searches for constrained maximal cliques to get all qualified biclusters (a bicluster is a set of genes and samples arranged in a matrix). This method can discover arbitrarily positioned and overlapping clusters of genetic data.

Finally, in "Using Semantic Dependencies to Mine Depressive Symptoms from Consultation Records," Chung-Hsien Wu, Liang-Chih Yu, and Fong-Lin Jang mine depressive symptoms from psychiatric-consultation records (text data). To discover the symptoms, their framework integrates a sentence's semantic dependencies and the strength of

Previous Work on Data Mining in Bioinformatics

Previous applications of data mining and machine learning to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein subcellular location prediction.

For example, we've used microarray technologies to predict a patient's outcome.[1] On the basis of patients' genotypic microarray data, our algorithm estimated their survival time and risk of tumor metastasis or recurrence. The results show that accurate prediction could help provide better treatment. The algorithm applies a support vector machine (SVM) method using some extreme cases. This shows that selecting only extreme patient samples for training can effectively improve prediction accuracy for different gene selection methods.

Yan Fu and his colleagues have applied machine learning to peptide identification through mass spectroscopy.[2] Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable. Fu and his colleagues used the correlative information to improve peptide identification accuracy through kernel computation. They have also applied a block-based SVM method to perform protein homology prediction, with good results.[3]

Soumya Ray and Mark Craven have applied machine learning to information extraction in text. They developed a method that automatically extracts information from text in biomedical research articles. To do this, it annotates a given protein with codes from the Gene Ontology project, using text from a biomedical research article.[4] Their method uses statistical methods to learn n-gram models for each gene ontology code and uses these models to hypothesize annotations.

References

1. H. Liu, J. Li, and L. Wong, "Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data," *Bioinformatics*, vol. 21, no. 16, 2005, pp. 3377–3384.
2. Yan Fu et al., "Exploiting the Kernel Trick to Correlate Fragment Ions for Peptide Identification via Tandem Mass Spectrometry," *Bioinformatics*, vol. 20, no. 12, 2004, pp. 1948–1954.
3. Yan Fu et al., "A Block-Based Support Vector Machine Approach to the Protein Homology Prediction Task in KDD Cup 2004," *ACM SIGKDD Explorations*, vol. 6, no. 2, 2004, pp. 120–124.
4. S. Ray and M. Craven, "Learning Statistical Models for Annotating Proteins with Function Information Using Biomedical Text," *BMC Bioinformatics*, vol. 6, suppl. 1, 2005, p. S18.

The Authors

Jinyan Li is the lead scientist and head of the Institute for Infocomm Research's Data Mining Lab. His research interests are machine learning, data mining, emerging patterns, bioinformatics (gene expression and proteomic-profiling-data analysis), clinical-data analysis, and decision systems. He received his PhD in computer science and software engineering from the University of Melbourne. Contact him at the Inst. for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613; jinyan@i2r.a-star.edu.sg; http://research.i2r.a-star.edu.sg/jinyan.

Limsoon Wong is a professor in the National University of Singapore's School of Computing and Faculty of Medicine. He works mostly on knowledge discovery technologies, especially their application to biomedicine. He serves on the editorial boards of several journals, including *Bioinformatics*, the *Journal of Bioinformatics and Computational Biology*, and *Drug Discovery Today*. He received his PhD in computer and information science from the University of Pennsylvania. Contact him at the School of Computing, Nat'l Univ. of Singapore, 3 Science Dr. 2, Bldg. S16, Room 06-05, Singapore 117543; wongls@comp.nus.edu.sg; www.comp.nus.edu.sg/~wongls.

Qiang Yang is a faculty member in the Hong Kong University of Science and Technology's Department of Computer Science. His research interests are AI planning, machine learning, case-based reasoning, and data mining. He's supported by Hong Kong Research Grants Council Central Allocation Grant HKUST CA 03/04 EG.01. He received his PhD from the University of Maryland, College Park. He's a senior member of the IEEE. Contact him at the Dept. of Computer Science, Hong Kong Univ. of Science and Technology, Clearwater Bay, Kowloon, Hong Kong; qyang@cs.ust.hk; www.cs.ust.hk/~qyang.

the lexical cohesion between sentences. It also uses a domain ontology to mine relations between the extracted symptoms.

Among the five accepted articles, gene finding, protein domain recognition, and secondary-structure prediction can be considered classic topics in bioinformatics. Microarray data analysis has become a hot topic in the past few years. Using ontologies to mine disease symptoms is, however, a relatively new topic. This constant broadening to include new topics and techniques makes bioinformatics a vibrant research area. Other emerging topics include noncoding genes, regulatory interactions, and cellular-level simulations. We hope to see another special issue of *IEEE Intelligent Systems* introducing these other exciting new topics in bioinformatics research. ■