# Guest Editors' Introduction to the Special Issue: Machine Learning for Bioinformatics—Part 1

Charles X. Ling, William Stafford Noble, and Qiang Yang

In recent years, rapid developments in genomics and proteomics have generated a large amount of data. Often, drawing conclusions from these data requires sophisticated computational analyses. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data.

A particularly active area of research in bioinformatics is the application and development of machine learning techniques to biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. Each of these tasks can be framed as a problem in machine learning. We therefore see a great potential to increase the interaction between machine learning and bioinformatics.

This special issue is aimed at facilitating that interaction. We believe that machine learning can provide powerful tools for analyzing, predicting, and understanding data from emerging genomic and proteomic technologies. The papers submitted to this special issue provide strong evidence that this is the case.

In total, more than 50 papers were submitted to the special issue. After extensive reviews and revisions, 13 papers were accepted into the special issue, which will be published in two parts. The quality and significance of the accepted papers are very high and the papers cover a wide variety of topics. Below, we provide a summary of the papers published in this journal issue.

- In the context of gene expression analysis, effective attribute selection and classification methods have been a focus of study in recent years. In the paper "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data" by Wai-Ho Au, Keith C.C. Chan, Andrew K.C. Wong, and Yang Wang, a new attribute clustering-based method for attribute selection and classification is proposed. The genes (attributes) clustered in the same groups are more correlated with one another, while the genes in different clusters are not. Top-ranked genes from each cluster are then chosen to build classifiers for prediction. Experiments show that this clustering-based gene selection and classification method predicts better (often much better) than using other attribute selection methods or using the entire gene set on two real-world cancer gene expression data sets.
- The paper "Joint Classification and Pairing of Human Chromosomes" by Pravesh Biyani, Xialoin Wu, and Abhijit Sinha addresses the problem of identifying and pairing human chromosomes from karyotype images. They propose heuristic methods for solving this NP-hard problem and demonstrate the validity of their approach using a large, clinical data set.
- In "Semisupervised Learning for Molecular Profiling," Cesare Furlanello, Maria Serafini, Stefano Merler, and Giuseppe Jurman explore a novel approach to identifying patterns in microarray expression data. They exploit the two-level cross-validation structure of the typical expression classifier experiment and show how the classifier validation runs in which a particular sample that is *not* included can be used to identify special, borderline, or outlier samples.
- The paper "Essential Latent Knowledge for Protein-Protein Interactions: Analysis by an Unsupervised Learning Approach" by Hiroshi Mamitsuka discusses a new probabilistic model for modeling protein-protein interactions. An EM-based algorithm is presented for learning the model.
- The paper "Markov Encoding for Detecting Signals in Genomic Sequences" by Jagath C. Rajapakse and Loi Sy Ho addresses how to use a Markov encoding of genomic sequences to enable a neural network to predict higher-order features in the sequences and has shown promising results.
- "The Latent Process Decomposition of cDNA Microarray Data Sets" by Simon Rogers, Mark Girolami, Colin Campbell, and Rainer Breitling presents a probabilistic clustering scheme for interpreting large expression data sets. Using a common set of latent variables, the method objectively assesses the number of classes in the data while allowing a given example to be assigned probabilistically to multiple classes.
- The paper "Fold Recognition by Predicted Alignment Accuracy" by Jinbo Xu deals with a template-ranking problem in a protein sequence using a support vector regression method which is shown to outperform the traditional Z-score-based method.

- *C.X. Ling is with the Department of Computer Science, The University of Western Ontario, Ontario, London ON N6A 5B7 Canada. E-mail: cling@csd.uwo.ca.*
- *W.S. Noble is with the Department of Genome Sciences, University of Washington, Box 357730, 1705 NE Pacific St #K-357, Seattle WA 98195-7730. E-mail: noble@gs.washington.edu.*
- *Q. Yang is with the Department of Computer Science, Hong Kong UST, Hong Kong. E-mail: qyang@cs.ust.hk.*

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

1545-5963/05/$20.00 © 2005 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

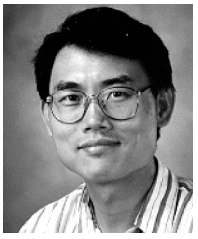
- The paper "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data" by Li Shen and Eng Chong Tan presents a method for dimensionality reduction using regression for cancer classification data. Dimensionality reduction is an important task in many bioinformatics problems due to the sheer size of the data.

## ACKNOWLEDGMENTS

Qiang Yang would like to thank Hong Kong RGC 6187/04E and CA03/04.EG01.HKBU2/03C for their support.

**Charles X. Ling** received the Msc and PhD degrees from the Department of Computer Science at the University of Pennsylvania in 1987 and 1989, respectively. Since then, he has been a faculty member in computer science at the University of Western Ontario. See http://www.csd.uwo.ca/faculty/cling for more information. His main research areas include machine learning (theory, algorithms, and its applications in Internet, e-business, and bioinformatics). He has published more than 80 research papers in journals (such as *Machine Learning*, the *Journal of Artificial Intelligence Research*, and *IEEE Transactions on Knowledge and Data Engineering*) and international conferences (such as IJCAI, and ICML).

**William Stafford Noble** received the PhD degree in computer science and cognitive science from the University of California, San Diego in 1998, where he studied with Charles Elkan. He then spent one year as a Sloan/DOE postdoctoral fellow with David Haussler at the University of California, Santa Cruz. From 1999 to 2002, he was an assistant professor in of the Department of Computer Science at Columbia University, with a joint appointment at the Columbia Genome Center. In 2002, he joined the faculty in the Department of Genome Sciences at the University of Washington, with adjunct appointments in the Department of Computer Science and Engineering and in the Division of Medicine. He is the recipient of a US National Science Foundation CAREER award and is a Sloan Research Fellow.



**Qiang Yang** received the bachelor's degree from Peking University in 1982 and the PhD degree from the University of Maryland, College Park in 1989. He was a faculty member at the University of Waterloo and Simon Fraser University in Canada between 1989 and 2001. At Simon Fraser University, he held an NSERC Industrial Chair from 1995 to 1999. He is currently a faculty member at the Hong Kong University of Science and Technology. He is interested in machine learning and data mining, AI planning, and case-based reasoning. He has published two books and more than 100 research articles in conferences and journals. He was the conference chair of the 2001 International Conference on Case-Based Reasoning, a program cochair for the 2000 Canadian AI Conference, and a tutorial cochair of AAAI 2005 Conference. He has been a guest editor for the *IEEE Transactions on Knowledge and Data Engineering*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *IEEE Intelligent Systems, Computational Intelligence Journal*, and *Applied Intelligence Journal*. He is also on the editorial board of the *IEEE Transactions on Knowledge and Data Engineering*, the *Knowledge and Information Intelligence Journal*, and the *Web Intelligence Journal*.