Pattern Recognition in Mining High-throughput Genomics/Proteomics Data: 
The New Challenges in Old Questions

Xuegong Zhang
Bioinformatics Division, TNLIST and Department of Automation, Tsinghua University
Beijing 100084, China

The current molecular biology and systems biology is featured by the rapid accumulation of high-throughput genomics and proteomics data like microarray and mass spectrometry (MS) data. Typical applications of these high-throughput biological data include the use of microarray or MS data for the molecular classification of complex human diseases like cancers, and for discovering the genes/protein markers (biomarkers) underlying the classification. Such tasks have been intensively studied as a new application field of pattern recognition in recent years. Almost all existing methods for classification and feature selection have been tried, and a large number of new or revised methods have been proposed. However, it is observed that these massive efforts have not significantly promoted the scientific discovery in the field. Instead, some scientists are starting to complain that they are puzzled by the too many choices for processing and mining the data, and the more and more sophisticated methods might introduce more noise than useful information to the scientific investigation. This question is becoming a focus of concern especially after some scientists have finished a systematic investigation on the reliability of microarray data in the recent MAQC (Microarray Quality Control) project and have concluded that the data themselves are ok. Instead of developing more pattern recognition methods for such applications, it is the time now to re-study the basic settings of the problems and the underlying fundamental questions.

Through our study on microarray and MS data, we have observed that the cancer classification and gene/biomarker selection task has many unique characteristics that distinguish itself from other standard pattern recognition tasks. Due to the extremely small sample size, the reliable assessment of the classification accuracy becomes a major question. For gene/biomarker selection, a key question is the significance of the selected genes/markers. Although most criteria for selecting differentially expressed genes are associated with some significance measure, the answer to this question is not available when the gene/marker selection step is wrapped with the classification procedure. We studied these questions with both simulated and real microarray and MS data. All existing classification methods only give a point estimation of error rates, no matter it is the apparent error or cross-validation error. We developed a perturbation-based method for estimating the distribution of error rates of a support vector machine classifier. The method can reveal a more complete picture of the data and the classifier, and can be used to guide the selection of models and parameters. For evaluating the statistical significance of gene lists selected by sophisticated machine learning methods, we defined the problem of rank significance of genes and developed a heuristic strategy for estimating this significance. These questions highlight two important aspects of the pattern recognition problems in high-throughput computational molecular biology. The awareness of such questions is a key for properly applying computational methods to practical data and for developing new methods that really target the scientific questions.
Brief Biography:

Dr. Xuegong Zhang earned his BS degree in Industrial Automation in 1989 and Ph.D. degree in Pattern Recognition and Intelligent Systems in 1994, both from Tsinghua University, Beijing, China. He joined the faculty of Tsinghua University, Department of Automation in 1994, where he has been an Assistant Professor, Associate Professor and Professor. Dr. Zhang worked at Harvard School of Public Health as a visiting scientist on computational biology in 2001-2002, and in Feb-March of 2006, both in the Biostatistics Department. Currently he is Professor of Pattern Recognition and Bioinformatics of Tsinghua University, and the Director of the Bioinformatics Division, Tsinghua National Laboratory for Information Science and Technology (TNLIST). His research interests include biological data mining, gene expression and regulation, alternative splicing, microRNA and RNA regulation, computational analysis of haplotypes and meiotic recombination hotspots, etc..

Email: zhangxg@tsinghua.edu.cn

Homepage: http://www.au.tsinghua.edu.cn/szll/XuegongZhangPage/XZhang_English.htm