MICROARRAY DATA MINING: RECENT ADVANCES

Gregory Piatetsky-Shapiro,
KDnuggets, USA

All organisms on Earth, except for viruses, consist of cells. Yeast, for example, has one cell, while humans have trillions of cells. All cells have a nucleus, and inside nucleus there is DNA, which encodes the "program" for making future organisms.

DNA Microarrays is a breakthrough technology [Sch95] that takes a "snapshot" of activity of many thousands of genes (potentially all genes in a cell) at the same time. It is creating a revolution in biology that has a potential to provide new and better diagnostics, lead to personalized treatments, and new cures for many diseases.

Analysis of microarrays presents a number of unique challenges for data mining [PT04a, PT04b]. Typical data mining applications in domains like banking or web, have a large number of records (thousands to millions) and much smaller number of fields (at most several hundred). In contrast, a typical microarray data analysis study may have only a small number of records (less than a hundred), while the number of fields, corresponding to the number of genes, is typically in thousands. Given the difficulty of collecting microarray samples, the number of samples is likely to remain small in many interesting cases. However, having so many fields relative to so few samples, creates a high likelihood of finding "false positives" that are due to chance – both in finding differentially expressed genes, and in building predictive models.

In this keynote talk we present new developments in microarray data analysis that help us to address these challenges. We first examine best practices in microarray data classification [PKR03], including feature selection approaches and examine the global feature selection bias.

We then discuss the technique of Gene Set Analysis [Mo03] and show how it can significantly increase the power of analysis.

Finally we present our results on global microarray data analysis, including pervasive lognormal distribution in microarray data and our proposal for the Comprehensive Microarray Test Suite.

References:

[Mo03] V. Mootha et al, PGC-1a-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, Nature Genetics, July 2003

[Sch95] M. Schena, et al Quantitative monitoring of gene expression patterns with a cDNA microarray. Science 270:467-470 (1995).

[PKR03] G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, Capturing Best Practice for Microarray Gene Expression Data Analysis, in Proceedings of KDD-2003 (ACM Conference on Knowledge Discovery and Data Mining), Washington, D.C., 2003.

[PT04a] G. Piatetsky-Shapiro and P. Tamayo, Microarray Data Mining: Facing the Challenges (PDF), , SIGKDD Explorations, Jan 2004.

[PT04b] G. Piatetsky-Shapiro and P. Tamayo, Guest Editors, SIGKDD Explorations Special Issue on Microarray Data Mining, Jan 2004
www.acm.org/sigs/sigkdd/explorations/issue5-2.htm