Microarray Data Mining: Facing the Challenges

Gregory Piatetsky-Shapiro
KDnuggets and U. Mass Lowell
gregory *at* kdnuggets.com

Pablo Tamayo
MIT / Broad Institute
tamayo *at* broad.mit.edu

1. MOLECULAR BIOLOGY AND DNA

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 has coding and non-coding segments, and coding segments, called “genes”, specify the structure of proteins, which are large molecules, like hemoglobin, that do the essential work in every organism. Practically all cells in the same organism have the same genes, but these genes can be expressed differently at different times and under different conditions. Genes make proteins in two steps. First, DNA is transcribed into messenger RNA or mRNA, which in turn is translated into proteins. The different patterns of gene expression following carefully tuned biological programs, according to tissue type, developmental stage, environment and genetic background account for the huge variety of different cells states and types. Virtually all major differences in cell state or type are correlated with changes in the mRNA levels of many genes.

2. MICROARRAYS: AN OVERVIEW

In recent years there has been an explosion in the rate of acquisition of biomedical data. Advances in molecular genetics technologies, such as DNA microarrays [1-8] allow us for the first time to obtain a "global" view of the cell. For example, we can now routinely investigate the biological molecular state of a cell measuring the simultaneous expression of tens of thousands of genes using DNA microarrays.

Different types of microarray use different technologies for measuring mRNA expression levels; detailed description of these technologies is beyond the scope of this paper. Here we will focus on the analysis of data from Affymetrix arrays, which are currently one of the most popular commercial arrays. However, the methodology for analysis of data from other arrays would be similar, but would use different technology-specific data preparation and cleaning steps.

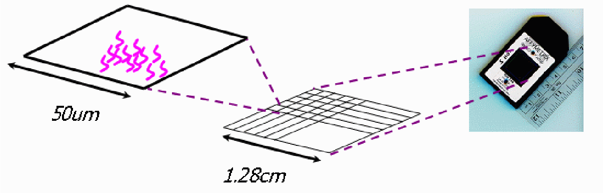
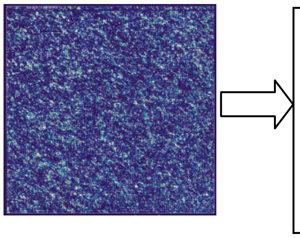


Figure 1: Affymetrix GeneChip® (right), its grid (center) and a cell in a grid (left).

This type of microarray is a silicon chip that can measure the expression levels of thousands of genes simultaneously**.** This is done by hybridizing a complex mixture of mRNAs (derived from tissue or cells) to microarrays that display probes for different genes tiled in a grid-like fashion. Hybridization events are detected using a fluorescent dye and a scanner that can detect fluorescence intensities. The scanners and associated software perform various forms of image analysis to measure and report raw gene expression values. This allows for a quantitative readout of gene expression on a gene-by-gene basis. As of 2003, there are one-chip microarrays that measure expression of over 30,000 genes, covering most of the human genome.



Gene	*Value*
D26018_at	193
D26067_at	-70
D26068_at	144
D26129_at	33

Figure 2: An example raw microarray image for a single sample (image courtesy of Affymetrix). The intensity of image on the left is translated by microarray software into numbers like the ones on the right.

Microarrays have opened the possibility of creating data sets of molecular information to represent many systems of biological or clinical interest. Gene expression profiles can be used as inputs to large-scale data analysis, for example, to serve as fingerprints to build more accurate molecular classification, to discover hidden taxonomies or to increase our understanding of normal and disease states.

The first generation of microarray analysis methodologies developed over the last 5 years has demonstrated that expression data can be used in a variety of class discovery or class prediction biomedical problems including those relevant to tumor classification [10-14]. Machine learning and statistical techniques applied to gene expression data have been used to address the questions of distinguishing tumor morphology, predicting post-treatment outcome, and finding molecular markers for disease. Today the microarray-based classification of different morphologies, lineages and cell histologies can be performed successfully in many instances. The performance in predicting treatment outcome or drug response has been more limited but some of the results are quite promising. Most results of microarray analysis still require further experimental validation and follow up study. Many current efforts are being directed in this direction. In a few cases the results of microarray analysis have found their way into more serious consideration in clinical use such as being part of clinical trials (e.g. the use of an outcome-specific, computationally selected gene marker such as PKC beta and associated inhibitor for Lymphoma treatment, [9]).

3. CHALLENGES FOR MICROARRAY DATA MINING

Analysis of microarrays presents a number of unique challenges for data mining. Typical data mining applications in domains like banking or web, have a large number of records (thousands and sometimes millions), while the number of fields is much smaller (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. We need especially robust methods to validate the models and assess their likelihood.

The main types of data analysis needed to for biomedical applications include:

- Gene Selection – in data mining terms this is a process of attribute selection, which finds the genes most strongly related to a particular class (see for example [15-21]).
- Classification – classifying diseases or predicting outcomes based on gene expression patterns, and perhaps even identifying the best treatment for given genetic signature (see for example [27-37]).
- Clustering – finding new biological classes or refining existing ones (see for example [22-26]).

Most papers in this special issue focus on these three topics, with some of the papers spanning several topics. Two papers use a number of techniques for biological discovery, and one paper presents a survey of approaches for applying machine learning to low-level analysis of microarray data.

3.1 Gene Selection

Attempts to find invariant or differential molecular behavior relevant to a given biological problem are also limited by the fact that in many cases little is known about the normal biological variation expected in a given tissue or biological state. To complicate matters a biological state is often defined only along very coarse-grained phenotypic lines. This issue of normal variation is directly addressed by V. Nadimpally and M. Zaki, where they analyze genes which show normal variance across genetically identical mice to characterize a set that could be used to identify false positives in differential gene expression studies, but also to provide insights into the processes underlying natural variation. Their analysis applied to six mouse tissues resulted in several genes which showed significant biological variations even among identical mice and provides a valuable compendium of normal variation in gene expression for mouse models.

Another approach to determining variability in small samples is taken by S. Mukherjee, P. Sykacek, S. Roberts, and S. Gurr, who propose a gene-ranking algorithm using bootstrapped P-values. This approach is especially beneficial for taking into account small-sample variability in observed values of the test statistic. They show that this method outperforms widely used two-sample T-test on artificial data and apply the method to two real datasets.

Most of the current gene selection methods in use today evaluate each gene in isolation and ignore the gene to gene correlations. From a biological perspective, however, we know that groups of genes working together as pathway components and reflecting the states of the cell are the real atomic units, or features, by which we might be more likely to predict the character or type of a particular sample and its corresponding biological state. It is these patterns of coherent gene expression that must form the input data on which sophisticated computational methods should operate (see for example [38-45]). In this context B. Hanczar, M. Courtine, A. Bennis, C. Hennegar, K. Clément, and J. Zucker propose to increase the accuracy of microarray classification by selecting appropriate "prototype" genes that represent a group of genes that share a profile and better represent the phenotypic class of interest. They present interesting results of the advantages of using prototype-based feature selection to classify adenocarcinomas.

3.2 Classification

Because the microarray dataset has many more features than records, the common statistical and machine learning procedures such as global feature selection can lead to false discoveries due to random chance. R. Simon highlights some of the common errors in identifying informative features and developing accurate classifiers, and shows the correct approach.

M. O'Connell presents a review of methods available in Insightful S+ArrayAnalyzer, which cover the full spectrum of microarray data analysis, including data preprocessing, experimental design, quality control, gene selection and differential expression analysis, classification, and clustering. (Note: S+ArrayAnalyzer includes much of the Bioconductor functionality (www.bioconductor.org), in addition to methods developed at Insightful.)

E. Bair and R. Tibshirani present a "nearest shrunken centroid" method that has been successfully used to detect clinically relevant differences in cancer patients. This method has been shown to be effective in dealing with noise inherent in microarrays and is implemented in their PAM system, available for researchers.

S. Dudoit, M. van der Laan, S. Keles, A. Molinaro, S. Sinisi, and S. Teng propose a unified loss-based methodology for estimator construction, selection, and performance assessment with cross-validation. They present new theoretical results that show that cross-validation selection can be used in intensive searches of large parameter spaces, even in finite sample situations. They also present a new D/S/A algorithm for classification, which makes Delete/Substitute/Add moves to find the best gene sets that minimize estimation error.

Other challenges to the deployment of molecular classification models come from the significant technical challenge in dealing with the variability due to the use of different technologies, platforms and heterogeneous sources of material. One would expect that different datasets representing the same biological system will display some amount of "invariant" biological characteristics independent of the idiosyncrasies or details of the sample sources, the preparation procedures and the technological platforms used to obtain the data. These invariant biological characteristics, when properly captured and exposed, can provide the basis to build more robust, general and accurate classification models. These models hopefully will be based more on reproducible biological behavior and less on biases, idiosyncrasies and technological details. The method of B. Y. M. Fung and V. T. Y. Ng to classify heterogeneous factors based on IFs (impact factors) addresses this problem. The IFs provide a way to measure the variations between individual classes in train and test samples and can be integrated into standard classifiers such as Weighted Voting or k-NN resulting in a significantly improvement in the accuracy for classifying heterogeneous samples.

3.3 Clustering and Visualization

One important clustering task is to identify groups of co-expressed genes recognize coherent expression patterns. However, the interpretation of co-expressed genes and coherent patterns strongly depends on the domain knowledge, which makes it difficult to fully automate. D. Jiang, J. Pei, and A. Zhang present an approach to this problem, based on interactive exploration of gene expression patterns. They develop a novel tool, called "coherent pattern index graph", which gives users visual feedback of strength and existence of coherent patterns.

Another aspect of clustering is finding gene networks and gene interactions. X. Wu, Y. Ye, and L. Zhang propose a graphical model based interaction analysis for this purpose. They apply graphical Gaussian model to discover pairwise gene interactions and use loglinear model to discover multi-gene interactions.

Scientific results are generally improved if we can look at the same events from different perspectives. There are many sources of biological knowledge that could be integrated into analysis of gene expression. P. Glenisson and J. Mathys explore how to combine expression data and literature extracted information to reveal biologically meaningful clusters that are not found from microarray data alone. This is an example of integrative genomics [46-47].

3.4 Biological Discovery

One of the main goals of microarray data analysis is discovery of biological knowledge, such as metabolic pathways. H. Mamitsuka, Y. Okuno, and A. Yamaguchi present an approach that builds a Markov model using the graph structure of a known pathway, and then estimates parameters using the microarray data.

A different approach is taken by M. Curran, H. Liu, F. Long, and N. Ge, who study methods for relating gene expression pattern to the pattern of transcription factor binding sites (TFBS). They use a linear model to fit the microarray data and identify potential up-regulated genes based on a specific biological hypothesis. Potential TFBS are then retrieved for the identified positive genes and randomly selected controls. Then, after removing similar binding sites, logistic regression is used to choose the best model to predict gene type (positive, control) based on the TFBS predictors.

3.5 Low-level analysis

The improvement of many analytical solutions to molecular classification problems is also conditional to the development of better and more powerful data analysis techniques, not only in high-level, but also in low level analysis. In low level analysis the focus is in providing better readouts, i.e. biological parameter and molecular probe estimates that are more accurate and faithfully reflect the actual biological state under study. B. Rubinstein, J. McAuliffe, S. Cawley, M. Palaniswami, K. Ramamohanarao, and T. Speed advocate the use of more sophisticated machine learning approaches particularly in low level analysis. An example of a low level analysis problem they addressed is the expression level summarization problem: given a probe set's intensities, after background correction and normalization, estimate the amount of target transcript present in the biological sample. They pointed out the potential of semi-supervised, heterogeneous and incremental learning for common microarray analysis settings such as those with partially labeled datasets, data from disparate domains, and sequential datasets. They identify the applicability of these machine learning approaches in other more general problems such as mass spectrometry, and fluorescence activated cell sorting.

4. Summary

Microarrays are a revolutionary new technology with great potential to provide accurate medical diagnostics, help find the right treatment and cure for many diseases and provide a detailed genome-wide molecular portrait of cellular states. The papers included in this issue are a good sample of second generation methodologies and techniques that are being used or under development today. As it can be seen from the results they are very promising and extend the possibilities of applying computational analysis and data mining to aid research in biology and medicine. We would like to close this introduction with a brief discussion to emphasize the large potential payoff of these analytical efforts but also pointing out the huge challenges ahead.

There is little doubt about the potential of computational and statistical analysis of molecular probes to improve the understanding of the cell and the possibilities of molecular medicine, Finding new insights into the molecular basis of biological processes and searching for new drugs and treatments is a problem of high complexity and where the techniques of molecular biology has been applied for many decades. The process is analogous to a large search of a few molecular entities, connections or relationships in a large sea of possibilities. One important goal of current and future computational analysis methods --short of reverse engineering the entire cell circuitry, which by the way it is still an intractable problem-- should be to reduce that search and help expose the most promising candidates (gene, proteins, drugs etc.) for further study. Current methods already succeed to some extent in exposing relationships and correlations and provide new hypothesis to be tested. Better accuracy, more robust models and estimators are clearly welcomed but sooner of later the biological interpretation of the computational or statistical results, e.g. gene selection, clustering, prediction, has to be done. For this reason one of the main challenges, besides finding relevant molecular features or building successful empirical models, is to reveal the biological mechanisms that are responsible for their success. Typically a computational researcher will apply his or her favorite algorithm to some microarray dataset and quickly obtain a voluminous set of results. These results are likely to be useful but only if they can be put in context and followed up with more detailed studies for example by a biologist or a clinical researcher. Often this follow up and interpretation is not done carefully enough because of the additional significant research involvement, the lack of domain expertise or proper collaborators, or due to the limitations of the computational analysis itself. This last is an important point as many approaches, despite being successful, left open the question regarding what the significant features or patterns mean from a biological perspective. Extracting knowledge from discovered patterns is a serious scientific bottleneck and a desirable goal of the next generation of molecular pattern recognition and data mining methods should be to provide a more integrated (e.g. along the lines of integrative genomics) and unified framework that not only builds models but also aids in the interpretation and understanding of them.

We hope that this special issue on Microarray Data Mining will make more researchers interested in the field and its challenges and will be a contribution towards realizing the potential of microarrays for biology and medicine.

5. REFERENCES

[1] Chipping Forecast 1999, 2002, The Chipping Forecast. Special Supplement. Nature Genet. 21, Jan. 1999.

[2] The Chipping Forecast II. Special Supplement. Nature Genet. 32, Dec. 2002

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

[4] DeRisi, J.L. et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686 (1997).

[5] Chu, S. et al. The transcriptional program of germ cell development in budding yeast. Science 282:699-705 (1998).

[6] Iyer, V.R. et al. The transcriptional program in the response of human fibroblasts to serum. Science 283: 83-87, (1999)

[7] DeRisi J, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996 Dec;14(4):457-60.

[8] Hegde P. et al. A concise guide to cDNA microarray analysis. Biotechniques. 2000 Sep;29(3):548-50, 552-4, 556.

[9] Shipp, M. et al 2001. "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning." Nature Medicine 8:1:68 – 74 (2002).

[10] Tamayo P. and S. Ramaswamy. "Cancer Genomics and Molecular Pattern Recognition" in Expression profiling of human tumors: diagnostic and research applications. Marc Ladanyi and William Gerald eds. Humana Press (2003).

[11] Butte A. 2002. The use and analysis of microarray data. NATURE REVIEWS DRUG DISCOVERY 1 (12): 951-960 DEC 2002.

[12] Xiang ZY et al. 2003. Microarray expression profiling: Analysis and applications. CURRENT OPINION IN DRUG DISCOVERY & DEVELOPMENT 6 (3): 384-395 MAY 2003

[13] Ramaswamy S. and T. R. Golub. DNA Microarrays in Clinical Oncology, Journal of Clinical Oncology 20, 1932-1941, 2002.

[14] Golub T.. Genome-Wide Views of Cancer, N Engl J Med 2001; 344:601-602.

[15] Marchal K et al Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays. JOURNAL OF BIOLOGICAL SYSTEMS 10 (4): 409-430 DEC (2002).

[16] Baldi P and AD Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 17: 509-519, (2001).

[17] Li C and WH Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology, (2001)/2/8/research/0032.

[18] Tusher VG et al. Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 98:5116-5121, (2001).

[19] Dudoit S et al. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12:111-139, (2002).

[20] Ideker, T. et al. Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. Journal of Computational Biology, 7, 805-817 (2000).

[21] Storey J. D. and R. Tibshirani. Statistical significance for genome wide studies. PNAS, August 5, 2003; 100(16): 9440 – 9445 (2003).

[22] Eisen M. et al. Cluster analysis and display of genome-wide expression patterns. PNAS, 95:14863-14868 (1998).

[23] Tamayo P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. PNAS, 96:2907-2912, (1999).

[24] Hastie T. et al. Supervised harvesting of expression trees. Genome Biology, 2(1) :research0003.1-0003.12, (2001).

[25] Li H and F. Hong. Cluster-Rasch models for microarray gene expression data. Genome Biology, 2(8)}:research0031.1-0031.13, (2001).

[26] Lin W. and C. Le Model-based cluster analysis of microarray gene expression data. Genome Biology, 3(2): research0009.1-0009.8, (2002).

[27] Golub T. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537, 1999.

[28] Alizadeh L. et al. Identification of clinically distinct types of diffuse large B-cell lymphoma based on gene expression patterns. Nature 403: 503-511 (2000).

[29] Bittner M. et al. Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. Nature 406: 536-540 (2000)

[30] Ramaswamy S. et al. Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures, PNAS 98: 15149-15154.

[31] Tibshirani R, et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression" PNAS 2002 99:6567-6572 (May 14).

[32] Ramaswamy S. et al. Evidence for a Molecular Signature of Metastasis in Primary Solid Tumors. Nature Genetics, vol. 33, January 2003, pp. 49-54.

[33] Khan J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, Volume 7, Number 6, June 2001.

[34] Hedenfalk I. et al. Gene Expression Profiles in Hereditary Breast Cancer. NEJM, 244:539-548. (2001).

[35] Chang HY et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol. 2004 Feb; 2(2): 1.

[36] Nutt CL. Et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 2003 Apr 1;63(7):1602-7.

[37] Lapointe J. et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. PNAS 2004 Jan 20; 101(3): 811.

[38] Cunliffe H.E. et al. The Gene Expression Response of Breast Cancer to Growth Regulators: Patterns and Correlation with Tumor Expression Profiles. Cancer Research, 63:7158-7166. (2003).

[39] Mootha VK. et al. PGC-1a Responsive Genes Involved in Oxidative Phosphorylation are Coordinately Downregulated in Human Diabetes. Nature Genet. 15 June 2003, vol. 34 no. 3 pp 267 – 273.

[40] Califano, A. et al Analysis of gene expression microarrays for phenotype classification. Proceedings of ISMB 2000.

[41] Cheng, Y and G.M. Church, Biclustering of expression data. Proceedings. of ISMB 2000.

[42] Alter, O. et al Singular value decomposition for genome-wide expression data processing and modeling. PNAS 97:10101–10106 (2000).

[43] Murali T.M. and S. Kasif. Extracting Conserved Gene Expression Motifs from Gene Expression Data. Proceedings of PSB 8:77-88(2003).

[44] Segal, E. Decomposing Gene Expression into Cellular Processes. Proceedings of PSB 8:89-100(2003).

[45] Brunet et al. Metagenes and Molecular Pattern Discovery using Matrix Factorization. PNAS 2004 (in press).

[46] Mootha et al. Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria. Cell 115: 629-640 (2003).

[47] Kohane I et al Microarrays for an Integrative Genomics

MIT Press, August 2002.