AI Methods for Analyzing Microarray Data

Amira Djebbari[1*], amira.djebbari@nrc-cnrc.gc.ca
Aedín C. Culhane[2,3], aedin@jimmy.harvard.edu
Alice J. Armstrong[4], piffle@gwu.edu
John Quackenbush[2,3], johnq@jimmy.harvard.edu

Amira Djebbari
Institute for Information Technology,
National Research Council Canada
1200 Montreal Road, Building M-50
Ottawa, ON, Canada K1A 0R6
(613) 949-0913 (phone)
(613) 952-0215(fax)
amira.djebbari@nrc-cnrc.gc.ca

Aedín C. Culhane
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
and Department of Biostatistics, Harvard School of Public Health
44 Binney St Smith 822
Boston, MA, 02115 USA
(617) 632 2468 (phone)
(617) 582 7760 (fax)
aedin@jimmy.harvard.edu

Alice J. Armstrong
Department of Computer Science, The George Washington University
801 22nd St NW Suite 704
Washington, DC, 20052 USA
(202) 994 7181 (phone)
(202) 994 4875 (fax)
piffle@gwu.edu

John Quackenbush
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
and Department of Biostatistics, Harvard School of Public Health
44 Binney St Smith 822A
Boston, MA, 02115 USA
(617) 582-8163 (phone)
(617) 632-2444 (fax)
johnq@jimmy.harvard.edu

* To whom correspondence should be addressed

AI Methods for Analyzing Microarray Data

Amira Djebbari[1*], amira.djebbari@nrc-cnrc.gc.ca
Aedín C. Culhane[2,3], aedin@jimmy.harvard.edu
Alice J. Armstrong[4], piffle@gwu.edu
John Quackenbush[2,3], johnq@jimmy.harvard.edu

[1] Institute for Information Technology, National Research Council, Ottawa, ON, Canada
[2] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Boston, MA, USA
[3] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[4] Department of Computer Science, The George Washington University, Washington, DC, USA
* To whom correspondence should be addressed

"Prediction is very difficult, especially about the future" – Niels Bohr

INTRODUCTION

Biological systems can be viewed as information management systems, with a basic instruction set stored in each cell's DNA as "genes." For most genes, their information is enabled when they are transcribed into RNA which is subsequently translated into the proteins that form much of a cell's machinery. Although details of the process for individual genes are known, more complex interactions between elements are yet to be discovered. What we do know is that diseases can result if there are changes in the genes themselves, in the proteins they encode, or if RNAs or proteins are made at the wrong time or in the wrong quantities.

Recent advances in biotechnology led to the development of DNA microarrays, which quantitatively measure the expression of thousands of genes simultaneously and provide a snapshot of a cell's response to a particular condition. Finding patterns of gene expression that provide insight into biological endpoints offers great opportunities for revolutionizing diagnostic and prognostic medicine and providing mechanistic insight in

data-driven research in the life sciences, an area with a great need for advances, given the urgency associated with diseases. However, microarray data analysis presents a number of challenges, from noisy data to the curse of dimensionality (large number of features, small number of instances) to problems with no clear solutions (*e.g.* real world mappings of genes to traits or diseases that are not yet known).

Finding patterns of gene expression in microarray data poses problems of class discovery, comparison, prediction, and network analysis which are often approached with AI methods. Many of these methods have been successfully applied to microarray data analysis in a variety of applications ranging from clustering of yeast gene expression patterns (Eisen *et al.*, 1998) to classification of different types of leukemia (Golub *et al.*, 1999). Unsupervised learning methods (*e.g.* hierarchical clustering) explore clusters in data and have been used for class discovery of distinct forms of diffuse large B-cell lymphoma (Alizadeh *et al.*, 2000). Supervised learning methods (*e.g.* artificial neural networks) utilize a previously determined mapping between biological samples and classes (*i.e.* labels) to generate models for class prediction. A k-nearest neighbor (k-NN) approach was used to train a gene expression classifier of different forms of brain tumors and its predictions were able to distinguish biopsy samples with different prognosis suggesting that microarray profiles can predict clinical outcome and direct treatment (Nutt *et al.*, 2003). Bayesian networks constructed from microarray data hold promise for elucidating the underlying biological mechanisms of disease (Friedman *et al.*, 2000).

BACKGROUND

Cells dynamically respond to their environment by changing the set and concentrations of active genes by altering the associated RNA expression. This "gene expression" is one of the main determinants of a cell's state, or phenotype. For example, we can investigate the differences between a normal cell and a cancer cell by examining their relative gene expression profiles.

Microarrays quantify gene expression levels in various conditions (such as disease *vs.* normal) or across time points. For *n* genes and *m* instances (biological samples), microarray measurements are stored in an *n* by *m* matrix where each row is a gene, each

column is a sample and each element in the matrix is the expression level of a gene in a biological sample, where samples are instances and genes are features describing those instances. Microarray data is available through many public online repositories (Table 1). In addition, the Kent-Ridge repository (http://sdmc.i2r.a-star.edu.sg/rp/) contains pre-formatted data ready to use with the well-known machine learning tool Weka (Witten & Frank, 2000).

Microarray data presents some unique challenges for AI such as a severe case of the curse of dimensionality due to the scarcity of biological samples (instances). Microarray studies typically measure tens of thousands of genes in only tens of samples. This low case to variable ratio increases the risk of detecting spurious relationships. This problem is exacerbated because microarray data contains multiple sources of within-class variability, both technical and biological. The high levels of variance and low sample size make feature selection difficult. Testing thousands of genes creates a multiple testing problem, which can result in underestimating the number of false positives. Given data with these limitations, constructing models becomes under-determined and therefore prone to over-fitting.

From biology, it is also clear that genes do not act independently. Genes interact in the form of pathways or gene regulatory networks. For this reason, we need models that can be interpreted in the context of pathways. Researchers have successfully applied AI methods to microarray data preprocessing, clustering, feature selection, classification, and network analysis.

Mining Microarray Data: Current Techniques, Challenges and Opportunities for AI

Data Preprocessing

After obtaining microarray data, normalization is performed to account for systematic measurement biases and to facilitate between-sample comparisons (Quackenbush, 2002). Microarray data may contain missing values that may be replaced by mean replacement or k-NN imputation (Troyanskaya *et al.*, 2001).

Feature Selection

The goal of feature selection is to find genes (features) that best distinguish groups of instances (*e.g.* disease *vs.* normal) to reduce the dimensionality of the dataset. Several statistical methods including t-test, significance analysis of microarrays (SAM) (Tusher *et al.*, 2001), and analysis of variance (ANOVA) have been applied to select features from microarray data.

In classification experiments, feature selection methods generally aim to identify relevant gene subsets to construct a classifier with good performance (Inza *et al.*, 2004). Features are considered to be relevant when they can affect the class; the strongly relevant are indispensable to prediction and the weakly relevant may only sometimes contribute to prediction.

Filter methods evaluate feature subsets regardless of the specific learning algorithm used. The statistical methods for feature selection discussed above as well as rankers like information gain rankers are filters for the features to be included. These methods ignore the fact that there may be redundant features (features that are highly correlated with each other and as such one can be used to replace the other) and so do not seek to find a set of features which could perform similarly with fewer variables while retaining the same predictive power (Guyon & Elisseeff, 2003). For this reason multivariate methods are more appropriate.

As an alternative, wrappers consider the learning algorithm as a black-box and use prediction accuracy to evaluate feature subsets (Kohavi & John, 1997). Wrappers are more direct than filter methods but depend on the particular learning algorithm used. The computational complexity associated with wrappers is prohibitive due to curse of dimensionality, so typically filters are used with forward selection (starting with an empty set and adding features one by one) instead of backward elimination (starting with all features and removing them one by one). Dimension reduction approaches are also used for multivariate feature selection.

Dimension Reduction Approaches

Principal component analysis (PCA) is widely used for dimension reduction in machine learning (Wall *et al.*, 2003). The idea behind PCA is quite intuitive: correlated objects can be combined to reduce data "dimensionality". Relationships between gene expression profiles in a data matrix can be expressed as a linear combination such that colinear variables are regressed onto a new set of coordinates. PCA, its underlying method Single Value Decomposition (SVD), related approaches such as correspondence analysis (COA), and multidimensional scaling (MDS) have been applied to microarray data and are reviewed by Brazma & Culhane (2005). Studies have reported that COA or other dual scaling dimension reduction approaches such as spectral map analysis may be more appropriate than PCA for decomposition of microarray data (Wouters *et al.*, 2003).

While PCA considers the variance of the whole dataset, clustering approaches examine the pairwise distance between instances or features. Therefore, these methods are complementary and are often both used in exploratory data analysis. However, difficulties in interpreting the results in terms of discrete genes limit the application of these methods.

Clustering

What we see as one disease is often a collection of disease subtypes. Class discovery aims to discover these subtypes by finding groups of instances with similar expression patterns. Hierarchical clustering is an agglomerative method which starts with a singleton and groups similar data points using some distance measure such that two data points that are most similar are grouped together in a cluster by making them children of a parent node in the tree. This process is repeated in a bottom-up fashion until all data points belong to a single cluster (corresponding to the root of the tree).

Hierarchical and other clustering approaches, including K-means, have been applied to microarray data (Causton *et al.*, 2003). Hierarchical clustering was applied to study gene expression in samples from patients with diffuse large B-cell lymphoma (DLBCL) resulting in the discovery of two subtypes of the disease. These groups were found by analyzing microarray data from biopsy samples of patients who had not been previously

treated. These patients continued to be studied after chemotherapy, and researchers found that the two newly discovered disease subtypes had different survival rates, confirming the hypothesis that the subtypes had significantly different pathologies (Alizadeh *et al.*, 2000).

While clustering simply groups the given data based on pair-wise distances, when information is known *a priori* about some or all of the data *i.e.* labels, a supervised approach can be used to obtain a classifier that can predict the label of new instances.

Classification (Supervised Learning)

The large dimensionality of microarray data means that all classification methods are susceptible to over-fitting. Several supervised approaches have been applied to microarray data including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and k-NNs among others (Hastie *et al.*, 2001).

A very challenging and clinically relevant problem is the accurate diagnosis of the primary origin of metastatic tumors. Bloom *et al.* (2004) applied ANNs to the microarray data of 21 tumor types with 88% accuracy to predict the primary site of origin of metastatic cancers with unknown origin. A classification of 84% was obtained on an independent test set with important implications for diagnosing cancer origin and directing therapy.

In a comparison of different SVM approaches, multicategory SVMs were reported to outperform other popular machine learning algorithms such as k-NNs and ANNs (Statnikov *et al.*, 2005) when applied to 11 publicly available microarray datasets related to cancer. It is worth noting that feature selection can significantly improve classification performance.

Cross-validation

Cross-validation (CV) is appropriate in microarray studies which are often limited by the number of instances (*e.g.* patient samples). In k-fold CV, the training set is divided into *k* subsets of equal size. In each iteration *k*-1 subsets are used for training and one subset is

used for testing. This process is repeated *k* times and the mean accuracy is reported. Unfortunately, some published studies have applied CV only partially, by applying CV on the creation of the prediction rule while excluding feature selection. This introduces a bias in the estimated error rates and over-estimates the classification accuracy (Simon *et al.*, 2003). As a consequence, results from many studies are controversial due to methodological flaws (Dupuy & Simon, 2007). Therefore, models must be evaluated carefully to prevent selection bias (Ambroise & McLachlan, 2002). Nested CV is recommended, with an inner CV loop to perform the tuning of the parameters and an outer CV to compute an estimate of the error (Varma & Simon, 2006).

Several studies which have examined similar biological problems have reported poor overlap in gene expression signatures. Brenton *et al.* (2005) compared two gene lists predictive of breast cancer prognosis and found only 3 genes in common. Even though the intersection of specific gene lists is poor, the highly correlated nature of microarray data means that many gene lists may have similar prediction accuracy (Ein-Dor *et al.*, 2004). Gene signatures identified from different breast cancer studies with few genes in common were shown to have comparable success in predicting patient survival (Buyse *et al.*, 2006).

Commonly used supervised learning algorithms yield black box models prompting the need for interpretable models providing insights about the underlying biological mechanism that produced the data.

Network Analysis

Bayesian networks (BNs), derived from an alliance between graph theory and probability theory, can capture dependencies among many variables (Pearl, 1988, Heckerman, 1996). Friedman *et al.* (2000) introduced a multinomial model framework for BNs to reverse-engineer networks and showed that this method differs from clustering in that it can discover gene interactions other than correlation when applied to yeast gene expression data. Spirtes *et al.* (2002) highlight some of the difficulties of applying this approach to microarray data. Nevertheless, many extensions of this research direction have been explored. Correlation is not necessarily a good predictor of interactions, and weak

interactions are essential to understand disease progression. Identifying the biologically meaningful interactions from the spurious ones is challenging, and BNs are particularly well-suited for modeling stochastic biological processes.

The exponential growth of data produced by microarray technology as well as other high-throughput data (*e.g.* protein-protein interactions) call for novel AI approaches as the paradigm shifts from a reductionist to a mechanistic systems view in the life sciences.

FUTURE TRENDS

Uncovering the underlying biological mechanisms that generate these data is harder than prediction and has the potential to have far reaching implications for understanding disease etiologies. Time series analysis (Bar-Joseph, 2004) is a first step to understanding the dynamics of gene regulation, but, eventually, we need to use the technology not only to observe gene expression data but also to direct intervention experiments (Pe'er *et al.*, 2001, Yoo *et al.*, 2002) and develop methods to investigate the fundamental problem of distinguishing correlation from causation.

CONCLUSION

We have reviewed AI methods for pre-processing, clustering, feature selection, classification and mechanistic analysis of microarray data. The clusters, gene lists, molecular fingerprints and network hypotheses produced by these approaches have already shown impact from discovering new disease subtypes and biological markers, predicting clinical outcome for directing treatment as well as unraveling gene networks. From the AI perspective, this field offers challenging problems and may have a tremendous impact on biology and medicine.

REFERENCES

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature, 403*(6769), 503-11.

Ambroise C., & McLachlan G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences,* 99(10), 6562-6.

Bar-Joseph Z. (2004). Analyzing time series gene expression data. *Bioinformatics, 20*(16), 2493-503.

Bloom G., Yang I.V., Boulware D., Kwong K.Y., Coppola D., Eschrich S., *et al.* (2004). Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology, 164*(1), 9-16.

Brenton J.D., Carey L.A., Ahmed A.A., & Caldas C. (2005). Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *Journal of Clinical Oncology, 23*(29), 7350-60.

Brazma A., & Culhane AC. (2005). Algorithms for gene expression analysis. In Jorde LB., Little PFR, Dunn MJ., Subramaniam S. (Eds.) *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.,* (3148 -3159) London: John Wiley & Sons.

Buyse, M., Loi S., Van't Veer L., Viale G., Delorenzi M., Glas A.M., *et al.* (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute, 98*, 1183-92.

Causton H.C., Quackenbush J., & Brazma A. (2003) Microarray Gene Expression Data Analysis: A Beginner's Guide. Oxford: Blackwell Science Limited.

Dupuy A., & Simon RM. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute, 99*(2), 147-57.

Ein-Dor L., Kela I., Getz G., Givol D., & Domany E. (2004). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics, 21*(2), 171-8.

Eisen M.B., Spellman P.T., Brown P.O., & Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences,* 95, 14863-14868.

Friedman N., Linial M., Nachman I., & Pe'er D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology, 7*(3-4), 601-20.

Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., *et al.* (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science, 286* (5439), 531.

Guyon, I., & Elisseff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Hastie T., Tibshirani R., & Friedman J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer Series in Statistics.

Heckerman D. (1996). A Tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-95-06. Microsoft Research.

Inza I., Larrañaga P., Blanco R., & Cerrolaza A.J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine, special issue in "Data mining in genomics and proteomics", 31*(2), 91-103.

Kohavi R., & John G.H. (1997). Wrappers for feature subset selection, *Artificial Intelligence, 97*(1-2), 273-324.

Nutt C.L., Mani D.R., Betensky R.A., Tamayo P., Cairncross J.G., Ladd C., *et al.* (2003). Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification. *Cancer Research,* 63, 1602-1607.

Pe'er D, Regev A, Elidan G, & Friedman N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics, 17 S1*, S215-24.

Pearl J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, San Mateo: Morgan Kaufmann Publishers.

Quackenbush J. (2002). Microarray data normalization and transformation, *Nature Genetics, 32*, 496–501.

Quackenbush J. (2006). Microarray Analysis and Tumor Classification. *The New England Journal of Medicine, 354*(23), 2463-72.

Simon R., Radmacher M.D., Dobbin K., & McShane L.M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute, 95*(1), 14-8.

Spirtes, P., Glymour, C., Scheines, R. Kauffman, S., Aimale, V., & Wimberly, F. (2001). Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data. *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology.*

Statnikov A., Aliferis C.F., Tsamardinos I., Hardin D., & Levy S. (2005). A comprehensive evaluation of multicategory classification methodsfor microarray gene expression cancer diagnosis. *Bioinformatics, 21*(5), 631-643

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., *et al.* (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520-5.

Tusher V.G., Tibshirani R., & Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, 98*(9), 5116-5121.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91

Witten, I. H. & Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers Inc.

Wall, M., Rechtsteiner, A., & Rocha, L. (2003). Singular value decomposition and principal component analysis. In D.P. Berrar, W. Dubitzky, M. Granzow (Eds.) *A Practical Approach to Microarray Data Analysis.* (91-109). Norwell: Kluwer.

Wouters, L., Gohlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G., & Lewi, P.J. (2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics, 59*, 1131-1139

Yoo C., Thorsson V., & Cooper G.F. (2002). Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. *Biocomputing: Proceedings of the Pacific Symposium, 7*, 498-509

TERMS AND DEFINITIONS

Microarray: A microarray is an experimental assay which measures the abundances of mRNA (intermediary between DNA and proteins) corresponding to gene expression levels in biological samples.

Curse of Dimensionality: A situation where the number of features (genes) is much larger than the number of instances (biological samples) which is known in statistics as p >> n problem.

Feature Selection: A problem of finding a subset (or subsets) of features so as to improve the performance of learning algorithms.

Supervised Learning: A learning algorithm that is given a training set consisting of feature vectors associated with class labels and whose goal is to learn a classifier that can predict the class labels of future instances.

Unsupervised Learning: A learning algorithm that tries to identify clusters based on similarity between features or between instances or both but without taking into account any prior knowledge.

Multiple testing problem: A problem that occurs when a large number of hypotheses are tested simultaneously using a user-defined α cut off p-value which may lead to rejecting a non-negligible number of null hypotheses by chance.

Over-fitting: A situation where a model learns spurious relationships and as a result can predict training data labels but not generalize to predict future data.

Table 1: Some public online repositories of microarray data

Name of the repository	URL
ArrayExpress at the European Bioinformatics Institute	http://www.ebi.ac.uk/arrayexpress/
Gene Expression Omnibus at the National Institutes of Health	http://www.ncbi.nlm.nih.gov/geo/
Stanford microarray database	http://smd.stanford.edu/
Oncomine	http://www.oncomine.org/main/index.jsp