AI Methods for Analyzing Microarray Data

Amira Djebbari
National Research Council Canada, Canada

Aedín C. Culhane
Harvard School of Public Health, USA

Alice J. Armstrong
The George Washington University, USA

John Quackenbush
Harvard School of Public Health, USA

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. Thus "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).

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

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 under-estimating 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