Journal of Informatics and Mathematical Sciences
Volume 3 (2011), Number 3, pp. 257–275

http://www.rgnpublications.com

Statistical Techniques for Microarray Technology

Bindu Punathumparambath, Sebastian George, and Kannan V.M.

Abstract. Microarray technology enable researchers to measure the expression levels for thousands of genes simultaneously. This paper describes the microarray technology and general methodology for the analysis of differential gene expression data from microarray experiments. First, we will provide a review of basic microarray concepts and an overview of some of the major statistical developments in microarray data analysis. In this work, we propose the two component Mixed Gaussian distribution as an approximation for the log-ratios of the measured gene expression across genes.

1. Introduction

Microarray technology developed in early 1990s facilitated measuring the expression level of thousands of genes in parallel. For thorough understanding of the data coming from microarray experiments, it is essential to have the knowledge of underlying biology and the setup of microarray experiment. Scientist in the computational and biological sciences together with significant contribution from statisticians are working on analysis and interpretation of huge data on gene expression levels.

Depending on the type of molecules used, microarrays are classified into different categories like cDNA microarray, protein microarray, antibody microarray etc. Microarrays are used for many purposes. In gene expression analysis the expression levels of genes of the test sample is compared with that of control sample. This is the most widely used type of microarray and is used to identify genes overexpressed or underexpressed in tissues from a patient to that of normal tissue. This will help in identifying genes related to the disease and those genes can be targeted for developing new therapeutics or for diagnosis/prognosis of the disease. In mutation analysis single base changes (single nucleotide polymorphism, SNP) in a sequence is analyzed. Single Nucleotide Polymorphisms are increasingly being implicated in drug reactions.

2010 *Mathematics Subject Classification.* 62N02, 62H30, 62P10, 92C40.
Key words and phrases. cDNA microarray; Gene expression data; mixed Guassian.

One of the important problems to be addressed in analysis of microarray data is the identification of differentially expressed gene for further investigation. Fold change is the simplest method for identifying differentially expressed genes (De Risi *et al*. [11] and Eisen *et al*. [16]. It is based on the observed ratio (or average of ratios) between two conditions. An arbitrary cut-off value (for example, 2 fold) is used to identify differentially expressed genes. The standard-t-test is usually used to identify significantly differentially expressed genes across two conditions, which assumes normally distributed data and equal variance within classes. Welch statistics for unequal variances, Wilcoxon nonparametric test, permutation test, penalized t-statistics, significance analysis of microarrays (SAM) (Tusher *et al*. [32]), regularized t-test (Cyber-T) (Baldi and Long [1]) and moderated t-statistics (Smyth [28]) are also applied for identifying differentially expressed genes in two conditions.

Potential applications of microarray technology are (1) Identification of change in expression of genes over time, between tissues, and disease states. (2) Identification of complex genetic diseases, drug discovery, toxicology studies, mutation/polymorphism detection (SNP's), and pathogen analysis, etc. Microarray methods were applied initially to disease states such as tumour classification but increasingly beeing applied to behavioural and pharmacological problems such as addiction and mental illness. Therefore, it is important that biologists and psychologists who study behaviour understand the technology and the types of issues involved in the analysis of data.

The paper is organized as follows. The cDNA microarray protocol is given in section 2. The summary of data pre-processing methods for the microarray experimental data are presented in section 3. In section 4 we discuss various classification tools. Various clustering algorithms are discussed in section 5. Different methods of analysis of comparative experiments for identifying the differential expression of the genes are discussed in section 6. Application of mixed Gaussian distribution is illustrated in section 7. Finally, we summarize in section 8.

2. **cDNA Microarray**

DNA microarray is a very powerful technology that measure the expression levels of thousands of genes in parallel. A cDNA microarray is a glass slide, to which single-stranded DNA molecules are attached at fixed locations (spots). DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription). The process of measuring gene expression via cDNA is called expression analysis or expression profiling. DNA microarray, or DNA chips are fabricated by high-speed robotics, generally on glass but sometimes on nylon substrates, thus allowing massively parallel gene expression and gene discovery studies. An experiment with a single DNA chip can provide researchers information on thousands of genes simultaneously.

Gene chips are usually categorized into one of two classes, based on the DNA actually arranged onto a support. An **Oligo** array is comprised of synthesized oligonucleotides, where a **cDNA** array contains cloned or PCR amplified complimentary DNA molecules. The purpose of either type of array is the same, specifically, to profile changes in the expression levels of many genes in a single experiment. Unlike oligonucleotide arrays, cDNA arrays contain longer cloned or PCR amplified cDNA and often the length of the cDNA species arranged onto a cDNA array varies considerably.

Also there are two types of microarrays depending on the number of channels or coloured dyes used. They are single channel microarrays or one colour microarrays and two-color microarrays or two-channel microarrays. In single-channel microarrays or one-color microarrays, the arrays are designed to give estimations of the absolute levels of gene expression. Therefore the comparison of two conditions requires two separate single-dye hybridization experiments. As only a single dye is used, the data collected represent absolute values of gene expression. These may be compared to other genes within a sample or to reference "normalizing" probes used to calibrate data across the entire array and across multiple arrays.

Two-color microarrays or two-channel microarrays are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus control tissue) and that are labeled with two different fluorophores. Fluorescent dyes commonly used for cDNA labelling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The green and red colour dye-labelled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

2.1. *The cDNA Microarray Protocol*

2.1.1. *Principle of cDNA microarray experiments.* The whole process is based on hybridization probing, a technique that uses fluorescently labeled nucleic acid molecules to identify complementary molecules, sequences that are able to base-pair with one another. Each single-stranded DNA fragment is made up of four different nucleotides, adenine (A), thymine (T), guanine (G), and cytosine (C), that are linked end to end. Adenine is the complement of, or will always pair with, thymine, and guanine is the complement of cytosine. Therefore, the complementary sequence to 5′-G-T-C-C-T-A-3′ will be 3′-C-A-G-G-A-T-5′. When two complementary sequences find each other, such as the immobilized probe DNA and

the mobile target DNA, cDNA, or mRNA, they will lock together, or hybridize. In this paper 'probe' refers to the DNA attached on to the glass slide and 'target' refers to the labelled DNA in solution.

2.1.2. *Preparing probe DNA samples for spotting.* Polymerase Chain Reaction (PCR) is used to amplify the gene or DNA. This probe DNA is attached to the glass slide. Microarrays spot probe DNA samples on a solid substrate, usually a glass slide, onto which thousands of spots of single-stranded DNA, in the form of cDNAs or oligonucleotide, are placed by a robot arrayer using contact or non contact printing methods.

2.1.3. *Preparation of target DNA.* RNA from tissues is prepared using standard protocols and mRNA isolated. cDNA synthesis from the mRNA is achieved by reverse transcription using reverse transcriptase enzyme. It is at this stage that a tag, a fluoresecently labeled nucleotide, is incorporated in the synthesized cDNA. mRNA can also be directly labeled and used in microarray.

2.1.4. *Hybridization of the target with the probe sequence on the array.* From control and test samples tissues cDNA or RNA is made and labeled with two fluorescent labels. For example, a red dye for RNA from the sample population and a green dye for that from the control population. Both extracts are mixed and washed over the microarray. Gene sequences from the extracts hybridize to their complementary sequences on the spots. If the RNA from the sample population is in abundance, the spot will be red; if the RNA from the control population is in abundance, the spot will be green; if sample and control bind equally, the spot will be yellow; if neither binds, the spot will appear black. Thus, the relative expression levels of the genes in the sample and in control populations can be estimated from the fluorescence intensities and colours of each spot.

2.1.5. *Detection of bound probe using laser.* For detection of the bound probe the spot is exposed to laser to excite the fluorescent dyes on the sample, then collect emitted light, and generate digital images of the fluorescent signal. Two general methods are used to acquire such images: laser excitation with a photomultiplier tube (PMT) detector, and filtered white-light excitation with a charge-coupled device (CCD) detector.

2.1.6. *Image analysis.* The raw data produced by microarray are in fact monochrome images. Transforming these images into gene expression matrix is a non-trivial process: the spots corresponding to genes on the microarray should be identified, their boundaries determined, the fluorescence intensity from each spot measured and compared to the background intensity and to these intensities for other channels. The software for this initial image processing is often provided with the image scanner, since it will depend on particular properties of the hardware. The ultimate goal of the image analysis is to automatically quantify

every individual array element (spot), providing information about the amount of DNA or RNA bound to a spot. The quality of image analysis process is crucial for accurate interpretation of data. Variety of algorithms and softwares are available, such as DeArray, ScanAlyze and CrazyQuant for the analysis.

Image analysis is an important aspect of microarray experiments. The main steps in the image analysis are gridding (spot recognition), segmentation and intensity extraction. Gridding is the process used to identify each spot on the array prior to extracting the information from it. The segmentation is the process of extracting information as to the pixel intensity within the spot once grids have been placed. Segmentation classifies the image pixels into foreground (pixels corresponds to spots of interest) and background (corresponds to the noise due to errors/contaminations). The intensity extraction step calculates the fluorescence intensities of the red and green spots, the back ground intensity and some quality measures. The background intensities are used to correct the foreground intensities.

3. **Data Pre-processing**

Microarray data can be quite noisy. Much of the variation in intensity levels can arise from technical rather than biological causes. Non-biological sources of variation can be introduced during sample preparation (e.g., dye effects), array manufacture (e.g., probe concentration), and hybridization (e.g., amount of sample) and in the measurement process (e.g., scanner inaccuracies). Preprocessing is the process of transforming the raw fluorescence signals detected by microarray scanner into a signals normalized for experimental errors. In order to accurately and precisely measure gene expression, it is important to consider random (experimental) and systematic variations. Hence preprocessing operations needs to be undertaken before analyzing the microarray data. Preprocessing of microarray data addresses following two issues: (1) To adjust for background intensities. (2) To transform the data into a scale suitable for analysis. Here we discuss important pre-processing techniques.

3.1. *Microarray normalization*

The process of minimizing the effects of systematic sources of variation, so that biological differences can be easily distinguished, is referred to as normalization. Normalization is performed both within each array and between arrays to make comparisons more meaningful. As pointed out by Yang *et al*. [38], [39] the purpose of normalization is to remove systematic variation in a microarray experiment which affects the measured gene expression levels. They summarized a number of normalization methods: (i) within-slide normalization, (ii) paired-slide normalization for dye-swap experiments, and (iii) multiple slide normalization. The normalization methods which have been presented in the literature can be

divided into two groups: linear methods and non-linear methods. The linear methods generally involve either estimating one or more global constants for a microarray, or fitting a linear regression to the $\log(R)$ versus $\log(G)$ data.

The non-linear methods that have been developed involve transforming the data onto the axes $((\log(R) + \log(G))/2$ versus $\log(R/G)$, and robustly fitting one or more robust lowess curves to the data and computing the residuals from the curve fit. Re-scaling of the points is done by dividing each final residual value by a robust estimate of the standard deviation of the residuals, the median absolute deviation. When a single curve is fitted, the method is referred to as slide normalization, and when a curve is fitted to the data for each individual array printer pin, the method is referred to as pin normalization.

3.2. *Filtering of genes*

The large number of gene expression profiles usually collected in microarray experiments can be drastically reduced since many of them exhibit near constant expression levels across samples. The aim of the filtering operation is to reduce variability by removing those genes whose measurements are not sufficiently accurate and also to decrease the dimensionality of the data by removing genes that are not sufficiently differentiated. Filtering methods, select a subset of genes independently of the classifier, are often used as pre-processing step to remove irrelevant features, i.e. those genes whose expression level is not correlated with the diagnosis. There are several techniques available to reduce data dimensionality and variability. A generally applicable filtering approach called the IQR filter proposed by Von Heydebreck, eliminate genes which are not sufficiently differentially expressed. This filter is implemented in the Bioconductor package. Comprehensive description of some filtering techniques are available in the book by Kohane *et al*. [18].

3.3. *Data transformation*

After normalization, the data must be transformed for statistical analysis. The type of transformation depends on the methods (parametric/nonparametric) used for the analysis. Since the corrected microarray intensity values are highly skewed many authors strongly recommended the log-transformation of the data (see, [41]). Some authors suggested cubic root transformation (see, Tusher *et al*. [32]) for microarray data.

4. **Classification methods**

Classification is known as discrimination in the statistical literature and it generates gene expression profiles which can discriminate between different known cell types or conditions. There is a distinction between classification and clustering. If the classes are pre-existing, then classification analysis is more appropriate than clustering analysis. Dudoit *et al*. [14] compared the performance

of various classification methods for classifying tumors based on gene expression profiles. Lee *et al*. [22] gave an extensive comparison of recent classification tools applied to microarray data. Here we discuss some of the important classification tools.

4.1. *Fisher's linear discriminant analysis (FLDA)*

FLDA is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. FLDA assumes observations have normal (Gaussian) distribution and equal group covariance. Additionally, variables cannot form linear expressions of one another. That is, they may not be perfectly correlated.

4.2. *Generalized partial least squares (GPLS))*

GPLS was originally designed for continuous response variable(s) and the method is advanced in the field of Chemometrics. Ding and Gentleman [12] applied generalized partial least squares approaches. Their functional ties are based on and extended to Iteratively Re Weighted Least Squares (IRWPLS) by fitting logit models for all C classes vs. baseline class separately with an option of Firth́s bias reduction procedure for two-group and multi-group classification proposed by Marx [25].

4.3. *Support Vector Machines (SVM)*

Support vector machine introduced by Vapnik [34] has attracted much research attention in recent years due its demonstrated improved generalization performance over other techniques in many real world applications including the analysis of microarrays [6], [33]. It has been used in classification as well as regression tasks. Advantage of this technique is that it minimizes the structural risk instead of the empirical risk. The principle is based on the fact that minimizing an upper bound on the generalization error rather than minimizing the training error is expected to perform better.

4.4. *Shrunken centroid method*

Tibshirani *et al*. [31] developed a shrunken centroid (SC) algorithm for classifying multiple cancer types. It is an integrated approach for feature selection and classification. Features are selected by considering one gene at a time, the difference between the class centroid (average expression level or ratio within a class) of a gene and the overall centroid (average expression level or ratio over all classes) of a gene is compared to the within-class standard deviation plus a shrinkage threshold, which is fixed for all genes. The genes with at least one class centroid that is significantly different from the overall centroid are selected as relevant genes. The size of the shrinkage threshold is determined by cross-validation on the training set to minimize classification errors.

5. **Clustering algorithms**

Clustering is a fundamental technique in exploratory data analysis and pattern discovery. Clustering microarray gene expression data yields valuable information into the gene function and reveal the underlying pattern in data sets. For example two functionally related genes are expressed in similar fashion. In addition, if a gene's function is unknown and it is clustered with gene's having known function, then the gene may share functionality with the genes of unknown function. Similarly, if the activity of genes in one cluster precedes the activity in a second cluster, then the genes in two clusters may be functionally related and genes in first cluster may regulate the activity of genes in the second cluster. Also clustering helps to identify co-regulated genes which is likely to be in common bio-chemical pathway. Following are commonly used clustering methods for microarray analysis.

5.1. *Hierarchical clustering*

An early and commonly used method is hierarchical clustering, an iterative procedure that selects at each iteration the two most similar vectors and merges them. This is perhaps the most commonly used clustering technique which presents the data as a list of genes in the form of dendrogram or tree view. At the initial level, all genes form their own clusters. Subsequently smaller groups of genes with smallest dissimilarity are grouped together producing a tree-like structure. The dissimilarity between two groups is taken to be the average of all the pairwise dissimilarity between the genes in the two groups. The tree is then cut at a predetermined height to determine the hard clusters. Similarity is defined as some distance measure (Euclidean, correlation, etc.). The output of the algorithm is a binary tree in which similar expression vectors are close together. Hierarchical clustering strategies have been applied to a fairly diverse set of experimental conditions to cluster genes with correlated expression patterns (Eisen *et al*. [16]).

5.2. *K-means*

In K-means clustering techniques objects are classified as belonging to one of k groups, k being chosen a priori. The cluster membership of an object is determined by calculating the centroid for each group (the multidimensional version of the mean) and assigning each object to the group with the closest centroid. This approach minimizes the overall within cluster dispersion by iterative reallocation of cluster members. The K-means clustering is sensitive to the chosen initial partition. There are two methods for determining an initial partition: a random based approach [2], [26] and an application-specific approach. The application-specific approach can determine an initial partition in several different ways. For example, we may set an initial partition for K-means clustering with the hierarchical clustering technique.

The K-means algorithm starts with a pre-specified cluster centers. The algorithm then assigns the observations into various clusters in order to minimize the within-class sum of squares. K-means, is an non hierarchical method, initially takes the number of data points on the microarray equal to the final required number of clusters. In this step itself, the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each point in the microarray and assigns it to one of the clusters depending on the minimum expression distance. The centroid position is recalculated every time a data point is added to the cluster and this continues until all the data points are grouped into the final required number of clusters. This method thus requires prior knowledge of the genes that are analyzed. One of the most significant limitations of this method is that you may get different results from the same data if the starting conditions are varied. Some other aspect of this method to note is that depending on the distance used for clustering, some data points may not converge and oscillate indefinitely, and the cluster solution can be influenced by the order of the cluster cases.

5.3. *Model based clustering*

The idea behind model based clustering is to regard the group labels as unknown parameters along with other model parameters which may be necessary to describe the data. The group levels are then selected by the methods like maximum likelihood. The method *mclust* (Banfield and Raftery, [3]) was used with the unconstrained option.

5.4. *Self-organizing maps (SOM)*

Self-organizing maps (SOM) are unsupervised learning algorithm. The application of SOM to expression data was pioneered in the Gene Cluster software developed at MIT and the Whitehead Institute (Tamayo *et al*. [30]). The SOM learning algorithm is an iterative procedure in which the vectors in the data set gravitate stochastically towards nodes in a pre-defined network of categories. The self-organizing map (SOM) is a method for producing ordered low dimensional representations of an input data space [30]. Typically, such input data is complex and high dimensional with data elements being related to each other in a nonlinear fashion. The self-organizing map method is ideally suited for explorative data analysis where prior information about the distribution of the data is not available. Also, the computational algorithms are relatively easy to implement, fast, and scalable to large data sets. The results are easy to visualize and interpret.

6. **Commonly used methods to find differentially expressed genes**

6.1. *Fold change*

Fold change is the simplest method for identifying differentially expressed genes. It evaluates the log ratio between two conditions (or the average of ratios

when there are replicates) and considers all genes that differ by more than an arbitrary cut-off value to be differentially expressed. The most important drawback of this method is that, because the threshold is chosen arbitrarily, it may often be inappropriate. For instance, if we want to select genes with at least 2 fold-change to be differentially expressed and the condition under study does not affect any gene to the point of inducing a 2 fold change, no genes will be selected resulting in zero sensitivity. On the other hand, if the condition is such that many genes are changing dramatically, the method will select too many genes and will have a low specificity. Also fold change does not take into account the variability in the data. This can lead to two problems. First, genes with low expression levels yet large fold changes and high variability may be identified as differentially expressed. Second, genes that display small but reproducible (i.e., low variability) changes in gene expression may be missed.

6.2. *The t-test*

The t-test is a simple, statistical method for detecting differentially expressed genes. When there are replicated samples under each condition the straight forward method is to adopt the traditional two sample t-test. Let X_1 and X_2 be two independent gene expression data under two conditions for a particular gene, the two sample t-statistic is computed as

$$t = \frac{\bar{X_1} - \bar{X_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.1}$$

where s^2, n_1 and n_2 are the pooled sample variance and the number of observations in each condition respectively and s^2 is computed as

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

where s_1^2 and s_2^2 are the sample variances. The t-statistic follows approximately a Student distribution, with $n_1 + n_2 - 2$ digress of freedom. The t-test utilize the variance of the samples hence it has the potential of addressing some of the shortcomings of fold change approach. However, the small sample sizes in microarray studies can affect the variance estimates.

6.3. *Modified t-tests*

More stable estimates can be obtained to find differentially expressed genes but these are subject to bias when the assumption of homogenous variance between genes is violated. In such situations, modified versions of the t-test are both more powerful and less subject to bias. The significance analysis of microarrays (SAM) is a modified version of the t-test (known as the S-test). To address the shortcomings

of the t-test several modifications have been applied.

$$t^* = \frac{\bar{M}}{(s+a)/\sqrt{n}} \tag{6.2}$$

where $\bar{M}$ is the mean of the M-values ($M = \log_2 \frac{R}{G}$, R is the intensity on the red channel and G is the intensity on the green channel) for any particular gene across the replicate arrays, 's' is the standard deviation of the M values across the replicates for the gene and n is the number of replicates. Any M-value that is an outlier will give rise to large standard deviations, which will usually prevent the gene in question from being spuriously identified as differentially expressed. With this modification, genes with small fold changes will not be selected as significant. Efron *et al*. [15], used a percentile of the distribution of sample standard deviations as 'a'. The SAM t-test (S-test) estimated 'a' from all the individual gene variances (Tusher [32]).

6.4. *Non-parametric approaches*

The Wilcoxon rank sum test (equivalent to Mann–Whitney test) for two groups or Kruskal-Wallis test for two or more groups has also been used as an alternative method in testing differential expression which can be applied especially when the data do not follow a normal distribution. Critical importance to possibly identify a small subset of genes that best discriminate between tissues under different conditions. The B statistic in (6.3) proposed by Lonnstedt and Speed [24] is log posterior odds ratio of differential expression versus non-differential expression

$$\mathrm{B} = \log \frac{\Pr\{\mathrm{DE}\}}{\Pr\{\text{not DE}\}}. \tag{6.3}$$

It allows for gene specific variances but it also combines information across many genes using an empirical Bayes approach and thus should be more stable than the t-statistic and it is equivalent to moderated t-statistics in terms of ranking of genes. The moderated t-statistic is shown to follow a t-distribution with augmented degrees of freedom. The moderated t inferential approach extends to accommodate tests involving two or more contrasts through the use of moderated F-statistics.

6.5. *ANOVA and multiple comparison*

ANOVA methods provide an automatic correction for the extraneous effects in a microarray experiment as an integral part of the data analysis. Changes in gene expression across experimental samples are captured in the variety x gene interaction terms of the ANOVA model. Kerr *et al*. [19] and G. Churchill [9] proposed the use of traditional ANOVA methods in analyzing microarray data (see, [20] and [21]). For a microarray gene expression data on several genes in various cell types (nominally two) fit a linear model on the log expression as the response. The model should include main effects and interaction due to gene and variety; it

could also include effects due to array and dye if appropriate data are available. Since the number of comparisons is enormous, the issue of a balance between pairwise and experiment wise error rate persists with this method as well. A typical ANOVA model can be given by,

$$\mathrm{y}_{ijkg} = \mu + \mathrm{A}_i + \mathrm{D}_j + \mathrm{AD}_{ij} + \mathrm{G}_g + \mathrm{AG}_{ig} + \mathrm{VG}_{kg} + \mathrm{DG}_{jg} + \epsilon_{ijkg} \tag{6.4}$$

where y_{ijkg} is the measured intensity from array i, dye j, variety k and gene g on appropriate scale (typically log scale). They called variety to refer to the mRNA samples under the study. The varieties may be treatment and control samples, cancer and normal cells or different time points of a biological process. In this model μ refers to the overall mean and the terms A, D and AD account for all aspects that are not gene specific. The gene effect G_g capture the average levels of expression of genes and the array-by-gene interaction (AG_{ig}) accounts for the difference due to varying sizes of spots on arrays. The dye-by-gene interaction (DG_{jg}) represent gene specific dye effects. None of these effects are of biological interest and helps in normalization of the data for ancillary source of variation. The effect of interest is the interaction between the gene and varieties (VG_{kg}), differences among these variety by gene interaction accounts for the relative gene expression. For example to estimate the relative gene expression of gene g in varieties 1 and 2, we should estimate $(\mathrm{VG}_{1g})-(\mathrm{VG}_{2g})$. The error terms ϵ_{ijkg} are assumed to be independent with mean 0 and variance σ^2.

6.6. *Regression model*

Linear models for microarray data analysis have been described by Kerr *et al.* [21], Wolfinger *et al*. [37]), Chu *et al*. [8] and Yang and Speed [40]. Kerr *et al*. [21] propose a single linear model for an entire microarray experiment. The single linear model approach assumes all equal variances across genes. Wolfinger *et al*. [37] fit separate models for each gene but model the individual channels of two color microarray data requiring the use of mixed linear models to accommodate the correlation between observations on the same spot. Chu *et al*. [8] propose mixed models for single channel oligonucleotide array experiments with multiple probes per gene. A general statistical model is represented as,

$$y_{gi} = \alpha_i + \beta_g x_i + \epsilon_{gi} \tag{6.5}$$

where $x_i = 1$, for $i = 1, 2, \ldots, n_1$ and $x_i = 0$, for $i = n_1 + 1, \ldots, n_2$.

The ϵ_{gi} are the random errors with mean 0. Hence, β_g represents the difference of expression levels of gene g across two conditions. The test for differential expression thus becomes testing for the null hypothesis in $H_0 : \beta_g = 0$ against $H_1 : \beta_g \neq 0$. The α_i and β_g estimated using weighted least square approach and estimated the variance of $\hat{\beta}_g$ using the robust or sandwich variance estimator.

6.7. *Multiple hypothesis testing*

The biological question of differential expression can be considered as a problem in multiple hypothesis testing in which m null hypotheses are simultaneously tested, where m can be considerably large. In such situations, false discoveries (true null hypothesis declared significant) are inevitable. Thus, it is important in any multiple testing problem to control the error rate of false discoveries. Multiple testing procedures consist of choosing a vector of cutoffs for the test statistics such that a suitably defined false positive rate is controlled at an apriori specified level.

A standard approach to the multiple testing problem consists of two aspects namely computing a test statistic T_j for each gene j and applying multiple testing procedures to determine which hypotheses to reject while controlling a suitably defined Type I error rate (Tusher *et al*. [32], Dudoit *et al*. [13] and Golub *et al.* [17]). The FDR-based multiple testing approaches, such as the Benjamini and Hochberg (BH) procedure [4] and the Benjamini and Liu procedure [5], have been developed for testing for a large family of hypotheses. FDR is the expected proportion of false positives among all rejected hypotheses. The Benjamini-Hochberg procedure has been shown to control the FDR under certain assumptions on the dependence structure of the gene expression levels. Unfortunately, there are many microarray studies not covered by the assumptions underlying the Benjamini-Hochberg algorithm. The Westfall and Young procedure [36] address the multiple test problem by controlling the family-wise error rate, which is the probability that at least one false positive occurs over the collective tests [16]. Storey and Tibshirani [29] developed a new measure, positive FDR (pFDR), that is an arguably more appropriate variation. Tusher *et al*. [32] developed a new FDR-based method, called Significance Analysis of Microarrays (SAM). SAM is very popular because it can identify genes with significant change of the level of expression and can estimate FDR based on permutations.

6.8. *Empirical Bayes Methods*

Bayesian approaches make assumptions about the parameters to be estimated (such as the differences between gene levels in treatment and control groups). Used intelligently, these assumptions can make use of prior experience with microarray data. A pure Bayes approach assumes specific distributions (prior distributions) for the mean differences of gene levels, and their standard deviations. The empirical Bayes approach assumes less, usually that the form of these distributions is known; the parameters of the prior distribution are estimated from data. As we get more detailed knowledge of the variability of individual genes, it should become possible to make detailed useful prior estimates based on past experience.

In the case of empirical Bayes studies, inference is usually made on some quantities related to the posterior distribution of the parameter of interest, or of a certain type of hypothesis. West *et al.* [35], considered an approach based on probit model and linear regression for characterizing differential gene expression. Another approach is to model the gene expression data through binary probit model for classification and use Bayesian variable selection methodology to select important genes (Brown *et al.* [7]). Lee *et al*. [23], proposed a hierarchical Bayesian model and employed latent variables to specialize the model to a regression setting and applied variable selection to select differentially expressed genes.

6.9. *Partial least squares*

Partial least squares is a statistical technique that is capable of modeling a large number of explanatory variables when only a few observations on each are present. An ordinary least squares regression would be impossible since the number of possible parameters will be lot more than what the available data can estimate. Since in a microarray experiment, expression levels of thousands of genes are measured, may be for different time points (or for a handful of samples), the use of partial least squares seems appropriate. Datta [10] investigated the question that how well the expression levels of a given gene can be predicted from the expression levels of the other genes using a simple model. The method of partial least squares was used for this purpose. Showed that this method may identify important gene relationships that warrant further biological investigation.

7. **Statistical model for microarray gene expression data**

In this section we use the two component mixed Gaussian distribution to model the distribution of a cDNA dual dye microarray gene expression data sets. We downloaded two cDNA dual dye microarray data sets (Experiment id-51401) from the Stanford Microarray Database. Each array chip contains approximately 42,000 human cDNA elements, representing over 30,000 unique genes. A type II experimental design was applied in which a universal human reference RNA (Stratagene) served as the common reference probe in all hybridisations. For each of these hybridisations, the universal human RNA was amplified and used to make cDNA probes labelled with green die Cy3 (Amersham). Amplified RNA from tissue samples was used to make cDNA probes labelled with Cy5 (red). Finally, after gridding, array data were uploaded to the Stanford Microarray Database. The dataset was normalised using quantile normalization method. This method is capable of removing intensity dependence in $\log_2(R_i/G_i)$ values. After normalisation, each distribution of the gene expression has a similar shape and exhibits heavier tails compared to a Gaussian distribution and a certain degree of

asymmetry. Box plot of $\log_2(R_i/G_i)$ for before and after quantile normalization are given below.

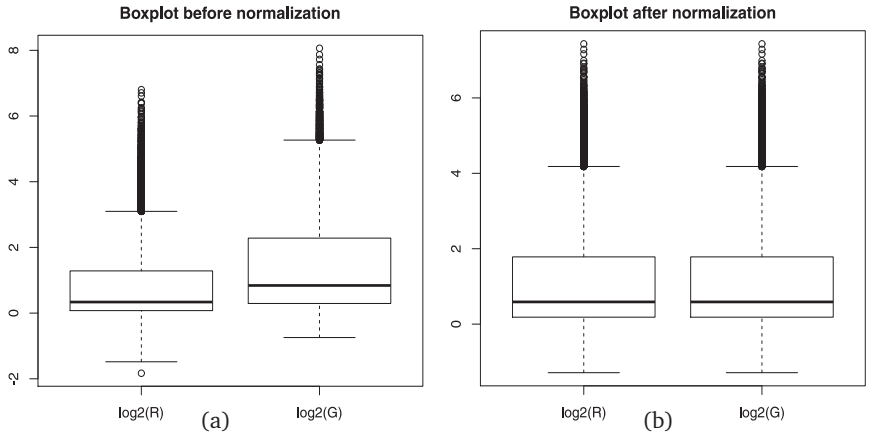

Figure 1. Box plot for microarray gene expression data experiment id-51401 (a) Before quantile normalization, (b) After quantile normalization

Mixture models are an interesting and flexible model family. The different uses of mixture models include for example generative component models, clustering and density estimation. Moreover, mixture models have been successfully used in various kinds of tasks such as modelling failure rate data, clustering teaching behaviour and in general for modelling large heterogeneous populations.

A two-component Gaussian mixture model can be defined by two Gaussian distributions, $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, and the probability that the random variable (the observable) arises from the first distribution is κ. The parameter in this model is the vector $\theta = (\kappa, \mu_1, \sigma_1, \mu_1, \sigma_2)$. Then the pdf of the mixture is

$$f(x;\theta) = \kappa f_1(\mu_1, \sigma_1) \; + \; (1-\kappa)\, f_2(\mu_2, \sigma_2) \tag{7.1}$$

where $f_i(\mu_i, \sigma_i)$ is the Gaussian pdf with parameters μ_i and σ_i. We fitted the mixed Gaussian distribution to the intensities $m = \log_2(R_i/G_i)$ for the data set (Experiment id-51401). Maximum likelihood estimators (MLE) of the parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ are obtained by using the *optim* function of the R statistical software, applying the BFGS algorithm (See R Development Core Team, [27]). The maximum likelihood estimates of the parameters, standard errors (SE) and 90% lower and upper confidence limits (LCL & UCL) are reported in Table 2. Figure 2, depicts a histogram of the gene expression data and the fitted probability density function evaluated at the MLEs.

We have proposed the mixed Gaussian family as a class of distributions that might be used to describe the pattern of differential gene expression for genes that are differentially expressed. This family was chosen primarily for its flexibility and tractability.

Table 1. Application - MLEs, standard errors (SE) and 90% confidence limits (LCL & UCL) for the parameters

	MLE	SE	LCL	UCL
μ_1	−0.040	0.010	−0.0570	−0.0232
μ_2	0.023	0.002	0.0194	0.0272
σ_1	1.232	0.010	1.2162	1.2481
σ_2	0.293	0.002	0.2883	0.2981
κ	0.368	0.005	0.3589	0.3764

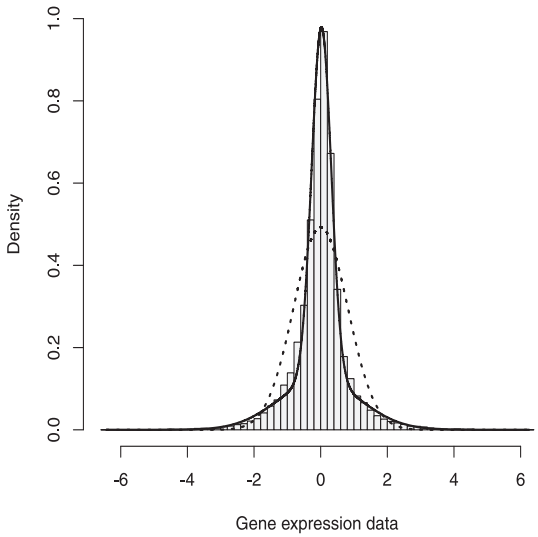

Figure 2. Histogram of microarray gene expression data. The lines represent distributions fitted using maximum likelihood estimation: mixed Gaussian probability density function (black line) and normal (black dotted line)

8. **Summary**

Microarrays are high throughput biological assays that allow the screening of thousands of genes for their expression. The main idea behind microarrays is to compute for each gene a unique signal that is directly proportional to the quantity of mRNA that was hybridized on the chip. Microarray experiment consists of large number of steps and errors can be introduced at any of these steps while performing the experiment. As many as possible of these errors should be taken into consideration while designing the layout of the experiment. A careful planning of the experiment before the actual execution would eventually minimize the effect of unwanted variations and maximize the precision of the estimates of the parameters of interest.

In this work, we have presented a new statistical model for the distribution of differential gene expression. The model provides the flexibility for modelling impulsiveness and skewness required for gene expression data. We believe that the statistical model presented in this paper will be very useful in estimation and detection problems involving microarray gene expression data. The mixed Gaussian distribution is often considered as a noise model in a number of signal processing applications.

The scope of statisticians to get involved in this exciting new area is tremendous, since the huge volume of data emerging from the microarray experiments need efficient and proper statistical methods for deriving valid conclusions. The tools and techniques described here are by no means comprehensive and many new algorithms and software tools are under development. This article makes a modest attempt to summarize the current state of affairs of statistical research in the area of microarray technology.

Acknowledgment

The first author is grateful to the Department of Science & Tehnology, Government of India, New Delhi, for financial support under the Women Scientist Scheme (WOS-A (2008)), Project No: SR/WOS-A/MS-09/2008.

References

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

[2] P. Baldi and G.W. Hatfield, *DNA Microarrays and Gene Expression*, Cambridge University Press (2002).

[3] J.D. Banfield and A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49** (1993), 803–822.

[4] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B* **57** (1995), 289–300.

[5] Y. Benjamini and W. Liu, A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence, *J. Stat. Plan. Inference* **82** (1995), 163–170.

[6] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C. Sugnet, M. Agnes, Jr. and D. Haussler, *Support vector machine classification of microarray gene expression data*, Technical report, University of California (Santa Cruz), (1999).

[7] P.J. Brown, M. Vannucci and T. Fearn, Multivariate Bayesian variable selection and prediction, *Journal of the Royal Statistical Society*: Series B (Statistical Methodology) **60** (1998), 627–641.

[8] T.M. Chu, B. Weir and R. Wolfinger, A systematic statistical linear modeling approach to oligonucleotide array experiments, *Mathematical Biosciences* **176** (2002), 35–51.

[9] G.A. Churchill, Fundamentals of experimental design for cDNA microarrays, *Nature Genet* **32** Supp. l (2002), 490–495.

[10] S. Datta, Exploring relationships in gene expressions: a partial least squares approach, *Gene Expression* **9**(2001), 257–264.

[11] J.L. De Risi, V.R. Iyer and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* **278**(1997), 680–685.
[12] B. Ding and R. Gentleman, *Classification Using Generalized Partial Least Squares*, Department of Biostatistics, Harvard University, 2003.
[13] S. Dudoit, Y.H. Yang, M.J. Callow and T.P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica* **12**(2002), 111–139.
[14] S. Dudoit, J. Fridlyand and P. Speed, Comparison of discrimination methods for classification of tumors using gene expression data, *J. Amer. Statist. Assoc.* **97**(2002) 77–87.
[15] B. Efron, R. Tibshirani, V. Goss and G. Chu, Microarrays and their use in a comparative experiment, *J. Amer. Stat. Assoc.* **96**(2001), 1151–1160.
[16] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Nat. Acad. Sci.* **95**(1998), 14863–14868.
[17] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard and M. Gaasenbeek *et al*., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(1999), 531–537.
[18] I.S, Kohane, A.T. Kho and A.J. Butte, *Microarrays for an Integrative Genomics*, MIT Press, Cambridge, MA, (2002).
[19] K.M. Kerr, M. Martin and G.A. Churchill, Analysis of variance for gene expression microarray data, *J. Comput. Biol.* **7**(2000), 819–837.
[20] K.M. Kerr and G.A. Churchill, experimental design issues for gene expression microarrays, *Biostatistics* **2**(2001), 183–201.
[21] K.M. Kerr, C.A. Afshari, L. Bennett, P. Bushel, N.W. Martinez and G.A. Churchill, Statistical analysis of a gene expression microarray experiment with replication, *Statistica Sinica* **12**(2002), 203–217.
[22] J.W. Lee, J.B. Lee, M. Park and S.H. Song, An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis* **48** (2005), 869–885.
[23] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci and B.K. Mallick, Gene selection: a Bayesian variable selection approach, *Bioinformatics* **19**(1)(2003), 90–97.
[24] I. Lonnstedt and T.P. Speed, Replicated microarray data, *Statistica Sinica* **12**(2002), 31–46.
[25] B.D. Marx, Iteratively reweighted partial least squares estimation for generalized linear regression, *Technometrics* **38** (1996), 374–381.
[26] W. Pan, J. Lin and C. Le, How many replicates of arrays are required to detect gene expression changes in microarray experiment? A mixture model approach, *Biostatistics*, University of MN Technical Report, 2002.
[27] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, `http://www.R-project.org/`, 2006.
[28] G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology* **3**(1) (2004), 3.
[29] J.D. Storey and R. Tibshirani, Statistical methods for identifying differentially expressed genes in DNA microarrays, *Methods Mol. Biol.* **224** (2003), 149–157.
[30] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander and T. Golub, Interpreting patterns of gene expression with self-organizing maps, *PNAS* **96** (1999), 2907–2912.

[31] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proc. Natl. Acad. Sci. USA* **99** (2002), 6567–6572.
[32] V.G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of National Academy of Science* **98** (2001), 5116–5121.
[33] G. Valentini, Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles, *Artificial Intelligence in Medicine* **26** (2002), 281–304.
[34] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
[35] M. West, J.R. Nevin, J.R. Marks, R. Spang and H. Zuzan, *Bayesian regression analysis in the Large p, small n paradigm with application in DNA microarray studies*, Technical Report, Duke University, 2000.
[36] P.H. Westfall and S.S. Young, *Resampling-based Multiple Testing*, John Wiley & Sons, New York, 1993.
[37] R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari and C.R.S. Paules, Assessing gene significance from cDNA microarray expression data via mixed models, *J. Comp. Biol.* **8**(2001), 625–637.
[38] Y.H. Yang, S. Dudoit, P. Luu and T.P. Speed, Normalization for cDNA microarray, In: M.L. Bittner, Y. Chen, A.N. Dorsel and E.R. Dougherty (eds), *Microarrays: Optical Technologies and Information*, SPIE, Soceity for Optical Engineering, San Jose, CA, 2001.
[39] Y.H. Yang, S. Dudoit, P. Luu, M.D. Lin, V. Peng, J. Ngai and T.P. Speed, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucl. Acids Res.* **30**(4)(2002), e15.
[40] Y.H. Yang and T.P. Speed, *Design and analysis of comparative microarray experiments, In: T.P. Speed (ed.)*, *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall: CRC Press, 35–91, 2003.
[41] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery and W.L. Ruzzo, Model based clustering and data transformations for gene expression data, *Bioinformatics* **17** (2001), 977–987.

Bindu Punathumparambath, *Department of Statistics, St.Thomas College, Pala, Kerala, India*.
E-mail: ppbindukannan@gmail.com

Sebastian George, *Department of Statistics, St.Thomas College, Pala, Kerala, India*.
E-mail: sthottom@gmail.com

Kannan V.M., *Department of Zoology, University of Calicut, Kerala, India*, (contributed section 1 and section 2).
E-mail: kannanvadakkadath@gmail.com

Received April 1, 2011
Accepted July 26, 2011